SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 1–20, January 1995

THE SPARSE BASIS PROBLEM AND MULTILINEAR ALGEBRA*

RICHARD A. BRUALDI†, SHMUEL FRIEDLAND‡, AND ALEX POTHEN§

Abstract. Let A be a $k \times n$ underdetermined matrix. The sparse basis problem for the row space W of A is to find a basis of W with the fewest number of nonzeros. Suppose that all the entries of A are nonzero, and that they are algebraically independent over the rational number field. Then every nonzero vector in W has at least $n - k + 1$ nonzero entries. Those vectors in W with exactly $n - k + 1$ nonzero entries are the elementary vectors of W. A simple combinatorial condition that is both necessary and sufficient for a set of k elementary vectors of W to form a basis of W is presented here. A similar result holds for the null space of A where the elementary vectors now have exactly $k + 1$ nonzero entries. These results follow from a theorem about nonzero minors of order m of the $(m - 1)$st compound of an $m \times n$ matrix with algebraically independent entries, which is proved using multilinear algebra techniques. This combinatorial condition for linear independence is a first step towards the design of algorithms that compute sparse bases for the row and null space without imposing artificial structure constraints to ensure linear independence.

Key words. elementary vector, matrix compound, null-space basis, row-space basis, sparse matrix, wedge product

AMS subject classifications. primary 65F50, 65K05, 15A69

1. Introduction.

Many situations in computational linear algebra and numerical optimization require the computation of a sparse basis for the row space or the null space of a sparse, underdetermined matrix A. The *sparse row-space basis problem* (hereafter the row-space problem) is to compute a basis for the row space of A with the fewest number of nonzeros. Similarly, the *sparse null-space basis problem* (hereafter the null-space problem) is to compute a basis for the null space of A with the fewest number of nonzeros. It turns out that both of these problems are computationally intractable: they are NP-hard [1], [8], [9]. Under a nondegeneracy assumption called the *matching property*, Hoffman and McCormick [5], [8] designed polynomial time algorithms to solve the row space problem. Sparsest null bases can be characterized by means of a matroid greedy algorithm [1], [9], yet the null space problem turned out to be harder than the row-space problem; heuristic algorithms to compute sparse null bases were designed and implemented in [2] and [4].

All algorithms known to us for computing sparse null bases have two components: a method to compute a sparse vector in the null space of the given matrix and a mechanism for ensuring linear independence when previously computed null vectors are augmented with the new null vector. To keep the time complexity of null basis algorithms low, the latter is achieved by insisting that the null basis be a trapezoidal matrix; that is, a matrix of the form $\begin{bmatrix} B_1 & L \end{bmatrix}$ where L is either an identity matrix

* Received by the editors April 22, 1992; accepted for publication (in revised form) by J. Gilbert, October 14, 1993. Part of this work was done while the authors were visiting the Institute for Mathematics and Its Applications (IMA) at the University of Minnesota.

† Department of Mathematics, University of Wisconsin, Madison, Wisconsin 53706 (brualdi@math.wisc.edu). This author's research was partially supported by National Science Foundation grant DMS-8901445 and by National Security Agency grant MDA904-89-H-2060.

‡ Department of Mathematics, University of Illinois at Chicago, Chicago, Illinois 60680 (u12735@uicvm.bitnet).

§ Department of Computer Science, University of Waterloo, Waterloo, Ontario, N2L 3G1 Canada (apothen@narnia.uwaterloo.ca, na.pothen@na-net.ornl.gov). This author was supported by National Science Foundation grant CCR-9024954 and by U.S. Department of Energy grant DE-FG02-91ER25095 at the Pennsylvania State University and by the Canadian Natural Sciences and Engineering Research Council grant OGP0008111 at the University of Waterloo.

or a lower triangular matrix with nonzero diagonal elements. However, this might be a severe restriction on the structure of the null basis since there may be sparser null bases that are not trapezoidal.

The fundamental question that we consider is the following: Given an underdetermined matrix A whose nonzero elements are algebraically independent, is there a combinatorial condition that characterizes a set of linearly independent vectors in the row space (or null space) of A? By a combinatorial condition we mean a condition that uses only the zero-nonzero structure of the set of vectors. This question was raised as an unsolved problem in [9].[1] A solution to this problem will enable us to design algorithms for computing sparse bases for the row and null space without imposing artificial structure constraints to ensure linear independence.

Since we are concerned only with sparse bases, we can restrict our attention to *elementary vectors* of the subspace (Fulkerson [3], Rockafellar [10], and Tutte [12]). (This restriction is necessary to obtain a nontrivial solution of the problem.) Accordingly we now turn to a discussion of elementary vectors. Let $x = (x_1, x_2, \ldots, x_n)$ be a vector in the n-dimensional real vector space $\mathbf{R}^n$. The *support* of x is the subset of $\{1, 2, \ldots, n\}$ given by $\text{supp}(x) = \{i : x_i \neq 0\}$. Now let W be a subspace of dimension k of $\mathbf{R}^n$. An *elementary vector* of W is a nonzero vector of W whose support is minimal, that is, does not properly contain the support of any nonzero vector of W. It is easy to verify that two elementary vectors of W with the same support are scalar multiples of each other and hence, up to scalar multiples, W has only finitely many elementary vectors. It is also easy to verify that the elementary vectors of W span W. It follows that a sparsest basis of W contains only elementary vectors. Thus it is natural to look for a basis of W among its elementary vectors.

Hence a more precise statement of the problem is to combinatorially characterize a set of linearly independent elementary vectors in the row space or the null space of an underdetermined matrix whose nonzero elements are algebraically independent. This problem turns out to be quite difficult, since the set of supports of the elementary vectors of a subspace W can have an intricate structure. However, we now consider a situation in which the set of supports of the elementary vectors has a simple structure, and in this case, we provide a combinatorial characterization of linear independence. Our proof of this result uses techniques from multilinear algebra.

Let A be a $k \times n$ matrix that is *nondegenerate* in the sense that every submatrix of A of order k is nonsingular. Then the support of each elementary vector in the row space of A has cardinality $n-k+1$ and each subset of $\{1, 2, \ldots, n\}$ of cardinality $n-k+1$ is the support of some elementary vector (see the next section). Similarly the support of each elementary vector in the null space of W has cardinality $k+1$ and each subset of $\{1, 2, \ldots, n\}$ of cardinality $k+1$ is the support of some elementary vector. Even in the restrictive case in which W is the row space or null space of a nondegenerate matrix, it seems difficult to determine if a set of elementary vectors of W is linearly independent. The linear independence of elementary vectors of such subspaces W does not generally depend only on the supports of the elementary vectors. Thus we need a more restrictive assumption than nondegeneracy.

A $k \times n$ matrix A is *generic* if all of its kn elements are nonzero and they form an algebraically independent set over the rational number field $\mathbf{Q}$. If A is generic over $\mathbf{Q}$, then obviously every submatrix of A of order k has a nonzero determinant. Hence, generic matrices are nondegenerate.

[1] We thank Steve Vavasis for rekindling our interest in this problem by raising it during the open problem session at the IMA workshop on Sparse Matrix Computations in October 1991.

In this paper we identify a necessary and sufficient condition that must be satisfied by the supports of the elementary vectors of the row space (respectively, null space) of a generic matrix in order that the elementary vectors be linearly independent. This condition leads to a polynomial algorithm for determining whether a set of elementary vectors in one of these two subspaces is a basis.

Let $\mathcal{J} = \{J_1, J_2, \ldots, J_t\}$ be t subsets of $\{1, 2, \ldots, n\}$ each of cardinality $m-1$. Then $\mathcal{J}$ satisfies the *m-intersection property* provided

$$|\cap_{i\in P} J_i| \leq m - |P| \quad (\forall P \subseteq \{1,2,\ldots,t\}, P \neq \emptyset). \tag{1.1}$$

The main results of this paper, as they apply to the row space and null space problems, are the following two theorems.

THEOREM 1.1. *Let A be a $k \times n$ matrix that is generic over $\mathbf{Q}$, and $\{I_1, I_2, \ldots, I_t\}$ denote a collection of $t \leq k$ subsets of $\{1,2,\ldots,n\}$ each of cardinality $n-k+1$. Then the elementary vectors $x(I_1), x(I_2), \ldots, x(I_t)$ with supports $I_1, I_2, \ldots, I_t$, respectively, of the row space of A are linearly independent if and only if the set $\{\overline{I}_1, \overline{I}_2, \ldots, \overline{I}_t\}$ consisting of the complements of their supports satisfies the k-intersection property, that is,*

$$|\cap_{i\in P} \overline{I}_i| \leq k - |P| \quad (\forall P \subseteq \{1,2,\ldots,t\}, P \neq \emptyset).$$

THEOREM 1.2. *Let A be a $k \times n$ matrix that is generic over $\mathbf{Q}$ and $\{I_1, I_2, \ldots, I_t\}$ denote a collection of $t \leq n-k$ subsets of $\{1,2,\ldots,n\}$ each of cardinality $k+1$. Then the elementary vectors $y(I_1), y(I_2), \ldots, y(I_t)$ with supports $I_1, I_2, \ldots, I_t$, respectively, of the null space of A are linearly independent if and only if the set $\{\overline{I}_1, \overline{I}_2, \ldots, \overline{I}_t\}$ consisting of the complements of their supports satisfies the $(n-k)$-intersection property, that is,*

$$|\cap_{i\in P} \overline{I}_i| \leq n - k - |P| \quad (\forall P \subseteq \{1,2,\ldots,t\}, P \neq \emptyset).$$

The combinatorial conditions given in these two theorems can be used to test the linear independence of a set of elementary vectors in polynomial time. We now show how this can be accomplished for the row space.

Let P be a nonempty subset of $\{1, \ldots, k\}$. The condition in Theorem 1.1 can be restated as

$$|\cup_{i\in P} I_i| \geq n - k + |P|,$$

since $|\cap_{i\in P} \overline{I_i}| + |\cup_{i\in P} I_i| = n$. Without loss of generality, assume that the rows in P are numbered $P = \{1, \ldots, p\}$. The last inequality yields

$$|\cup_{i\in P\setminus\{p\}} I_i \setminus I_p| \geq |P| - 1 \quad (\forall P \subseteq \{1,2,\ldots,p\}, p \in P).$$

If we let X denote the $k \times n$ matrix with rows $x(I_1), x(I_2), \ldots, x(I_k)$, then this is the set of Philip Hall conditions for the submatrix $X[\{1, \ldots, p-1\}, \overline{I_p}]$ to have a row-perfect matching.

We can use the above condition to test the linear independence of a set of elementary vectors in the row space when a partial basis of $p-1$ rows is augmented by a newly computed row p. We assume inductively that the partial basis satisfies the k-intersection property. Now when the pth row is added to the partial basis, we check whether the submatrix in the preceding paragraph has a row-perfect matching.

If it does, then clearly every set $P' \subseteq P$ that includes p satisfies the k-intersection property. Also, every set $P' \subseteq P$ that does not include p satisfies the k-intersection property by the inductive hypothesis. Hence the k-intersection property for row space bases can be checked by solving k maximum matching problems. The matchings can be computed in $\mathcal{O}(k^{1.5}e)$ time, where e is the number of nonzeros in the sparse row basis.

Theorems 1.1 and 1.2 are consequences of a theorem (Theorem 2.1) about compound matrices, and we briefly review this matrix construction. Let X be a $p \times q$ matrix and let r be a positive integer with $r \leq p, q$. Let $\mathcal{S}_{r,p}$ denote the sequence of all subsets of $\{1, 2, \ldots, p\}$ of cardinality r ordered lexicographically. Similarly, let $\mathcal{S}_{r,q}$ denote the sequence of all subsets of $\{1, 2, \ldots, q\}$ of cardinality r ordered lexicographically. The rth-*compound* of X is the $\binom{p}{r} \times \binom{q}{r}$ matrix $C_r(X)$ with rows indexed by $\mathcal{S}_{r,p}$ and columns indexed by $\mathcal{S}_{r,q}$ whose entry in the position corresponding to K in $\mathcal{S}_{r,p}$ and L in $\mathcal{S}_{r,q}$ is the determinant $\det X[K, L]$ of the submatrix of X with row indices in K and column indices in L. An important fact about compounds is that the multiplicative property $C_r(XY) = C_r(X)C_r(Y)$ holds. In particular, if X is a square nonsingular matrix of order n and $Y = X^{-1}$, then $C_r(X)C_r(X^{-1}) = C_r(I_n) = I_N$, where $N \equiv \binom{n}{r}$, and hence $C_r(X)$ is nonsingular. Notice that if X is a square matrix of order n, then $C_{n-1}(X)$ is, up to multiplication of some of its rows and columns by -1, the adjoint of X.

The rest of this paper is organized as follows. In §2, first we show that the problem of linear independence of a set of elementary vectors (of the row space and null space) of a $k \times n$ nondegenerate matrix A is equivalent to the problem of determining whether the determinant of a certain submatrix of the $(k-1)$th compound matrix $C_{k-1}(A)$ of A is not zero. The entries of $C_{k-1}(A)$ are the determinants of all the submatrices of A of order $k-1$ arranged in lexicographical order of their set of row indices and of their set of column indices. If the determinant of this submatrix of $C_{k-1}(A)$ is nonzero, then we show that k-intersection property must be satisfied. However, to prove the converse for generic matrices, we must show that the k-intersection property implies that this determinant is not identically zero. Since the determinant of a submatrix of $C_{k-1}(A)$ is an expression involving determinants of submatrices of A of order $k-1$, we are faced with the task of showing that it is *not* a determinantal identity.[2] We conclude §2 by stating our main result (Theorem 2.1) about compound matrices. In §3 we discuss certain concepts in multilinear algebra, namely, tensor spaces and exterior vector spaces that are needed to obtain our results. In §4 we state our main theorem (Theorem 4.1) in multilinear algebra, and in §5 we apply this theorem to prove Theorem 2.1. In §6 we give the proof of the main theorem. In §7 we make a few concluding remarks and state a conjecture.

2. Elementary vectors and matrix compounds. Let A be a $k \times n$ nondegenerate, real matrix and let W be the row space of A. Then each elementary vector of W contains exactly $k-1$ zeros and $n-k+1$ nonzeros. Moreover, given any subset I of $\{1, 2, \ldots, n\}$ of cardinality $n-k+1$, there is an elementary vector $x(I)$ of W

[2] One could argue that our task would have been a lot simpler if we had only to verify that a certain expression involving determinants of submatrices of A was a determinantal identity, that is, was equal to zero no matter what real values were substituted for the indeterminate entries of A. To show that an expression is not a determinantal identity, one must verify that one can choose real values for the indeterminate entries in order that the expression is not zero. One cannot expect to be able to construct these real values, but only to show that they must exist.

whose support equals I. The nonzero coordinates of the vector $x(I)$ are given by

$$x(I)_j = (-1)^{p_j+1} \det A[:, \overline{I} \cup \{j\}] \quad (j \in I), \tag{2.1}$$

where p_j equals the number of integers r in $\overline{I}$ that are less than j. Here $\overline{I}$ is the complement of I in $\{1, 2, \ldots, n\}$ and $A[:, \overline{I} \cup \{j\}]$ denotes the full-rowed submatrix of A of order k determined by the columns indexed by the integers in $\overline{I} \cup \{j\}$. To see that this defines a vector in the row space of A whose support is I, we expand the determinant in (2.1) by column j of A and obtain

$$x(I)_j = \sum_{i=1}^{k} (-1)^i \det A[\overline{i}, \overline{I}] a_{ij} \quad (j \in I), \tag{2.2}$$

where $\overline{i}$ denotes the complement of $\{i\}$ in $\{1, 2, \ldots, k\}$ and $A[\overline{i}, \overline{I}]$ is the submatrix of A determined by the rows and columns indexed by the integers in $\overline{i}$ and $\overline{I}$, respectively. For j in I, $x(I)_j$ is a linear combination of the elements in column j of A by (2.2). For j in $\overline{I}$, $x(I)_j$ is zero by (2.1), since it is the determinant of a matrix in which column j of A occurs twice. Thus $x(I)$ is a linear combination of the rows of A and hence belongs to the row space of A.

Let $x(I_1), x(I_2), \ldots, x(I_t)$ be t elementary vectors of W. For each vector $x(I_j)$ there exists a unique vector $y(I_j)$ in $\mathbf{R}^k$ such that

$$x(I_j) = y(I_j)A.$$

Moreover, since the rank of A is k, $x(I_1), x(I_2), \ldots, x(I_t)$ are linearly independent vectors in $\mathbf{R}^n$ if and only $y(I_1), y(I_2), \ldots, y(I_t)$ are linearly independent vectors in $\mathbf{R}^k$. Since $x(I_j)_i = 0$ for i in $\overline{I_j}$, the vector $y(I_j)$ is the unique (up to scalar multiples) nontrivial solution z in $\mathbf{R}^k$ of the $k-1$ equations

$$zA[:, \overline{I_j}] = 0.$$

Thus by Cramer's rule

$$y(I_j)_i = (-1)^i \det A[\overline{i}, \overline{I_j}] \quad (i = 1, 2, \ldots, k), \tag{2.3}$$

where, as before, $\overline{i}$ is the complement of $\{i\}$ in $\{1, 2, \ldots, k\}$. Hence

$$\begin{bmatrix} y(I_1)^T & y(I_2)^T & \cdots & y(I_t)^T \end{bmatrix}$$

is a $k \times t$ submatrix of the $(k-1)$st compound $C_{k-1}(A)$ of A. (More precisely, it is a $k \times t$ submatrix of $C_{k-1}(A)$ with row i multiplied by $(-1)^i$ for $i = 1, 2, \ldots, k$.) Note that $C_{k-1}(A)$ is a $k \times \binom{n}{k-1}$ matrix. Summarizing, we have what follows.

(i) The elementary vectors $x(I_1), x(I_2), \ldots, x(I_t)$ of the row space W of the $k \times n$ nondegenerate matrix A are linearly independent if and only if the $k \times t$ submatrix $C_{k-1}(A)[:, \{\overline{I_1}, \overline{I_2}, \ldots, \overline{I_t}\}]$ of $C_{k-1}(A)$ determined by its columns indexed by $\overline{I_1}, \overline{I_2}, \ldots, \overline{I_t}$ has rank equal to t. Equivalently, the elementary vectors $x(I_1), x(I_2), \ldots, x(I_t)$ are linearly independent if and only if not all of the determinants

$$\det C_{k-1}(A)[\{\overline{i_1}, \overline{i_2}, \ldots, \overline{i_t}\}, \{\overline{I_1}, \overline{I_2}, \ldots, \overline{I_t}\}],$$
$$(1 \le i_1 < i_2 < \cdots < i_t \le k)$$

vanish.

If we assume that the matrix A is generic over $\mathbf{Q}$, then by taking $t = k$ we see that the problem of determining whether a set of k elementary vectors of the subspace W (the row space of the $k \times n$ generic matrix A over $\mathbf{Q}$) is a basis of W is equivalent to the problem of determining whether the determinant of a submatrix of order k of the $(k-1)$st compound $C_{k-1}(A)$ does not vanish identically (that is, is *not* an identity satisfied by the determinants of the submatrices of order $k-1$ of $k \times n$ real matrices).

Considerations similar to the above apply to the null space U of the matrix A. Assume again that A is nondegenerate. Then the supports of elementary vectors of U are exactly the subsets I of $\{1, 2, \ldots, n\}$ of cardinality $k+1$. Indeed by Cramer's rule again, it follows that for each subset I of $\{1, 2, \ldots, n\}$ of cardinality $k+1$ the elementary vector $y(I)$ of U with support I satisfies

$$y(I)_i = (-1)^i \det A[:, I \setminus \{i\}] \quad (i \in I).$$

Let $y(I_1), y(I_2), \ldots, y(I_t)$ be t elementary vectors of U. There exists an $n-k \times n$ matrix B with rank equal to $n-k$ such that the row space of B equals U. Suppose that some submatrix of B of order $n-k$ has a zero determinant. Then after elementary row operations we may assume that some row of B has at least $n-k$ zeros. Since $AB^T = O$ this implies that some set of k columns of A is linearly dependent contradicting the nondegeneracy of A. We conclude that the matrix B is also nondegenerate. Let $z(I)$ be the unique vector in $\mathbf{R}^{n-k}$ such that $y(I) = z(I)B$. The vectors $y(I_1), y(I_2), \ldots, y(I_t)$ are linearly independent if and only if $z(I_1), z(I_2), \ldots, z(I_t)$ are linearly independent. The vector $z(I_j)$ is the unique (up to scalar multiples) nontrivial solution v of

$$vB[:, \overline{I_j}] = 0.$$

Using Cramer's rule as above we make the following conclusion.

(ii) The elementary vectors $y(I_1), y(I_2), \ldots, y(I_t)$ of the null space U of the k by n nondegenerate matrix A are linearly independent if and only if the $n-k \times t$ submatrix of $C_{n-k-1}(B)$ determined by its columns indexed by $\{\overline{I_1}, \overline{I_2}, \ldots, \overline{I_t}\}$ has rank equal to t. Equivalently, the elementary vectors $y(I_1), y(I_2), \ldots, y(I_t)$ are linearly independent if and only if not all of the determinants

$$\det C_{n-k-1}(B)[\{\overline{i_1}, \overline{i_2}, \ldots, \overline{i_t}\}, \{\overline{I_1}, \overline{I_2}, \ldots, \overline{I_t}\}],$$
$$(1 \le i_1 < i_2 < \cdots < i_t \le n-k)$$

vanish.

If A is generic over $\mathbf{Q}$, then by taking $t = n-k$, we see that the problem of determining whether a set of $n-k$ elementary vectors of the null space U of A is a basis of U is equivalent to the problem of determining whether the determinant of a full-rowed submatrix of order $n-k$ of the $(n-k-1)$st compound of the matrix B does not vanish identically.

Now let A denote an $m \times n$ real matrix. Let $J_1, J_2, \ldots, J_t$ be $t \le m$ subsets of $\{1, 2, \ldots, n\}$ each of cardinality $m-1$. We consider the $m \times t$ (full-rowed) submatrix

$$C_{m-1}(A)[:, \{J_1, J_2, \ldots, J_t\}] \tag{2.4}$$

of the $(m-1)$st compound of A. If for some $i \neq j$ we have $J_i = J_j$, then two columns of (2.4) are identical and hence the matrix has linearly dependent columns. If $t > m$ then (2.4) has more columns than rows and hence has linearly dependent columns.

We generalize these observations by showing that if the t sets $J_1, J_2, \ldots, J_t$ do not satisfy the m-intersection property, then the columns of (2.4) are linearly dependent.

Assume that $p \geq 2$ of the sets, say $J_1, J_2, \ldots, J_p$, satisfy

$$J_1 \cap J_2 \cap \cdots \cap J_p = J \quad \text{where } |J| = q \geq m - p + 1.$$

First suppose that the columns of A with index in J are linearly dependent. Then the matrix $A[:, J_1]$ has linearly dependent columns and hence its rank is at most $m - 2$. We may multiply A with nonsingular matrices corresponding to elementary row operations without changing linearly independent sets of columns of A. By the multiplicative property of compounds, the same observation can be made for compound matrices of A. Hence we may assume that the last two rows of $A[:, J_1]$ are zero rows. This implies that the column of $C_{m-1}(A)$ with index J_1 is a zero column and hence (2.4) has linearly dependent columns.

Now suppose that the columns of A with index in J are linearly independent. Using the multiplicative property of compounds again we may assume that

$$A = \left[\begin{array}{c|c} I_q & X \\ \hline O & F \end{array}\right],$$

where I_q is the identity matrix of order q, O is an $m - q \times q$ zero matrix, and F is an $m - q \times n - q$ matrix. Let Z be the $m \times p$ submatrix of (2.4) corresponding to the index sets $J_1, J_2, \ldots, J_p$. Let $J_i' = J_i \setminus J$ $(i = 1, 2, \ldots, p)$. The submatrix of Z determined by its last $m - q$ rows equals

$$C_{m-1}(A)[\{\overline{q+1}, \ldots, \overline{m}\}, \{J_1, J_2, \ldots, J_p\}] = C_{m-q-1}(F)[:, \{J_1', J_2', \ldots, J_p'\}].$$

By the Laplace expansion for determinants along a set of rows, it follows that for each j between 1 and q, the row of Z indexed by $\overline{j}$ is a linear combination of its last $m - q$ rows. Hence the rank of Z is at most

$$m - q \leq m - (m - p + 1) = p - 1.$$

Thus the columns of Z, and hence the columns of (2.4), are linearly dependent if $J_1, J_2, \ldots, J_t$ do not satisfy the m-intersection property.

Our main result about compound matrices asserts that for generic matrices, the converse holds as well.

THEOREM 2.1. *Let A be an $m \times n$ matrix that is generic over $\mathbf{Q}$. Let $J_1, J_2, \ldots, J_t$ be t subsets of $\{1, 2, \ldots, n\}$ each of cardinality $m - 1$. Then the rank of the $m \times t$ submatrix of the $(m-1)$st compound $C_{m-1}(A)$ given by*

$$C_{m-1}(A)[:, \{J_1, J_2, \ldots, J_t\}] \tag{2.5}$$

equals t if and only if $J_1, J_2, \ldots, J_t$ satisfy the m-intersection property.

In the next section we discuss the multilinear algebra that we use to show that if A is generic over $\mathbf{Q}$ and $J_1, J_2, \ldots, J_t$ are subsets of cardinality $m - 1$ that satisfy the m-intersection property

$$|\cap_{i \in P} J_i| \leq m - |P| \quad (P \subseteq \{1, 2, \ldots, t\}), \tag{2.6}$$

then the columns of (2.5) are linearly independent.

Theorem 2.1 is proved in §5.

3. Tensor and exterior spaces. We refer the reader to Marcus ([6] and [7]) for the basic multilinear algebra discussed in this section. As already pointed out, our task is made more complicated by the fact that we must show that a certain expression is not a determinantal identity. The multilinear algebra is needed (apparently) to show the existence of certain numbers without actually being able to construct them.

Let W be an n-dimensional vector space over $\mathbf{R}$. The *tensor product* of W with itself is the n^2-dimensional real vector space $W \otimes W$ spanned by the decomposable tensors $x \otimes y$ with x and y in W. The tensor product is an abstract algebraic construction. If W equals $\mathbf{R}^n$ and $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ are vectors in W, then a concrete realization of $x \otimes y$ is the outer product $x^T y$. In this case, $W \otimes W$ is the vector space spanned by the outer products of vectors in W.

The mth *tensor power* of W is the n^m-dimensional real vector space

$$\otimes^m W = W \otimes \cdots \otimes W \ (m \ W\text{'s})$$

spanned by all of the decomposable tensors $w_1 \otimes \cdots \otimes w_m$ where $\{w_1, \ldots, w_m\} \subseteq W$. The essential facts to keep in mind about the tensor power $\otimes^m W$ are as follows.

(1) The map

$$(w_1, \ldots, w_m) \to w_1 \otimes \cdots \otimes w_m$$

is multilinear: for instance,

$$(cw_1' + dw_1'') \otimes w_2 \otimes \cdots \otimes w_m = c(w_1' \otimes w_2 \otimes \cdots \otimes w_m) + d(w_1'' \otimes w_2 \otimes \cdots \otimes w_m)$$

for all real numbers c and d and all vectors $w_1', w_1'', w_2, \ldots, w_m$ in W.

(2) $c(w_1 \otimes w_2 \otimes \cdots \otimes w_m) = (cw_1) \otimes w_2 \otimes \cdots \otimes w_m = \cdots = w_1 \otimes w_2 \otimes \cdots \otimes (cw_m)$ for all real numbers c and all vectors $w_1, w_2, \ldots, w_m$ in W.

(3) If $\{x_1, x_2, \ldots, x_n\}$ is a basis of W then the set of n^m vectors

$$\{x_{i_1} \otimes \cdots \otimes x_{i_m} : 1 \leq i_1, \ldots, i_m \leq n\}$$

is a basis of $\otimes^m W$.

An inner product $(\cdot, \cdot)$ on W induces an inner product on $\otimes^m W$ by defining

$$(w_1 \otimes \cdots \otimes w_m, v_1 \otimes \cdots \otimes v_m) = \prod_{i=1}^{m} (w_i, v_i)$$

and extending linearly.[3]

The *wedge product* of vectors $w_1, \ldots, w_m$ is the element of $\otimes^m W$ defined by

$$w_1 \wedge \cdots \wedge w_m = \sum_{\sigma} \operatorname{sign}(\sigma) w_{\sigma(1)} \otimes \cdots \otimes w_{\sigma(m)},$$

where the summation extends over all permutations σ of $\{1, 2, \ldots, m\}$ and $\operatorname{sign}(\sigma)$ is $+1$ if σ is an even permutation and -1 otherwise. If $w_1, w_2, \ldots, w_m$ are the row vectors of an $m \times n$ matrix B, then $C_m(B)$ is a concrete realization of $w_1 \wedge \cdots \wedge w_m$. The subspace of $\otimes^m W$ spanned by all the wedge products of m vectors of W is the mth *exterior space*[4] over W and is denoted by $\wedge^m W$. The essential facts to keep in mind about the exterior space $\wedge^m W$ are as follows.

[3] All of this applies to the complex number field provided we use a unitary inner product.

[4] It is also called the mth *Grassmann space* over W and the mth *skew-symmetric space* over W.

(i) If $\{y_1, y_2, \ldots, y_n\}$ is a basis of W, then the set of vectors

$$\{y_{i_1} \wedge \cdots \wedge y_{i_m} : 1 \leq i_1 < \cdots < i_m \leq n\}$$

is a basis of $\wedge^m W$ (in particular, these vectors are linearly independent) and

$$\dim \wedge^m W = \binom{n}{m}.$$

(ii) $w_1 \wedge \cdots \wedge w_m = 0$ if and only if the vectors $w_1, \ldots, w_m$ are linearly dependent.

(iii) If U is a subspace of W of dimension m with a basis $u_1, \ldots, u_m$, then $\{u_1 \wedge \cdots \wedge u_m\}$ is the subspace $\wedge^m U$ of $\wedge^m W$ of dimension one.

Using the definition of wedge product, we calculate that the induced inner product on the exterior space $\wedge^m W$ satisfies

$$\begin{aligned}(u_1 \wedge \cdots \wedge u_m, v_1 \wedge \cdots \wedge v_m) &= m!(u_1 \otimes \cdots \otimes u_m, v_1 \wedge \cdots \wedge v_m) \\ &= m! \det \begin{bmatrix} (u_1, v_1) & \cdots & (u_1, v_m) \\ \vdots & \ddots & \vdots \\ (u_m, v_1) & \cdots & (u_m, v_m) \end{bmatrix}.\end{aligned} \tag{3.1}$$

Hereafter we shall denote any matrix of the form as the one appearing in (3.1) by specifying its (i, j)th element:

$$\begin{bmatrix} (u_i, v_j) \end{bmatrix} \quad (\text{for } i, j = 1, \ldots, m).$$

If U and V are two subspaces of W of dimension m, then it follows from (ii) and (iii) that for bases $\{u_1, \ldots, u_m\}$ of U and $\{v_1, \ldots, v_m\}$ of V, whether or not $(u_1 \wedge \cdots \wedge u_m, v_1 \wedge \cdots \wedge v_m)$ equals zero is independent of the choice of the bases $\{u_1, \ldots, u_m\}$ and $\{v_1, \ldots, v_m\}$ of V. *For convenience we denote any of these inner products* $(u_1 \wedge \cdots \wedge u_m, v_1 \wedge \cdots \wedge v_m)$ *by* $[U, V]$. The *orthogonal complement* of a subspace V of W is denoted by $V^\perp$.

LEMMA 3.1. *Let U and V be subspaces of W of dimension m. Then the following are equivalent:*

(a) $[U, V] \neq 0$,

(b) $U^\perp \cap V = \{0\}$,

(c) $U \cap V^\perp = \{0\}$.

Proof. Let $u_1, \ldots, u_m$ be a basis of U. If there were a nonzero vector v_1 in $U^\perp \cap V$, then extending v_1 to a basis $v_1, \ldots, v_m$ of V we see that the determinant in (3.1) is zero and hence $[U, V] = 0$. Therefore (a) implies (b). Now assume that (b) holds and consider the vector space $U_i = \langle u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_m \rangle$ spanned by all but the ith basis vector u_i. It follows from (b) that $\dim U_i^\perp \cap V = 1$, since we can obtain a vector in this subspace by subtracting appropriate multiples of the vectors in a basis of U_i from u_i. Let v_i be any nonzero vector in $U_i^\perp \cap V$. By (b) again we conclude that $(u_i, v_i) \neq 0$. We can repeat this argment to conclude that $(u_i, v_i) \neq 0$, for $i = 1, \ldots, m$. Hence the determinant in (3.1), $\prod_{j=1}^m (u_j, v_j) \neq 0$ and thus (a) holds. Since $[U, V] = [V, U]$, (b) and (c) are equivalent and the lemma follows. □

4. A theorem in multilinear algebra. We now formulate a theorem concerning exterior spaces that enables us to solve our original problems concerning bases for the row space and null space of a generic matrix. In the next section we show how this

theorem and a combinatorial lemma can be used to prove Theorems 1.1 and 1.2. In the final section we prove the multilinear algebra theorem. It will be convenient to use the language of projective geometry and algebraic varieties to describe the theorem.

We obtain an equivalence relation on points in $\mathbf{R}^{N+1}$ by defining two points $x = (x_0, \ldots, x_N)$ and $x' = (x_0', \ldots, x_N')$ to be *equivalent* if there is a real constant λ such that $x = \lambda x'$. Then N-dimensional projective space over the real field $\mathbf{P}^N(\mathbf{R})$ is the set of equivalence classes of this relation on $\mathbf{R}^{N+1} \setminus \{0\}$, and $(x_0, \ldots, x_N)$ are the homogeneous coordinates of x. Note that the projective dimension is one less than the number of coordinates.

Let

$$U_0 = \{p : p = (x_0, \ldots, x_N) \in \mathbf{P}^N(\mathbf{R}) \text{ and } x_0 \neq 0\}.$$

Then the map Φ taking $(x_1, \ldots, x_N) \in \mathbf{R}^N$ to $(1, x_1, \ldots, x_N) \in \mathbf{P}^N(\mathbf{R})$ is a one-to-one correspondence between $\mathbf{R}^N$ and U_0 because given $p = (x_0, \ldots, x_N) \in U_0$, we can multiply by $(1/x_0)$ to obtain an equivalent point and then compute the inverse map from U_0 to $\mathbf{R}^N$. Thus we can identify U_0 with $\mathbf{R}^N$. If $H = \{p \neq 0 : p = (0, x_1, \ldots, x_N)\}$ "the hyperplane at infinity," then N-dimensional projective space has the representation $\mathbf{P}^N(\mathbf{R}) = U_0 \cup H$, i.e., it consists of $\mathbf{R}^N$ augmented with the hyperplane at infinity.

A *variety* is the solution set of a system of multivariate polynomials $p_1 = 0, \ldots, p_s = 0$ in the variables $x_0, \ldots, x_N$. It is a *projective variety* if each p_i is a homogeneous polynomial, i.e., each term in p_i has the same total degree.

Let W be an inner product space of dimension n over $\mathbf{R}$. Let m be an integer with $1 \leq m \leq n$. The set of all subspaces X of W of dimension m are the points of a projective variety $\mathcal{W}_m$. Choose an $m \times n$ matrix E whose rows form a basis of X, and consider the map $X \mapsto C_m(E)$ that maps the subspace X to the set of $\binom{n}{m}$ determinants of all submatrices of order m of E. This is a well-defined, injective map from the set of m-dimensional subspaces of W to real projective space $\mathcal{P}$ of (projective) dimension $\binom{n}{m} - 1$. The $\binom{n}{m}$ homogeneous coordinates are called the *Plücker coordinates* of X and they satisfy certain quadratic relations called the *Plücker relations*. If we choose another matrix F whose rows form a basis of X, then the effect is to multiply the *Plücker coordinates* of X by a common nonzero scale factor. The projective variety $\mathcal{W}_m$ consists of all points that satisfy the *Plücker relations* and is known as the *Grassmann variety.*

A *subvariety* of $\mathcal{W}_m$ is a variety that is a nonempty subset of the subspaces in $\mathcal{W}_m$. A subvariety of $\mathcal{W}_m$ is *proper* provided that it does not contain at least one subspace of W.

Let X denote a subspace of W of dimension m. By property (i) of exterior spaces, $\wedge^{m-1}X$ is a subspace of $\otimes^{m-1}W$ of dimension m. By property (ii) $\wedge^m(\wedge^{m-1}X)$ is a subspace of $\otimes^{m(m-1)}W$ of dimension one. Let $U_1, U_2, \ldots, U_m$ be m subspaces of W of dimension $m-1$. Then each $\wedge^{m-1}U_i$ is a subspace of $\otimes^{m-1}W$ of dimension one, and $(\wedge^{m-1}U_1) \wedge (\wedge^{m-1}U_2) \wedge \cdots \wedge (\wedge^{m-1}U_m)$ is a subspace of $\otimes^{m(m-1)}W$ of dimension zero or one. The subspaces $U_1, U_2, \ldots, U_m$ satisfy the *dimension m-intersection property* provided that

$$\dim \cap_{i \in P} U_i \leq m - |P| \quad (\forall P \subseteq \{1, 2, \ldots, m\}, P \neq \emptyset). \tag{4.1}$$

Clearly the dimension m-intersection property is the analogue for subspaces of the m-intersection property for subsets.

We now come to the main theorem, the proof of which is given in the final section.

THEOREM 4.1. *Let W be an inner product space over $\mathbf{R}$ of dimension n, let m be an integer with $2 \leq m \leq n$, and let $U_1, U_2, \ldots, U_m$ be m subspaces of W of dimension $m-1$. Define $\mathcal{W}_m(U_1, U_2, \ldots, U_m)$ to be the set of all subspaces X of W of dimension m satisfying*

$$[\wedge^m(\wedge^{m-1}X), (\wedge^{m-1}U_1) \wedge (\wedge^{m-1}U_2) \wedge \cdots \wedge (\wedge^{m-1}U_m)] = 0. \tag{4.2}$$

Then $\mathcal{W}_m(U_1, U_2, \ldots, U_m)$ is a proper subvariety of $\mathcal{W}_m$ if and only if $U_1, \ldots, U_m$ satisfy the dimension m-intersection property.

In other words, the theorem states that there exists an m-dimensional subspace X of W for which (4.2) is not satisfied if and only if $U_1, \ldots, U_m$ satisfy the dimension m-intersection property.

Let X have a basis $x_1, \ldots, x_m$, and for $i = 1, \ldots, m$, let X_i be a subspace of X spanned by $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m$. Then making use of (3.1), we can express the inner product in (4.2) as

$$\begin{aligned} &[(\wedge^{m-1}X_1) \wedge (\wedge^{m-1}X_2) \wedge \cdots \wedge (\wedge^{m-1}X_m), \\ &\quad (\wedge^{m-1}U_1) \wedge (\wedge^{m-1}U_2) \wedge \cdots \wedge (\wedge^{m-1}U_m)] \\ &= \det \left[\ [X_i, U_j] \ \right] \quad (\text{for } i, j = 1, \ldots, m). \end{aligned} \tag{4.3}$$

We will make use of this representation of the inner product in the remaining sections of the paper.

5. Proofs of Theorems 1.1, 1.2, and 2.1. Before applying Theorem 4.1 to compound matrices, we prove the following lemma that may be of interest in its own right.

LEMMA 5.1. *Let $I_1, I_2, \ldots, I_t$ be $t < m$ subsets of $\{1, 2, \ldots, n\}$ each of cardinality $m-1$, and assume that the m-intersection property*

$$|\cap_{i \in P} I_i| \leq m - |P| \tag{5.1}$$

holds for all nonempty subsets P of $\{1, 2, \ldots, t\}$. Then there exist $m-t$ subsets $I_{t+1}, \ldots, I_m$ of $\{1, 2, \ldots, n\}$ of cardinality $m-1$ such that (5.1) holds for all nonempty subsets P of $\{1, 2, \ldots, m\}$.

Proof. It suffices to show that there exists a subset I_{t+1} of $\{1, 2, \ldots, n\}$ of cardinality $m-1$ such that (5.1) holds for all nonempty subsets P of $\{1, 2, \ldots, t+1\}$. If $|\cap_{i \in P} I_i| < m - |P|$ for all subsets P of $\{1, 2, \ldots, t\}$ with $|P| \geq 2$, then we may choose I_{t+1} to be any subset of $\{1, 2, \ldots, n\}$ of cardinality $m-1$ different from $I_1, I_2, \ldots, I_t$.

Hence consider the situation when there exists a subset P with $|P| \geq 2$ that satisfies the m-intersection property (5.1) as an equality. We show that then $\{1, 2, \ldots, t\}$ can be partitioned into maximal subsets that satisfy (5.1) as equalities.

Let P and Q be two *nondisjoint* subsets of $\{1, 2, \ldots, t\}$ satisfying $|\cap_{i \in P} I_i| = m - |P|$ and $|\cap_{i \in Q} I_i| = m - |Q|$, respectively. Write $X \equiv \cap_{i \in P} I_i$ and $Y \equiv \cap_{i \in Q} I_i$. Then applying the identity $|X \cap Y| = |X| + |Y| - |X \cup Y|$ we obtain

$$|\cap_{i \in P \cup Q} I_i| = |\cap_{i \in P} I_i| + |\cap_{i \in Q} I_i| - |(\cap_{i \in P} I_i) \cup (\cap_{i \in Q} I_i)|.$$

Since $\cap_{i \in P} I_i, \cap_{i \in Q} I_i \subseteq \cap_{i \in P \cap Q} I_i$, we see that

$$(\cap_{i \in P} I_i) \cup (\cap_{i \in Q} I_i) \subseteq \cap_{i \in P \cap Q} I_i.$$

Putting it all together, we obtain

$$\begin{aligned} m - |P \cup Q| &\geq |\cap_{i\in P\cup Q} I_i| \\ &= |\cap_{i\in P} I_i| + |\cap_{i\in Q} I_i| - |(\cap_{i\in P} I_i) \cup (\cap_{i\in Q} I_i)| \\ &= m - |P| + m - |Q| - |(\cap_{i\in P} I_i) \cup (\cup_{i\in Q} I_i)| \\ &\geq m - |P| + m - |Q| - |(\cap_{i\in P\cap Q} I_i)| \\ &\geq m - |P| + m - |Q| - (m - |P \cap Q|) \\ &= m - |P \cup Q|. \end{aligned}$$

Therefore

$$|\cap_{i\in P\cup Q} I_i| = m - |P \cup Q|.$$

It follows that there exists a partition $\{P_1, P_2, \ldots, P_\ell\}$ of $\{1, 2, \ldots, t\}$ into $\ell \geq 1$ sets such that (5.1) holds with equality for each P_i and

$$|\cap_{i\in Q} I_i| = m - |Q| \text{ implies that } Q \subseteq P_i \text{ for some } i.$$

We proceed to show how the set I_{t+1} may be chosen in this situation. Let x be any element of $\cap_{i\in P_1} I_1$, and choose I_{t+1} to be any subset of $m-1$ elements of $\{1, 2, \ldots, n\}$ such that

$$I_{t+1} \cap (\cap_{i\in P_1} I_i) = \cap_{i\in P_1} I_i \setminus \{x\}.$$

Since $|I_{t+1}| = m-1$ and by the choice of I_{t+1} we have $|I_{t+1} \cap (\cap_{i\in P_1} I_i)| = m-|P_1|-1$, I_{t+1} contains exactly $|P_1|$ elements not in $\cap_{i\in P_i} I_i$. To prove that (5.1) holds for all nonempty subsets P of $\{1, 2, \ldots, t+1\}$, it suffices to show that for each nonempty subset Q of $\{1, 2, \ldots, t\}$ for which $|\cap_{i\in Q} I_i| = m - |Q|$, we have

$$\cap_{i\in Q} I_i \not\subseteq I_{t+1}. \tag{5.2}$$

Case 1. $Q \subseteq P_1$. Then

$$x \in \cap_{i\in P_1} I_i \subseteq \cap_{i\in Q} I_i \quad \text{and} \quad x \notin I_{t+1}$$

imply that (5.2) holds.

Case 2. $Q \subseteq P_j$ for some $j \neq 1$. Then using (5.1) and the fact that P_1 is maximal with respect to the property that $|\cap_{i\in P_1} I_i| = m - |P_1|$, we obtain

$$q := |(\cap_{i\in P_1} I_i) \cap (\cap_{i\in Q} I_i| = |\cap_{i\in P_1\cup Q} I_i| \leq m - |P_1| - |Q| - 1.$$

Hence

$$|(\cap_{i\in Q} I_i) \setminus (\cap_{i\in P_1} I_i)| = m - |Q| - q \geq |P_1| + 1.$$

Now by construction, I_{t+1} contains exactly $|P_1|$ elements not in $\cap_{i\in P_1} I_i$. Since $\cap_{i\in Q} I_i$ contains at least $|P_1| + 1$ elements not in $\cap_{i\in P_1} I_i$, there exists an element y in $\cap_{i\in Q} I_i$ that is not an element of I_{t+1}. This completes the proof. □

Proof of Theorem 2.1. In §2 we showed that the m-intersection property is a necessary condition for the matrix (2.5) to have full row rank. Now suppose that the m-intersection property holds. It follows from Lemma 5.1 that it suffices to prove that the rank of the matrix (2.5) equals m when $t = m$. Thus assume that

$t = m$, that is, that (2.5) is a square matrix. Since the entries of A are algebraically independent over $\mathbf{Q}$ and since the determinant of the matrix (2.5) is a polynomial in the entries of A with integer coefficients, it suffices to show that this determinant is not identically zero. Let $e_1, e_2, \ldots, e_n$ be the standard basis of $\mathbf{R}^n$ and let U_k denote the subspace spanned by $\{e_i : i \in J_k\}$ $(k = 1, 2, \ldots, t)$. We write the standard basis of U_k as $\{e_1^k, \ldots, e_{m-1}^k\}$. Since $\{J_1, J_2, \ldots, J_m\}$ satisfy the m-intersection property, it follows easily that $\{U_1, U_2, \ldots, U_m\}$ satisfy the dimension m-intersection property. By Theorem 4.1 there exists a subspace X of $\mathbf{R}^n$ of dimension m such that (4.2) does not hold.

Let B be an $m \times n$ matrix whose rows $x_1, \ldots, x_m$ form a basis of X. Now

$$\begin{aligned}
&[X_i, U_k] \\
&= \left(x_1 \wedge \cdots \wedge x_{i-1} \wedge x_{i+1} \wedge \cdots \wedge x_m, e_1^k \wedge \cdots \wedge e_{m-1}^k\right) \\
&= \det \left[\ (x_j, e_\ell^k) \ \right] \quad (\text{for } j = 1, \ldots, i-1, i+1, \ldots, m, \quad \ell = 1, \ldots, m-1) \\
&= C_{m-1}(B)[\bar{i}, J_k].
\end{aligned}$$

Hence from (4.2) and (4.3), we have

$$\begin{aligned}
&[(\wedge^{m-1} X_1) \wedge (\wedge^{m-1} X_2) \wedge \cdots \wedge (\wedge^{m-1} X_m), \\
&\quad (\wedge^{m-1} U_1) \wedge (\wedge^{m-1} U_2) \wedge \cdots \wedge (\wedge^{m-1} U_m)] \\
&= \det \left[\ [X_i, U_j] \ \right] \quad (\text{for } i, j = 1, \ldots, m) \\
&= \det C_{m-1}(B)[:, \{J_1, J_2, \ldots, J_m\}] \neq 0. \qquad \square
\end{aligned}$$

Proofs of Theorems 1.1 *and* 1.2. The proof of Theorem 1.1 follows immediately from Theorem 2.1 and the calculations of §2. The necessity of the $(n-k)$-intersection property for the linear independence of the elementary vectors of the null space of A, $y(I_1), y(I_2), \ldots, y(I_t)$, is an immediate consequence of the calculations of §2.

An argument is needed to derive the converse of Theorem 1.2 from Theorem 2.1, since the assumption that the matrix A is generic does not imply that the matrix B (defined in §2), whose row space is the null space of A, is generic. But we shall overcome this by first choosing a generic B and then defining A.

Assume first only that A is a nondegenerate matrix and the sets $\{\bar{I}_1, \bar{I}_2, \ldots, \bar{I}_t\}$ satisfy the $(n-k)$-intersection property. Since the entries of each elementary vector are polynomials in the entries of A, it follows that the elementary vectors in the null space of A, $y(I_1), y(I_2), \ldots, y(I_t)$, are linearly dependent if and only if the determinantal polynomial vanishes identically for every submatrix of order t of the $t \times n$ matrix Y formed by these elementary vectors. The theorem follows if we can show that there exists at least one nondegenerate $k \times n$ matrix A of rank k for which $y(I_1), y(I_2), \ldots, y(I_t)$ are linearly independent, for then at least one of these determinantal polynomials does not vanish identically. Let B be an $n-k \times n$ generic matrix. Let $x(I_1), x(I_2), \ldots, x(I_t)$ be elementary vectors of the row space of B with supports $I_1, I_2, \ldots, I_t$, respectively. Choose A to be any $k \times n$ matrix of rank k such that $AB^T = O$. Since $BA^T = O$ the arguments in §2 show that A is nondegenerate. Since the $(n-k)$-intersection property holds, we conclude from Theorem 1.1 that $x(I_1), x(I_2), \ldots, x(I_t)$ are linearly independent elementary vectors in the row space of B. We now take the vectors $x(I_1), x(I_2), \ldots, x(I_t)$ as the elementary vectors $y(I_1), y(I_2), \ldots, y(I_t)$ in the null space of A. This completes the proof. $\square$

6. Proof of the main theorem. In this section we give the proof of Theorem 4.1. The following two elementary lemmas used in our proof concern vector

spaces generated by certain operations on subspaces of a vector space; we review these operations now. If V_1 and V_2 are subspaces of a finite dimensional vector space W, then their *union*

$$V_1 \cup V_2 = \{v : v \in V_1\} \cup \{v : v \in V_2\}$$

is generally not a vector space, since it is not necessarily closed under vector addition. The *sum*

$$V_1 + V_2 = \{v_1 + v_2 : v_1 \in V_1, v_2 \in V_2\}$$

and *intersection*

$$V_1 \cap V_2 = \{v : v \in V_1 \cap V_2\}$$

are vector spaces, and it is easy to verify that the sum is the smallest vector space that contains the vectors in $V_1 \cup V_2$.

LEMMA 6.1. *Let k be a positive integer and let $V, V_1, \ldots, V_k$ be subspaces of a finite dimensional vector space W over $\mathbf{R}$. Then $V \subseteq V_1 \cup \cdots \cup V_k$ if and only if $V \subseteq V_i$ for some i.*

Proof. Let $V_i' = V_i \cap V$ for $i = 1, \ldots, k$. Then each V_i' is a subspace of V. If each V_i' is a proper subspace of V, then $V \setminus \cup_{i=1}^k V_i = V \setminus \cup_{i=1}^k V_i'$ is a set of positive Lebesgue measure of dimension $\dim V$. □

LEMMA 6.2. *Let $k \geq 2$ be an integer and let $V_1, \ldots, V_k$ be subspaces of a finite dimensional inner product space W over $\mathbf{R}$. Then*

$$\cap_{i=1}^k V_i = (V_1^\perp + \cdots + V_k^\perp)^\perp.$$

Proof. First suppose that $k = 2$. Then the proof follows by choosing an orthonormal basis B_{12} of $V_1 \cap V_2$, extending to orthonormal bases $B_{12} \cup B_1$ of V_1 and $B_{12} \cup B_2$ of V_2, and then extending to an orthonormal basis $B_{12} \cup B_1 \cup B_2 \cup B$ of W. Then $B \cup B_1 \cup B_2$ is an orthonormal basis of $V_1^\perp + V_2^\perp$, and it follows that B_{12} is an orthonormal basis of $(V_1^\perp + V_2^\perp)^\perp$. We now assume that $k > 2$ and use induction on k. Using the inductive assumption twice, we obtain

$$\begin{aligned} \cap_{i=1}^k V_i = (\cap_{i=1}^{k-1} V_i) \cap V_k &= ((\cap_{i=1}^{k-1} V_i)^\perp + V_k^\perp)^\perp \\ &= ((V_1^\perp + \cdots V_{k-1}^\perp) + V_k^\perp)^\perp. \end{aligned}$$ □

Proof of Theorem 4.1. Let $U_1, U_2, \ldots, U_m$ be m subspaces of W of dimension $m-1$. Then $\mathcal{W}_m(U_1, U_2, \ldots, U_m)$ is clearly a subvariety of $\mathcal{W}_m$. Thus the theorem is only concerned with whether or not it equals $\mathcal{W}_m$.

This proof is technically the most demanding part of the paper, and hence we provide a sketch of our proof technique before we embark on proving the theorem. The necessity of the dimension intersection property is the easier part of the proof. We use dimension-counting arguments to show that certain subspaces occurring in the determinantal representation (4.3) of the inner product (4.2) have nontrivial intersection, leading to a large zero submatrix that makes the determinant zero. Sufficiency is harder, and is proved by induction on m, by showing that when the dimension intersection property is satisfied there exists a subspace X of dimension m, constructed using $U_1^\perp, \ldots, U_m^\perp$, such that (4.2) does not hold.

First assume that the dimension m-intersection property (4.1) does not hold. Without loss of generality, assume that $V \equiv \cap_{i=1}^p U_i$ satisfies

$$\dim V = m - p + 1, \tag{6.1}$$

where p is an integer with $2 \leq p \leq m$. Let X be any subspace of W of dimension m. We have $\dim X^\perp = n - m$, and by (6.1), $\dim V^\perp = n - m + p - 1$; hence there exists a subspace F, contained in both $V^\perp$ and X, of dimension $p-1$. Choose a set $\{x_1, x_2, \ldots, x_{p-1}\}$ of $p-1$ linearly independent vectors spanning F. For $j = 1, 2, \ldots, p$, by Lemma 6.2

$$U_j^\perp \subseteq (U_1^\perp + \cdots + U_p^\perp) = V^\perp.$$

Since F and $U_j^\perp$ are subspaces of $V^\perp$, and

$$\dim F + \dim U_j^\perp = (p-1) + (n-m+1) = n - m + p > \dim V^\perp = n - m + p - 1,$$

we have $F \cap U_j^\perp \neq \{0\}$ for each $j = 1, 2, \ldots, p$. We now extend $x_1, x_2, \ldots, x_{p-1}$ to a basis $x_1, x_2, \ldots, x_m$ of X and let X_i be the subspace of X with basis $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m$ $(i = 1, 2, \ldots, m)$. By the above it follows that

$$X_i \cap U_j^\perp \neq \{0\} \quad (i = p, p+1, \ldots, m; j = 1, 2, \ldots, p),$$

since such subspaces X_i contain F and $F \cap U_j^\perp \neq \{0\}$. Hence by Lemma 3.1

$$[X_i, U_j] = 0 \quad (i = p, p+1, \ldots, m; j = 1, 2, \ldots, p).$$

Therefore the matrix in (4.3) whose (i,j)-entry equals $[X_i, U_j]$ $(i, j = 1, 2, \ldots, m)$ has an $m - p + 1 \times p$ zero submatrix with $(m-p+1)+p = m+1$, and it follows from the Frobenius–König theorem that its determinant equals zero. This implies that (4.2) holds for every subspace X of W of dimension m and hence $\mathcal{W}_m(U_1, U_2, \ldots, U_m) = \mathcal{W}_m$.

Now we prove sufficiency of the dimension intersection property. Assume that $U_1, U_2, \ldots, U_m$ are subspaces of W of dimension $m-1$ satisfying the dimension m-intersection property (4.1); in particular, no two of $U_1, U_2, \ldots, U_m$ are equal. We prove by induction on m that there exists a subspace X of W of dimension m for which (4.2) does not hold.

First we consider the base case $m = 2$. Then U_1 and U_2 are distinct subspaces of W of dimension one and we choose X to be the subspace of dimension two spanned by u_1 and u_2 where u_1 is a basis for U_1 and u_2 is a basis for U_2. Then $u_1 \wedge u_2 \in (\wedge^2(\wedge^1 X)) \cap ((\wedge^1 U_1) \wedge (\wedge^1 U_2))$ and $(u_1 \wedge u_2, u_1 \wedge u_2) \neq 0$. Hence (4.2) does not hold.

Now suppose that $m > 2$. If U is a subspace of W of dimension $m-1$, then we define a subvariety $\mathcal{F}(U)$ of $\mathcal{W}_m$ by

$$\mathcal{F}(U) = \{X : X \in \mathcal{W}_m, \dim X \cap U^\perp \geq 2\}.$$

Let X be a subspace in $\mathcal{W}_m \setminus \mathcal{F}(U)$. Since $\dim X = m$ and $\dim U^\perp = n - m + 1$, by the choice of X, we have $\dim X \cap U^\perp = 1$. Thus the subspace $X^* = X \cap (X \cap U^\perp)^\perp$ is in $\mathcal{W}_{m-1}$. The map

$$\phi_U : \mathcal{W}_m \setminus \mathcal{F}(U) \to \mathcal{W}_{m-1}, \quad \text{where } \phi_U(X) = X^*,$$

is a rational map.

We proceed to construct a subspace X of dimension m, $m-1$ subspaces of X of dimension $m-2$, and $m-1$ subspaces U_i' of dimension $m-2$ to set up the inductive step in the proof.

Since $U_1, U_2, \ldots, U_m$ are distinct subspaces of the same dimension $m-1$, it follows from Lemma 6.1 that

$$U_m^{\perp} \not\subseteq U_1^{\perp} \cup U_2^{\perp} \cup \cdots \cup U_{m-1}^{\perp}.$$

Let x_m be any vector satisfying

$$x_m \in U_m^{\perp} \setminus (U_1^{\perp} \cup \cdots \cup U_{m-1}^{\perp}). \tag{6.2}$$

Let X be a subspace in $\mathcal{W}_m \setminus \mathcal{F}(U_m)$ containing x_m and let $\{x_1, \ldots, x_{m-1}, x_m\}$ be an orthogonal basis of X containing x_m. Let X_i be the subspace of X that is spanned by $\{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m\}$, $(i = 1, 2, \ldots, m)$. We have $X_m = \phi_{U_m}(X)$. Since the vector x_m belongs to $U_m^{\perp}$ and $X_1, \ldots, X_{m-1}$, by Lemma 3.1 we have

$$[X_i, U_m] = 0 \quad (i = 1, 2, \ldots, m-1). \tag{6.3}$$

Furthermore, since $\dim X \cap U_m^{\perp} = 1$,

$$[X_m, U_m] \neq 0. \tag{6.4}$$

By (4.3), (4.2) is not zero if and only if

$$\det \left[\ [X_i, U_j] \ \right] \quad (\text{for } i, j = 1, \ldots, m)$$

is nonzero, and hence by (6.3) and (6.4), if and only if

$$\det \left[\ [X_i, U_j] \ \right] \quad (\text{for } i, j = 1, \ldots, m-1)$$

is nonzero. Thus (4.2) does not hold if and only if

$$\begin{aligned} &[(\wedge^{m-1} X_1) \wedge (\wedge^{m-1} X_2) \wedge \cdots \wedge (\wedge^{m-1} X_{m-1}), \\ &(\wedge^{m-1} U_1) \wedge (\wedge^{m-1} U_2) \wedge \cdots \wedge (\wedge^{m-1} U_{m-1})] \neq 0. \end{aligned} \tag{6.5}$$

We now reduce the dimensions of $U_1, \ldots, U_{m-1}$ by one to apply the inductive assumption.

By (6.2), x_m does not belong to the subspaces $U_1^{\perp}, \ldots, U_{m-1}^{\perp}$, and hence for each $i = 1, 2, \ldots, m-1$, there exists a basis $\{u_1^i, u_2^i, \ldots, u_{m-1}^i\}$ of U_i with

$$(x_m, u_j^i) = 0 \ (j = 1, 2, \ldots, m-2) \quad \text{and} \quad (x_m, u_{m-1}^i) = 1. \tag{6.6}$$

Let subspaces of W be defined by

$$U_i' = U_i \cap \{x_m\}^{\perp} \quad (i = 1, 2, \ldots, m-1).$$

By (6.2) we have

$$\dim U_i' = m-2 \quad (i = 1, 2, \ldots, m-1).$$

We now use the bases of X_i and U_j, Lemma 3.1, and the determinantal formula in (3.1) to compute $[X_i, U_j]$ for $i, j = 1, 2, \ldots, m-1$. Let X_i' be the subspace of X_i with basis $\{x_j : j = 1, \ldots, i-1, i+1, \ldots, m-1\}, (i = 1, 2, \ldots, m-1)$. Using the Laplace expansion of the determinant in (3.1) by the last row (which is the vector $(0, \ldots, 0, 1)$ by (6.6)), we see that each $[X_i, U_j] = m[X_i', U_j']$. Hence (6.5) equals

$$\begin{aligned} &[(\wedge^{m-2} X_1') \wedge (\wedge^{m-2} X_2') \wedge \cdots \wedge (\wedge^{m-2} X_{m-1}'), \\ &(\wedge^{m-2} U_1') \wedge (\wedge^{m-2} U_2') \wedge \cdots \wedge (\wedge^{m-2} U_{m-1}')]. \end{aligned} \tag{6.7}$$

It now follows that (4.2) is not identically zero provided (6.7) is not zero. By the induction hypothesis, (6.7) is not zero provided $U_1', U_2', \ldots, U_{m-1}'$ satisfy the dimension $(m-1)$-intersection property. Our proof is complete if we show that these subspaces satisfy the required dimension intersection property for some choice of x_m.

Assume to the contrary that, for any admissible choice of x_m in $U_m^\perp \setminus (U_1^\perp \cup \cdots \cup U_{m-1}^\perp$, there exists an integer k with $2 \leq k \leq m-1$ and a subset of $\{1, 2, \ldots, m-1\}$ (both depending on x_m) of cardinality k, say the subset $\{1, 2, \ldots, k\}$, such that

$$\dim \cap_{i=1}^k U_i' \geq (m-1) - k + 1 = m - k. \tag{6.8}$$

Since $\cap_{i=1}^k U_i' \subseteq \cap_{i=1}^k U_i$, we have

$$\dim \cap_{i=1}^k U_i \geq m - k,$$

and since $U_1, U_2, \ldots, U_m$ satisfy the dimension m-intersection property, we have

$$\dim \cap_{i=1}^k U_i = m - k. \tag{6.9}$$

Hence there exists a set $Z \subseteq U_m^\perp \setminus (U_1^\perp \cup \cdots \cup U_{m-1}^\perp)$ of positive Lebesgue measure in $U_m^\perp$ such that

$$\cap_{i=1}^k U_i = \cap_{i=1}^k U_i'. \tag{6.10}$$

We now show that (6.10) leads to a contradiction of the dimension m-intersection property (4.1).

If (6.10) holds for all $x_m \in Z$, we claim that

$$U_m^\perp \subseteq U_1^\perp + \cdots + U_k^\perp. \tag{6.11}$$

(Note that now we are considering the *sums* of the vector spaces, and not the *unions* considered in (6.2).) The proof of the claim is also by contradiction. If the claim were not true, then $(U_1^\perp + \cdots + U_k^\perp) \cap U_m^\perp$ is a proper subspace of $U_m^\perp$ and hence we may choose the vector $x_m \in Z$ in (6.2) so that x_m is in $U_m^\perp \setminus (U_1^\perp + \cdots + U_k^\perp)$. Let V be the subspace of W spanned by x_m. Then using the definitions of the subspaces U_i', we have

$$\cap_{i=1}^k U_i' = V^\perp \cap (\cap_{i=1}^k U_i) = (V + U_1^\perp + \cdots + U_k^\perp)^\perp \subset (U_1^\perp + \cdots + U_k^\perp)^\perp = \cap_{i=1}^k U_i,$$

where we have used Lemma 6.2 twice. The containment relation we have obtained contradicts (6.10). We conclude that (6.11) is true whenever (6.10) holds.

Writing (6.11) in the form

$$U_m^\perp \subseteq U_1^\perp + \cdots + U_k^\perp = (\cap_{i=1}^k U_i)^\perp,$$

we find

$$\cap_{i=1}^k U_i \subseteq U_m.$$

Therefore

$$U_m \cap (\cap_{i=1}^k U_i) = \cap_{i=1}^k U_i. \tag{6.12}$$

But now (6.9) and (6.12) contradict the dimension m-intersection property (4.1). This completes the inductive proof of sufficiency and the proof of the theorem. □

7. Coda. Theorem 4.1 implies a sufficient condition for a collection of vectors in the wedge product of a vector space to be linearly independent.

COROLLARY 7.1. *Let W be an inner product space over $\mathbf{R}$ of dimension n and let m be an integer with $2 \leq m \leq n$. If $U_1, U_2, \ldots, U_m$ are m subspaces of W of dimension $m-1$ that satisfy the dimension m-intersection property, then $\wedge^{m-1}U_1$, $\wedge^{m-1}U_2, \ldots, \wedge^{m-1}U_m$ as vectors in $\wedge^{m-1}W$ are linearly independent.*

Proof. Assume that $U_1, U_2, \ldots, U_m$ are subspaces of W of dimension $m-1$ satisfying the dimension m-intersection property. Recall that each $\wedge^{m-1}U_i$ is a subspace of $\wedge^{m-1}W$ of dimension one and thus can be regarded as a nonzero vector of $\wedge^{m-1}W$. It follows from Theorem 4.1 that there exists a choice of subspaces $X_1, X_2, \ldots, X_m$ of W of dimension $m-1$ such that

$$[\wedge^{m-1}X_1 \wedge \wedge^{m-1}X_2 \wedge \cdots \wedge \wedge^{m-1}X_m, \wedge^{m-1}U_1 \wedge \wedge^{m-1}U_2 \wedge \cdots \wedge \wedge^{m-1}U_m] \neq 0.$$

Let $\chi_i = \wedge^{m-1}X_i$ and $\omega_i = \wedge^{m-1}U_i$ for $i = 1, 2, \ldots, m$. It follows from (3.1) and elementary column operations that if $\omega_1, \omega_2, \ldots, \omega_m$ are linearly dependent, then $[\chi_1 \wedge \chi_2 \wedge \cdots \wedge \chi_m, \omega_1 \wedge \omega_2 \wedge \cdots \wedge \omega_m] = 0$. Hence it holds that $\omega_1, \omega_2, \ldots, \omega_m$ are linearly independent. □

We remark that the converse of Corollary 7.1 is not true in general. For example, let $n = 4$ and $m = 3$ and let e_1, e_2, e_3, e_4 be the standard basis of $W = \mathbf{R}^4$. Also let U_1, U_2, and U_3 be the subspaces of W spanned by $\{e_1, e_4\}, \{e_2, e_4\}$, and $\{e_3, e_4\}$, respectively. Then U_1, U_2, U_3 do not satisfy the dimension 3-intersection property, since $U_1 \cap U_2 \cap U_3 \neq \{0\}$. Using the concrete realization of the wedge product, we see that $\wedge^2 U_1, \wedge^2 U_2$, and $\wedge^2 U_3$ are spanned by $e_1 \wedge e_4 = (0,0,1,0,0,0), e_2 \wedge e_4 = (0,0,0,0,1,0)$, and $e_3 \wedge e_4 = (0,0,0,0,0,1)$, respectively, and hence are linearly independent.

However the converse of Corollary 7.1 is true if $m = n$.

COROLLARY 7.2. *Let $U_1, U_2, \ldots, U_t$ be subspaces of $\mathbf{R}^m$ of dimension $m-1$. Then $\wedge^{m-1}U_1, \wedge^{m-1}U_2, \ldots, \wedge^{m-1}U_t$ are linearly independent if and only if $U_1, U_2, \ldots, U_t$ satisfy the dimension m-intersection property.*

Proof. Let $J \subseteq \{1, 2, \ldots, t\}$. Since $U_1, U_2, \ldots, U_t$ are $(m-1)$-dimensional subspaces of an m-dimensional space, the dimension of $\cap_{j \in J} U_j$ is at least $m - |J|$. By Lemma 6.2,

$$(\cap_{j \in J} U_j)^{\perp} = \left(\sum_{j \in J} U_j^{\perp} \right)^{\perp}.$$

Hence $\dim(\cap_{j \in J} U_j) \leq m - |J|$ (and so equals $m - |J|$) if and only if the vectors $\wedge^{m-1}U_1, \wedge^{m-1}U_2, \ldots, \wedge^{m-1}U_t$ are linearly independent. □

The following lemma identifies the support of the elementary vectors of an arbitrary subspace (taken as the row space of a matrix) of $\mathbf{R}^n$.

LEMMA 7.3. *Let A be an $m \times n$ real matrix of* rank m. *Let I be a subset of $\{1, 2, \ldots, n\}$. Then there exists an elementary vector of the row space of A with support I if and only if* (i) *the rank of $A[:, \overline{I}]$ equals $m-1$ and* (ii) *the rank of $A[:, \overline{I} \cup \{j\}]$ equals m for each $j \in I$.*

Proof. First assume that there is an elementary vector $x(I)$ with support I. If the rank of $A[:, \overline{I}]$ equals m, then any linear combination of the rows of A that vanishes on $\overline{I}$ is a trivial linear combination. If the rank of $A[:, \overline{I} \cup \{j\}]$ is less than m for some $j \in I$, then there is a nontrivial linear combination of the rows of A that vanishes on $\overline{I} \cup \{j\}$ and hence $x(I)$ is not an elementary vector. Assertions (i) and (ii) now follow.

Now assume that (i) and (ii) hold. Let K be any subset of $\overline{I}$ of cardinality $m-1$ such that the rank of $A[:,K]$ equals $m-1$. Then with $\overline{I}$ replaced by K, (2.1) defines an elementary vector $x(I)$ in the row space of A with support I. □

We now give a criterion for a set of elementary vectors of a subspace of $\mathbf{R}^n$ (again taken as the row space of a matrix) to be a basis.

THEOREM 7.4. *Let A be an $m \times n$ real matrix of* rank m *and let $x(I_1)$, $x(I_2), \ldots,$ $x(I_m)$ be elementary vectors in the row space W of A. Let U_j be the subspace of $\mathbf{R}^m$ spanned by the columns of $A[:,\overline{I_j}]$ $(j = 1,2,\ldots,m)$. Then $\{x(I_1), x(I_2), \ldots, x(I_m)\}$ is a basis of W if and only if $U_1, U_2, \ldots, U_m$ satisfy the dimension m-intersection property.*

Proof. By Lemma 7.3 each of the subspaces U_j has dimension $m-1$. For each $j = 1,2,\ldots,m$ there exists a vector $y(I_j)$ such that $x(I_j) = y(I_j)A$. The vectors $x(I_1), x(I_2), \ldots, x(I_m)$ are linearly independent if and only if $y(I_1), y(I_2), \ldots, y(I_m)$ are. By Lemma 7.3, there exists $K_j \subseteq \overline{I_j}$ such that $|K_j| = m-1$ and the rank of $A[:,K_j]$ equals $m-1$. We can identify the vector $y(I_j)$ with the vector $\wedge^{m-1}U_j$ (cf. (2.3)). If $y(I_1), y(I_2), \ldots, y(I_m)$ are linearly dependent, then $\wedge^{m-1}U_1 \wedge \cdots \wedge \wedge^{m-1}U_m = 0$ and hence by Theorem 4.1, $U_1, U_2, \ldots, U_m$ do not satisfy the dimension m-intersection property.

Conversely, suppose that $U_1, U_2, \ldots, U_m$ do not satisfy the dimension m-intersection property. Then, as remarked in the proof of Theorem 4.1, we have

$$[\wedge^{m-1}U_1 \wedge \cdots \wedge \wedge^{m-1}U_m, \wedge^{m-1}U_1 \wedge \cdots \wedge \wedge^{m-1}U_m] = 0$$

and hence $\wedge^{m-1}U_1, \ldots, \wedge^{m-1}U_m$ are linearly dependent. □

In the case that A is generic over $\mathbf{Q}$, we have shown that the subspaces $U_1, \ldots, U_m$ satisfy the dimension m-intersection property if and only if the sets $\overline{I_1}, \ldots, \overline{I_m}$ satisfy the m-intersection property. More generally we make the following conjecture.

CONJECTURE. If A is an $m \times n$ matrix whose *nonzero* elements are algebraically independent over $\mathbf{Q}$, then the elementary vectors $x(I_1), x(I_2), \ldots, x(I_m)$ form a basis of the row space of A (that is, by Theorem 7.4, the subspaces $U_1, U_2, \ldots, U_m$ satisfy the dimension m-intersection property) if and only if

$$\text{rank } A[:, \cap_{i \in P} \overline{I_i}] \leq m - |P| \quad (\forall P \subseteq \{1,2,\ldots,m\}, P \neq \emptyset).$$

REFERENCES

[1] T. F. COLEMAN AND A. POTHEN, *The null space problem* I. *Complexity*, SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 527–537.

[2] ———, *The null space problem* II. *Algorithms*, SIAM J. Algebraic Discrete Methods, 8 (1987), pp. 544–563.

[3] D. R. FULKERSON, *Networks, frames, and blocking systems*, Mathematics of the Decision Sciences, Vol. 2, G.B. Dantzig and A.F. Veinott, eds., American Mathematical Society, Providence, RI, 1968, pp. 303–334.

[4] J. R. GILBERT AND M. T. HEATH, *Computing a sparse basis for the null space*, SIAM J. Algebraic Discrete Methods, 8 (1987), pp. 446–459.

[5] A. J. HOFFMAN AND S. T. MCCORMICK, *A fast algorithm for making matrices optimally sparse*, Progress in Combinatorial Optimization, W.R. Pulleyblank, ed., Academic Press, New York, 1984, pp. 185–196.

[6] M. MARCUS, *Finite Dimensional Multilinear Algebra, Part* I, Marcel Dekker, New York, 1973.

[7] ———, *Finite Dimensional Multilinear Algebra, Part* II, Marcel Dekker, New York, 1973.

[8] S. T. MCCORMICK, *A Combinatorial Approach to Some Sparse Matrix Problems*, Ph.D. thesis, Stanford University, Stanford, CA, 1983. (Tech. Report 83-5, Stanford Optimization Lab.)

[9] A. POTHEN, *Sparse Null Bases and Marriage Theorems*, Ph.D. thesis, Cornell University, Ithaca, New York, 1984.

[10] R. T. ROCKAFELLAR, *The elementary vectors of subspaces of R^n*, Combinatorial Mathematics and its Applications, R. C. Bose and T. A. Dowling, eds., University of North Carolina Press, Chapel Hill, NC, 1969, pp. 104–127.

[11] J. M. STERN AND S. A. VAVASIS, *Nested dissection for sparse null space bases*, Tech. Report 90-1173, Computer Science, Cornell University, Ithaca, NY, December 1990.

[12] W. T. TUTTE, *Introduction to the Theory of Matroids*, Elsevier, New York, 1971.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 21–28, January 1995

002

DUALITY AND BLACK BOX INTERPOLATION I: THE ONE VARIABLE NONDEROGATORY CASE *

VAIDYANATH MANI† AND ROBERT E. HARTWIG†

Abstract. The one variable black box interpolation problem is solved for the case of a nonderogatory linear operator on a vector space over a closed field.

Key words. interpolation, black box, nonderogatory

AMS subject classifications. Primary 15A21, 15A15, 41-10, 34-36

1. Introduction. Consider a linear map $\mathbb{A}$: $\mathcal{V} \to \mathcal{V}$ on a possibly infinite dimensional vector space $\mathcal{V}$ over an algebraically closed field $\mathbb{F}$. Suppose that a vector $f \in \mathcal{V}$ admits an annihilating polynomial over $\mathbb{F}$ and thus has a *minimal annihilating polynomial* (m.a.p.) $\eta(\lambda)$ such that $\eta(\mathbb{A})f = 0$. It is clear that if $\eta(\lambda)$ is the minimal polynomial for f, then its degree $\partial\eta = m$ if and only if $f, \mathbb{A}f, \ldots, \mathbb{A}^k f$ are linearly independent for $k = m - 1$ and linearly dependent for $k = m$ as vectors in $\mathcal{V}$. The existence of the minimal polynomial is crucial in our consideration, since in essence it turns our problem into one that is finite dimensional and hence a matrix problem. Suppose further that η has the form

$$\eta(\lambda) = \prod_{k=1}^{s} (\lambda - \lambda_k)^{q_k}, \tag{1.1}$$

where $q_1 + \cdots + q_s = m$ and λ_i are distinct.

As a consequence of Euclid's algorithm we have "duality," i.e., the identity

$$\bigoplus_{k=1}^{s} N[(\mathbb{A} - \lambda_k I)^{q_k}] = N\left[\prod_{k=1}^{s} (\mathbb{A} - \lambda_k)^{q_k}\right] \tag{1.2}$$

is valid. For now let us set

$$G = \bigoplus_{k=1}^{s} G_k.$$

Last, we assume that these generalized eigenspaces $G_k = N[(\mathbb{A} - \lambda_k I)^{q_k}]$ are *finite* dimensional with dimension q_k and bases $\{\phi_k^{(j)}\}, k = 1, 2, \ldots, s, j = 0, \ldots, q_k - 1$. In this case, needless to say, duality gives that

$$\left[\prod_{k=1}^{s} (\mathbb{A} - \lambda_k I)^{q_k}\right] f = 0 \Leftrightarrow f = \sum_{k=1}^{s} \sum_{j=0}^{q_k - 1} c_k^{(j)} \phi_k^{(j)}. \tag{1.3}$$

In other words we assume that the restricted operator $\mathbb{A}^{\sim} = \mathbb{A}|_G$ has one Jordan block per eigenvalue, i.e., $\mathbb{A}^{\sim}$ is *nonderogatory.* The following are important examples where duality holds.

* Received by the editors July 28, 1992; accepted for publication (in revised form) by A. Berman, August 24, 1993.

† North Carolina State University, Raleigh, North Carolina 27695-8205 (hartwig@math.ncsu.edu, vmani@unity.ncsu.edu).

a. Dim $(\mathcal{V})$ is finite, e.g., for matrices where $\mathbb{A} = A$ and $\mathcal{V} = \mathbb{F}^n$.

b. $\mathbb{A} = \mathbb{D} = d/dx$ and $\mathcal{V} = c[a, b]$ (the continuous functions on $[a, b]$) as well as

c. $\mathbb{A} = \mathbb{S}$ and $\mathbb{S}(f(x)) = f(px)(p$ is a prime) and $\mathcal{V} = \mathbb{F}[x]$.

Consider $\mathcal{V}$ as a vector space of functions. In the black box interpolation problem the aim is to expand a given function $f \in \mathcal{V}$ in terms of the eigenfunctions of $\mathbb{A}$ subject to the constraint that one is only allowed to use up to a prescribed number of terms, say at most s. That is, we are given the following:

a. Operator $\mathbb{A}$, its spectrum, and its generalized eigenvectors, possibly infinite in number;

b. Function f;

c. The sparsity of the problem, i.e., either the exact (finite) number of terms in the basis expansion that one is willing to use or, in a weaker version, an upper bound for this number;

d. A black box

which generates a set of output values, say $\mathbb{A}^i f(a_j), i = 0, 1, \ldots$, for one or more input values a_j, which we refer to as nodes. We denote these output values by $f^{(i)}(a_j)$ and refer to them as "derivatives" of f at a_j. Clearly these values will serve as initial conditions for the interpolation problem. From these output values we want to find out the following:

a. which finite subset of generalized eigenspaces $G_k, k = 1, \ldots, s$ must be selected;

b. which basis vectors $\phi_k^{(j)}$ in these eigenspaces must be used;

c. coefficients $c_k^{(j)}$ such that our given f can be expanded as

$$f = \sum_{k=1}^{s} \sum_{j=0}^{q_k - 1} c_k^{(j)} \phi_k^{(j)}. \tag{1.4}$$

The difference here from the usual interpolation problem is that the basis vectors (i.e., the choice of eigenfunctions) as well as the exponents q_k are *unknown* beforehand (see [1], [2]).

Remark 1. In the matrix case, when $f(a) = \underline{a}^T \underline{f}$ or $\langle \underline{a} | \underline{f} \rangle$, the duality relation (1.2) may also be derived from the existence of the Drazin inverse A^D of A. Indeed, if Z_i denote the principal idempotents of A, then $f = Z_1 f + \cdots + Z_s f$ since $Z_1 + \cdots + Z_s = I$. To show that $Z_k f \in N(\mathbb{A} - \lambda_k I)^{q_k}$, we first observe that

$$\eta(A) f = 0 \Rightarrow \eta(A) Z_1 f = \prod_{k=1}^{s} (\mathbb{A} - \lambda_k I)^{q_k} \cdot Z_1 f = 0.$$

Thence $U(\mathbb{A} - \lambda_1 I)^{q_1} Z_1 f = 0$, where

$$U = \prod_{k=2}^{s} (\mathbb{A} - \lambda_k I)^{q_k} Z_1 + (Z_2 + \cdots + Z_s)$$

is invertible with inverse

$$U^{-1} = \prod_{k=2}^{s} [(\mathbb{A} - \lambda_k I)^D]^{q_k} Z_1 + (Z_2 + \cdots + Z_s).$$

Consequently, $(\mathbb{A} - \lambda_1 I)^{q_1} Z_1 f = 0$ and likewise $Z_k f \in G_k$, as desired.

Remark 2. For the case of the derivative operator $\mathbb{D}$, duality is usually proven via the fundamental theorem for initial value problems.

Remark 3. In the case when $\mathbb{A} = \mathbb{S}$ one may show duality directly by using the fact that $\mathcal{V}$ is made up of polynomials.

Remark 4. In the above three cases, the generalized eigenspaces and their bases are as follows.

Case 2. $G_k = \{f; (\mathbb{D} - \lambda_k I)^{q_k} f(t) = 0\}$, which has basis

$$\{t^j e^{\lambda_k t}; j = 0, 1, \ldots, q_k - 1\}. \tag{1.5}$$

Case 3. $G_k = \{f; (\mathbb{S} - p^{\lambda_k} I)^{q_k} f = 0\}$, which has basis

$$\{(\log x)^j x^{\lambda_k}; j = 0, 1, \ldots, q_k - 1\}. \tag{1.6}$$

Case 1.

$$G_k = N\left[\prod_{k=1}^{s} (\mathbb{A} - \lambda_k)^{q_k}\right]$$

which has as a basis the links in the Jordan chain of length q_k, associated with eigenvalue λ_k, i.e.,

$$\{\underline{x}_1, (A - \lambda_k I)\underline{x}_1, \ldots, (A - \lambda_k I)^{q_k - 1}\underline{x}_1\}. \tag{1.7}$$

Remark 5. It should be noted in Remarks 2 and 3 that $\dim G_k$ is in fact equal to q_k, *without* further assumption, while in case 1, the matrix case, we must still *assume* that A is nonderogatory. Indeed, using an integrating factor and induction it follows at once that *every* solution to $(\mathbb{D} - \alpha I)^m y = 0$ must be of the form $\sum_{j=0}^{m-1} c_j t^j e^{\alpha t}$, so that the functions $t^j e^{\alpha t}$ span G_k. Since linear independence is trivial, they form a basis.

Remark 6. The m.a.p. $x = e^t$ converts the basis in Remark 2 into a basis for Remark 3 as illustrated in [4] for the Cauchy equation. This is really a consequence of the isomorphism between the underlying Lie algebras.

Remark 7. Except in the one-dimensional case, the degree m of the minimal polynomial η, *need not be equal to the sparsity* of the vector! For example, if $f = te^t$, then $\eta(\lambda) = (\lambda - 1)^2$ relative to $\mathbb{D}$, yet f is *only* one-sparse. This illustrates that we must be more careful in the higher dimensional cases.

Remark 8. We may think of (1.2) as a local version of the primary decomposition theorem or as an operator version of the Chinese remainder theorem. We stress the fact that equality in (1.2) holds *whenever* we have Euclid's algorithm at our disposal.

Recall that if $\eta_j(\lambda) = \prod_{k\neq j}^{s}(\lambda - \lambda_k)^{q_k}$, then $\gcd(\eta_1(\lambda), \ldots, \eta_s(\lambda)) = 1$, and hence by Euclid's algorithm there exist polynomials $g_i(\lambda)$ such that $1 = g_1\eta_1 + \cdots + g_s\eta_s$. If we now set $Z_i = g_i(\mathbb{A})\eta_i(\mathbb{A})$. Then $I = Z_1 + \cdots + Z_s$ and hence

$$f = Z_1 f + \cdots + Z_s f, \tag{1.8}$$

where $Z_i f = g_i(\mathbb{A})\eta_i(\mathbb{A})f \in G_i$. Moreover it follows that

$$Z_i Z_j f = 0 \quad \text{for } i \neq j, \qquad Z_i^2 f = Z_i f \quad \text{for } i = 1, \ldots, s. \tag{1.9}$$

The operator Z_j is the spectral components of $\mathbb{A}$ associated with λ_j.

Remark 9. In general, we want to select our bases to make our black box interpolation as easy as possible, i.e., with as few conditions as possible.

Consider G_k where $\lambda_k = \alpha$ and $q_k = q$. Since G_k is cyclic, we may either select a cyclic basis

$$u_i\colon \{\phi_o, \mathbb{A}\phi_o, \mathbb{A}^2\phi_o, \ldots, \mathbb{A}^{q-1}\phi_o\} \tag{1.10}$$

where the minimal polynomial for ϕ_o is $(\lambda - a)^q$, a Jordan chain basis

$$v_j\colon \{(\mathbb{A} - \alpha I)^{q-1}\phi_o, (\mathbb{A} - \alpha I)^{q-2}\phi_o, \ldots, (\mathbb{A} - \alpha I)\phi_o, \phi_o\}, \tag{1.11}$$

which starts with the eigenvector $v_o = (\mathbb{A} - \alpha I)^{q-1}\phi_o$, or a reverse Jordan basis

$$w_k\colon \{\phi_o, (\mathbb{A} - \alpha I)\phi_o, \ldots, (\mathbb{A} - \alpha I)^{q-1}\phi_o\}, \tag{1.12}$$

where the leader ϕ_o is a generalized eigenvector of grade q, i.e., $(\mathbb{A} - \alpha I)^q\phi_o = 0 \neq (\mathbb{A} - \alpha I)^{q-1}\phi_o$. It is easily seen that if $U = [u_o, \ldots, u_{q-1}], V = [v_o, \ldots, v_{q-1}]$, and $W = [w_o, w_1, \ldots, w_{q-1}]$, then

$$\mathbb{A}U = UL[(\lambda - \alpha)^q], \quad \mathbb{A}V = VJ_q(\alpha), \quad \text{and} \quad AW = WJ_q(\alpha)^T, \tag{1.13}$$

where $L = L[f(\lambda)] = [\underline{e}_2, \underline{e}_3, \ldots, \underline{e}_n, -\underline{f}^T]$ is the companion matrix of the monic polynomial $f(\lambda) = f_o + f_1\lambda + \cdots + \lambda^n$ with $\underline{f} = [f_o, \ldots, f_{n-1}]^T$ and

$$J_n(\alpha) = \begin{bmatrix} \alpha & 1 & & & 0 \\ & \alpha & 1 & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & 1 \\ 0 & & & & \alpha \end{bmatrix}_{n\times n} = \alpha I_n + N \tag{1.14}$$

is the standard Jordan block. Needless to say, $\mathbb{A}^k U = UL^k$ and $\mathbb{A}^k V = VJ^k$. The bases in (1.10) and (1.11) are further related via the equation

$$U = V\Omega_n(\alpha), \tag{1.15}$$

where the $n \times n$ binomial matrix $\Omega_n(\alpha)$ is given by

$$(\Omega_n(\alpha))_{ij} = \binom{i}{j}\alpha^{i-j}, \qquad i, j = 0, \ldots, n-1. \tag{1.16}$$

It should be noted that the following are true.

(i) Row k contains the binomial coefficients from $(\alpha+1)^k$.
(ii) The entries in Ω_n indeed vanish for $i<j$.
(iii) As a matter of fact, $\Omega(\alpha)$ is the Wronskian of $\{t^j e^{\alpha t}/j!\}$ evaluated at $t=0$.
(iv) $\Omega_n(\alpha)^{-1}=\Omega_n(-\alpha)$.
(v) $\mathbb{A}^k U=\mathbb{A}^k V\Omega(\alpha)=VJ^k\Omega(\alpha)$.

For later use let us denote the submatrix of Ω_n, which is made up of its first k columns by Ω_k^n.

Let us now go through the two stages of the black box problem for a nonderogatory $\mathbb{A}$. We shall see that the case of *one* data node, say value a, allows us to solve the *nonderogatory* black box interpolation problem.

2. The nonderogatory case. Our procedure consists of two phases. In Phase I we use initial conditions $f^{(j)}(a)=\mathbb{A}^j f(a)$ at node a to find operator $\mathbb{L}=\eta(\mathbb{A})$. Then by factoring $\eta(\lambda)$, we find the eigenvalues λ_k and $\mathbb{A}$ and their multiplicities q_k in η. Subsequently, the associated Jordan bases $v_k^{(j)}$ are known.

In Phase II we again use initial conditions to find the coefficients in the basis expansion. That is, we decide which of the basis vectors are actually used to represent f. An interesting aspect of this approach is that Phase II is needed to complete Phase I!

Suppose that we are given f and its derivatives and we want to express f as a linear combination *at most* m of its generalized eigenvectors. This means that $\eta(\mathbb{A})f=0$ for some polynomial $\eta(\lambda)$. Indeed, we may assume that the minimal polynomial of f has exact degree m, say, $\eta(\lambda)=b_o+b_1\lambda+\cdots+\lambda^m$. However, unlike degree m, coefficients b_j are *unknown*. The minimality of m assures that $f,\mathbb{A}f,\ldots,\mathbb{A}^{m-1}f$ are linearly independent as vectors in $\mathcal{V}$. Now $\mathbb{A}^j\eta(\mathbb{A})f=0$ for all $j=0,1,\ldots,m$, which we may evaluate at point a to give $0=\sum_{i=o}^m b_i\mathbb{A}^{i+j}f(a)$. In matrix form this becomes $H\underline{b}=-\underline{h}$ or

$$
(2.1)\qquad
\begin{bmatrix}
f(a) & \mathbb{A}f(a) & \cdots & \mathbb{A}^{m-1}f(a)\\
\mathbb{A}f(a) & \mathbb{A}^2f(a) & \cdots & \mathbb{A}^{m}f(a)\\
\mathbb{A}^2f(a) & \mathbb{A}^3f(a) & \cdots & \mathbb{A}^{m+1}f(a)\\
\vdots & & & \vdots\\
\mathbb{A}^{m-1}f(a) & \mathbb{A}^{m}f(a) & \cdots & \mathbb{A}^{2m-2}f(a)
\end{bmatrix}
\begin{bmatrix} b_o\\ b_1\\ \vdots\\ b_{m-1}\end{bmatrix}
=(-)\begin{bmatrix}\mathbb{A}^m f(a)\\ \vdots\\ \mathbb{A}^{2m-1}f(a)\end{bmatrix}.
$$

We shall show shortly that the invertibility of this Hankel matrix H is really equivalent to the following problem, which actually can be rephrased as the *homogeneous* case of Phase II!

Given $\eta(\mathbb{A})f=0$, what additional conditions of the form $A^i f(a_j)=0$ are needed on f to ensure that f vanishes?

Assuming for now that the Hankel matrix is invertible, we may solve (2.1) to find coefficients b_k and hence $\eta(x)$. Given that $\mathbb{F}$ is closed, we now factor $\eta(x)$ as $\prod_{k=1}^m(x-\lambda_k)^{q_k}$, which shows *which* of the eigenvalues from the point spectrum $\sigma(\mathbb{A})$ have actually been used and what their indices are in the minimal polynomial $\eta(\lambda)$ for f. As a last step, we compute the associated generalized-eigenvector bases for G_k and proceed to Phase II.

Phase II. The only problem remaining is that of finding the m Fourier coefficients. In theory this is a standard procedure analogous to the imposition of initial conditions on the solutions of a differential equation.

To find coefficients $c_k^{(j)}$ in (1.4), we again use the fact that the derivatives $f^{(j)}=$

$\mathbb{A}^j f(a)$ are *all* known. With aid of the Jordan bases, we start by writing (1.4) as

$$f(x) = \sum_{k=1}^{s} \sum_{j=o}^{q_k-1} c_k^{(j)} v_k^{(j)}(x) = \sum_{k=1}^{s} V_k \underline{c}_k = V(x)\underline{c}, \tag{2.2}$$

where V_k is as in (1.13) and $V = [V_1, \ldots, V_s]$. It then follows that for all $j = 0, 1, \ldots,$

$$\mathbb{A}^j f = \sum_{k=1}^{s} \mathbb{A}^i V_k \underline{c}_k = \sum_{k=1}^{s} V_k [J_{q_k}(\lambda_k)]^j \underline{c}_{k_j}. \tag{2.3}$$

For simplicity consider next the case $\lambda_k = \alpha, q = q_k$ and recall that

$$[J_q(\alpha)]^j = \sum_{r=0}^{j} \binom{j}{r} \alpha^r N^{j-r}.$$

Now recall that for a given $q \leq m, \Omega_q^m(\alpha)$ denotes the $m \times q$ submatrix of Ω_m made up of its *first* q columns, i.e., $\Omega_m = [\Omega_q^m(\alpha), *]$. We now come to the crucial identity

$$\begin{bmatrix} V \\ VJ \\ \vdots \\ V(J_q)^{m-1} \end{bmatrix} = \Omega_m(\alpha) \begin{bmatrix} V \\ VN \\ \vdots \\ V(N)^{m-1} \end{bmatrix} = \Omega_q^m(\alpha) \begin{bmatrix} V \\ VN \\ \vdots \\ V(N)^{q-1} \end{bmatrix}, \tag{2.4}$$

in which the last matrix has Toeplitz form

$$T = \begin{bmatrix} v_0 & v_1 & & \cdots & v_{q-1} \\ 0 & v_o & v_1 & \cdots & v_{q-2} \\ \vdots & & & \ddots & \vdots \\ & & & & v_1 \\ 0 & & & & v_o \end{bmatrix}. \tag{2.5}$$

In general, when $\alpha = \lambda_k$ we indicate $q_k \times q_k$ matrix T by $T_k(x)$. Substituting in (2.3) we arrive at

$$\begin{bmatrix} f \\ \mathbb{A}f \\ \vdots \\ \mathbb{A}^{m-1} f \end{bmatrix} = [\Omega_{q_1}^m(\lambda_1); \Omega_{q_2}^m(\lambda_2); \ldots; \Omega_{q_s}^m(\lambda_s)] \begin{bmatrix} T_1 & & & 0 \\ & T_2 & & \\ & & \ddots & \\ 0 & & & T_s \end{bmatrix} \begin{bmatrix} \underline{c}_1 \\ \underline{c}_2 \\ \vdots \\ \underline{c}_s \end{bmatrix} \tag{2.6}$$

or $\underline{f} = \Omega T \underline{c}$. In this example, the matrix

$$\Omega = [\Omega_{q_1}^m(\lambda_1); \Omega_{q_2}^m(\lambda_2); \ldots; \Omega_{q_s}^m(\lambda_s)]$$

is the Wronskian matrix of the functions

$$\{x^j e^{\lambda_k x}/j!\}, j = 0, \ldots, q-1 \quad \text{at} \quad x = 0$$

and hence is *always* invertible!

We note in passing that the transpose of this matrix enters in the proof of the spectral theorem [3].

Evaluating (2.6) at $x = a$, we see that to obtain an invertible matrix $T = \text{diag}\,[T_1, \ldots, T_s]$, we must require that eigenvectors $w_o^{(k)}$ corresponding to the roots of $\eta(\lambda)$ cannot vanish at node a, i.e., $w_o^{(k)}(a) \neq 0$ for all $k = 1, \ldots, s$. We then obtain the unique solution $\underline{c} = T^{-1}\Omega^{-1}\underline{f}$ thus completing the proof of Phase II.

We make the following remarks before continuing with the proof of Phase I.

Remark 1. The above solution shows that the exponential function also plays a dominant role in black box interpolation as seen in the spectral theorem and the theory of functions of a matrix.

Remark 2. We solved the more general black box interpolation problem in which we only gave an upper bound m for the number of terms to be used.

Remark 3. Matrix Ω_n is a special case of the more general block derivative matrix defined as follows.

Let $A(\lambda) = [A_{ij}(\lambda)]$ be a block matrix in which each block $A_{ij}(\lambda)$ is differentiable. Define $n \times n$ block matrix Ω_n by

$$(\Omega_n[A(\lambda)])_{ij} = \binom{i}{j} \mathbb{D}^{i-j}[A(\lambda)]. \tag{2.7}$$

If $A(\lambda) = \exp(\alpha\lambda)$ we recover $\Omega(\alpha)$. It can be shown that

$$\Omega_n[A(\lambda)B(\lambda)] = \Omega_n[A(\lambda)]\Omega_n[B(\lambda)], \tag{2.8}$$

and that if $A(\lambda)B(\lambda) = B(\lambda)A(\lambda)$, then $\Omega_n[A(\lambda)]$ and $\Omega_n[B(\lambda)]$ commute.

Let us now return to the question of the invertibility of Hankel matrix H in (2.1), again under the nonderogatory assumption. Consider the set of m functions $\{f_1, \ldots, f_m\}$ from vector space $\mathcal{V}$ and their associated $m \times m$ Wronskian matrix

$$H(x) = H[f_1(x), \ldots, f_m(x)] = \begin{bmatrix} f_1 & f_2 & \cdots & f_m \\ \mathbb{A}f_1 & \mathbb{A}f_2 & & \mathbb{A}f_m \\ \vdots & & & \\ \mathbb{A}^{m-1}f_1 & \mathbb{A}^{m-1}f_2 & & \mathbb{A}^{m-1}f_m \end{bmatrix}.$$

The key question is how does the invertibility of $H(x)$ relate to the independence of vectors $f_i(x)$. Following the theory of differential equations we have Lemma 1.

LEMMA 1. *If $H(x)$ is invertible for at least one choice of vector x (say at $x = a$ in set $\mathcal{D}$), then $f_1, f_2, \ldots, f_m$ are linearly independent over $\mathcal{D}$.*

Proof. As always let $\sum_{i=1}^m c_i f_i = 0$. Then $\sum_{i=1}^m c_i \mathbb{A}^j f_i = 0$ for all $j = 0, 1, \ldots$, and hence $H(x)\underline{c} = \underline{0}$. At $x = a$, we arrive at $H(a)\underline{c} = 0$, which, because of the invertibility of $H(a)$, yields $\underline{c} = \underline{0}$. In other words, f_j are linearly independent. □

The converse is generally *not* true and requires the additional fact that f_i are solutions to an operator equation.

LEMMA 2. *Suppose that $\mathbb{A}$ is nonderogatory and $\eta(\mathbb{A})f = 0$ where*

$$\eta(\lambda) = \prod_{k=1}^{s} (\lambda - \lambda_k)^{q_k} \quad \textit{and} \quad q_1 + \cdots + q_s = m.$$

Then the initial conditions $\mathbb{A}^j f(a) = 0, j = 0, \ldots, m-1$ force $f = 0$, provided that the eigenvectors $v_o^{(k)}$ of $\mathbb{A}$ corresponding to λ_k do not vanish at node a.

Proof. By duality $\eta(\mathbb{A})f = 0$ gives $f = \sum_{k=1}^{s} V_k \underline{c}_k$ as in Phase II. From (2.6) we see that $\Omega T \underline{c} = \underline{f} = \underline{0}$, and hence $\underline{c} = \underline{0}$ provided that $\mathbb{A}$ is nonderogatory and the eigenvectors satisfy $v_o^{(k)}(a) \neq 0$. □

We now may apply this proof to give the desired converse.

LEMMA 3. *If $f_1, \ldots, f_m$ are linearly independent (on $\mathcal{D}$) and $\eta(\mathbb{A})f_i = 0$ then $H(x)$ is invertible for all x (in $\mathcal{D}$).*

Proof. If this is not true, let $H(a)$ be singular. Then there exists $\underline{c} \neq \underline{0}$ so that $H(a)\underline{c} = \underline{0}$. With scalars c_i, we next define the function $g = c_1 f_1 + \cdots + c_m f_m$. Then $\eta(\mathbb{A})g = 0$, while $H(a)\underline{c} = \underline{0}$ ensures that $\mathbb{A}^j g(a) = 0$ for $j = 0, \ldots, m-1$. Invoking Lemma 2 we then may conclude that $g = 0$, which in turn implies that $f_1, f_2, \ldots, f_m$ are linearly dependent, a contradiction. □

It should be clear that Lemma 2 is the "kingpin" in the whole interpolation story as well as in the question of solvability! Returning to Phase I and (2.1), we may therefore conclude that if $\eta(\lambda)$ is the minimal polynomial for f, then the functions $\{\mathbb{A}^j f\}, j = 0, \ldots, m-1$ are linearly independent solutions to $\eta(\mathbb{A})f = 0$ and hence by Lemma 3 the Hankel matrix $H = [\mathbb{A}^{i+j} f(a)]$ is indeed invertible for all a.

We remark in closing that, since the differential operator $\mathbb{D}$ is nonderogatory, the above approach gives a constructive proof to the existence of a unique solution to an nth order linear differential equation $\eta(\mathbb{D})f = 0$, with constant coefficients, *without* using the fundamental theorem for initial value problems! This illustrates the fact that it is an algebraic result and not an analytic result.

Acknowledgment. The authors thank Dr. Michael Singer for posing this problem and several stimulating discussions.

REFERENCES

[1] M. BEN-OR AND P. TIWARI, *A deterministic algorithmic for sparse multivariate polynomial interpolation*, Proc. 20th ACM STOC, 1989, pp. 301–309.

[2] D. Y. GRIGORIEV, M. KARPINSKI, AND M. SINGER, *The interpolation problem for k-sparse sums of eigenfunctions of operators*, Adv. Appl. Math., 12 (1991), pp. 76–81.

[3] R. E. HARTWIG, *Applications of the Wronskian and the Gram matrices of $\{t^j e^{\lambda_k t}\}$*, Linear Algebra Appl. 43 (1982), pp. 229–241.

[4] A. L. RABENSTEIN, *Introduction to Ordinary Differential Equations*, 2nd ed., Academic Press, New York, 1972, p. 88.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 29–39, January 1995

003

EXISTENCE AND UNIQUENESS OF OPTIMAL MATRIX SCALINGS*

V. BALAKRISHNAN† AND S. BOYD‡

Abstract. The problem of finding a diagonal similarity scaling to minimize the scaled singular value of a matrix arises frequently in robustness analysis of control systems. It is shown here that the set of optimal diagonal scalings is nonempty and bounded if and only if the matrix that is being scaled is irreducible. For an irreducible matrix, a sufficient condition is derived for the uniqueness of the optimal scaling.

Key words. diagonal similarity scalings, scaled singular value minimization, irreducible matrices

AMS subject classifications. 65F35, 15A60, 15A12, 47A55

Notation. $\mathbf{R}$ ($\mathbf{C}$) denotes the set of real (complex) numbers. $\mathbf{R}_+$ stands for the set of positive real numbers. For $z \in \mathbf{C}$, $\operatorname{Re} z$ is the real part of z. The set of $m \times n$ matrices with real (complex) entries is denoted $\mathbf{R}^{m\times n}$ ($\mathbf{C}^{m\times n}$). I stands for the identity matrix with size determined from context. For a matrix $P \in \mathbf{C}^{m\times n}$, P^T stands for the transpose and P^* stands for the complex conjugate of P^T. $\|P\|$ is the spectral norm (maximum singular value) of P given by the square root of the maximum eigenvalue of P^*P. (For a vector $v \in \mathbf{C}^n$, $\|v\|$ is just the Euclidean norm.) For $P \in \mathbf{C}^{n\times n}$, $\operatorname{Tr} P$ stands for the trace, that is, the sum of the diagonal entries of P.

1. Introduction.

Given a complex matrix $M \in \mathbf{C}^{n\times n}$ and a nonsingular diagonal matrix $D \in \mathbf{C}^{n\times n}$, the *similarity-scaled singular value* of M corresponding to scaling D is defined as

$$f(M, D) = \left\|DMD^{-1}\right\|.$$

The optimal diagonal scaling problem is to minimize $f(M, D)$ over all diagonal nonsingular matrices D:

$$(1)\quad f_{\min}(M) = \inf\left\{\left\|DMD^{-1}\right\| \;\middle|\; D \in \mathbf{C}^{n\times n},\ D \text{ is diagonal and nonsingular}\right\}.$$

We refer to $f_{\min}(M)$ as the *optimally scaled singular value* of M.

Problem (1) arises in the robustness analysis of control systems with structured uncertainties. For further details, see [11] and [4]. Much research has focused on the related problem of finding optimal (with various criteria for optimality) diagonal preconditioners for use in iterative algorithms; see, for example, [5] and [7].

Reformulation as a convex optimization problem. We note that $f(M, |D|) = f(M, D)$; we also observe that $f(M, D)$ is homogeneous of degree zero in D, that is, $f(M, \alpha D) = f(M, D)$ for all nonzero $\alpha \in \mathbf{C}$. Therefore, we may rewrite (1) as

$$(2)\qquad f_{\min}(M) = \inf\left\{\left\|e^D M e^{-D}\right\| \;\middle|\; D \in \mathbf{R}^{n\times n},\ D \text{ is diagonal}, \operatorname{Tr} D = 0\right\}.$$

* Received by the editors August 10, 1992; accepted for publication (in revised form) by L. Kaufman, September 30, 1993. This research was supported in part by Air Force Office of Scientific Research under contract F49620-92-J-0013.

† School of Electrical Engineering, Purdue University, West Lafayette, Indiana 47907 (ragu@ecn.purdue.edu).

‡ Information Systems Laboratory, Department of Electrical Engineering, Stanford University, Stanford, California 94305 (boyd@isl.stanford.edu).

The reason for rewriting (1) as (2) is that $\|e^D M e^{-D}\|$ is a *convex* function of D—this fact will prove important in the sequel—while $\|DMD^{-1}\|$ is not [12], [13]. For convenience, we let

$$\mathcal{D} = \{D \mid D \in \mathbf{R}^{n\times n},\ D \text{ is diagonal, Tr } D = 0\}\ .$$

In this paper, we do not concern ourselves with the solution of (2). We instead investigate the set of *minimizers* for (2), that is, the set of *optimal* scalings $\mathcal{D}_{\text{opt}}$ defined by

$$\mathcal{D}_{\text{opt}} \triangleq \{D \mid D \in \mathcal{D}, \|e^D M e^{-D}\| = f_{\min}(M)\}\ . \tag{3}$$

In the process, we provide a sufficient condition for $\mathcal{D}_{\text{opt}}$ to be nonempty (which means the infimum in (2) is *achieved*) and a sufficient condition for $\mathcal{D}_{\text{opt}}$ to be a singleton (which means that there is a *unique* optimal scaling).

2. Boundedness of $\mathcal{D}_{\text{opt}}$. We start with a few definitions.

DEFINITION 1. A *permutation* matrix P is a real, orthogonal $n \times n$ matrix (i.e., $PP^T = P^T P = I$) with entries that are either one or zero. We let $\mathcal{P}$ denote the set of $n \times n$ permutation matrices.

DEFINITION 2. A complex matrix M is said to be *reducible* if there exists some $P \in \mathcal{P}$ such that PMP^T is block upper triangular, that is,

$$PMP^T = \begin{bmatrix} M_{11} & M_{12} \\ 0 & M_{22} \end{bmatrix},$$

where M_{11}, M_{22} are square matrices of appropriate sizes [6], [1]. A matrix that is not reducible is termed *irreducible*.

Remark. For any permutation matrix P,

$$\|e^D M e^{-D}\| = \|Pe^D P^T PMP^T Pe^{-D}P^T\|\ .$$

Note that $Pe^D P^T$ is diagonal and corresponds to just a reordering of the diagonal entries of e^D. Therefore, as far as the scaling problem is concerned, if a matrix M is reducible, we may assume without loss of generality that

$$M = \begin{bmatrix} M_{11} & M_{12} \\ 0 & M_{22} \end{bmatrix},$$

bearing in mind that a reordering of the entries of the scaling D might be necessary. In the sequel, the phrase "within a permutation" refers to such a reordering of the entries of D and the corresponding permutation similarity transformation on M.

Let $\mathcal{D}_\gamma$ denote the *sublevel* set

$$\{D \in \mathcal{D} \mid \|e^D M e^{-D}\| < \gamma\}\ .$$

The following theorem relates the irreducibility of M to the boundedness of the sublevel sets.

THEOREM 2.1. *For any $\gamma > f_{\min}(M)$, the sublevel set $\mathcal{D}_\gamma$ is bounded if and only if M is irreducible.*

Proof. We first note the following lemma.

LEMMA 2.2. *It holds that*

$$\left\| \begin{bmatrix} M_{11} & M_{12} \\ 0 & M_{22} \end{bmatrix} \right\| \geq \left\| \begin{bmatrix} M_{11} & 0 \\ 0 & M_{22} \end{bmatrix} \right\|,$$

where M_{11}, M_{12}, and M_{22} are matrices of appropriate sizes.

Proof. The proof is left to the reader.

We first assume that M is reducible. Then to within a permutation,

$$M = \begin{bmatrix} M_{11} & M_{12} \\ 0 & M_{22} \end{bmatrix},$$

where $M_{11} \in \mathbf{C}^{r\times r}$ with $r < n$.

Then given any $\gamma > f_{\min}(M)$ and $D \in \mathcal{D}_\gamma$ (note that $\mathcal{D}_\gamma$ is nonempty), partition D conformally with the block upper triangular structure of M above as

$$D = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}.$$

Consider now a sequence of scaling matrices $D^{(i)}$ of the form

$$D^{(i)} = \begin{bmatrix} D_1 - i\,(n-r) & 0 \\ 0 & D_2 + i\,r \end{bmatrix}, \qquad i = 1, 2, \ldots.$$

(Note that $D^{(i)} \in \mathcal{D}$ for $i = 1, 2, \ldots$.)

For such a sequence of scalings, $\left\| e^{D^{(i)}} M e^{-D^{(i)}} \right\|$ converges to

$$\max\left(\left\| e^{D_1} M_{11} e^{-D_1} \right\|, \left\| e^{D_2} M_{22} e^{-D_2} \right\| \right),$$

which is less than or equal to

$$\left\| e^{D^{(i)}} M e^{-D^{(i)}} \right\|$$

for every i, from Lemma 2.2. Thus for every $\gamma > f_{\min}(M)$, the set $\mathcal{D}_\gamma$ is unbounded.

To prove the converse, let us assume that for some $\gamma > f_{\min}(M)$, $\mathcal{D}_\gamma$ is not bounded. Then there is a sequence of scalings $D^{(i)}$ in $\mathcal{D}_\gamma$ with some of the elements of the diagonal scaling matrix $D^{(i)}$ with absolute value tending to infinity. Then, there exists a subsequence $D^{(n_i)}$, which can be partitioned to within a permutation as

$$D^{(n_i)} = \begin{bmatrix} D_{1,n_i} & 0 \\ 0 & D_{2,n_i} \end{bmatrix},$$

where every element of D_{1,n_i} diverges to $-\infty$ with i, while every element of D_{2,n_i} is bounded below. (In fact, at least one of the elements of D_{2,n_i} must diverge to ∞, but we will not use this fact.)

Thus the maximum singular value of

$$M = \begin{bmatrix} e^{D_{1,n_i}} M_{11} e^{-D_{1,n_i}} & e^{D_{1,n_i}} M_{12} e^{-D_{2,n_i}} \\ e^{D_{2,n_i}} M_{21} e^{-D_{1,n_i}} & e^{D_{2,n_i}} M_{22} e^{-D_{2,n_i}} \end{bmatrix},$$

remains bounded with every element of D_{1,n_i} diverging to $-\infty$ while the elements of D_{2,n_i} are bounded below. This immediately means that $M_{21} = 0$, which shows that M must be reducible. □

COROLLARY 2.3. *$\mathcal{D}_{\text{opt}}$ is nonempty and bounded if M is irreducible.*

Proof. If M is irreducible, the sublevel set $\mathcal{D}_\gamma$ is bounded for every $\gamma > f_{\min}(M)$; since $\left\| e^D M e^{-D} \right\|$ is a continuous function of D over $\mathcal{D}_\gamma$, the infimum in (2) is achieved.

Thus $\mathcal{D}_{\text{opt}}$ is nonempty if M is irreducible. Boundedness of $\mathcal{D}_{\text{opt}}$ follows from an argument similar to the one in the proof of Theorem 2.1. □

We note that this sufficient condition for the existence of optimal scalings can also be found in [3, Prop. 4].

Remark. Thus, irreducibility of M is a sufficient condition for the existence of optimal matrix scalings. If M is reducible, two cases are possible: $\mathcal{D}_{\text{opt}}$ may be empty or it may be nonempty and unbounded. The following examples illustrate this.

Example 1 ($\mathcal{D}_{\text{opt}}$ empty).

$$M = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

It is shown in the Appendix that $\mathcal{D}_{\text{opt}}$ is empty. The optimally scaled singular value is the limit of the sequence of scaled singular values corresponding to scalings $D(d)$ with $d \downarrow -\infty$:

$$D(d) = \begin{bmatrix} d & 0 & 0 \\ 0 & d & 0 \\ 0 & 0 & -2d \end{bmatrix}.$$

Example 2 ($\mathcal{D}_{\text{opt}}$ nonempty and unbounded).

$$M = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}.$$

It is shown in the Appendix that

$$\mathcal{D}_{\text{opt}} = \left\{ D \;\middle|\; D = \begin{bmatrix} d & 0 & 0 \\ 0 & d & 0 \\ 0 & 0 & -2d \end{bmatrix},\ d \in (-\infty, \log(3/2)/6] \right\}.$$

3. $\mathcal{D}_{\text{opt}}$ for irreducible matrices. We next derive a sufficient condition for $\mathcal{D}_{\text{opt}}$ to be a singleton.

We first state without proof a condition for optimality of a scaling D.

THEOREM 3.1. *Suppose the maximum singular value of $e^D M e^{-D}$ is* isolated, *i.e., of unit multiplicity. Then D is an optimal scaling for Problem* 2 *if and only if there exist vectors u and v, with $\|u\| = \|v\| = 1$, such that*

$$\begin{aligned} e^D M e^{-D}\, v &= f_{\min}(M)\, u \\ e^{-D} M^* e^D\, u &= f_{\min}(M)\, v, \end{aligned} \qquad \text{and} \qquad \left|u^{(i)}\right| = \left|v^{(i)}\right|,\ i = 1, 2, \ldots, n,$$

where $u^{(i)}$ and $v^{(i)}$, $i = 1, 2, \ldots, n$ are the components of u and v, respectively.

Theorem 3.1, which is a "magnitude-matching" condition on the components of the left and right singular vectors of the scaled matrix, follows immediately from simple gradient calculations (see, for example, [9]).

We also need the following theorem about the analyticity properties of the singular values of a complex matrix that depends on a real parameter (see [2], [10], [8]).

THEOREM 3.2. *Let $A(x)$ be a (complex) $m \times n$ matrix, the entries of which are analytic functions of a real parameter x. There are real analytic functions $f_i : \mathbf{R} \to \mathbf{R}, i = 1, \ldots, \min(m,n)$ such that, for all $x \in \mathbf{R}$,*

$$\{\sigma_i(A(x)), i = 1, \ldots, \min(m,n)\} = \{|f_i(x)|, i = 1, \ldots, \min(m,n)\}, \tag{4}$$

where $\sigma_i(A(x))$ stands for the ith singular value of $A(x)$. (Thus, the f_i's are the unordered and unsigned singular value functions of $A(x)$.)

For convenience, we let $\gamma = f_{\min}(M)$. With D being an optimal scaling, suppose that (i) γ is the isolated maximum singular value of $e^D M e^{-D}$ and (ii) the left and right singular vectors of $e^D M e^{-D}$ (i.e., u and v in Theorem 3.1) belong to the same *coordinate subspace*, i.e., a subspace of the form $\bigcup_{i \in \mathbf{I}} \operatorname{span}\{e_i\}$, where $\mathbf{I}$ is a proper subset of the set of indices $\{1, \ldots, n\}$ and $\{e_i,\ i = 1, \ldots, n\}$ are coordinate vectors (i.e., unit vectors of $\mathbf{R}^n$ in the standard basis). We will show that this means that $\mathcal{D}_{\text{opt}}$ is not a singleton.

First note that to within a permutation, we have

$$u = \begin{bmatrix} u_1 \\ 0 \end{bmatrix}, \qquad v = \begin{bmatrix} v_1 \\ 0 \end{bmatrix}, \tag{5}$$

where $u_1, v_1 \in \mathbf{C}^r$ with $1 \le r < n$; we then partition $e^D M e^{-D}$ as

$$e^D M e^{-D} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix},$$

where $M_{11} \in \mathbf{C}^{r \times r}$. Of course, $\left|u_1^{(i)}\right| = \left|v_1^{(i)}\right|$, $i = 1, \ldots, r$, and γ is the optimally scaled maximum singular value of M_{11}. Now, with

$$D(\lambda) = \begin{bmatrix} \lambda(n-r) I_1 & 0 \\ 0 & -\lambda r I_2 \end{bmatrix} + D,$$

where I_1 is the $r \times r$ identity matrix, consider

$$e^{D(\lambda)} M e^{-D(\lambda)} = \begin{bmatrix} M_{11} & e^{\lambda n} M_{12} \\ e^{-\lambda n} M_{21} & M_{22} \end{bmatrix}.$$

For every $\lambda \in \mathbf{R}$, γ is a singular value of $e^{D(\lambda)} M e^{-D(\lambda)}$, with u and v in (5) being the corresponding left and right singular vectors. Moreover, every entry of $e^{D(\lambda)} M e^{-D(\lambda)}$ is an analytic function of λ. Then, using Theorem 3.2 and the assumption that the maximum singular value of $e^D M e^{-D}$ is isolated, we conclude that the maximum singular value of $e^{D(\lambda)} M e^{-D(\lambda)}$ is isolated, and hence a real analytic function of λ for $\lambda \in [-\epsilon, \epsilon]$, where $\epsilon > 0$ is sufficiently small. It follows immediately that for $\lambda \in [-\epsilon, \epsilon]$, γ is the maximum singular value of $e^{D(\lambda)} M e^{-D(\lambda)}$. In other words, $D(\lambda)$ is also an optimal scaling for M, for $\lambda \in [-\epsilon, \epsilon]$.

Conversely, let us assume that $\mathcal{D}_{\text{opt}}$ is not a singleton, so that there exist $D_1, D_2 \in \mathcal{D}_{\text{opt}}$, with $D_1 \ne D_2$. Then, from the convexity of $\mathcal{D}_{\text{opt}}$, $D(\lambda) = \lambda D_1 + (1-\lambda) D_2 \in \mathcal{D}_{\text{opt}}$ for every $\lambda \in [0,1]$. Moreover, let us assume that γ is the isolated maximum singular value of $e^{D(\lambda)} M e^{-D(\lambda)}$ for $\lambda \in [0,1]$.

Since $D_1 \ne D_2$, to within a permutation,

$$D_1 - D_2 = \begin{bmatrix} d_1 I_1 & 0 & \cdots & 0 \\ 0 & d_2 I_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_p I_p \end{bmatrix},$$

where $p > 1$, $d_1 > d_2 > \cdots > d_p$, and $I_1, I_2, \ldots, I_p$ are identity matrices of sizes $n_1, n_2, \ldots, n_p$ respectively. Of course, $\sum_{i=1}^{p} n_i = n$, and $\sum_{i=1}^{p} n_i d_i = 0$. Note that every entry of $e^{D(\lambda)} M e^{-D(\lambda)}$ is an analytic function of λ, more specifically equal to a ratio of polynomials of (the components of) $z = \left[e^{\lambda d_1} \; e^{\lambda d_2} \cdots e^{\lambda d_p}\right]$. Then, using Theorem 3.2, we conclude that since γ is the maximum singular value of $e^{D(\lambda)} M e^{-D(\lambda)}$ for $\lambda \in [0,1]$, it must be a singular value of $e^{D(\lambda)} M e^{-D(\lambda)}$ for *all* $\lambda \in \mathbf{R}$.

Next, let $u(\lambda)$ and $v(\lambda)$ be the left and right singular vectors of $e^{D(\lambda)} M e^{-D(\lambda)}$ corresponding to the singular value γ, so that

$$\begin{aligned} e^{D(\lambda)} M \, e^{-D(\lambda)} \, v(\lambda) &= \gamma \, u(\lambda), \\ e^{-D(\lambda)} M^* \, e^{D(\lambda)} \, u(\lambda) &= \gamma \, v(\lambda), \end{aligned} \tag{6}$$

with $\|u(\lambda)\| = \|v(\lambda)\| = 1$. Then, by a direct calculation, $u(\lambda)$ and $v(\lambda)$ can be chosen as analytic functions of λ whose every entry can be expressed as a ratio of a polynomial of z and the square root of a polynomial of z. Therefore, the limits, as $\lambda \to \pm\infty$, of $u(\lambda)$ and $v(\lambda)$ exist. Next, from Theorem 3.1 we have $\left|u^{(i)}(\lambda)\right| = \left|v^{(i)}(\lambda)\right|$ for $i = 1, 2, \ldots, n$ and $\lambda \in [0,1]$, and therefore $\left|u^{(i)}(\lambda)\right| = \left|v^{(i)}(\lambda)\right|$ for $i = 1, 2, \ldots, n$ and for *all* $\lambda \in \mathbf{R}$.

Partitioning $e^{D_2} M e^{-D_2}$, $u(\lambda)$ and $v(\lambda)$ as

$$e^{D_2} M e^{-D_2} = \begin{bmatrix} M_{11} & M_{12} & \ldots & M_{1p} \\ M_{21} & M_{22} & \ldots & M_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ M_{p1} & M_{p2} & \ldots & M_{pp} \end{bmatrix}, \quad u = \begin{bmatrix} u_1(\lambda) \\ u_2(\lambda) \\ \vdots \\ u_p(\lambda) \end{bmatrix}, \quad v = \begin{bmatrix} v_1(\lambda) \\ v_2(\lambda) \\ \vdots \\ v_p(\lambda) \end{bmatrix},$$

where $M_{ii} \in \mathbf{C}^{n_i \times n_i}$, and $u_i(\lambda)$ and $v_i(\lambda) \in \mathbf{C}^{n_i}$ for $i = 1, 2, \ldots, p$, we now show that γ is the optimally scaled maximum singular value of M_{11} or M_{22} or $\ldots$ M_{pp}.

Consider the following equation, taken from (6).

$$e^{-\lambda d_1} M_{11} v_1(\lambda) + e^{-\lambda d_2} M_{12} v_2(\lambda) + \cdots + e^{-\lambda d_p} M_{1p} v_p(\lambda) = \gamma e^{-\lambda d_1} u_1(\lambda).$$

Letting $\lambda \to -\infty$ in the above equation, we get

$$M_{11} v_1(-\infty) = \gamma u_1(-\infty).$$

Since $v_1(-\infty)^* v_1(-\infty) = u_1(-\infty)^* u_1(-\infty)$ (this follows from $\left|u^{(i)}(\lambda)\right| = \left|v^{(i)}(\lambda)\right|$ for $i = 1, 2, \ldots, n$ and for $\lambda \in \mathbf{R}$), we conclude that either γ is the optimally scaled maximum singular value of M_{11} or $u_1(-\infty) = v_1(-\infty) = 0$. Continuing similarly, it follows that γ is the optimally scaled maximum singular value of M_{ii}, for some $i = 1, \ldots, p$. (Recall our assumption that γ is the isolated maximum singular value of $e^{D(\lambda)} M e^{-D(\lambda)}$ for $\lambda \in [0,1]$, so that only one of $M_{11}, \ldots, M_{pp}$ can have a maximum singular value of γ.)

Remark. Suppose $u_1(-\infty) \neq 0 \neq v_1(-\infty)$. Then, by replacing λ by $\lambda + \eta$ (where $\eta \in \mathbf{R}$ is fixed) in the preceding argument, we may show that $[u_1(-\infty)^* \; 0 \cdots 0]^*$ and $[v_1(-\infty)^* \; 0 \cdots 0]^*$ are left and right singular vectors of $e^{D(\eta)} M e^{-D(\eta)}$ corresponding to a singular value γ for *every* $\eta \in \mathbf{R}$, where $D(\eta) = \eta D_1 + (1 - \eta) D_2$.

Remark. If the entries of $D_1 - D_2$ are distinct, then there exist left and right singular vectors of $e^{D_2} M e^{-D_2}$ corresponding to the maximum singular value that both equal the same coordinate vector.[1]

[1] We thank Reviewer 1 for drawing our attention to this remark.

In summary, we have shown that there exist two different optimal scalings D_1 and D_2, with the optimally scaled maximum singular value being isolated for all $D(\lambda) = D_2 + \lambda(D_1 - D_2)$, $\lambda \in [0,1]$, if and only if there exist left and right singular vectors of $e^{D_2} M e^{-D_2}$ (indeed, of $e^{D(\lambda)} M e^{-D(\lambda)}$, $\lambda \in [0,1]$) corresponding to the isolated maximum singular value, belonging to the same coordinate subspace.

We thus arrive at the following sufficient condition for the optimal scaling to be unique.

THEOREM 3.3. *For an irreducible matrix M, let D be an optimal scaling, and let the maximum singular value of $e^D M e^{-D}$ be isolated. Then D is the unique optimal scaling if and only if there exists no pair of vectors u and v, with $\|u\| = \|v\| = 1$ satisfying*

$$\begin{aligned} e^D M e^{-D} v &= \gamma u, \\ e^{-D} M^* e^D u &= \gamma v \end{aligned} \tag{7}$$

that belong to the same coordinate subspace.

Remark. With D being an optimal scaling, if the maximum singular value of $e^D M e^{-D}$ is not isolated, then there always exist v and u with $\|u\| = \|v\| = 1$, satisfying (7) and belonging to the same subspace. In this case, the optimal scaling may or may not be unique as the following two examples illustrate.

Example 3 ($\mathcal{D}_{\text{opt}}$ is a singleton).

$$M = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & 1 \\ -1 & 1 & 1 \end{bmatrix}.$$

It is shown in the Appendix that the unique optimal scaling is zero, i.e., the "identity" scaling, though $[1/\sqrt{2} \;\; 1/\sqrt{2} \;\; 0]^T$ is both a left and right singular vector corresponding to the maximum singular value of two. Note that the maximum singular value at the optimal scaling is not isolated.

Example 4 ($\mathcal{D}_{\text{opt}}$ is not a singleton).

$$M = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & 1 \\ -1 & 1 & \dfrac{1}{\sqrt{2}} \end{bmatrix}.$$

It is shown in the Appendix that $\mathcal{D}_{\text{opt}}$ is given by

$$\mathcal{D}_{\text{opt}} = \left\{ D \;\middle|\; D = \begin{bmatrix} d & 0 & 0 \\ 0 & d & 0 \\ 0 & 0 & -2d \end{bmatrix}, \; d \in [-d_* \, , d_*] \right\},$$

where

$$d_* = \left(\frac{1}{6}\right) \log \frac{9 + \sqrt{17}}{8}.$$

For every $D \in \mathcal{D}_{\text{opt}}$, $[1/\sqrt{2} \;\; 1/\sqrt{2} \;\; 0]^T$ is both a left and right singular vector of $e^D M e^{-D}$ corresponding to the maximum singular value of two. Note that the maximum singular value at the optimal scaling

$$\begin{bmatrix} d_* & 0 & 0 \\ 0 & d_* & 0 \\ 0 & 0 & -2d_* \end{bmatrix}$$

is not isolated, as with Example 3.

4. Conclusion. We have derived sufficient conditions for existence and uniqueness of optimal diagonal similarity scalings for scaled singular value minimization. These conditions can be extended to the other structured scaling problems such as block diagonal similarity scaling.

Appendix. More on the examples. *Example* 1. Let d_1, d_2, and d_3 be the diagonal entries of D, with $d_1 + d_2 + d_3 = 0$. Then,

$$M = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad e^D M e^{-D} = \begin{bmatrix} 1 & e^{d_1-d_2} & e^{d_1-d_3} \\ e^{d_2-d_1} & 1 & e^{d_2-d_3} \\ 0 & 0 & 1 \end{bmatrix}.$$

We observe that if $d_1 \neq d_2$, then $\|e^D M e^{-D}\| > 2$, since the maximum singular value of the principal 2×2 block exceeds 2. With $d_1 = d_2 = d$, $\|e^D M e^{-D}\| > 2$ once again, since

$$\left(e^D M e^{-D}\right)^* e^D M e^{-D} = \begin{bmatrix} 2 & 2 & 2e^{3d} \\ 2 & 2 & 2e^{3d} \\ 2e^{3d} & 2e^{3d} & 1 + 2e^{6d} \end{bmatrix}$$

is a matrix with positive entries, and therefore its spectral radius (the maximum magnitude of its eigenvalues) is strictly greater than four, which is the spectral radius of its principal 2×2 block (see, for example, [1]). Therefore, it follows that $\|e^D M e^{-D}\| > 2$ for every scaling D.

Finally, we note that with $d_1 = d_2 = d$, as $d \to -\infty$, $\|e^D M e^{-D}\| \to 2$.

A plot of the singular values of $e^D M e^{-D}$ as a function of d is shown in Fig. 1.

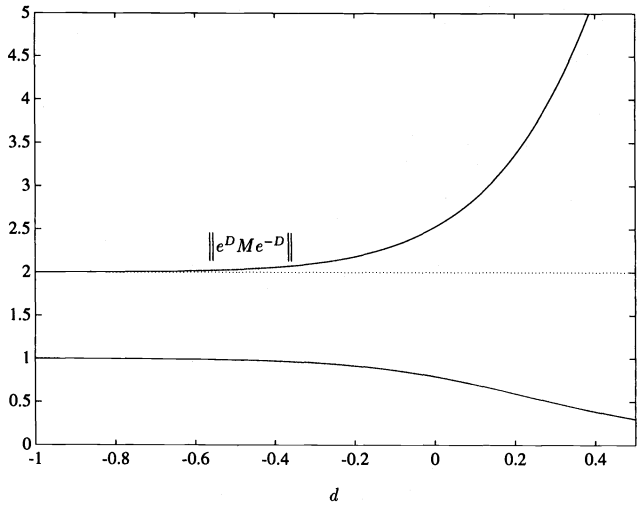


FIG. 1. *Example* 1.

Example 2. We have

$$M = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad e^D M e^{-D} = \begin{bmatrix} 1 & e^{d_1-d_2} & e^{d_1-d_3} \\ e^{d_2-d_1} & 1 & -e^{d_2-d_3} \\ 0 & 0 & 1 \end{bmatrix},$$

with $d_1 + d_2 + d_3 = 0$.

Once again, if $d_1 \neq d_2$, then $\|e^D M e^{-D}\| > 2$. However, in contrast with Example 1, with $d_1 = d_2 = d$, $\|e^D M e^{-D}\|$ is only greater than *or equal to* two. Since

$$\left(e^D M e^{-D}\right)^* e^D M e^{-D} = \begin{bmatrix} 2 & 2 & 0 \\ 2 & 2 & 0 \\ 0 & 0 & 1+2e^{6d} \end{bmatrix},$$

the singular values of $e^D M e^{-D}$ are $\sqrt{1+2e^{6d}}$, 2, and 0. Therefore if $d \leq d_* = \log(3/2)/6$, $\|e^D M e^{-D}\| = 2$.

A plot of the singular values of of $e^D M e^{-D}$ as a function of d is shown in Fig. 2.

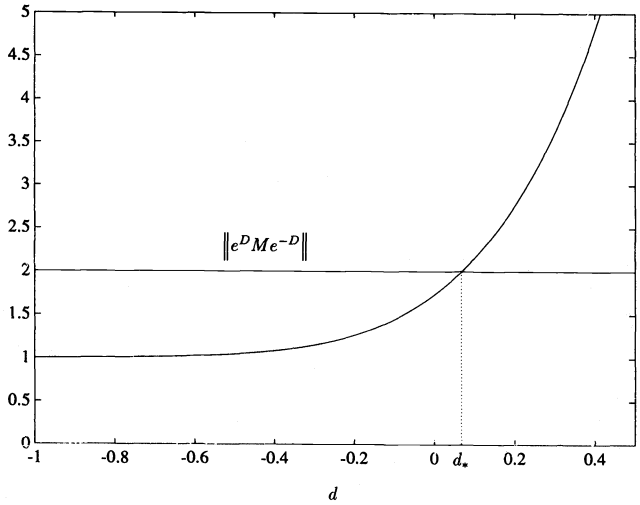


FIG. 2. *Example* 2.

Example 3. We have

$$M = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & 1 \\ -1 & 1 & 1 \end{bmatrix} \quad \text{and} \quad e^D M e^{-D} = \begin{bmatrix} 1 & e^{d_1-d_2} & -e^{d_1-d_3} \\ e^{d_2-d_1} & 1 & e^{d_2-d_3} \\ -e^{d_3-d_1} & e^{d_3-d_2} & 1 \end{bmatrix},$$

with $d_1 + d_2 + d_3 = 0$.

Once again, if $d_1 \neq d_2$, then $\|e^D M e^{-D}\| > 2$. With $d_1 = d_2 = d$, consider

$$\left(e^D M e^{-D}\right)^* e^D M e^{-D} = \begin{bmatrix} 2+e^{-6d} & 2-e^{-6d} & -e^{-3d} \\ 2-e^{-6d} & 2+e^{-6d} & e^{-3d} \\ -e^{-3d} & e^{-3d} & 1+2e^{6d} \end{bmatrix}.$$

The eigenvalues of this matrix are

$$4, \quad \frac{1}{2}\left((1+2e^{6d}+2e^{-6d}) \pm \sqrt{(1+2e^{6d}+2e^{-6d})^2 - 16}\right).$$

Therefore the maximum singular value of $e^D M e^{-D}$ exceeds two if $d \neq 0$, and equals two if $d = 0$. In other words, the unique optimal scaling is zero, i.e., the "identity" scaling. Note that the maximum singular value at the optimal scaling is not isolated.

A plot of the singular values of of $e^D M e^{-D}$ as a function of d is shown in Fig. 3.

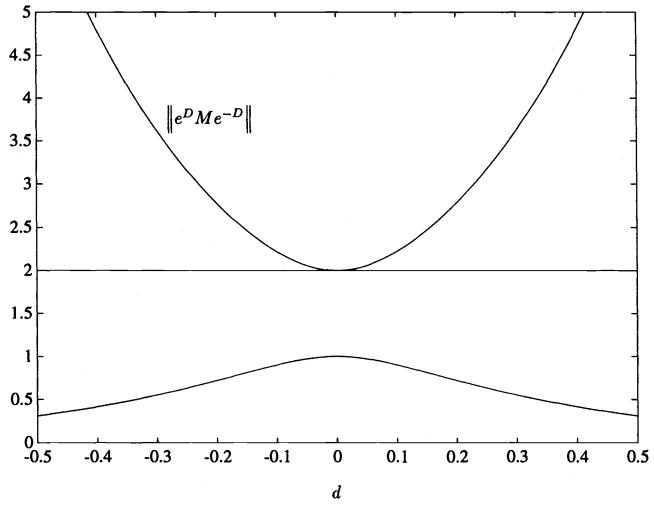


FIG. 3. *Example* 3.

Example 4. We have

$$M = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & 1 \\ -1 & 1 & \dfrac{1}{\sqrt{2}} \end{bmatrix} \quad \text{and} \quad e^D M e^{-D} = \begin{bmatrix} 1 & e^{d_1-d_2} & -e^{d_1-d_3} \\ e^{d_2-d_1} & 1 & e^{d_2-d_3} \\ -e^{d_3-d_1} & e^{d_3-d_2} & \dfrac{1}{\sqrt{2}} \end{bmatrix},$$

with $d_1 + d_2 + d_3 = 0$.

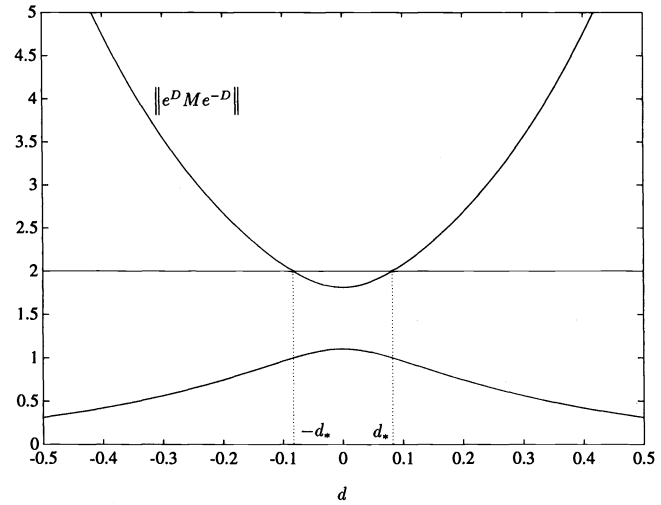


FIG. 4. *Example* 4.

Once again, if $d_1 \neq d_2$, then $\|e^D M e^{-D}\| > 2$. With $d_1 = d_2 = d$, consider

$$\left(e^D M e^{-D}\right)^* e^D M e^{-D} = \begin{bmatrix} 2 + e^{-6d} & 2 - e^{-6d} & -(1/\sqrt{2})e^{-3d} \\ 2 - e^{-6d} & 2 + e^{-6d} & (1/\sqrt{2})e^{-3d} \\ -(1/\sqrt{2})e^{-3d} & (1/\sqrt{2})e^{-3d} & 1/2 + 2e^{6d} \end{bmatrix}.$$

The eigenvalues of this matrix are

$$4, \quad \frac{1}{2}\left((1/2 + 2e^{6d} + 2e^{-6d}) \pm \sqrt{(1/2 + 2e^{6d} + 2e^{-6d})^2 - 16}\right).$$

From this, it follows that the maximum singular value of $e^D M e^{-D}$ equals two if $d \in [-d_* , d_*]$, where

$$d_* = (1/6) \log \frac{9 + \sqrt{17}}{8}.$$

Note that the maximum singular value of $e^D M e^{-D}$ is isolated for $d \in (-d_* , d_*)$.

A plot of the singular values of of $e^D M e^{-D}$ as a function of d is shown in Fig. 4.

REFERENCES

[1] A. BERMAN AND R. J. PLEMMONS, *Nonnegative matrices in the mathematical sciences*, Comput. Sci. Appl. Math., Academic Press, New York, 1979.

[2] S. BOYD AND V. BALAKRISHNAN, *A regularity result for the singular values of a transfer matrix and a quadratically convergent algorithm for computing its* $\mathbf{L}_\infty$*-norm*, Systems Control Let., 15 (1990), pp. 1–7.

[3] M. K. FAN AND A. L. TITS, m*-form numerical range and the computation of the structured singular value*, IEEE Trans. Automat. Control, 33 (1988), pp. 284–289.

[4] M. K. H. FAN, A. L. TITS, AND J. C. DOYLE, *Robustness in the presence of mixed parametric uncertainty and unmodeled dynamics*, IEEE Trans. Automat. Control, 36 (1991), pp. 25–38.

[5] G. E. FORSYTHE AND E. G. STRAUS, *On best conditioned matrices*, Proc. Amer. Math. Soc., 6 (1955), pp. 340–345.

[6] F. R. GANTMACHER, *The Theory of Matrices*, Vol. 2, Chelsea, New York, 1959.

[7] A. GREENBAUM AND G. H. RODRIGUE, *Optimal preconditioners of a given sparsity pattern*, BIT, 29 (1989), pp. 610–634.

[8] T. KATO, *A Short Introduction to Perturbation Theory for Linear Operators*, Springer-Verlag, New York, Berlin, 1982.

[9] N. M. KHRAISHI AND A. EMAMI-NAEINI, *A characterization of optimal scaling for structured singular value computation*, Systems Control Lett., 15 (1990), pp. 105–109.

[10] B. D. MOOR AND S. BOYD, *Analytic properties of singular values and vectors*, Tech. Report ESAT-SISTA Report 1989-28, Department of Electrical Engineering, Katholieke Universiteit Leuven, Belgium, December 1989.

[11] M. G. SAFONOV, *Stability margins of diagonally perturbed multivariable feedback systems*, IEE Proc., 129-D (1982), pp. 251–256.

[12] R. SEZGINER AND M. OVERTON, *The largest singular value of* $e^X A_0 e^{-X}$ *is convex on convex sets of commuting matrices*, IEEE Trans. Automat. Control, 35 (1990), pp. 229–230.

[13] N.-K. TSING, *Convexity of the largest singular value of* $e^D M e^{-D}$*: a convexity lemma*, IEEE Trans. Automat. Control, 35 (1990), pp. 748–749.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 40–57, January 1995

004

ON THE STABILITY OF THE BAREISS AND RELATED TOEPLITZ FACTORIZATION ALGORITHMS*

A. W. BOJANCZYK†, R. P. BRENT‡, F. R. DE HOOG§, AND D. R. SWEET¶

Abstract. This paper contains a numerical stability analysis of factorization algorithms for computing the Cholesky decomposition of symmetric positive definite matrices of displacement rank 2. The algorithms in the class can be expressed as sequences of *elementary downdating* steps. The stability of the factorization algorithms follows directly from the numerical properties of algorithms for realizing elementary downdating operations. It is shown that the Bareiss algorithm for factorizing a symmetric positive definite Toeplitz matrix is in the class and hence the Bareiss algorithm is stable. Some numerical experiments that compare behavior of the Bareiss algorithm and the Levinson algorithm are presented. These experiments indicate that generally (when the reflection coefficients are not all of the same sign) the Levinson algorithm can give much larger residuals than the Bareiss algorithm.

Key words. Toeplitz matrices, Bareiss algorithm, Levinson algorithm, numerical stability

AMS subject classifications. 65F05, 65G05, 47B35, 65F30

1. Introduction.

We consider the numerical stability of algorithms for solving a linear system

$$Tx = b, \tag{1.1}$$

where T is an $n \times n$ positive definite Toeplitz matrix and b is an $n \times 1$ vector. We assume that the system is solved in floating point arithmetic with relative precision ϵ by first computing the Cholesky factor of T. Hence the emphasis of the paper is on factorization algorithms for the matrix T.

Roundoff error analyses of Toeplitz systems solvers have been given by Cybenko [10] and Sweet [22]. Cybenko showed that the Levinson–Durbin algorithm produces a residual which, under the condition that all reflection coefficients are positive, is of comparable size to that produced by the well-behaved Cholesky method. He hypothesised that the same is true even if the reflection coefficients are not all positive. If correct, this would indicate that numerical quality of the Levinson–Durbin algorithm is comparable to that of the Cholesky method.

In his Ph.D. thesis [22], Sweet presented a roundoff error analysis of a variant of the Bareiss algorithm [2] and concluded that the algorithm is numerically stable (in the sense specified in §7). In this paper we strengthen and generalize these early results on the stability of the Bareiss algorithm. In particular, our approach via elementary downdating greatly simplifies roundoff error analysis and makes it applicable to a larger-than-Toeplitz class of matrices.

After introducing the notation and the concept of *elementary downdating* in §§2 and 3, in §4 we derive matrix factorization algorithms as a sequence of elementary

* Received by the editors October 21,1991; accepted for publication (in revised form) by J. R. Bunch, September 21, 1993.

† School of Electrical Engineering, Cornell University, Ithaca, New York 14853-5401 (adamb@toeplitz.ee.cornell.edu).

‡ Computer Sciences Laboratory, Australian National University, Canberra, ACT 0200, Australia (rpb@cslab.anu.edu.au).

§ Division of Mathematics and Statistics, Commonwealth Scientific and Industrial Research Organization, Canberra, ACT 2601, Australia, (frank@dmscanb.cbr.dms.csiro.au).

¶ Electronics Research Laboratory, Defense Science and Technology Organization, Salisbury, SA 5108, Australia, (dougs@ewd.dsto.gov.au).

downdating operations (see also [4]). In §5 we present a first order analysis by bounding the first term in an asymptotic expansion for the error in powers of ϵ. By analyzing the propagation of first order error in the sequence of downdatings that define the algorithms, we obtain bounds on the perturbations of the factors in the decompositions. We show that the computed upper triangular factor $\tilde{U}$ of a positive definite Toeplitz matrix T satisfies

$$T = \tilde{U}^T\tilde{U} + \Delta T\,, \quad ||\Delta T|| \leq c(n)\epsilon||T||\,,$$

where $c(n)$ is a low order polynomial in n and is independent of the condition number of T. Many of the results of §§2–5 were first reported in [5], which also contains some results on the stability of Levinson's algorithm.

In §6 we discuss the connection with the Bareiss algorithm and conclude that the Bareiss algorithm is stable for the class of symmetric positive definite matrices. Finally, in §7 we report some interesting numerical examples that contrast the behaviour of the Bareiss algorithm with that of the Levinson algorithm. We show numerically that, in cases where the reflection coefficients are not all of the same sign, the Levinson algorithm can give much larger residuals than the Bareiss or Cholesky algorithms.

2. Notation. Unless it is clear from the context, all vectors are real and of dimension n. Likewise, all matrices are real and their default dimension is $n \times n$. If $\mathbf{a} \in \Re^n$, $\|\mathbf{a}\|$ denotes the usual Euclidean norm, and if $T \in \Re^{n\times n}$, $\|T\|$ denotes the induced matrix norm:

$$\|T\| = \max_{\|\mathbf{a}\|=1} \|T\mathbf{a}\|\,.$$

Our primary interest is in a symmetric positive definite Toeplitz matrix T whose i,jth entry is

$$t_{ij} = t_{|i-j|}.$$

We denote by $\mathbf{e}_k$, $k = 1,\ldots,n$, the unit vector whose kth element is 1 and whose other elements are 0. We use the following special matrices:

$$Z \equiv \sum_{i=1}^{n-1} e_{i+1}e_i^T = \begin{pmatrix} 0 & & \cdots & \cdots & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & \ddots & & \vdots & \vdots \\ \vdots & & \ddots & 0 & 0 \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix},$$

$$J \equiv \sum_{i=1}^{n} e_{n-i+1}e_i^T = \begin{pmatrix} 0 & \cdots & \cdots & 0 & 1 \\ \vdots & & \cdot & 1 & 0 \\ \vdots & \cdot & \cdot & \cdot & \vdots \\ 0 & 1 & \cdot & & \vdots \\ 1 & 0 & \cdots & \cdots & 0 \end{pmatrix}.$$

The matrix Z is known as a *shift-down* matrix. We also make use of powers of the matrix Z, for which we introduce the following notation:

$$Z_k = \begin{cases} I & \text{if } k = 0, \\ Z^k & \text{if } k > 0. \end{cases}$$

The antidiagonal matrix J is called a *reversal* matrix, because the effect of applying J to a vector is to reverse the order of components of the vector:

$$J \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_n \\ x_{n-1} \\ \vdots \\ x_1 \end{bmatrix}.$$

The *hyperbolic rotation* matrix $H(\theta) \in \Re^{2\times 2}$ is defined by

$$H(\theta) = \frac{1}{\cos\theta} \begin{bmatrix} 1 & -\sin\theta \\ -\sin\theta & 1 \end{bmatrix}. \tag{2.1}$$

The matrix $H(\theta)$ satisfies the relation

$$H(\theta) \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} H(\theta) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

and it has eigenvalues $\lambda_1(\theta)$, $\lambda_2(\theta)$ given by

$$\lambda_1(\theta) = \lambda_2^{-1}(\theta) = \sec\theta - \tan\theta. \tag{2.2}$$

For a given pair of real numbers a and b with $|a| > |b|$, there exists a hyperbolic rotation matrix $H(\theta)$ such that

$$H(\theta) \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sqrt{a^2 - b^2} \\ 0 \end{bmatrix}. \tag{2.3}$$

The angle of rotation θ is determined by

$$\sin\theta = \frac{b}{a}, \qquad \cos\theta = \frac{\sqrt{a^2 - b^2}}{a}. \tag{2.4}$$

3. Elementary downdating. In this section we introduce the concept of elementary downdating. The elementary downdating problem is a special case of a more general downdating problem that arises in Cholesky factorization of a positive definite difference of two outer product matrices [1], [6], [7], [12]. In §4, factorization algorithms are derived in terms of a sequence of downdating steps. The numerical properties of the algorithms are then related to the properties of the sequence of elementary downdating steps.

Let $\mathbf{u}_k$, $\mathbf{v}_k \in \Re^n$ have the following form:

$$\begin{array}{lccccccccc} & & & & & & k & & \\ & & & & & & \downarrow & & \\ \mathbf{u}_k^T = [0 & \ldots & 0 & \times & \times & \times & \ldots & \times\,], \\ \mathbf{v}_k^T = [0 & \ldots & 0 & 0 & \times & \times & \ldots & \times\,], \\ & & & & & & \uparrow & & \\ & & & & & & k+1 & & \end{array}$$

that is

$$\mathbf{e}_j^T \mathbf{u}_k = 0,\ j < k, \quad \text{and} \quad \mathbf{e}_j^T \mathbf{v}_k = 0,\ j \le k.$$

Applying the shift-down matrix Z to $\mathbf{u}_k$, we have

$$\begin{array}{lcccccccc} & & & & & & & k+1 & \\ & & & & & & & \downarrow & \\ \mathbf{u}_k^T Z^T = [0 & \ldots & 0 & 0 & \times & \times & \ldots & \times] , \\ \mathbf{v}_k^T = [0 & \ldots & 0 & 0 & \times & \times & \ldots & \times] . \\ & & & & & & & \uparrow & \\ & & & & & & & k+1 & \end{array}$$

Suppose that we wish to find $\mathbf{u}_{k+1}$, $\mathbf{v}_{k+1} \in \Re^n$ to satisfy

$$\mathbf{u}_{k+1}\mathbf{u}_{k+1}^T - \mathbf{v}_{k+1}\mathbf{v}_{k+1}^T = Z\mathbf{u}_k\mathbf{u}_k^T Z^T - \mathbf{v}_k\mathbf{v}_k^T, \tag{3.1}$$

where

$$\begin{array}{lcccccccc} & & & & & & & k+1 & \\ & & & & & & & \downarrow & \\ \mathbf{u}_{k+1}^T = [0 & \ldots & 0 & 0 & \times & \times & \ldots & \times] , \\ \mathbf{v}_{k+1}^T = [0 & \ldots & 0 & 0 & 0 & \times & \ldots & \times] , \\ & & & & & & & \uparrow & \\ & & & & & & & k+2 & \end{array}$$

that is

$$\mathbf{e}_j^T \mathbf{u}_{k+1} = 0, \; j < k+1, \quad \text{and} \quad \mathbf{e}_j^T \mathbf{v}_{k+1} = 0, \; j \leq k+1. \tag{3.2}$$

We refer to the problem of finding $\mathbf{u}_{k+1}$ and $\mathbf{v}_{k+1}$ to satisfy (3.1), given $\mathbf{u}_k$ and $\mathbf{v}_k$, as the *elementary downdating* problem. It can be rewritten as follows:

$$[\mathbf{u}_{k+1}\ \mathbf{v}_{k+1}] \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \mathbf{u}_{k+1}^T \\ \mathbf{v}_{k+1}^T \end{bmatrix} = [Z\mathbf{u}_k\ \mathbf{v}_k] \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \mathbf{u}_k^T Z^T \\ \mathbf{v}_k^T \end{bmatrix} .$$

From (2.1), (2.3), and (2.4), it is clear that the vectors $\mathbf{u}_{k+1}$ and $\mathbf{v}_{k+1}$ can be found by using a hyperbolic rotation $H(\theta_k)$ defined by the following relations:

$$\sin\theta_k = \mathbf{e}_{k+1}^T \mathbf{v}_k / \mathbf{e}_k^T \mathbf{u}_k , \tag{3.3a}$$

$$\cos\theta_k = \sqrt{1 - \sin^2\theta_k} , \tag{3.3b}$$

and

$$\begin{bmatrix} \mathbf{u}_{k+1}^T \\ \mathbf{v}_{k+1}^T \end{bmatrix} = H(\theta_k) \begin{bmatrix} \mathbf{u}_k^T Z^T \\ \mathbf{v}_k^T \end{bmatrix} . \tag{3.4}$$

The elementary downdating problem has a unique solution (up to sign changes) if

$$|\mathbf{e}_k^T \mathbf{u}_k| > |\mathbf{e}_{k+1}^T \mathbf{v}_k| .$$

The calculation of $\mathbf{u}_{k+1}$, $\mathbf{v}_{k+1}$ via (3.4) can be performed in the obvious manner. Following common usage, algorithms that perform downdating in this manner are referred to as *hyperbolic* downdating algorithms.

Some computational advantages may be obtained by rewriting (3.1) as follows:

$$[\mathbf{u}_{k+1}\ \mathbf{v}_k] \begin{bmatrix} \mathbf{u}_{k+1}^T \\ \mathbf{v}_k^T \end{bmatrix} = [Z\mathbf{u}_k\ \mathbf{v}_{k+1}] \begin{bmatrix} \mathbf{u}_k^T Z^T \\ \mathbf{v}_{k+1}^T \end{bmatrix} .$$

Consider now an orthogonal rotation matrix $G(\theta_k)$,

$$G(\theta_k) = \begin{bmatrix} \cos\theta_k & \sin\theta_k \\ -\sin\theta_k & \cos\theta_k \end{bmatrix},$$

where $\cos\theta_k$ and $\sin\theta_k$ are defined by (3.3b) and (3.3a), respectively. Then it is easy to check that

$$G(\theta_k) \begin{bmatrix} \mathbf{u}_{k+1}^T \\ \mathbf{v}_k^T \end{bmatrix} = \begin{bmatrix} \mathbf{u}_k^T Z^T \\ \mathbf{v}_{k+1}^T \end{bmatrix}, \tag{3.5}$$

or, equivalently,

$$\begin{bmatrix} \mathbf{u}_{k+1}^T \\ \mathbf{v}_k^T \end{bmatrix} = G(\theta_k)^T \begin{bmatrix} \mathbf{u}_k^T Z^T \\ \mathbf{v}_{k+1}^T \end{bmatrix}. \tag{3.6}$$

Thus, we may rewrite (3.6) as

$$\mathbf{v}_{k+1} = (\mathbf{v}_k - \sin\theta_k Z\mathbf{u}_k)/\cos\theta_k, \tag{3.7a}$$

$$\mathbf{u}_{k+1} = -\sin\theta_k \mathbf{v}_{k+1} + \cos\theta_k Z\mathbf{u}_k. \tag{3.7b}$$

Note that (3.7a) is the same as the second component of (3.4). However, (3.7b) differs from the first component of (3.4) as it uses $\mathbf{v}_{k+1}$ in place of $\mathbf{v}_k$ to define $\mathbf{u}_{k+1}$. It is possible to construct an alternative algorithm by using the first component of (3.5) to define $\mathbf{u}_{k+1}$. This leads to the following formulas:

$$\mathbf{u}_{k+1} = (Z\mathbf{u}_k - \sin\theta_k \mathbf{v}_k)/\cos\theta_k, \tag{3.8a}$$

$$\mathbf{v}_{k+1} = -\sin\theta_k \mathbf{u}_{k+1} + \cos\theta_k \mathbf{v}_k. \tag{3.8b}$$

We call algorithms based on (3.7a)–(3.7b) or (3.8a)–(3.8b) *mixed* elementary downdating algorithms. The reason for considering mixed algorithms is that they have superior stability properties to hyperbolic algorithms in the following sense.

Let $\tilde{\mathbf{u}}_k$, $\tilde{\mathbf{v}}_k$ be the values of $\mathbf{u}_k$, $\mathbf{v}_k$ that are computed in floating point arithmetic with relative machine precision ϵ. The computed values $\tilde{\mathbf{u}}_k$, $\tilde{\mathbf{v}}_k$ satisfy a perturbed version of (3.1), that is,

$$\tilde{\mathbf{u}}_{k+1}\tilde{\mathbf{u}}_{k+1}^T - \tilde{\mathbf{v}}_{k+1}\tilde{\mathbf{v}}_{k+1}^T = Z\tilde{\mathbf{u}}_k\tilde{\mathbf{u}}_k^T Z^T - \tilde{\mathbf{v}}_k\tilde{\mathbf{v}}_k^T + \epsilon G_k + O(\epsilon^2), \tag{3.9}$$

where the second order term $O(\epsilon^2)$ should be understood as a matrix whose elements are bounded by a constant multiple of $\epsilon^2||G_k||$. The norm of the perturbation G_k depends on the precise specification of the algorithm used. It can be shown [6] that the term G_k satisfies

$$\|G_k\| \leq c_m \left(\|Z\mathbf{u}_k\|^2 + \|\mathbf{v}_k\|^2 + \|\mathbf{u}_{k+1}\|^2 + \|\mathbf{v}_{k+1}\|^2\right) \tag{3.10}$$

when a mixed downdating strategy is used (here c_m is a positive constant). When hyperbolic downdating is used, the term G_k satisfies

$$\|G_k\| \leq c_h \|H(\theta_k)\| \left(\|Z\mathbf{u}_k\| + \|\mathbf{v}_k\|\right) \left(\|\mathbf{u}_{k+1}\| + \|\mathbf{v}_{k+1}\|\right), \tag{3.11}$$

where c_h is a positive constant [6]. (The constants c_m and c_h are dependent on implementation details, but are of order unity and independent of n.) Note the

presence of the multiplier $\|H(\theta_k)\|$ in the bound (3.11) but not in (3.10). In view of (2.2), $\|H(\theta_k)\|$ could be large. The significance of the multiplier $\|H(\theta_k)\|$ depends on the context in which the downdating arises. We consider the implications of the bounds (3.10) and (3.11) in §5 after we make a connection between downdating and the factorization of Toeplitz matrices.

It is easily seen that a single step of the hyperbolic or mixed downdating algorithm requires $4(n-k)+O(1)$ multiplications. A substantial increase in efficiency can be achieved by considering the following modified downdating problem. Given $\alpha_k, \beta_k \in \Re$ and $\mathbf{w}_k, \mathbf{x}_k \in \Re^n$ that satisfy

$$\mathbf{e}_j^T \mathbf{w}_k = 0\ ,\ j < k \quad \text{and} \quad \mathbf{e}_j^T \mathbf{x}_k = 0\ ,\ j \le k,$$

find α_{k+1}, β_{k+1} and $\mathbf{w}_{k+1}, \mathbf{x}_{k+1} \in \Re^n$ that satisfy

$$\alpha_{k+1}^2 \mathbf{w}_{k+1}\mathbf{w}_{k+1}^T - \beta_{k+1}^2 \mathbf{x}_{k+1}\mathbf{x}_{k+1}^T = \alpha_k^2 Z\mathbf{w}_k\mathbf{w}_k^T Z^T - \beta_k^2 \mathbf{x}_k\mathbf{x}_k^T,$$

with

$$\mathbf{e}_j^T \mathbf{w}_k = 0,\ j < k \quad \text{and} \quad \mathbf{e}_j^T \mathbf{x}_k = 0,\ j \le k.$$

If we make the identification

$$\mathbf{u}_k = \alpha_k \mathbf{w}_k \quad \text{and} \quad \mathbf{v}_k = \beta_k \mathbf{x}_k,$$

then we find that the modified elementary downdating problem is equivalent to the elementary downdating problem. However, the extra parameters can be chosen judiciously to eliminate some multiplications. For example, if we take $\alpha_k = \beta_k$, $\alpha_{k+1} = \beta_{k+1}$, then from (3.3a), (3.3b), and (3.4),

$$\sin\theta_k = \mathbf{e}_{k+1}^T \mathbf{x}_k / \mathbf{e}_k^T \mathbf{w}_k, \tag{3.12a}$$

$$\alpha_{k+1} = \alpha_k / \cos\theta_k, \tag{3.12b}$$

and

$$\mathbf{w}_{k+1} = Z\mathbf{w}_k - \sin\theta_k \mathbf{x}_k, \tag{3.13a}$$

$$\mathbf{x}_{k+1} = -\sin\theta_k Z\mathbf{w}_k + \mathbf{x}_k. \tag{3.13b}$$

Equations (3.12a)–(3.13b) form a basis for a *scaled hyperbolic* elementary downdating algorithm that requires $2(n-k)+O(1)$ multiplications. This is about half the number required by the unscaled algorithm based on (3.4). (The price is an increased likelihood of underflow or overflow, but this can be avoided if suitable precautions are taken in the code.)

Similarly, from (3.7a) and (3.7b) we can obtain a *scaled mixed* elementary downdating algorithm via

$$\begin{aligned} \sin\theta_k &= \beta_k \mathbf{e}_{k+1}^T \mathbf{x}_k / \alpha_k \mathbf{e}_k^T \mathbf{w}_k, \\ \alpha_{k+1} &= \alpha_k \cos\theta_k, \\ \beta_{k+1} &= \beta_k / \cos\theta_k, \end{aligned}$$

and

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{x}_k - \frac{\sin\theta_k \alpha_k}{\beta_k} Z\mathbf{w}_k, \\ \mathbf{w}_{k+1} &= -\frac{\sin\theta_k \beta_{k+1}}{\alpha_{k+1}} \mathbf{x}_{k+1} + Z\mathbf{w}_k. \end{aligned}$$

The stability properties of scaled mixed algorithms are similar to those of the corresponding unscaled algorithms [12].

4. Symmetric factorization. We adopt the following definition from [18].

DEFINITION 4.1. *An $n \times n$ symmetric matrix T has displacement rank 2 if and only if there exist vectors $\mathbf{u}, \mathbf{v} \in \Re^n$ such that*

$$T - ZTZ^T = \mathbf{u}\mathbf{u}^T - \mathbf{v}\mathbf{v}^T. \tag{4.1}$$

The vectors $\mathbf{u}$ and $\mathbf{v}$ are called the generators of T and determine the matrix T uniquely. Whenever we want to stress the dependence of T on $\mathbf{u}$ and $\mathbf{v}$ we write $T = T(\mathbf{u}, \mathbf{v})$.

In the sequel we will be concerned with a subset $\mathcal{T}$ of all matrices satisfying (4.1). The subset is defined as follows.

DEFINITION 4.2. *A matrix T is in $\mathcal{T}$ if and only if the following apply:*

(a) *T is positive definite;*

(b) *T satisfies (4.1) with generators $\mathbf{u}$ and $\mathbf{v}$;*

(c) *$\mathbf{v}^T\mathbf{e}_1 = 0$, i.e., the first component of $\mathbf{v}$ is zero.*

It is well known that positive definite $n \times n$ Toeplitz matrices form a subset of $\mathcal{T}$. Indeed, if $T = (t_{|i-j|})_{i,j=0}^{n-1}$, then

$$T - ZTZ^T = \mathbf{u}\mathbf{u}^T - \mathbf{v}\mathbf{v}^T,$$

where

$$\mathbf{u}^T = (t_0, t_1, \ldots, t_{n-1})/\sqrt{t_0}\,,$$
$$\mathbf{v}^T = (0, t_1, \ldots, t_{n-1})/\sqrt{t_0}\,.$$

The set $\mathcal{T}$ also contains matrices that are not Toeplitz, as the following example shows.

Example. Let

$$T = \begin{bmatrix} 25 & 20 & 15 \\ 20 & 32 & 29 \\ 15 & 29 & 40 \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} 5 \\ 4 \\ 3 \end{bmatrix}, \quad \text{and} \quad \mathbf{v} = \begin{bmatrix} 0 \\ 3 \\ 1 \end{bmatrix}.$$

It is easy to check that T is positive definite. Moreover,

$$T - ZTZ^T = \begin{bmatrix} 25 & 20 & 15 \\ 20 & 7 & 9 \\ 15 & 9 & 8 \end{bmatrix} = \begin{bmatrix} 25 & 20 & 15 \\ 20 & 16 & 12 \\ 15 & 12 & 9 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 9 & 3 \\ 0 & 3 & 1 \end{bmatrix} = \mathbf{u}\mathbf{u}^T - \mathbf{v}\mathbf{v}^T.$$

Hence $T = T(\mathbf{u}, \mathbf{v}) \in \mathcal{T}$, but T is not Toeplitz. □

We now establish a connection between the elementary downdating problem and symmetric factorizations of a matrix from the set $\mathcal{T}$.

Let $T = T(\mathbf{u}, \mathbf{v}) \in \mathcal{T}$. Set

$$\mathbf{u}_1 = \mathbf{u}, \qquad \mathbf{v}_1 = \mathbf{v}$$

and, for $k = 1, \ldots, n-1$, solve the elementary downdating problem defined by (3.1),

$$\mathbf{u}_{k+1}\mathbf{u}_{k+1}^T - \mathbf{v}_{k+1}\mathbf{v}_{k+1}^T = Z\mathbf{u}_k\mathbf{u}_k^T Z^T - \mathbf{v}_k\mathbf{v}_k^T,$$

which we assume for the moment has a solution for each k. On summing over $k = 1, \ldots, n-1$ we obtain

$$\sum_{k=1}^{n-1} \mathbf{u}_{k+1}\mathbf{u}_{k+1}^T - \sum_{k=1}^{n-1} \mathbf{v}_{k+1}\mathbf{v}_{k+1}^T = \sum_{k=1}^{n-1} Z\mathbf{u}_k\mathbf{u}_k^T Z^T - \sum_{k=1}^{n-1} \mathbf{v}_k\mathbf{v}_k^T.$$

If we now observe that, from (3.2),

$$Z\mathbf{u}_n = \mathbf{v}_n = 0,$$

we arrive at the following relation:

$$\sum_{k=1}^{n} \mathbf{u}_k\mathbf{u}_k^T - Z\left(\sum_{k=1}^{n} \mathbf{u}_k\mathbf{u}_k^T\right)Z^T = \mathbf{u}_1\mathbf{u}_1^T - \mathbf{v}_1\mathbf{v}_1^T, \tag{4.2}$$

which implies that $\sum_{k=1}^{n} \mathbf{u}_k\mathbf{u}_k^T \in \mathcal{T}$. Moreover, as matrices having the same generators are identical, we obtain

$$T = \sum_{k=1}^{n} \mathbf{u}_k\mathbf{u}_k^T = U^TU,$$

where

$$U = \sum_{k=1}^{n} \mathbf{e}_k\mathbf{u}_k^T$$

is upper triangular, and hence is the Cholesky factor of T. We have derived, albeit in a rather indirect manner, the basis of an algorithm for calculating the Cholesky decomposition of a matrix from the set $\mathcal{T}$.

We now return to the question of existence of a solution to the elementary downdating problem for each $k = 1, \ldots, n-1$. It is easy to verify that, if $T \in \mathcal{T}$, then $|\mathbf{e}_1^T\mathbf{u}_1| > |\mathbf{e}_2^T\mathbf{v}_1|$. Using (4.2) and (3.1), it can be shown by induction on k that

$$|\mathbf{e}_k^T\mathbf{u}_k| > |\mathbf{e}_{k+1}^T\mathbf{v}_k|, \qquad k = 2, \ldots, n-1.$$

Consequently, $|\sin\theta_k| < 1$ in (3.3a), and the elementary downdating problem has a solution for each $k = 1, \ldots, n-1$.

To summarize, we have the following algorithm for factorizing a matrix:

$$T = T(\mathbf{u}\,,\,\mathbf{v}) \in \mathcal{T}.$$

Algorithm FACTOR(T):

Set $\mathbf{u}_1 = \mathbf{u}$, $\mathbf{v}_1 = \mathbf{v}$.

For $k = 1, \ldots, n-1$ calculate $\mathbf{u}_{k+1}$, $\mathbf{v}_{k+1}$ such that

$$\mathbf{u}_{k+1}\mathbf{u}_{k+1}^T - \mathbf{v}_{k+1}\mathbf{v}_{k+1}^T = Z\mathbf{u}_k\mathbf{u}_k^TZ^T - \mathbf{v}_k\mathbf{v}_k^T,$$
$$\mathbf{e}_{k+1}^T\mathbf{v}_{k+1} = 0\,.$$

Then $T = U^TU$, where $U = \sum_{k=1}^{n} \mathbf{e}_k\mathbf{u}_k^T$.

In fact we have not one algorithm, but a class of factorization algorithms, where each algorithm corresponds to a particular way of realizing the elementary downdating steps. For example, the connection with the scaled elementary downdating problem is straightforward. On making the identification

$$\mathbf{u}_k = \alpha_k\mathbf{w}_k \quad \text{and} \quad \mathbf{v}_k = \beta_k\mathbf{x}_k, \tag{4.3}$$

we obtain

$$T = W^TD^2W,$$

where

$$W = \sum_{k=1}^{n} \mathbf{e}_k \mathbf{w}_k^T,$$
$$D = \sum_{k=1}^{n} \alpha_k \mathbf{e}_k \mathbf{e}_k^T.$$

It is clear from §3 that Algorithm FACTOR(T) requires $2n^2 + O(n)$ multiplications when the unscaled version of elementary downdating is used, and $n^2 + O(n)$ multiplications when the scaled version of elementary downdating is used. However, in the sequel we do not dwell on the precise details of algorithms. Using (4.3), we can relate algorithms based on the scaled elementary downdating problem to those based on the unscaled elementary downdating problem. Thus, for simplicity, we consider only the unscaled elementary downdating algorithms.

5. Analysis of factorization algorithms. In this section we present a numerical stability analysis of the factorization of $T \in \mathcal{T}$ via Algorithm FACTOR(T). The result of the analysis is applied to the case when the matrix T is Toeplitz.

Let $\tilde{\mathbf{u}}_k$, $\tilde{\mathbf{v}}_k$ be the values of $\mathbf{u}_k$, $\mathbf{v}_k$ that are computed in floating point arithmetic with relative machine relative precision ϵ. The computed quantities $\tilde{\mathbf{u}}_k$ and $\tilde{\mathbf{v}}_k$ satisfy the relations

$$\tilde{\mathbf{u}}_k = \mathbf{u}_k + O(\epsilon), \qquad \tilde{\mathbf{v}}_k = \mathbf{v}_k + O(\epsilon), \tag{5.1}$$

and the aim of this section is to provide a first order analysis of the error. By a first order analysis we mean that the error can be bounded by a function that has an asymptotic expansion in powers of ϵ, but we only consider the first term of this asymptotic expansion. One should think of $\epsilon \to 0+$ while the problem remains fixed [19]. Thus, in this section (except for Corollary 5.5) we omit functions of n from the "O" terms in relations such as (5.1) and (5.2).

The computed vectors $\tilde{\mathbf{u}}_k$, $\tilde{\mathbf{v}}_k$ satisfy a perturbed version (3.9) of (3.1). On summing (3.9) over $k = 1, \ldots, n-1$ we obtain

$$\tilde{T} - Z\tilde{T}Z^T = \tilde{\mathbf{u}}_1\tilde{\mathbf{u}}_1^T - \tilde{\mathbf{v}}_1\tilde{\mathbf{v}}_1^T - (Z\tilde{\mathbf{u}}_n\tilde{\mathbf{u}}_n^T Z^T - \tilde{\mathbf{v}}_n\tilde{\mathbf{v}}_n^T) + \epsilon \sum_{k=1}^{n-1} G_k + O(\epsilon^2),$$

where

$$\tilde{T} = \tilde{U}^T\tilde{U},$$
$$\tilde{U} = \sum_{k=1}^{n} \mathbf{e}_k \tilde{\mathbf{u}}_k^T.$$

Since

$$Z\tilde{\mathbf{u}}_n = O(\epsilon), \qquad \tilde{\mathbf{v}}_n = O(\epsilon),$$

we find that

$$\tilde{T} - Z\tilde{T}Z^T = \tilde{\mathbf{u}}_1\tilde{\mathbf{u}}_1^T - \tilde{\mathbf{v}}_1\tilde{\mathbf{v}}_1^T + \epsilon \sum_{k=1}^{n-1} G_k + O(\epsilon^2). \tag{5.2}$$

Now define

$$\tilde{E} = \tilde{T} - T. \tag{5.3}$$

Then, using (4.1), (5.2), and (5.3),

$$\tilde{E} - Z\tilde{E}Z^T = \tilde{\mathbf{u}}_1\tilde{\mathbf{u}}_1^T - \mathbf{u}\mathbf{u}^T + \tilde{\mathbf{v}}_1\tilde{\mathbf{v}}_1^T - \mathbf{v}\mathbf{v}^T + \epsilon\sum_{k=1}^{n-1} G_k + O(\epsilon^2)\,.$$

In a similar manner we obtain expressions for $Z_j\tilde{E}Z_j^T - Z_{j+1}\tilde{E}Z_{j+1}^T$, $j = 0, \ldots, n-1$. Summing over j gives

$$\begin{aligned}\tilde{E} =& \sum_{j=0}^{n-1} Z_j\Big((\tilde{\mathbf{u}}_1\tilde{\mathbf{u}}_1^T - \mathbf{u}_1\mathbf{u}_1^T) + (\tilde{\mathbf{v}}_1\tilde{\mathbf{v}}_1^T - \mathbf{v}_1\mathbf{v}_1^T)\Big)Z_j^T \\ &+ \epsilon\sum_{j=0}^{n-1}\sum_{k=1}^{n-1} Z_jG_kZ_j^T + O(\epsilon^2)\,.\end{aligned} \tag{5.4}$$

We see from (5.4) that the error consists of two parts—the first part associated with initial errors and the second part associated with the fact that (5.2) contains an inhomogeneous term. Now

$$\begin{aligned}\|\tilde{\mathbf{u}}_1\tilde{\mathbf{u}}_1^T - \mathbf{u}\mathbf{u}^T\| &\le 2\|\mathbf{u}\|\,\|\tilde{\mathbf{u}}_1 - \mathbf{u}\| + O(\epsilon^2), \\ \|\tilde{\mathbf{v}}_1\tilde{\mathbf{v}}_1^T - \mathbf{v}\mathbf{v}^T\| &\le 2\|\mathbf{v}\|\,\|\tilde{\mathbf{v}}_1 - \mathbf{v}\| + O(\epsilon^2).\end{aligned}$$

Furthermore, from (4.1),

$$Tr(T) - Tr(ZTZ^T) = \|\mathbf{u}\|^2 - \|\mathbf{v}\|^2 > 0,$$

and hence

$$\begin{aligned}&\Big\|\sum_{j=0}^{n-1} Z_j(\tilde{\mathbf{u}}_1\tilde{\mathbf{u}}_1^T - \mathbf{u}\mathbf{u}^T + \tilde{\mathbf{v}}_1\tilde{\mathbf{v}}_1^T - \mathbf{v}\mathbf{v}^T)Z_j^T\Big\| \\ &\le 2n\|\mathbf{u}\|\Big(\|\tilde{\mathbf{u}}_1 - \mathbf{u}\| + \|\tilde{\mathbf{v}}_1 - \mathbf{v}\|\Big) + O(\epsilon^2).\end{aligned} \tag{5.5}$$

This demonstrates that initial errors do not propagate unduly. To investigate the double sum in (5.4) we require a preliminary result.

LEMMA 5.1. *For $k = 1, 2, \ldots, n-1$, and $j = 0, 1, 2, \ldots$,*

$$\|Z_j\mathbf{v}_k\| \le \|Z_{j+1}\mathbf{u}_k\|\,.$$

Proof. Let

$$T_k = T - \sum_{l=1}^{k} \mathbf{u}_l\mathbf{u}_l^T = \sum_{l=k+1}^{n} \mathbf{u}_l\mathbf{u}_l^T.$$

It is easy to verify that

$$T_k - ZT_kZ^T = Z\mathbf{u}_k\mathbf{u}_k^TZ^T - \mathbf{v}_k\mathbf{v}_k^T$$

and, since T_k is positive semidefinite,

$$Tr\Big(Z_j T_k Z_j^T - Z_{j+1} T_k Z_{j+1}^T\Big) = \|Z_{j+1}\mathbf{u}_k\|^2 - \|Z_j\mathbf{v}_k\|^2 \geq 0. \qquad \square$$

We now demonstrate stability when the mixed version of elementary downdating is used in Algorithm FACTOR(T). In this case the inhomogeneous term G_k satisfies a shifted version of (3.10), that is

$$\|Z_j G_k Z_j^T\| \leq c_m\Big(\|Z_{j+1}\mathbf{u}_k\|^2 + \|Z_j\mathbf{v}_k\|^2 + \|Z_j\mathbf{u}_{k+1}\|^2 + \|Z_j\mathbf{v}_{k+1}\|^2\Big)\,, \tag{5.6}$$

where c_m is a positive constant.

THEOREM 5.2. *Assume that* (3.9) *and* (5.6) *hold. Then*

$$\|T - \tilde{U}^T\tilde{U}\| \leq 2n\|\mathbf{u}\|\Big(\|\tilde{\mathbf{u}}_1 - \mathbf{u}\| + \|\tilde{\mathbf{v}}_1 - \mathbf{v}\|\Big) + 4\epsilon c_m \sum_{j=0}^{n-1} Tr(Z_j T Z_j^T) + O(\epsilon^2).$$

Proof. Using Lemma 5.1,

$$\|Z_j G_k Z_j^T\| \leq 2c_m\Big(\|Z_{j+1}\mathbf{u}_k\|^2 + \|Z_j\mathbf{u}_{k+1}\|^2\Big).$$

Furthermore, since

$$Tr(Z_j T Z_j^T) = \sum_{k=1}^{n} \|Z_j\mathbf{u}_k\|^2,$$

it follows that

$$\Big\|\sum_{j=0}^{n-1}\sum_{k=1}^{n} Z_j G_k Z_j^T\Big\| \leq 4c_m \sum_{j=0}^{n-1} Tr(Z_j T Z_j^T). \tag{5.7}$$

The result now follows from (5.4), (5.5), and (5.7). $\square$

For the hyperbolic version of the elementary downdating algorithms a shifted version of the weaker bound (3.11) on G_k holds (see [6]), namely,

$$\|Z_j G_k Z_j^T\| \leq c_h\|H(\theta_k)\|(\|Z_{j+1}\mathbf{u}_k\| + \|Z_j\mathbf{v}_k\|)(\|Z_j\mathbf{u}_{k+1}\| + \|Z_j\mathbf{v}_{k+1}\|). \tag{5.8}$$

By Lemma 5.1, this simplifies to

$$\|Z_j G_k Z_j^T\| \leq 4c_h\|H(\theta_k)\|\,\|Z_{j+1}\mathbf{u}_k\|\,\|Z_j\mathbf{u}_{k+1}\|. \tag{5.9}$$

The essential difference between (3.10) and (3.11) is the occurrence of the multiplier $\|H(\theta_k)\|$, which can be quite large. This term explains numerical difficulties in applications such as the downdating of a Cholesky decomposition [6]. However, because of the special structure of the matrix T, it is of lesser importance here, in view of the following result.

LEMMA 5.3. *For* $k = 1, 2, \ldots, n-1$, *and* $j = 0, 1, \ldots, n-k$,

$$\|H(\theta_k)\|\,\|Z_j\mathbf{u}_{k+1}\| \leq 2(n-k-j)\|Z_{j+1}\mathbf{u}_k\|.$$

Proof. It is easy to verify from (3.4) that

$$\frac{1 \mp \sin\theta_k}{\cos\theta_k}\big(\mathbf{u}_{k+1} \mp \mathbf{v}_{k+1}\big) = Z\mathbf{u}_k \mp \mathbf{v}_k\,,$$

and from (2.1) that

$$\|H(\theta_k)\| = \frac{1+|\sin\theta|}{\cos\theta}.$$

Thus,

$$\begin{aligned}\|H(\theta_k)\|\,\|Z_j\mathbf{u}_{k+1}\| &\le \|H(\theta_k)\|\,\|Z_j\mathbf{v}_{k+1}\| + \|Z_{j+1}\mathbf{u}_k\| + \|Z_j\mathbf{v}_k\| \\ &\le \|H(\theta_k)\|\,\|Z_{j+1}\mathbf{u}_{k+1}\| + 2\|Z_{j+1}\mathbf{u}_k\|,\end{aligned}$$

where the last inequality was obtained using Lemma 5.1. Thus

$$\|H(\theta_k)\|\,\|Z_j\mathbf{u}_{k+1}\| \le 2\sum_{l=j+1}^{n-k}\|Z_l\mathbf{u}_k\|,$$

and the result follows. □

Remark. Lemma 5.3 does not hold for the computed quantities unless we introduce an $O(\epsilon)$ term. However, in a first order analysis we only need it to hold for the exact quantities.

THEOREM 5.4. *Assume that* (3.9) *and* (5.8) *hold. Then*

$$\|T-\tilde{U}^T\tilde{U}\| \le 2n\|\mathbf{u}\|\big(\|\tilde{\mathbf{u}}_1-\mathbf{u}\|+\|\tilde{\mathbf{v}}_1-\mathbf{v}\|\big) + 8\epsilon c_h\sum_{j=1}^{n-1}(n-j)Tr(Z_jTZ_j^T) + O(\epsilon^2).$$

Proof. Applying Lemma 5.3 to (5.9) gives

$$\|Z_jG_kZ_j^T\| \le 8c_h(n-j-1)\|Z_{j+1}\mathbf{u}_k\|^2,$$

and hence

$$\begin{aligned}\Big\|\sum_{j=0}^{n-1}\sum_{k=1}^{n-1}Z_jG_kZ_j^T\Big\| &\le 8c_h\sum_{j=1}^{n-1}\sum_{k=1}^{n-1}(n-j)\|Z_j\mathbf{u}_k\|^2 \\ &\le 8c_h\sum_{j=1}^{n-1}(n-j)Tr(Z_jTZ_j^T).\end{aligned} \tag{5.10}$$

The result now follows from (5.4), (5.5), and (5.10). □

Note that, when T is Toeplitz,

$$Tr(Z_jTZ_j^T) = (n-j)t_0.$$

Hence, from Theorems 5.2 and 5.4, we obtain our main result on the stability of the factorization algorithms based on Algorithm FACTOR(T) for a symmetric positive definite Toeplitz matrix.

COROLLARY 5.5. *The factorization algorithm* FACTOR(T) *applied to a symmetric positive definite Toeplitz matrix T produces an upper triangular matrix $\tilde{U}$ such that*

$$T = \tilde{U}^T\tilde{U} + \Delta T,$$

where

$$\|\Delta T\| = O(\epsilon t_0 n^2)$$

when mixed downdating is used, and

$$\|\Delta T\| = O(\epsilon t_0 n^3)$$

when hyperbolic downdating is used.

6. The connection with the Bareiss algorithm. In 1969, Bareiss [2] proposed an $O(n^2)$ algorithm for solving Toeplitz linear systems. For a symmetric Toeplitz matrix T, the algorithm, called a *symmetric Bareiss algorithm* in [22], can be expressed as follows. Start with a matrix $A^{(0)} := T$ and partition it in two ways:

$$A^{(0)} = \begin{pmatrix} U^{(0)} \\ T^{(1)} \end{pmatrix} = \begin{pmatrix} T^{(-1)} \\ L^{(0)} \end{pmatrix},$$

where $U^{(0)}$ is the first row of T and $L^{(0)}$ is the last row of T. Now, starting from $A^{(0)}$, compute successively two matrix sequences $\{A^{(i)}\}$ and $\{A^{(-i)}\}$, $i = 1, \ldots, n-1$, according to the relations

$$\begin{aligned} A^{(i)} &= A^{(i-1)} - \alpha_{i-1} Z_i A^{(-i+1)}, \\ A^{(-i)} &= A^{(-i+1)} - \alpha_{-i+1} Z_i^T A^{(i-1)}. \end{aligned} \tag{6.1}$$

For $1 \le i \le n-1$, partition $A^{(i)}$ and $A^{(-i)}$ as follows:

$$A^{(i)} = \begin{pmatrix} U^{(i)} \\ T^{(i+1)} \end{pmatrix}, \qquad A^{(-i)} = \begin{pmatrix} T^{(-i-1)} \\ L^{(i)} \end{pmatrix},$$

where $U^{(i)}$ denotes the first $i+1$ rows of $A^{(i)}$ and $L^{(i)}$ denotes the last $i+1$ rows of $A^{(-i)}$. It is shown in [2] that the following are true.

(a) $T^{(i+1)}$ and $T^{(-i-1)}$ are Toeplitz.

(b) For a proper choice of α_{i-1} and α_{-i+1}, the matrices $L^{(i)}$ and $U^{(i)}$ are lower and upper trapezoidal, respectively.

(c) With the choice of α_{i-1} and α_{-i+1} as in (b), the Toeplitz matrix $T^{(-i-1)}$ has zero elements in positions $2, \ldots, i+1$ of its first row, while the Toeplitz matrix $T^{(i+1)}$ has zero elements in positions $n-1, \ldots, n-i$ of its last row.

Pictorially,

$$A^{(i)} = \left(\frac{U^{(i)}}{T^{(i+1)}}\right) = \begin{pmatrix} \times & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \times \\ 0 & \times & & & & & & \times \\ \vdots & \ddots & \times & \times & & & & \vdots \\ 0 & \cdots & 0 & \times & \cdots & \cdots & \cdots & \times \\ \hline \times & 0 & \cdots & 0 & \times & \cdots & \cdots & \times \\ \vdots & \ddots & & & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & & & \ddots & \ddots & \vdots \\ \times & \cdots & \cdots & \times & 0 & \cdots & 0 & \times \end{pmatrix}$$

and

$$A^{(-i)} = \left(\frac{T^{(-i-1)}}{L^{(i)}}\right) = \begin{pmatrix} \times & 0 & \cdots & 0 & \times & \cdots & \cdots & \times \\ \vdots & \ddots & & & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & & & \ddots & \ddots & \vdots \\ \times & \cdots & \cdots & \times & 0 & \cdots & 0 & \times \\ \hline \times & \cdots & \cdots & \times & \times & 0 & \cdots & 0 \\ \vdots & & & & & \ddots & \ddots & \vdots \\ \vdots & & & & & & \ddots & 0 \\ \times & \times & \cdots & \cdots & \cdots & \cdots & \cdots & \times \end{pmatrix}.$$

After $n-1$ steps, the matrices $A^{(n-1)}$ and $A^{(-n+1)}$ are lower and upper triangular, respectively. At step i only rows $i+1,\dots,n$ of $A^{(i)}$ and rows $1,2,\dots,n-i$ of $A^{(-i)}$ are modified; the remaining rows stay unchanged. Moreover, Bareiss [2] noticed that, because of the symmetry of T,

$$T^{(i+1)} = J_{i+1}T^{(-i-1)}J_n \quad \text{and} \quad \alpha_{i-1} = \alpha_{-i+1}, \tag{6.2}$$

Here, J_{i+1} and J_n are the reversal matrices of dimension $(i+1)\times(i+1)$ and $n\times n$, respectively.

Now, taking into account (6.2), it can be seen that the essential part of a step of the Bareiss algorithm (6.1) can be written as follows:

$$\begin{pmatrix} t_{i+2}^{(i+1)} & t_{i+3}^{(i+1)} & \dots & t_n^{(i+1)} \\ 0 & t_{i+3}^{(-i-1)} & \dots & t_n^{(-i-1)} \end{pmatrix} = \begin{pmatrix} 1 & -\alpha_{i-1} \\ -\alpha_{i-1} & 1 \end{pmatrix} \begin{pmatrix} t_{i+2}^{(i)} & t_{i+3}^{(i)} & \dots & t_n^{(i)} \\ t_{i+2}^{(-i)} & t_{i+3}^{(-i)} & \dots & t_n^{(-i)} \end{pmatrix}, \tag{6.3}$$

where $(t_{i+2}^{(-i)}, t_{i+2}^{(-i)}, \dots, t_n^{(-i)})$ are the last $(n-i-1)$ components of the first row of $T^{(-i)}$, and $(t_{i+2}^{(i)}, t_{i+3}^{(i)}, \dots, t_n^{(i)})$ are the last $(n-i-1)$ components of the first row of $T^{(i)}$.

Note that (6.3) has the same form as (3.13a)–(3.13b), and hence a connection between the Bareiss algorithm and algorithm FACTOR(T) is evident. That such a connection exists was observed by Sweet [22], and later by Delosme and Ipsen [11]. Sweet [22] related a step of the Bareiss algorithm (6.3) to a step of Bennett's downdating procedure [3]. Next, he derived the LU factorization of a Toeplitz matrix as a sequence of Bennett's downdating steps. Finally, he estimated the forward error in the decomposition using Fletcher and Powell's methodology [12]. This paper generalizes and presents new derivations of the results obtained in [22].

7. Numerical examples. We adopt from [17] the following definitions of forward and backward stability.

DEFINITION 7.1. *An alogrithm for solving* (1.1) *is* forward stable *if the computed solution* $\tilde{x}$ *satisfies*

$$||x-\tilde{x}|| \le c_1(n)\,\epsilon\,\mathrm{cond}(T)||\tilde{x}||,$$

where $\mathrm{cond}(T) = ||T||\,||T^{-1}||$ *is the condition number of* T, *and* $c_1(n)$ *may grow at most as fast as a polynomial in* n, *the dimension of the system.*

DEFINITION 7.2. *An algorithm for solving* (1.1) *is* backward stable *if the computed solution* $\tilde{x}$ *satisfies*

$$||T\tilde{x}-b|| \le c_2(n)\epsilon||T||\,||\tilde{x}||,$$

where $c_2(n)$ *may grow at most as fast as a polynomial in* n, *the dimension of the system.*

It is known that an algorithm (for solving a system of linear equations) is backward stable if and only if there exists a matrix ΔT such that

$$(T+\Delta T)\tilde{x} = b, \qquad ||\Delta T|| \le c_3(n)\epsilon||T||,$$

where $c_3(n)$ may grow at most as fast as a polynomial in n.

Note that our definitions do not require the perturbation ΔT to be Toeplitz, even if the matrix T is Toeplitz. The case that ΔT is Toeplitz is discussed in [13], [24]. The reader is referred to [9], [14], [19] for a detailed treatment of roundoff analysis for general matrix algorithms.

It is easy to see that backward stability implies forward stability, but not vice versa. This is manifested by the size of the residual vector.

Cybenko [10] showed that the L_1 norm of the inverse of a $n \times n$ symmetric positive definite Toeplitz matrix T_n is bounded by

$$\max\left\{\frac{1}{\prod_{i=1}^{n-1}\cos^2\theta_i}, \frac{1}{\prod_{i=1}^{n-1}(1+\sin\theta_i)}\right\} \le \|T_n^{-1}\|_1 \le \prod_{i=1}^{n-1}\frac{1+|\sin\theta_i|}{1-|\sin\theta_i|},$$

where $\{-\sin\theta_i\}_{i=1}^{n-1}$ are quantities called *reflection coefficients*. It is not difficult to pick the reflection coefficients in such a way that the corresponding Toeplitz matrix T_n satisfies

$$\mathrm{cond}(T_n) \approx 1/\epsilon.$$

One possible way of constructing a Toeplitz matrix with given reflection coefficients $\{-\sin\theta_i\}_{i=1}^{n-1}$ is by tracing the elementary downdating steps backwards.

An example of a symmetric positive definite Toeplitz matrix that can be made poorly conditioned by suitable choice of a parameter is the *prolate* matrix [21], [23], defined by

$$t_k = \begin{cases} 2\omega & \text{if } k = 0, \\ \sin(2\pi\omega k)/(\pi k) & \text{otherwise,} \end{cases}$$

where $0 \le \omega \le \frac{1}{2}$. For small ω the eigenvalues of the prolate matrix cluster around 0 and 1.

We performed numerical tests in which we solved systems of Toeplitz linear equations using variants of the Bareiss and Levinson algorithms and (for comparison) the standard Cholesky method. The relative machine precision was $\epsilon = 2^{-53} \approx 10^{-16}$. We varied the dimension of the system from 10 to 100, the condition number of the matrix from 1 to ϵ^{-1}, the signs of reflection coefficients, and the right-hand side so that the magnitude of the norm of the solution vector varied from 1 to ϵ^{-1}. In each test we monitored the errors in the decomposition, in the solution vector, and the size of the residual vector.

Let x_B and x_L denote the solutions computed by the Bareiss and Levinson algorithms. Also, let $r_B = Tx_B - b$ and $r_L = Tx_L - b$. Then for the Bareiss algorithms we always observed that the scaled residual

$$s_B \equiv \frac{\|r_B\|}{\epsilon\|x_B\|\|T\|}$$

was of order unity, as small as would be expected for a backward stable method. However, we were not able to find an example that would demonstrate the superiority of the Bareiss algorithm based on mixed downdating over the Bareiss algorithm based on hyperbolic downdating. In fact, the Bareiss algorithm based on hyperbolic downdating often gave slightly smaller errors than the Bareiss algorithm based on mixed downdating. In our experiments with Bareiss algorithms, neither the norm of the

TABLE 7.1

Prolate matrix, $n = 21$, $\omega = 0.25$, *cond* $= 3.19 \cdot 10^{14}$.

	decomp. error	soln. error	resid. error
Cholesky	$5.09 \cdot 10^{-1}$	$7.67 \cdot 10^{-3}$	$1.25 \cdot 10^{0}$
Bareiss(hyp)	$3.45 \cdot 10^{0}$	$1.40 \cdot 10^{-2}$	$8.72 \cdot 10^{-1}$
Bareiss(mixed)	$2.73 \cdot 10^{0}$	$1.41 \cdot 10^{0}$	$1.09 \cdot 10^{0}$
Levinson	–	$5.30 \cdot 10^{0}$	$4.57 \cdot 10^{3}$

TABLE 7.2

Reflection coefficients of the same magnitude $|K|$ *but alternating signs,* $|K| = 0.8956680108101296$, $n = 41$, *cond* $= 8.5 \cdot 10^{15}$.

	decomp. error	soln. error	resid. error
Cholesky	$1.72 \cdot 10^{-1}$	$6.84 \cdot 10^{-2}$	$3.11 \cdot 10^{-1}$
Bareiss(hyp)	$2.91 \cdot 10^{0}$	$2.19 \cdot 10^{-1}$	$1.15 \cdot 10^{-1}$
Bareiss(mixed)	$3.63 \cdot 10^{0}$	$2.48 \cdot 10^{-1}$	$2.47 \cdot 10^{-1}$
Levinson	–	$5.27 \cdot 10^{-1}$	$1.47 \cdot 10^{5}$

TABLE 7.3

Reflection coefficients of the same magnitude $|K|$ *but alternating signs,* $|K| = 0.9795872473975045$, $n = 92$, *cond* $= 2.77 \cdot 10^{15}$.

	decomp. error	soln. error	resid. error
Cholesky	$8.51 \cdot 10^{-1}$	$3.21 \cdot 10^{-2}$	$4.28 \cdot 10^{-1}$
Bareiss(hyp)	$8.06 \cdot 10^{0}$	$1.13 \cdot 10^{-1}$	$2.28 \cdot 10^{-1}$
Bareiss(mixed)	$6.71 \cdot 10^{0}$	$1.16 \cdot 10^{-1}$	$3.20 \cdot 10^{-1}$
Levinson	–	$2.60 \cdot 10^{-1}$	$1.06 \cdot 10^{5}$

error matrix in the decomposition of T nor the residual error in the solution seemed to depend in any clear way on n, although a quadratic or cubic dependence would be expected from the worst-case error bounds of Theorems 5.2, 5.4, and Corollary 5.5.

For well-conditioned systems, the Bareiss and Levinson algorithms behaved similarly and gave results comparable to results produced by a general stable method (the Cholesky method). Differences between the Bareiss and Levinson algorithms were noticeable only for very ill-conditioned systems and special right-hand side vectors.

For the Levinson algorithm, when the matrix was very ill conditioned and the norm of the solution vector was of order unity (that is, when the norm of the solution vector did not reflect the ill conditioning of the matrix), we often observed that the scaled residual

$$s_L \equiv \frac{\|r_L\|}{\epsilon \|x_L\| \|T\|},$$

was as large as 10^5. Varah [23] was the first to observe this behavior of the Levinson algorithm on the prolate matrix. Higham and Pickering [16] used a search method proposed in [15] to generate Toeplitz matrices for which the Levinson algorithm yields large residual errors. However, the search never produced s_L larger than $5 \cdot 10^5$. It is plausible that s_L is a slowly increasing function of n and $1/\epsilon$.

Tables 7.1–7.3 show typical behavior of the Cholesky, Bareiss, and Levinson algorithms for ill-conditioned Toeplitz systems of linear equations when the norm of the solution vectors is of order unity. The decomposition error was measured for the Cholesky and Bareiss algorithms by the quotient $||T - L \cdot L^T||/(\epsilon \cdot ||T||)$, where

L was the computed factor of T. The solution error was measured by the quotient $||x_{\text{comp}} - x||/||x||$, where x_{comp} was the computed solution vector. Finally, the residual error was measured by the quotient $||T \cdot x_{\text{comp}} - b||/(||T|| \cdot ||x_{\text{comp}}|| \cdot \epsilon)$.

8. Conclusions. This paper generalizes and presents new derivations of results obtained earlier by Sweet [22]. The bound in Corollary 5.5 for the case of mixed downdating is stronger than that given in [22]. The applicability of the Bareiss algorithms based on elementary downdating steps is extended to a class of matrices, satisfying Definition 4.2 that includes symmetric positive definite Toeplitz matrices. The approach via elementary downdating greatly simplifies roundoff error analysis. Lemmas 5.1 and 5.3 appear to be new. The stability of the Bareiss algorithms follows directly from these lemmas and the results on the roundoff error analysis for elementary downdating steps given in [6].

The approach via downdating can be extended to the symmetric factorization of positive definite matrices of displacement rank $k \geq 2$ (satisfying additional conditions similar to those listed in Definition 4.2); see [18]. For matrices of displacement rank k the factorization algorithm uses elementary rank-k downdating via hyperbolic Householder or mixed Householder reflections [8], [20].

We conclude by noting that the Bareiss algorithms guarantee small residual errors in the sense of Definition 7.2, but the Levinson algorithm can yield residuals at least five orders of magnitude larger than those expected for a backward stable method. This result suggests that, if the Levinson algorithm is used in applications where the reflection coefficients are not known in advance to be positive, the residuals should be computed to see if they are acceptably small. This can be done in $O(n \log n)$ arithmetic operations (using the fast Fourier transform).

It is an interesting open question whether the Levinson algorithm can give scaled residual errors that are arbitrarily large (for matrices that are numerically nonsingular). A related question is whether the Levinson algorithm, for positive definite Toeplitz matrices T without a restriction on the reflection coefficients, is stable in the sense of Definitions 7.1 and 7.2.

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] E. H. Bareiss, *Numerical solution of linear equations with Toeplitz and vector Toeplitz matrices*, Numer. Math., 13 (1969), pp. 404–424.

[3] J. M. Bennett, *Triangular factorization of modified matrices*, Numer. Math., 7 (1965), pp. 217–221.

[4] A. W. Bojanczyk, R. P. Brent, and F. R. de Hoog, *QR factorization of Toeplitz matrices*, Numer. Math., 49 (1986), pp. 81–94.

[5] ———, *Stability analysis of fast Toeplitz linear system solvers*, Report CMA-MR17-91, Centre for Mathematical Analysis, The Australian National University, Canberra, August 1991.

[6] A. W. Bojanczyk, R. P. Brent, P. van Dooren, and F. R. de Hoog, *A note on downdating the Cholesky factorization*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 210–220.

[7] A. W. Bojanczyk and A. O. Steinhardt, *Matrix downdating techniques for signal processing*, Proc. SPIE, Vol. 975: Advanced Algorithms and Architectures for Signal Processing III, 1988, pp. 68–75.

[8] ———, *Stabilized hyperbolic Householder transformations*, IEEE Trans. Acoustics, Speech and Signal Processing, ASSP–37 (1989), pp. 1286–1288.

[9] J. R. Bunch, *The weak and strong stability of algorithms in numerical linear algebra*, Linear Algebra Appl., 88/89 (1987), pp. 49–66.

[10] 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.

[11] J-M. DELOSME AND I. C. F. IPSEN, *From Bareiss's algorithm to the stable computation of partial correlations*, J. Comput. Appl. Math., 27 (1989), pp. 53–91.

[12] R. FLETCHER AND M. J. D. POWELL, *On the modification of LDL^T factorizations*, Math. Comp., 28 (1974), pp. 1067–1087.

[13] I. GOHBERG, I. KOLTRACHT, AND D. XIAO, *On the solution of the Yule–Walker equation*, Proc. SPIE, Vol. 1566: Advanced Algorithms and Architectures for Signal Processing IV, 1991.

[14] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, second ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

[15] N. J. HIGHAM, *Optimization by direct search in matrix computations*, SIAM J. Matrix Anal. Appl., 14 (1994), pp. 317–333.

[16] N. J. HIGHAM AND R. L. PICKERING, private communication, April 1992.

[17] M. JANKOWSKI AND H. WOZNIAKOWSKI, *Iterative refinement implies numerical stability*, BIT, 17 (1977), pp. 303–311.

[18] T. KAILATH, S. Y. KUNG, AND M. MORF, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979), pp. 395–407.

[19] W. MILLER AND C. WRATHALL, *Software for Roundoff Analysis of Matrix Algorithms*, Academic Press, New York, 1980.

[20] C. M. RADER AND A. O. STEINHARDT, *Hyperbolic Householder transformations*, IEEE Trans. Acoustics, Speech and Signal Processing, ASSP–34 (1986), pp. 1584–1602.

[21] D. SLEPIAN, *Prolate spheroidal wave functions, Fourier analysis, and uncertainty V: the discrete case*, Bell Systems Tech. J., 57 (1978), pp. 1371–1430.

[22] D. R. SWEET, *Numerical Methods for Toeplitz Matrices*, Ph.D. thesis, University of Adelaide, Adelaide, Australia, 1982.

[23] J. M. VARAH, *The prolate matrix: a good Toeplitz test example*, SIAM Conference on Control, Signals and Systems, San Francisco, 1990. Also, Linear Algebra Appl., 187 (1993), pp. 269–278.

[24] ———, *Backward error estimates for Toeplitz and Vandermonde systems*, preprint, 1992. Also, Tech. Report 91-20, Univ. of British Columbia, Sept. 1991.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 58–78, January 1995

DISPLACEMENT STRUCTURE AND COMPLETION PROBLEMS *

TIBERIU CONSTANTINESCU†, ALI H. SAYED‡, AND THOMAS KAILATH§

Abstract. A general result is proven concerning time-variant displacement equations with positive solutions in a general operatorial setting. It is then shown that the solutions of several completion problems, recently considered in connection with classical interpolation and moment theory, follow as special cases of the main result. The main purpose of this paper is to show that under supplementary finite-dimensionality conditions, a so-called generalized Schur algorithm, which naturally arises in connection with displacement equations, can be used to prove the above-mentioned result. The associated transmission-line interpretation is also discussed in terms of a cascade of elementary sections with intrinsic blocking properties.

Key words. displacement structure, generalized Schur algorithm, interpolation, completion problems

AMS subject classifications. 15A23, 93A20, 47A20

1. Introduction.

We prove a general result concerning time-variant displacement equations with positive solutions. More specifically, we show that a contractive upper-triangular operator S can always be associated with a Pick solution $R(t)$ of a time-variant Lyapunov (or displacement) equation. This result is actually a variation of the commutant lifting theorem of Sarason–Sz. Nagy–Foias and many other formulations have been considered in the literature (see, e.g., [2], [13], [16], [20]–[22]). Under supplementary finite-dimensionality and nondegeneracy conditions, we further derive a so-called generalized Schur algorithm and discuss an associated system-theoretic interpretation in terms of a cascade of elementary sections with intrinsic blocking properties. These considerations lead to a constructive proof of the previous result about displacement equations. Several classical algorithms proposed in the literature for the solution of interpolation problems, such as Schur, Nevanlinna–Pick, and extensions thereof, follow as special cases of the general framework presented here. We also extend the content of our companion paper [26] where several other applications of the algorithm are presented.

The paper is organized as follows. In §2 we introduce our notation and define the class of time-variant structured matrices. We also prove the main result concerning the existence of an upper-triangular operator S in connection with Pick solutions of time-variant Lyapunov equations. In §3 we show that several moment, interpolation, and completion problems, and extensions thereof, follow as special cases of the main

* Received by the editors September 24, 1992; accepted for publication by P. Van Dooren, October 13, 1993. This work was supported in part by Army Research Office grant DAAH04-93-G-0029, the Advanced Research Projects Agency of the Department of Defense and was monitored by the Air Force Office of Scientific Research grant F49620-93-1-0085.

† Information Systems Laboratory, Stanford University, Stanford, California, 94305. This work was performed while this author was with the Information Systems Laboratory, Stanford University, Stanford, California and on leave from the Institute of Mathematics, Bucharest, Romania. He is currently with the Programs in Mathematical Sciences, University of Texas at Dallas, Richardson, Texas 75083.

‡ Information Systems Laboratory, Stanford University, Stanford, California 94305. This work was performed while this author was a Research Associate in the Information Systems Laboratory, Stanford University, Stanford, California and on leave from Escola Politécnica da Universidade de São Paulo, Brazil. His work was also supported by a fellowship from FAPESP, Brazil. He is currently an Assistant Professor in the Department of Electrical and Computer Engineering, University of California, Santa Barbara, California 93106.

§ Information Systems Laboratory, Stanford University, Stanford, California 94305.

theorem of §2. In §4 we derive a computationally oriented recursive procedure that leads to a constructive realization of all possible solutions S in terms of a cascade of elementary sections with certain intrinsic blocking properties. In §5 we further elaborate on possible simplifications and describe the associated Schur parameters.

(An early account of the results of this paper was announced in [9]. We further remark that after submitting this paper, a closely related result to Theorem 2.2 was independently derived in [11].)

2. Displacement structure and abstract interpolation. We first introduce our notation. The symbol $\mathbb{Z}$ denotes the set of integers, and for two Hilbert spaces $\mathcal{H}$ and $\mathcal{H}'$, we write $\mathcal{L}(\mathcal{H},\mathcal{H}')$ to denote the set of bounded linear operators acting from $\mathcal{H}$ into $\mathcal{H}'$. We further consider three families $\{\mathcal{U}(t),\mathcal{V}(t),\mathcal{R}(t)\}_{t\in\mathbb{Z}}$ of Hilbert spaces depending on the parameter $t\in\mathbb{Z}$, two families of bounded linear operators $G(t)\in\mathcal{L}(\mathcal{U}(t)\oplus\mathcal{V}(t),\mathcal{R}(t))$ and $F(t)\in\mathcal{L}(\mathcal{R}(t-1),\mathcal{R}(t))$, and we define the symmetry $J(t)=(I_{\mathcal{U}(t)}\oplus -I_{\mathcal{V}(t)})$ acting on $\mathcal{U}(t)\oplus\mathcal{V}(t)$, where $I_{\mathcal{U}(t)}$ denotes the identity operator on the space $\mathcal{U}(t)$. We partition $G(t)=\begin{bmatrix} U(t) & V(t)\end{bmatrix}$, where $U(t)\in\mathcal{L}(\mathcal{U}(t),\mathcal{R}(t))$ and $V(t)\in\mathcal{L}(\mathcal{V}(t),\mathcal{R}(t))$. We also use the symbol * to denote the adjoint operator and we write $F^*(t)=(F(t))^*$.

DEFINITION 2.1. *A family of operators $\{R(t)\in\mathcal{L}(\mathcal{R}(t))\}_{t\in\mathbb{Z}}$ is said to have a time-variant displacement structure with respect to $\{F(t),G(t)\}_{t\in\mathbb{Z}}$ if $\{R(t)\}_{t\in\mathbb{Z}}$ is uniformly bounded, viz., there exists $r>0$ such that $\|R(t)\|\le r$ for all $t\in\mathbb{Z}$, and $R(t)$ satisfies the time-variant Lyapunov (or displacement) equation*

$$R(t)-F(t)R(t-1)F^*(t)=G(t)J(t)G^*(t). \tag{1}$$

The cardinal number $r(t)=\dim\mathcal{U}(t)+\dim\mathcal{V}(t)$ is called the displacement rank of $R(t)$ in (1). We say that (1) has a Pick solution if $R(t)$ is positive-semidefinite for every $t\in\mathbb{Z}$.

(For more discussion on the application of time-variant structured matrices in adaptive filtering, matrix factorization, and interpolation problems; the reader is referred to the companion papers [25], [26], [29].)

We further introduce some assumptions that will guarantee the existence of a unique family with time-variant displacement structure with respect to a given set of generators $\{F(t),G(t)\}_{t\in\mathbb{Z}}$. To this effect, we consider the infinite (block) matrices

$$\mathbf{U}(t)=\begin{bmatrix}\dots & F(t)F(t-1)U(t-2) & F(t)U(t-1) & U(t)\end{bmatrix},$$
$$\mathbf{V}(t)=\begin{bmatrix}\dots & F(t)F(t-1)V(t-2) & F(t)V(t-1) & V(t)\end{bmatrix},$$

and assume that for each $t\in\mathbb{Z}$ and $h\in\mathcal{R}(t)$, we have

$$F^*(t-n)F^*(t-n+1)\dots F^*(t-1)F^*(t)h\ \to 0\ \text{ as }\ n\to\infty, \tag{2a}$$

(2b) $\mathbf{U}(t)$ and $\mathbf{V}(t)$ are well-defined bounded linear operators,

$$\mathbf{U}(t)\in\mathcal{L}(\bigoplus_{j\le t}\mathcal{U}(j),\mathcal{R}(t))\,,\qquad \mathbf{V}(t)\in\mathcal{L}(\bigoplus_{j\le t}\mathcal{V}(j),\mathcal{R}(t)).$$

(2c) $\{\mathbf{U}(t),\mathbf{V}(t)\}_{t\in\mathbb{Z}}$ are uniformly bounded families:

$$\exists\, c_u>0 \text{ and } c_v>0 \text{ such that } \|\mathbf{U}(t)\|\le c_u \text{ and } \|\mathbf{V}(t)\|\le c_v \text{ for all } t\in\mathbb{Z}.$$

The above assumptions imply that (1) has a unique uniformly bounded solution given by

$$R(t) = \mathbf{U}(t)\mathbf{U}^*(t) - \mathbf{V}(t)\mathbf{V}^*(t). \tag{3}$$

We should remark that assumptions (2a)–(2c) are automatically satisfied in some special, though frequent, cases such as the following.

1. $\{G(t)\}_{t\in\mathbb{Z}}$ is a uniformly bounded family, viz., there exists $c_g > 0$ such that $\|G(t)\| \leq c_g$, and $F(t) = 0$ for $|t|$ sufficiently large.

2. $\{G(t)\}_{t\in\mathbb{Z}}$ is a uniformly bounded family, viz., there exists $c_g > 0$ such that $\|G(t)\| \leq c_g$, and $\{F(t)\}_{t\in\mathbb{Z}}$ is a *stable* family, i.e., there exists $c_f > 0$ such that $\|F(t)\| \leq c_f < 1$ for all t.

3. $\{F(t)\}_{t\in\mathbb{Z}}$ is a uniformly bounded family and $G(t) = 0$ for $|t|$ sufficiently large.

The following result shows that the existence of a Pick solution of (1) is equivalent to the existence of an upper-triangular contraction relating $\mathbf{U}(t)$ and $\mathbf{V}(t)$, which will play a fundamental role in subsequent sections.

THEOREM 2.2. *The displacement equation* (1) *has a Pick solution* $R(t)$ *if and only if there exists an upper-triangular contraction* $S \in \mathcal{L}(\oplus_{t\in\mathbb{Z}}\mathcal{V}(t), \oplus_{t\in\mathbb{Z}}\mathcal{U}(t))$, $(\|S\| \leq 1)$, *such that*

$$\mathbf{V}(t) = \mathbf{U}(t)P_{\mathcal{U}}(t)S/\bigoplus_{j\leq t}\mathcal{V}(j) \quad \textit{for every } t \in \mathbb{Z}, \tag{4}$$

where $P_{\mathcal{U}}(t)$ *denotes the orthogonal projection of* $\oplus_{t\in\mathbb{Z}}\mathcal{U}(t)$ *onto* $\oplus_{j\leq t}\mathcal{U}(j)$.

Proof. One implication is immediate. If an upper-triangular contraction S exists such that (4) holds, then the solution given by (3) is a Pick solution. The converse is a consequence of a commutant lifting theorem. Thus assume (1) has a Pick solution. Then $R(t) = \mathbf{U}(t)\mathbf{U}^*(t) - \mathbf{V}(t)\mathbf{V}^*(t)$ are positive operators for all $t \in \mathbb{Z}$. Hence, there exist contractive operators $\bar{S}(t)$ (i.e., $\|\bar{S}(t)\| \leq 1$),

$$\bar{S}(t) \in \mathcal{L}(\bigoplus_{j\leq t}\mathcal{V}(j), \overline{\mathcal{R}(\mathbf{U}^*(t))}),$$

such that $\mathbf{V}(t) = \mathbf{U}(t)\bar{S}(t)$ for all $t \in \mathbb{Z}$, where $\overline{\mathcal{R}(\mathbf{U}^*(t))}$ denotes the closure of the range of $\mathbf{U}^*(t)$. Let us define, for every $t \in \mathbb{Z}$, the shift (or marking) operator $M_{\mathcal{U}}(t) : \oplus_{j\leq t-1}\mathcal{U}(j) \longrightarrow \oplus_{j\leq t}\mathcal{U}(j)$,

$$M_{\mathcal{U}}(t) = \begin{bmatrix} \ddots & \ddots & & \\ & \mathbf{0} & I & \\ & & \mathbf{0} & I \\ & & & \mathbf{0} \end{bmatrix}.$$

It is easy to check that for all $t \in \mathbb{Z}$, $\mathbf{U}(t)M_{\mathcal{U}}(t) = F(t)\mathbf{U}(t-1)$ and $\mathbf{V}(t)M_{\mathcal{V}}(t) = F(t)\mathbf{V}(t-1)$. Hence,

$$\begin{aligned} M_{\mathcal{V}}^*(t)\bar{S}^*(t)\mathbf{U}^*(t) &= M_{\mathcal{V}}^*(t)\mathbf{V}^*(t) = \mathbf{V}^*(t-1)F^*(t) \\ &= \bar{S}^*(t-1)\mathbf{U}^*(t-1)F^*(t) = \bar{S}^*(t-1)M_{\mathcal{U}}^*(t)\mathbf{U}^*(t). \end{aligned}$$

We now use the comutant lifting theorem of Sarason–Sz.Nagy–Foias [13], [23] in its "time-variant" formulation in [4]—actually we use a slight variation in [8]—to conclude that there exists a family $\{\hat{S}(t) \in \mathcal{L}(\oplus_{j\leq t}\mathcal{V}(j), \oplus_{j\leq t}\mathcal{U}(j))\}_{t\in\mathbb{Z}}$ of contractions,

with the properties: $\bar{S}^*(t) = \hat{S}^*(t)/\overline{\mathcal{R}(\mathbf{U}^*(t))}$, and $\hat{S}(t)M_{\mathcal{V}}(t) = M_{\mathcal{U}}(t)\hat{S}(t-1)$. This is a rather standard argument by now (see [21], the proof of Theorem VIII-2.2 in [13], or Theorem 5.C.4 in [16]). We then conclude from the last equality that there exists an upper-triangular contraction $S \in \mathcal{L}(\oplus_{t\in\mathbb{Z}}\mathcal{V}(t), \oplus_{t\in\mathbb{Z}}\mathcal{U}(t))$ such that $\bar{S}(t) = P_{\mathcal{U}}(t)S/\oplus_{j\le t}\mathcal{V}(j)$. This can be viewed as a time-variant version of Lemma V-3.5 in [31]. Consequently, S satisfies (4) and the proof is complete. □

We have thus shown that an upper-triangular contraction S can always be associated with a Pick solution of time-variant displacement equations of the form (1). This is a general result that includes, as special cases, solutions of several interpolation, completion, and moment problems considered in the literature. In fact, it will become clear throughout our discussion that the solutions of these problems correspond to determining the appropriate contraction S that is associated with the Pick solution $R(t)$ of (1) for specific choices of $F(t), G(t)$, and $J(t)$. Some examples to this effect are discussed in the next section. It should be noted though that the argument used in the above proof only assures the existence of S. It does not show how to construct such an S. Later in §4 we shall, however, describe a recursive algorithm that, under suitable finite-dimensionality conditions, leads to a constructive proof of Theorem 2.2.

3. Connections with completion problems. In this section we illustrate the application of Theorem 2.2 to the solution of some moment and completion problems (and extensions thereof).

3.1. A positive completion problem. We begin by considering the following moment-type problem. We fix a positive integer p and a family $\{\mathcal{E}(n)\}_{n\in\mathbb{Z}}$ of Hilbert spaces.

PROBLEM 3.1. Given a family $\{\tilde{Q}_{ij}/\ i,j\in\mathbb{Z},\ \ |j-i|\le p\}$ of operators such that $\tilde{Q}_{ij} = \tilde{Q}_{ji}^*$ and $\tilde{Q}_{ij} \in \mathcal{L}(\mathcal{E}(j),\mathcal{E}(i))$, it is required to find conditions for the existence of a positive definite kernel $M = [Q_{ij}]_{i,j\in\mathbb{Z}}$ such that for $i,j\in\mathbb{Z}$ and $|j-i|\le p$ we have $Q_{ij} = \tilde{Q}_{ij}$.

By a positive-definite kernel we mean an application $M = [Q_{ij}]_{i,j\in\mathbb{Z}}$ on $\mathbb{Z}\times\mathbb{Z}$ such that for $i,j\in\mathbb{Z}$ we have $Q_{ij}\in\mathcal{L}(\mathcal{E}(j),\mathcal{E}(i))$ and $\sum_{i,j=-n}^{n} < Q_{ij}h_j, h_i > \ge 0$, for every integer $n>0$ and every set of vectors $\{h_{-n}, h_{-n+1},\dots,h_n\}$, $h_k\in\mathcal{E}(k)$, $|k|\le n$. We show here how to solve the above problem by using Theorem 2.2 and connections with displacement structure theory.

We can assume, without loss of generality, that $\tilde{Q}_{ii} = I$ for all $i\in\mathbb{Z}$. We also define the Hilbert spaces

$$\mathcal{R}(t) = \bigoplus_{k=0}^{p}\mathcal{E}(-t+k), \mathcal{U}(t) = \mathcal{V}(t) = \mathcal{E}(-t),$$

and the operators

$$U(t) = \begin{bmatrix} I \\ \tilde{Q}_{-t+1,-t} \\ \tilde{Q}_{-t+2,-t} \\ \vdots \\ \tilde{Q}_{-t+p,-t} \end{bmatrix}, \qquad V(t) = \begin{bmatrix} \mathbf{0} \\ \tilde{Q}_{-t+1,-t} \\ \tilde{Q}_{-t+2,-t} \\ \vdots \\ \tilde{Q}_{-t+p,-t} \end{bmatrix}.$$

We further consider the operators $J(t) = (I_{\mathcal{U}(t)} \oplus -I_{\mathcal{V}(t)})$, $G(t) = \begin{bmatrix} U(t) & V(t) \end{bmatrix}$,

and

$$F(t) = \begin{bmatrix} \mathbf{0} & & & \\ I & \mathbf{0} & & \\ & \ddots & \ddots & \\ & & I & \mathbf{0} \end{bmatrix}. \tag{5}$$

The elements $\{F(t), G(t), J(t)\}_{t\in\mathbb{Z}}$, as defined above, specify a displacement structure of the form (1). Using the result of Theorem 2.2 we conclude the following.

THEOREM 3.2. *Problem* 3.1 *has solutions if and only if the displacement equation* (1), *associated with the data* $\{F(t), G(t), J(t)\}_{t\in\mathbb{Z}}$ *defined above, has a Pick solution.*

Proof. Using the defined operators $\{F(t), G(t), J(t)\}_{t\in\mathbb{Z}}$, we readily check that the solution $R(t)$ of the corresponding (1) can be written as

$$R(t) = \mathbf{U}(t)\mathbf{U}^*(t) - \mathbf{V}(t)\mathbf{V}^*(t) = \begin{bmatrix} I & Q_{-t,-t+1} & & Q_{-t,-t+p} \\ Q_{-t+1,-t} & I & & Q_{-t+1,-t+p} \\ \vdots & & \ddots & \vdots \\ Q_{-t+p,-t} & Q_{-t+p,-t+1} & \cdots & I \end{bmatrix}.$$

We thus conclude that if Problem 3.1 has positive-definite solutions M then $R(t)$ is a positive block matrix for all $t \in \mathbb{Z}$. Conversely, if (1) has a Pick solution $R(t)$ then, by Theorem 2.2, there exists an upper triangular contraction $S = [S_{ij}]_{i,j\in\mathbb{Z}} \in \mathcal{L}(\oplus_{t\in\mathbb{Z}}\mathcal{V}(t),\ \oplus_{t\in\mathbb{Z}}\mathcal{U}(t))$ such that $\mathbf{V}(t) = \mathbf{U}(t)P_{\mathcal{U}}(t)S/\oplus_{j\le t}\mathcal{V}(j)$. If we take the structure of $\mathbf{U}(t)$ and $\mathbf{V}(t)$ into account we then conclude that $S_{tt} = \mathbf{0}$ for all $t \in \mathbb{Z}$. We define $Q_{ij} = \sum_{k=j+1}^{i-1} Q_{i,k}S_{-k,-j} + S_{-i,-j}$ for $i > j$, $|j-i| > p$, $Q_{ij} = Q_{ji}^*$ for $i < j$, $|j-i| > p$, and $Q_{ij} = \tilde{Q}_{ij}$ for $|j-i| \le p$, and consider the kernel $M = [Q_{ij}]_{i,j\in\mathbb{Z}}$. We now check that M is indeed a positive definite solution of Problem 3.1. For this purpose, we consider a positive integer $N > p$ and define the operators

$$U_N(t) = \begin{bmatrix} I \\ Q_{-t+1,-t} \\ \vdots \\ Q_{-t+N,-t} \end{bmatrix}, \qquad V_N(t) = \begin{bmatrix} 0 \\ Q_{-t+1,-t} \\ \vdots \\ Q_{-t+N,-t} \end{bmatrix}.$$

Then

$$\begin{bmatrix} I & Q_{-t,-t+1} & & Q_{-t,-t+N} \\ Q_{-t+1,-t} & I & & Q_{-t+1,-t+N} \\ \vdots & & & \vdots \\ Q_{-t+N,-t} & Q_{-t+N,-t+1} & \cdots & I \end{bmatrix} = \mathbf{U}_N(t)\mathbf{U}_N^*(t) - \mathbf{V}_N(t)\mathbf{V}_N^*(t)$$

$$= \mathbf{U}_N(t)\left[I - S_tS_t^*\right]\mathbf{U}_N^*(t),$$

where $S_t = P_{\mathcal{U}}(t)S/\oplus_{j\le t}\mathcal{V}(j)$. Consequently, M is positive-definite. □

We further remark that the well-known trigonometric moment problem [1] corresponds to the special case $\tilde{Q}_{ij} = \tilde{Q}_{|j-i|}$ (i.e., the entries of the specified band exhibit a Toeplitz structure). The case $\mathcal{E}(n) = \mathbf{0}$, for $|n|$ large enough, was considered and solved in [12].

3.2. The Hermite–Fejér interpolation problem. The solution of the positive completion problem corresponds to the special choice of the operator $F(t)$ as in (5). The statement of Theorem 2.2, however, allows for other choices of $F(t)$. Indeed, choosing $F(t)$ in a general Jordan form leads to the solution of a general interpolation problem of the Hermite–Fejér type as we now elaborate.

We first introduce some notation. We consider three families of Hilbert spaces $\{\mathcal{U}(t), \mathcal{V}(t), \mathcal{F}\}_{t\in\mathbb{Z}}$ and bounded upper-triangular operator $\mathbf{T} \in \mathcal{L}(\oplus_{t\in\mathbb{Z}}\mathcal{V}(t) \oplus_{t\in\mathbb{Z}}\mathcal{U}(t))$, $\mathbf{T} = [T_{ij}]_{i,j\in\mathbb{Z}}$, $T_{ij} \in \mathcal{L}(\mathcal{U}(j), \mathcal{V}(i))$,

$$\mathbf{T} = \begin{bmatrix} \ddots & \ddots & & & & \\ & T_{-1,-1} & \boxed{T_{-1,0}} & T_{-1,1} & \cdots & \\ & & \boxed{T_{00}} & T_{01} & T_{02} & \cdots \\ & \mathbf{O} & & T_{11} & T_{12} & T_{13} \\ & & & & \ddots & \ddots \end{bmatrix},$$

where $\boxed{T_{00}}$ denotes the $(0,0)$ entry of $\mathbf{T}$. We further consider a *stable* family of operators $\{f(t) \in \mathcal{L}(\mathcal{F})\}_{t\in\mathbb{Z}}$, viz., there exists $c > 0$ such that $|f(t)| < c < 1$ for all $t \in \mathbb{Z}$. We also introduce the symmetric functions $s_k^{(n)}$ of n variables (taken k at a time). That is, $s_0^{(n)} = 1$ and

$$s_k^{(n)}(x_1, x_2, \ldots, x_n) = \sum_{1\le i_1<\cdots<i_k\le n} x_{i_1}x_{i_2}\ldots x_{i_k}.$$

For example, $n = 3$ corresponds to $s_0^{(3)} = 1$, $s_1^{(3)}(x_1, x_2, x_3) = x_1 + x_2 + x_3$, $s_2^{(3)}(x_1, x_2, x_3) = x_1x_2 + x_1x_3 + x_2x_3$, and $s_3^{(3)}(x_1, x_2, x_3) = x_1x_2x_3$.

For a uniformly bounded family of operators $\{u(t) \in \mathcal{L}(\mathcal{V}(t), \mathcal{F})\}_{t\in\mathbb{Z}}$, viz., there exists $\bar{c} > 0$ such that $\|u(t)\| < \bar{c}$ for all t, we define the operator $u(t) \bullet \mathbf{T}(f(t))$ as follows:

$$u(t) \bullet \mathbf{T}(f(t)) = u(t)T_{tt} + f(t)u(t-1)T_{t-1,t} + f(t)f(t-1)u(t-2)T_{t-2,t} + \cdots.$$

This corresponds to a time-variant tangential evaluation along the "direction" defined by $u(t)$. More generally, we define the operators (for $p \ge 0$)

$$u(t) \bullet \frac{1}{p!}\mathbf{T}^{(p)}(f(t)) \equiv \sum_{m=0}^{\infty} s_m^{(m+p)}\left[f(t), f(t-1), \ldots, f(t-m-p+1)\right] u(t-m-p)T_{t-m-p,t}\,.$$

We shall also use the compact notation $\begin{bmatrix} u_1(t) & u_2(t) \end{bmatrix} \bullet H_{\mathbf{T}}^2(f(t))$ to denote the operator $\begin{bmatrix} u_1(t) \bullet \mathbf{T}(f(t)) & u_1(t) \bullet \frac{1}{1!}\mathbf{T}^{(1)}(f(t)) + u_2(t) \bullet \mathbf{T}(f(t)) \end{bmatrix}$, which we also write as

$$\begin{bmatrix} u_1(t) & u_2(t) \end{bmatrix} \bullet \begin{bmatrix} \mathbf{T}(f(t)) & \frac{1}{1!}\mathbf{T}^{(1)}(f(t)) \\ & \mathbf{T}(f(t)) \end{bmatrix}.$$

More generally, we write $\begin{bmatrix} u_1(t) & u_2(t) & \ldots & u_r(t) \end{bmatrix} \bullet H_{\mathbf{T}}^r(f(t)) =$

$$\begin{bmatrix} u_1(t) & u_2(t) & \ldots & u_r(t) \end{bmatrix} \bullet \begin{bmatrix} \mathbf{T}(f(t)) & \frac{1}{1!}\mathbf{T}^{(1)}(f(t)) & \cdots & \frac{1}{(r-1)!}\mathbf{T}^{(r-1)}(f(t)) \\ & \mathbf{T}(f(t)) & \cdots & \frac{1}{(r-2)!}\mathbf{T}^{(r-2)}(f(t)) \\ & & \ddots & \vdots \\ & \mathbf{O} & & \frac{1}{1!}\mathbf{T}^{(1)}(f(t)) \\ & & \mathbf{T}(f(t)) & \mathbf{T}(f(t)) \end{bmatrix}.$$

To formulate the Hermite–Fejér problem, we again consider three families of Hilbert spaces $\{\mathcal{U}(t), \mathcal{V}(t), \mathcal{F}_i\}_{t\in\mathbb{Z}, 0\le i<m}$, and m stable families $\{\alpha_i(t) \in \mathcal{L}(\mathcal{F}(t))\}_{t\in\mathbb{Z}}$ for $i = 0, 1, \ldots, m-1$. We associate with each $\alpha_i(t)$ a positive integer $r_i \ge 1$ and uniformly bounded families of operators $\mathbf{a}_i(t)$ and $\mathbf{b}_i(t)$ partitioned as follows:

$$\mathbf{a}_i(t) = \begin{bmatrix} u_1^{(i)}(t) & u_2^{(i)}(t) & \ldots & u_{r_i}^{(i)}(t) \end{bmatrix}, \quad \mathbf{b}_i(t) = \begin{bmatrix} v_1^{(i)}(t) & v_2^{(i)}(t) & \ldots & v_{r_i}^{(i)}(t) \end{bmatrix},$$

where $u_j^{(i)}(t) \in \mathcal{L}(\mathcal{U}(t), \mathcal{F}_i)$ and $v_j^{(i)}(t) \in \mathcal{L}(\mathcal{V}(t), \mathcal{F}_i), j = 1, \ldots, r_i$, are uniformly bounded families of operators.

PROBLEM 3.3. Given m stable families $\{\alpha_i(t)\}$ with the associated uniformly bounded data $\mathbf{a}_i(t)$ and $\mathbf{b}_i(t)$, as described above, it is required to find necessary and sufficient conditions for the existence of upper-triangular contractive operators S ($\|S\|_\infty \le 1$) that satisfy

$$\mathbf{b}_i(t) = \mathbf{a}_i(t) \bullet H_S^{r_i}(\alpha_i(t)) \quad \textit{for} \quad 0 \le i \le m-1 \quad \textit{and} \quad t \in \mathbb{Z}. \tag{6}$$

The first step in the solution consists in constructing three operators $F(t), G(t)$, and $J(t)$ directly from the interpolation data: $F(t)$ contains the information relative to the operators $\{\alpha_i(t)\}$ and the dimensions $\{r_i\}$, $G(t)$ contains the information relative to the direction operators $\{\mathbf{a}_i(t), \mathbf{b}_i(t)\}$, and $J(t) = (I_{\mathcal{U}(t)} \oplus -I_{\mathcal{V}(t)})$. Define, for $i = 0, 1, \ldots, m-1$, $\mathcal{R}_i(t) = \mathcal{F}_i \oplus \mathcal{F}_i \oplus \cdots \oplus \mathcal{F}_i$ (r_i times), and

$$\mathcal{R}(t) = \bigoplus_{i=0}^{m-1} \mathcal{R}_i(t).$$

The operators $F(t)$ and $G(t)$ are then constructed as follows: we associate with each $\alpha_i(t)$ an operator in Jordan form $\bar{F}_i(t) \in \mathcal{L}(\mathcal{R}_i(t-1), \mathcal{R}_i(t))$ $(= \mathcal{L}(\mathcal{R}_i(t)))$, in this case),

$$\bar{F}_i(t) = \begin{bmatrix} \alpha_i(t) & & & \\ 1 & \alpha_i(t) & & \\ & \ddots & \ddots & \\ & & 1 & \alpha_i(t) \end{bmatrix},$$

and two operators $U_i(t)$ and $V_i(t)$, respectively, which are composed of the operators associated with $\alpha_i(t)$, viz.,

$$U_i(t) = \begin{bmatrix} u_1^{(i)}(t) \\ u_2^{(i)}(t) \\ \vdots \\ u_{r_i}^{(i)}(t) \end{bmatrix}, \qquad V_i(t) = \begin{bmatrix} v_1^{(i)}(t) \\ v_2^{(i)}(t) \\ \vdots \\ v_{r_i}^{(i)}(t) \end{bmatrix}.$$

Then $F(t) = \text{ diagonal } \{\bar{F}_0(t), \bar{F}_1(t), \ldots, \bar{F}_{m-1}(t)\}$ and

$$G(t) = \begin{bmatrix} U_0(t) & V_0(t) \\ U_1(t) & V_1(t) \\ \vdots & \vdots \\ U_{m-1}(t) & V_{m-1}(t) \end{bmatrix} \equiv \begin{bmatrix} \mathbf{U}(t) & \mathbf{V}(t) \end{bmatrix}.$$

We shall denote the diagonal entries of $F(t)$ by $\{f_i(t)\}_{i=0}^{n-1}$ (for example, $f_0(t) = f_1(t) = \cdots = f_{r_0-1}(t) = \alpha_0(t)$). We have thus specified all the elements of a displacement equation as in (1).

THEOREM 3.4. *The tangential Hermite–Fejér Problem* 3.3 *is solvable if and only if the displacement equation* (1), *associated with the interpolation data above, has a Pick solution.*

Proof. The result follows by showing that the interpolation conditions (6) follow from Theorem 2.2. The assumptions made in the statement of Problem 3.3 guarantee that conditions (2a)–(2c) are satisfied. If $R(t)$ is a Pick solution then there exists an upper-triangular contraction S that satisfies (4). Now, by comparing terms on both sides of (4) and by invoking the special constructions of $\{F(t), G(t)\}$ as above, we conclude that expression (4) can be rewritten as

$$\mathbf{b}_i(t) = \mathbf{a}_i(t) \bullet H_S^{r_i}(\alpha_i(t)), \qquad \text{for } i = 0, 1, \ldots, m-1,$$

which is the desired interpolation property (6). Conversely, assume there exists an interpolating solution S that satisfies (6). Then, by comparing terms on both sides of (6), we conclude that the ith entry of $\mathbf{U}(t)P_{\mathcal{U}}(t)S/\oplus_{j\le t}\mathcal{V}(j)$ is the ith entry of $\mathbf{V}(t)$. Hence, S satisfies $\mathbf{V}(t) = \mathbf{U}(t)P_{\mathcal{U}}(t)S/\oplus_{j\le t}\mathcal{V}(j)$ for every $t \in \mathbb{Z}$. Consequently, $R(t)$ is a Pick solution. □

3.3. A special case: The Carathéodory–Fejér problem. The Hermite-Fejér problem includes as a special case the following so-called Carathéodory–Fejér problem.

PROBLEM 3.5. Given families of Hilbert spaces $\{\mathcal{U}(t), \mathcal{V}(t)\}_{t\in\mathbb{Z}}$, and n families $\{\beta_i(t)\}$, $i = 0, 1, \ldots, n-1$, of operators $\beta_i(t) \in \mathcal{L}(\mathcal{V}(t), \mathcal{U}(t-n+1))$, it is required to find necessary and sufficient conditions for the existence of an upper-triangular contraction $S \in \mathcal{L}(\oplus_{t\in\mathbb{Z}}\mathcal{V}(t), \oplus_{t\in\mathbb{Z}}\mathcal{U}(t))$, $S = [S_{ij}]_{i,j\in\mathbb{Z}}$, such that for all $t \in \mathbb{Z}$ we have $S_{t-i,t} = \beta_i(t)$ for $i = 0, 1, \ldots, n-1$.

The classical Carathéodory–Fejér–Schur interpolation problem [1] corresponds to the special case $\beta_i(t) = \beta_i$ for all $t \in \mathbb{Z}$ and $i = 0, 1, \ldots, n-1$. Several other contractive completion problems, such as those considered in [4], also follow as special cases of Problem 3.5 by choosing $\beta_i(t) = 0$ for sufficiently large values of t.

To put the above problem into our framework, as described in the previous section, we construct the operators

$$U(t) = \begin{bmatrix} I \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}, \qquad V(t) = \begin{bmatrix} \beta_0(t) \\ \beta_1(t) \\ \vdots \\ \beta_{n-1}(t) \end{bmatrix},$$

as well as $J(t) = (I_{\mathcal{U}(t)} \oplus -I_{\mathcal{V}(t)})$, $G(t) = \begin{bmatrix} U(t) & V(t) \end{bmatrix}$, and

$$F(t) = \begin{bmatrix} \mathbf{0} & & & \\ I & \mathbf{0} & & \\ & \ddots & \ddots & \\ & & I & \mathbf{0} \end{bmatrix}. \tag{7}$$

COROLLARY 3.6. *Problem* 3.5 *has solutions if and only if* $\|M(t)\| \le 1$ *for all*

$t \in \mathbb{Z}$, where

$$M(t) = \begin{bmatrix} \beta_0(t) & & & \\ \beta_1(t) & \beta_0(t-1) & & \\ \vdots & & \ddots & \\ \beta_{n-1}(t) & \cdots & & \beta_0(t-n+1) \end{bmatrix}.$$

3.4. A special case: The Nevanlinna–Pick problem. The Hermite–Fejér problem also includes as a special case, the following so-called time-variant version of Nevanlinna–Pick introduced in [10] and further studied and extended in [5].

PROBLEM 3.7. Given families of Hilbert spaces $\{\mathcal{F}_i, \mathcal{U}(t), \mathcal{V}(t)\}_{t\in\mathbb{Z}}$, and n stable families of operators $\{\alpha_i(t)\}$, $i = 0, 1, \ldots, n-1$, $\alpha_i(t) \in \mathcal{L}(\mathcal{F}_i)$, with two uniformly bounded families of operators $\{u_i(t), v_i(t)\}_{t\in\mathbb{Z}}$, $i = 0, 1, \ldots, n-1$, $u_i(t) \in \mathcal{L}(\mathcal{U}(t), \mathcal{F}_i)$, $v_i(t) \in \mathcal{L}(\mathcal{V}(t), \mathcal{F}_i)$, it is required to find necessary and sufficient conditions for the existence of an upper-triangular contraction $S \in \mathcal{L}(\oplus_{t\in\mathbb{Z}} \mathcal{V}(t), \oplus_{t\in\mathbb{Z}} \mathcal{U}(t))$ such that for all $t \in \mathbb{Z}$ we have $u_i(t) \bullet S(\alpha_i(t)) = v_i(t)$, $\quad i = 0, 1, \ldots, n-1$.

The classical Nevanlinna–Pick problem [1] corresponds to the special case $\alpha_i(t) = \alpha_i$, $u_i(t) = u_i$, and $v_i(t) = v_i$ for all $t \in \mathbb{Z}$ and $i = 0, 1, \ldots, n-1$. Following the construction given in the previous section we get

$$F(t) = \begin{bmatrix} \alpha_0(t) & & & \\ & \alpha_1(t) & & \\ \mathrm{O} & & \ddots & \mathrm{O} \\ & & & \alpha_{n-1}(t) \end{bmatrix}, \qquad G(t) = \begin{bmatrix} u_0(t) & v_0(t) \\ u_1(t) & v_1(t) \\ \vdots & \vdots \\ u_{n-1}(t) & v_{n-1}(t) \end{bmatrix}.$$

COROLLARY 3.8. *The Nevanlinna–Pick Problem* 3.7 *has solutions if and only if*

$$\left[\{u_i(t)u_j^*(t) - v_i(t)v_j^*(t)\} \bullet N_{\alpha_j^*}(\alpha_i(t))\right]_{i,j=0}^{n-1} \geq 0 \quad \textit{for all } t \in \mathbb{Z},$$

where, for a stable family $\{\alpha(t)\}_{t\in\mathbb{Z}}$, the upper-triangular operator N_α is defined by $(N_\alpha)_{tt} = I$, and $(N_\alpha)_{t-j,t} = \alpha(t-j+1)\alpha(t-j+2)\ldots\alpha(t)$ for $j \geq 1$. (The stability of $\{\alpha(t)\}_{t\in\mathbb{Z}}$ assures that N_α is a well-defined bounded operator.)

4. A recursive solution. The examples considered in the previous sections reveal the significance of Theorem 2.2 in the solution of moment and interpolation problems. However, the result of Theorem 2.2 only provides an existence statement, i.e., it only assures the existence of an upper-triangular contraction S that satisfies (4). It does not show how to construct or find such an S. The several applications mentioned above though, motivate the need for an alternative route that would also lead to a construction of S. In this section, following the arguments in [24]–[27], we describe a recursive procedure that leads us to what we call a *generalized Schur algorithm*, and which allows us to provide an implementation for S in terms of a cascade of elementary sections. The derivation that follows, however, is only applicable to the case where the involved Hilbert spaces are finite-dimensional. More specifically, we consider displacement equations as in (1), viz.,

$$R(t) - F(t)R(t-1)F^*(t) = G(t)J(t)G^*(t),$$

where $R(t) \in \mathcal{L}(\mathcal{R}(t))$, $G(t) \in \mathcal{L}(\mathcal{U}(t) \oplus \mathcal{V}(t), \mathcal{R}(t))$, $F(t) \in \mathcal{L}(\mathcal{R}(t-1), \mathcal{R}(t))$, and $J(t) = (I_{\mathcal{U}(t)} \oplus -I_{\mathcal{V}(t)})$, and assume the following *finite-dimensionality* conditions.

(8a) There exists a finite positive integer n *such that* $\mathcal{R}(t) = \overset{n-1}{\underset{i=0}{\oplus}} \mathcal{R}_i(t)$, for all t;

(8b) $\dim \mathcal{R}_i(t)$ are all equal and finite for all $t \in \mathbb{Z}$ and $i = 0, 1, \ldots, n-1$;

(8c) $\dim \mathcal{U}(t) = p(t) < \infty$ and $\dim \mathcal{V}(t) = q(t) < \infty$ for all $t \in \mathbb{Z}$.

We further assume that the following are true.

(8d) $\{F(t)\}$ is a uniformly bounded family of lower triangular operators with stable families of diagonal entries $\{f_i(t)\}_{i=0}^{n-1}$;

(8e) $\{G(t)\}$ is a uniformly bounded family.

By condition (8a) we can write $R(t) = [r_{ij}(t)]_{i,j=0}^{n-1}$, with block entries $r_{ij}(t) \in \mathcal{L}(\mathcal{R}_j(t), \mathcal{R}_i(t))$.

4.1. A time-variant embedding relation. A major tool in our analysis is a so-called embedding result for displacement equations. This result was derived in [14] in the time-invariant case and further explored and discussed in [19]. Its relevance to rational interpolation problems was detailed in [24], [25], [28] and in connection with time-variant interpolation problems in [5], [25], [26]. Here we discuss the general time-variant case following the pattern developed in [25], [26].

For this purpose, we consider again the time-variant displacement equation (1) and, in addition, assume that $\{R(t)\}_{t\in\mathbb{Z}}$ is also uniformly bounded from below, viz.,

(8f) $\exists\ r_1 > 0$ such that $0 < r_1 I \leq R(t)$ for all $t \in \mathbb{Z}$.

THEOREM 4.1. *Suppose* (8a)–(8f) *hold, then there exist uniformly bounded families of operators* $\{H(t)\}_{t\in\mathbb{Z}}$ *and* $\{K(t)\}_{t\in\mathbb{Z}}$,

$$H(t) \in \mathcal{L}(\mathcal{R}(t-1), \mathcal{U}(t) \oplus \mathcal{V}(t)), \qquad K(t) \in \mathcal{L}(\mathcal{U}(t) \oplus \mathcal{V}(t)),$$

such that the following time-variant embedding relation is satisfied

$$\begin{bmatrix} F(t) & G(t) \\ H(t) & K(t) \end{bmatrix} \begin{bmatrix} R(t-1) & \mathbf{0} \\ \mathbf{0} & J(t) \end{bmatrix} \begin{bmatrix} F(t) & G(t) \\ H(t) & K(t) \end{bmatrix}^* = \begin{bmatrix} R(t) & \mathbf{0} \\ \mathbf{0} & J(t) \end{bmatrix}. \tag{9}$$

Proof. It is easy to check (as in [19], [24], [25]) that the following choices for $H(t)$ and $K(t)$ satisfy the embedding relation (we use the notation $\Theta^{-1}(t)$ to mean $(\Theta(t))^{-1}$):

$$H(t) = \Theta^{-1}(t) J(t) G^*(t) \left[R^{\frac{*}{2}}(t) - \tau(t) R^{\frac{*}{2}}(t-1) F^*(t)\right]^{-1} \left[\tau(t) R^{-\frac{1}{2}}(t-1) - R^{-\frac{1}{2}}(t) F(t)\right],$$

$$K(t) = \Theta^{-1}(t) \left\{ I - J(t) G^*(t) \left[R^{\frac{*}{2}}(t) - \tau(t) R^{\frac{*}{2}}(t-1) F^*(t)\right]^{-1} R^{-\frac{1}{2}}(t) G(t) \right\},$$

for an arbitrary $J(t)$-unitary operator $\Theta(t)$ and an arbitrary unitary operator $\tau(t)$, whenever the inverse of $R^{\frac{*}{2}}(t) - \tau(t) R^{\frac{*}{2}}(t-1) F^*(t)$ exists. Here, $R^{\frac{1}{2}}(t)$ denotes the operator defined by $R(t) = R^{\frac{1}{2}}(t) R^{\frac{*}{2}}(t)$. (The finite-dimensionality conditions

guarantee that it is always possible to choose a unitary matrix $\tau(t)$ so as to assure the invertibility of $R^{\frac{*}{2}}(t) - \tau(t)R^{\frac{*}{2}}(t-1)F^*(t)$.)

We now show that we can choose $\Theta(t)$ and $\tau(t)$ adequately so as to guarantee the uniform boundedness of the families $\{H(t), K(t)\}_{t\in\mathbb{Z}}$. By our hypothesis, there exist $r_1 > 0$ and $r_2 > 0$, independent of t, such that $0 < r_1 I \leq R(t) \leq r_2 I$ for all $t \in \mathbb{Z}$. It follows that we can always find $\tau(t)$ such that

$$\left[R^{*/2}(t) - \tau(t)R^{*/2}(t-1)F^*(t)\right]\left[R^{*/2}(t) - \tau(t)R^{*/2}(t-1)F^*(t)\right]^* \geq \epsilon I > 0,$$

for some $\epsilon > 0$. Indeed, define $A(t) = R^{\frac{1}{2}}(t-1)F^*(t)R^{-\frac{1}{2}}(t)$. If $A(t) = 0$ then the claim is obvious, otherwise write $A(t) = (A_1(t) \oplus \mathbf{0})$ with respect to the decompositions $\mathcal{R}(A^*(t)) \oplus \mathcal{K}\text{er}\ A(t)$ and $\mathcal{R}(A(t)) \oplus \mathcal{K}\text{er}\ A^*(t)$ of $\mathcal{R}(t)$ and $\mathcal{R}(t-1)$, respectively. We readily conclude that $A_1(t)$ is invertible. If we define $\tau(t) = (-A_1^*(t)\,[A_1(t)A_1^*(t)]^{-\frac{1}{2}} \oplus B(t))$, with respect to the above decompositions, and for an arbitrary unitary operator $B(t)$, then it follows that $[\tau^*(t) - A(t)]\,[\tau(t) - A^*(t)] \geq I$. Therefore, $R^{\frac{*}{2}}(t) - \tau(t)R^{\frac{*}{2}}(t-1)F^*(t)$ is invertible and the family

$$\left\{\left[R^{*/2}(t) - \tau(t)R^{*/2}(t-1)F^*(t)\right]^{-1}\right\}_{t\in\mathbb{Z}}$$

is uniformly bounded. Taking $\Theta(t) = I$ for all $t \in \mathbb{Z}$ leads to uniformly bounded families $\{H(t), K(t)\}_{t\in\mathbb{Z}}$. □

4.2. Generalized Schur algorithm. We now use the embedding result of Theorem 4.1 to derive a generalized Schur algorithm for block matrices $R(t) = [r_{ij}(t)]_{i,j=0}^{n-1}$ along the lines presented in [25]–[27]. More precisely, we focus on the time-variant displacement equation (1) and show that it allows the successive computation of the Schur complements of $R(t)$ to be reduced to a computationally efficient recursive procedure applied to the so-called generator matrix $G(t)$.

Let $R_i(t)$ denote the Schur complement of the leading $i \times i$ submatrix of $R(t)$. If $l_i(t)$ and $d_i(t)$ stand for the the first column and the $(0,0)$ entry of $R_i(t)$, respectively, then (the positive-definiteness of $R(t)$ guarantees $d_i(t) > 0$ for all i)

$$R_i(t) - l_i(t)d_i^{-1}(t)l_i^*(t) = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & R_{i+1}(t) \end{bmatrix} \equiv \tilde{R}_{i+1}(t). \tag{10}$$

Hence, starting with an $n \times n$ matrix $R(t)$ and performing n consecutive Schur complement steps, we obtain the triangular factorization of $R(t)$, viz.,

$$R(t) = l_0(t)d_0^{-1}(t)l_0^*(t) + \begin{bmatrix} 0 \\ l_1(t) \end{bmatrix} d_1^{-1}(t) \begin{bmatrix} 0 \\ l_1(t) \end{bmatrix}^* + \cdots = L(t)D^{-1}(t)L^*(t),$$

where $D(t) = \text{diag}\{d_0(t), \ldots, d_{n-1}(t)\}$ ($D^{-1}(t)$ stands for $(D(t))^{-1}$) and the (nonzero parts of the) columns of the lower triangular matrix $L(t)$ are $\{l_0(t), \ldots, l_{n-1}(t)\}$. The point, however, is that this procedure can be speeded up for matrices $R(t)$ that exhibit a time-variant displacement structure as in (1). In this case, the above (Gauss/Schur) reduction procedure can be shown to reduce to a recursion on the elements of the generator matrix $G(t)$. The computational advantage then follows from the fact that the column dimension of $G(t)$, viz., $r(t) = p(t) + q(t)$, is usually small when compared to the dimension of $R(t)$. The following theorem shows that the triangular factor at

time $(t-1)$, viz., $L(t-1)$, can be time-updated to $L(t)$ via a recursive procedure on $G(t)$.

THEOREM 4.2. *The Schur complements $R_i(t)$ are also structured with generator matrices $G_i(t)$, viz., $R_i(t) - F_i(t)R_i(t-1)F_i^*(t) = G_i(t)J(t)G_i^*(t)$, where $G_i(t)$ is a matrix that satisfies, along with $l_i(t)$, the following generator recursion: $G_0(t) = G(t)$, $F_0(t) = F(t)$,*

$$\begin{bmatrix} l_i(t) & \mathbf{0} \\ & G_{i+1}(t) \end{bmatrix} = \begin{bmatrix} F_i(t)l_i(t-1) & G_i(t) \end{bmatrix} \begin{bmatrix} f_i^*(t) & h_i^*(t)J(t) \\ J(t)g_i^*(t) & J(t)k_i^*(t)J(t) \end{bmatrix},$$

where $g_i(t)$ is the top (block) row of $G_i(t)$, $F_i(t)$ is the submatrix obtained after deleting the first (block) row and column of $F_{i-1}(t)$, and $h_i(t)$ and $k_i(t)$ are arbitrary matrices chosen so as to satisfy the time-variant embedding relation

$$\begin{bmatrix} f_i(t) & g_i(t) \\ h_i(t) & k_i(t) \end{bmatrix} \begin{bmatrix} d_i(t-1) & \mathbf{0} \\ \mathbf{0} & J(t) \end{bmatrix} \begin{bmatrix} f_i(t) & g_i(t) \\ h_i(t) & k_i(t) \end{bmatrix}^* = \begin{bmatrix} d_i(t) & \mathbf{0} \\ \mathbf{0} & J(t) \end{bmatrix}, \tag{11}$$

where $d_i(t)$ satisfies the time-update $d_i(t) = f_i(t)d_i(t-1)f_i^*(t) + g_i(t)J(t)g_i^*(t)$.

Proof. We prove the result for $i = 0$. The same argument is valid for $i \geq 1$. Let $d_0(t), l_0(t)$, and $g_0(t)$, denote the $(0,0)$ (block) entry of $R(t)$, the first (block) column of $R(t)$, and the first (block) row of $G(t)$, respectively. It then follows from the displacement equation (1) that $l_0(t) = F(t)l_0(t-1)f_0^*(t) + G(t)J(t)g_0^*(t)$ and $d_0(t) = f_0(t)d_0(t-1)f_0^*(t) + g_0(t)J(t)g_0^*(t)$. Let $F_1(t)$ be the submatrix obtained after deleting the first (block) row and column of $F(t)$. Using the expressions for $l_0(t), d_0(t)$, and (10), it is straightforward to check that we can write $\tilde{R}_1(t) - F(t)\tilde{R}_1(t-1)F^*(t)$

$$\begin{aligned} &= G(t)J(t)\left\{J(t) - g_0^*(t)d_0^{-1}(t)g_0(t)\right\}J(t)G^*(t) \\ &\quad - F(t)l_0(t-1)f_0^*(t)d_0^{-1}(t)g_0(t)J(t)G^*(t) \\ &\quad - G(t)J(t)g_0^*(t)d_0^{-1}(t)f_0(t)l_0^*(t-1)F^*(t) \\ &\quad - F(t)l_0(t-1)\left[d_0^{-1}(t-1) - f_0^*(t)d_0^{-1}(t)f_0(t)\right]l_0^*(t-1)F^*(t). \end{aligned} \tag{12a}$$

We now verify that the right-hand side of the above expression can be put into the form of a *perfect square* by introducing some auxiliary quantities. Consider a (block) column vector $h_0(t)$ and a matrix $k_0(t)$ that are defined to satisfy the following relations (in terms of the quantities that appear on the right-hand side of the above expression, this is always possible as explained ahead)

$$h_0^*(t)J(t)h_0(t) = d_0^{-1}(t-1) - f_0^*(t)d_0^{-1}(t)f_0(t)\,, \tag{12b}$$

$$k_0^*(t)J(t)k_0(t) = J(t) - g_0^*(t)d_0^{-1}(t)g_0(t), \qquad k_0^*(t)J(t)h_0(t) = g_0^*(t)d_0^{-1}(t)f_0(t)\,.$$

Using $\{h_0(t), k_0(t)\}$, we can factor the right-hand side of (12a) as $\tilde{G}_1(t)J(t)\tilde{G}_1^*(t)$, where $\tilde{G}_1(t) = F(t)l_0(t-1)h_0^*(t)J(t) + G(t)J(t)k_0^*(t)J(t)$. Recall that the first (block) row and column of $\tilde{R}_1(t)$ are zero. Hence, the first (block) row of $\tilde{G}_1(t)$ is zero, $\tilde{G}_1(t) = \begin{bmatrix} \mathbf{0} & G_1^{\mathbf{T}}(t) \end{bmatrix}^{\mathbf{T}}$. Moreover, it follows from (12b) (and from the expression for $d_0(t)$) that

$$\begin{bmatrix} f_0(t) & g_0(t) \\ h_0(t) & k_0(t) \end{bmatrix}^* \begin{bmatrix} d_0^{-1}(t) & \mathbf{0} \\ \mathbf{0} & J(t) \end{bmatrix} \begin{bmatrix} f_0(t) & g_0(t) \\ h_0(t) & k_0(t) \end{bmatrix} = \begin{bmatrix} d_0^{-1}(t-1) & \mathbf{0} \\ \mathbf{0} & J(t) \end{bmatrix},$$

which is equivalent to (11) for $i = 0$. □

The existence of uniformly bounded families $\{h_i(t), k_i(t)\}_{t\in\mathbb{Z}}$ that satisfy (11) follow as a special case of Theorem 4.1, since $d_i(t)$ satisfies a time-variant displacement equation, viz.,

$$d_i(t) = f_i(t)d_i(t-1)f_i^*(t) + g_i(t)J(t)g_i^*(t),$$

the finite-dimensionality conditions stated prior to Theorem 4.1 are satisfied, and the families $\{d_i(t), g_i(t)\}_{t\in\mathbb{Z}}$ are uniformly bounded as shown next.

LEMMA 4.3. *The sequences $\{d_i(t)\}_{t\in\mathbb{Z}}$ and $\{g_i(t)\}_{t\in\mathbb{Z}}$ obtained through the recursive Schur reduction procedure are uniformly bounded. More specifically, there exist real numbers b_d, c_d, and c_v (independent of t) such that*

$$0 < b_d I < d_i(t) < c_d I, \qquad \|g_i(t)\| < c_v \quad \textit{for all } t.$$

Proof. It is clear that $\{d_0(t)\}_{t\in\mathbb{Z}}$ is uniformly bounded from above since $\{f_0(t)\}_{t\in\mathbb{Z}}$ is stable and $\{g_0(t)J(t)g_0^*(t)\}_{t\in\mathbb{Z}}$ is uniformly bounded. A similar argument shows that $\{l_0(t)\}_{t\in\mathbb{Z}}$ is also uniformly bounded. It further follows from $0 < r_1 I \le R(t)$ that the sequence $\{d_0(t)\}_{t\in\mathbb{Z}}$ is uniformly bounded from below, viz., $d_0(t) \ge r_1 I > 0$ for all t. Hence, by Theorem 4.1, we can always choose uniformly bounded sequences $\{h_0(t)\}_{t\in\mathbb{Z}}$ and $\{k_0(t)\}_{t\in\mathbb{Z}}$ so as to satisfy the embedding relation (11). From the generator recursion we get $g_1(t) = e_1F(t)l_0(t-1)h_0^*(t)J(t) + e_1G(t)J(t)k_0^*(t)J(t)$. It then follows that $\{g_1(t)\}_{t\in\mathbb{Z}}$ is also uniformly bounded. Repeating this argument we conclude, by induction, that there exist real numbers $c_d > 0$ and $c_v > 0$ such that $d_i(t) < c_d I$ and $\|g_i(t)\| < c_v$ for all $t \in \mathbb{Z}$.

To show that the sequence $\{d_i(t)\}_{t\in\mathbb{Z}}$ is uniformly bounded from *below*, we use the fact that the successive Schur complements $R_i(t)$ also satisfy relations similar to (1). For this purpose, we rewrite each step of the Schur reduction procedure (10) in the form

(13)

$$R_i(t) = \begin{bmatrix} l_i(t)d_i^{-1}(t) & \mathbf{0} \\ \mathbf{0} & I_{n-i-1} \end{bmatrix} \begin{bmatrix} d_i(t) & \mathbf{0} \\ \mathbf{0} & R_{i+1}(t) \end{bmatrix} \begin{bmatrix} l_i(t)d_i^{-1}(t) & \mathbf{0} \\ \mathbf{0} & I_{n-i-1} \end{bmatrix}^*,$$

which exhibits a congruence relation. We define, for notational simplicity,

$$A_i(t) \;\equiv\; \begin{bmatrix} l_i(t)d_i^{-1}(t) & \mathbf{0} \\ \mathbf{0} & I_{n-i-1} \end{bmatrix},$$

which is an invertible lower triangular matrix. Assume that $R_i(t) > \epsilon_i I$ for some $\epsilon_i > 0$ independent of t ($\epsilon_0 = r_1$ since $0 < r_1 I \le R(t)$). Then clearly $d_i(t) > \epsilon_i I$ and $A_i(t)$ is uniformly bounded. For any nonzero column vector $\mathbf{y}$, we can always write $\mathbf{y} = A_i^*(t)\mathbf{x}$ for some nonzero column vector $\mathbf{x}$, since $A_i(t)$ has full rank. Therefore,

$$\begin{aligned} \mathbf{y}^* \begin{bmatrix} d_i(t) & \mathbf{0} \\ \mathbf{0} & R_{i+1}(t) \end{bmatrix} \mathbf{y} = \mathbf{x}^* A_i(t) \begin{bmatrix} d_i(t) & \mathbf{0} \\ \mathbf{0} & R_{i+1}(t) \end{bmatrix} A_i^*(t)\mathbf{x} \;&=\; \mathbf{x}^* R_i(t)\mathbf{x} \\ > \epsilon_i\|\mathbf{x}\|^2 \;=\; \epsilon_i\|A_i^{-*}(t)\mathbf{y}\|^2 \;&\equiv\; \epsilon_{i+1}\|\mathbf{y}\|^2, \end{aligned}$$

where in the last equality we defined ϵ_{i+1} and used the fact that $\{A_i^{-1}(t)\}_{t\in\mathbb{Z}}$ is uniformly bounded. Consequently, $d_{i+1}(t) > \epsilon_{i+1}$ and we can choose $b_d = \min_{0\le i\le n-1}\epsilon_i$. We thus conclude that $\{d_i(t)\}_{t\in\mathbb{Z}}$ is uniformly bounded from below.

We finally remark that we can also conversely show that if $\{d_i(t)\}_{t\in\mathbb{Z}}$ is uniformly bounded from below, then $\{R(t)\}_{t\in\mathbb{Z}}$ is also uniformly bounded from below. To see this, we apply the same argument and use (14) backwards starting with $R_{n-1}(t) = d_{n-1}(t)$ down to $R_0(t) = R(t)$. □

4.3. Recursive construction of S. The question now is: How does the recursive algorithm in Theorem 4.2 relate to the result of Theorem 2.2? The relevant fact to note here is that each recursive step gives rise to a linear discrete-time system (in state-space form)

$$\begin{bmatrix} f_i^*(t) & h_i^*(t)J(t) \\ J(t)g_i^*(t) & J(t)k_i^*(t)J(t) \end{bmatrix},$$

which appears on the right-hand side of the generator recursion in Theorem 4.2. This can be thought of as the (state-space) transition matrix of a linear system as follows

$$(14) \qquad \begin{bmatrix} \mathbf{x}_i(t+1) & \mathbf{y}_i(t) \end{bmatrix} = \begin{bmatrix} \mathbf{x}_i(t) & \mathbf{w}_i(t) \end{bmatrix} \begin{bmatrix} f_i^*(t) & h_i^*(t)J(t) \\ J(t)g_i^*(t) & J(t)k_i^*(t)J(t) \end{bmatrix},$$

where $\mathbf{x}_i(t)$ denotes the state, $\mathbf{w}_i(t)$ denotes an input vector, and $\mathbf{y}_i(t)$ denotes an output vector at time t.

The second important observation, which we shall verify very soon, is that each such section exhibits an intrinsic blocking property. The cascade of n sections would then exhibit certain global blocking properties, which will be shown to be equivalent to the desired result (4). Interesting enough, these blocking properties simply follow from the fact that each step of the Schur reduction procedure yields a matrix with a new zero row and column (as in (10)), which translates to a generator matrix with a new zero row as in Theorem 4.2.

4.3.1. Properties of the first-order sections. Let

$$T_i = \begin{bmatrix} T_{lj}^{(i)} \end{bmatrix}_{l,j=-\infty}^{\infty}$$

denote the upper-triangular transfer operator associated with (14), where the $\{T_{lj}^{(i)}\}$ denote the time-variant Markov parameters of T_i and are given by

$$T_{ll}^{(i)} = J(l)k_i^*(l)J(l), \qquad T_{l,l+1}^{(i)} = J(l)g_i^*(l)h_i^*(l+1)J(l+1),$$
$$T_{lj}^{(i)} = J(l)g_i^*(l)f_i^*(l+1)f_i^*(l+2)\dots f_i^*(j-1)h_i^*(j)J(j) \quad \text{for } j > l+1.$$

The output and input sequences of T_i are clearly related by

$$\begin{bmatrix} \dots & \mathbf{y}_i(-1) & \boxed{\mathbf{y}_i(0)} & \mathbf{y}_i(1) & \dots \end{bmatrix} = \begin{bmatrix} \dots & \mathbf{w}_i(-1) & \boxed{\mathbf{w}_i(0)} & \mathbf{w}_i(1) & \dots \end{bmatrix} T_i.$$

After n recursive steps (recall that $G(t)$ has n rows) we obtain a cascade of sections $\mathbf{T} = T_0T_1 \dots T_{n-1}$, which may be regarded as a generalized transmission line.

LEMMA 4.4. *Each first-order section T_i is a bounded upper-triangular linear operator.*

Proof. We already know that $\{f_i(t)\}_{t\in\mathbb{Z}}$ and $g_i(t)\}_{t\in\mathbb{Z}}$ are stable and uniformly bounded sequences, respectively, and that $\{h_i(t), k_i(t)\}_{t\in\mathbb{Z}}$ can always be chosen to be uniformly bounded sequences as well. It is then a standard result that the corresponding transfer operator T_i is bounded (see, e.g., [15]). □

Moreover, if we define the direct sum $\mathbf{J} = \oplus_{t\in\mathbb{Z}} J(t)$, it then follows that each T_i also satisfies the following $\mathbf{J}$-losslessness property.

LEMMA 4.5. *Each first-order section T_i satisfies $T_i\mathbf{J}T_i^* = \mathbf{J}$ and $T_i^*\mathbf{J}T_i = \mathbf{J}$.*

Proof. The proof is a direct consequence of the embedding construction (11), which leads to the relations

$$\begin{aligned} f_i^*(t)d_i^{-1}(t)f_i(t) + h_i^*(t)J(t)h_i(t) &= d_i^{-1}(t-1). \\ f_i^*(t)d_i^{-1}(t)g_i(t) + h_i^*(t)J(t)k_i(t) &= \mathbf{0}. \\ g_i^*(t)d_i^{-1}(t)g_i(t) + k_i^*(t)J(t)k_i(t) &= J(t). \end{aligned}$$

Therefore, we can expand $d_i^{-1}(t)$ and write

$$\begin{aligned} d_i^{-1}(t) = {} & h_i^*(t+1)J(t+1)h_i(t+1) \\ & + f_i^*(t+1)h_i^*(t+2)J(t+2)h_i(t+2)f_i(t+1) \\ & + f_i^*(t+1)f_i^*(t+2)h_i^*(t+3)J(t+3)h_i(t+3)f_i(t+2)f_i(t+1) + \cdots. \end{aligned}$$

Now the tth element on the main diagonal of $T_i\mathbf{J}T_i^*$ (denoted by λ_{tt}) is given by

$$\begin{aligned} \lambda_{tt} = {} & J(t)[k_i^*(t)J(t)k_i(t) + g_i^*(t)h_i^*(t+1)J(t+1)h_i(t+1)g_i(t) \\ & + g_i^*(t)f_i^*(t+1)h_i^*(t+2)J(t+2)h_i(t+2)f_i(t+1)g_i(t) + \cdots]J(t). \end{aligned}$$

Using the expression for $d_i^{-1}(t)$, we obtain

$$\lambda_{tt} = J(t) - J(t)g_i^*(t)\left[d_i^{-1}(t) - d_i^{-1}(t)\right]g_i(t)J(t) = J(t).$$

The same argument can be used to show that the off-diagonal elements of $T_i\mathbf{J}T_i^*$ are zero. For proving that $T_i^*\mathbf{J}T_i = \mathbf{J}$ we use a similar procedure. □

Furthermore, each section T_i satisfies an important blocking property in the following sense.

THEOREM 4.6. *Each first-order section T_i satisfies*

$$\begin{bmatrix} \cdots & f_i(t)f_i(t-1)g_i(t-2) & f_i(t)g_i(t-1) & g_i(t) & ? \end{bmatrix} T_i = \begin{bmatrix} \mathbf{0} & ? \end{bmatrix},$$

where $g_i(t)$ is at the tth position of the row vector. Consequently, $g_i(t) \bullet T_i(f_i(t)) = \mathbf{0}$.

Proof. This follows directly from the embedding result (11) (as well as from the fact that each step of the generator recursion in Theorem 2.2 produces a new zero row). The output of T_i at time t is given by

$$\begin{aligned} \mathbf{y}_i(t) &= \cdots + f_i(t)f_i(t-1)g_i(t-2)T_{t-2,t} + f_i(t)g_i(t-1)T_{t-1,t} + g_i(t)T_{tt} \\ &= [-d_i(t-1) + d_i(t-1)]\, f_i(t)h_i^*(t)J(t) = \mathbf{0}, \end{aligned}$$

where we substituted the expressions for the Markov parameters $\{T_{jt}\}_{j \le t}$ and used

$$\begin{aligned} d_i(t) = {} & g_i(t)J(t)g_i^*(t) + f_i(t)g_i(t-1)J(t-1)g_i^*(t-1)f_i^*(t) \\ & + f_i(t)f_i(t-1)g_i(t-2)J(t-2)g_i^*(t-2)f_i^*(t-1)f_i^*(t) + \cdots. \end{aligned}$$

The same argument holds for the previous outputs. □

In general terms, the blocking property means that when $g_i(t)$ (which is the first row of $G_i(t)$) is applied to T_i we obtain a zero output at $f_i(t)$ at time t. We say that $f_i(t)$ is a time-variant *transmission-zero* of T_i and $g_i(t)$ is the associated time-variant *left-zero direction.* We remark that the concepts of transmission zeros and blocking directions are central to many problems in network theory and linear systems [17].

We can now put together the two main pieces proved so far: the Schur reduction procedure and the blocking properties of the elementary sections. This leads to

the following constructive proof of Theorem 2.2, assuming finite-dimensional spaces $\{\mathcal{R}(t),\mathcal{U}(t),\mathcal{V}(t)\}_{t\in\mathbb{Z}}$ and the supplementary nondegeneracy condition $\mathbf{U}(t)\mathbf{U}^*(t) \geq \mu > 0$ for all $t \in \mathbb{Z}$, where μ is a fixed constant.

THEOREM 4.7. *Assuming finite-dimensional spaces $\{\mathcal{R}(t),\mathcal{U}(t),\mathcal{V}(t)\}_{t\in\mathbb{Z}}$ and the nondegeneracy condition $\mathbf{U}(t)\mathbf{U}^*(t) \geq \mu > 0$ for all t, the time-variant displacement equation (1) has a Pick solution $R(t)$ such that $R(t) > \epsilon I > 0$ for a constant ϵ and for all t if and only if there exists an upper-triangular strict contraction S ($\|S\| < 1$), $S \in \mathcal{L}(\oplus_{t\in\mathbb{Z}}\mathcal{V}(t),\ \oplus_{t\in\mathbb{Z}}\mathcal{U}(t))$, such that*

$$\mathbf{V}(t) = \mathbf{U}(t)P_{\mathcal{U}}(t)S/\bigoplus_{j\leq t}\mathcal{V}(j) \quad \text{for every } t \in \mathbb{Z}. \tag{15}$$

Proof. One implication is immediate. We now give a constructive proof of the converse statement. So assume the displacement equation (1) has a Pick solution $R(t)$ such that $R(t) > \epsilon I > 0$ for a constant ϵ and for all t. Then applying the Schur reduction procedure (or the generalized Schur algorithm) to $\{F(t), G(t)\}$ leads to a cascade of elementary sections, viz., $\mathbf{T} = T_0T_1 \dots T_{n-1}$. Following an argument similar to that presented in [19] for the time-invariant case and in [24], [26] for the time-variant case, we readily conclude that the entire cascade admits the following state-space description:

$$\begin{bmatrix} \mathbf{x}(t+1) & \mathbf{y}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{x}(t) & \mathbf{w}(t) \end{bmatrix} \begin{bmatrix} F^*(t) & H^*(t)J(t) \\ J(t)G^*(t) & J(t)K^*(t)J(t) \end{bmatrix},$$

where $\{H(t), K(t)\}_{t\in\mathbb{Z}}$ are, due to our assumptions, uniformly bounded operators that satisfy the embedding relation

$$\begin{bmatrix} F(t) & G(t) \\ H(t) & K(t) \end{bmatrix} \begin{bmatrix} R(t-1) & \mathbf{0} \\ \mathbf{0} & J(t) \end{bmatrix} \begin{bmatrix} F(t) & G(t) \\ H(t) & K(t) \end{bmatrix}^* = \begin{bmatrix} R(t) & \mathbf{0} \\ \mathbf{0} & J(t) \end{bmatrix}. \tag{16}$$

Moreover, it follows from the blocking properties of the sections T_i that the entire cascade $\mathbf{T}$ satisfies the global blocking relation

$$\begin{bmatrix} \dots F(t)F(t-1)G(t-2) & F(t)G(t-1) & G(t) & \mathbf{0} \dots \end{bmatrix} \mathbf{T} = \begin{bmatrix} \mathbf{0} & ? \end{bmatrix}, \tag{17}$$

where $G(t)$ appears in the tth position.

We further partition the matrix entries T_{lj} of the cascade $\mathbf{T}$ accordingly with $J(l)$ and $J(j)$,

$$T_{lj} = \begin{bmatrix} T_{11}^{lj} & T_{12}^{lj} \\ T_{21}^{lj} & T_{22}^{lj} \end{bmatrix},$$

and consider the triangular operators

$$\mathbf{T}_{11} = \left[T_{11}^{lj}\right], \quad \mathbf{T}_{21} = \left[T_{21}^{lj}\right], \quad \mathbf{T}_{12} = \left[T_{12}^{lj}\right], \quad \mathbf{T}_{22} = \left[T_{22}^{lj}\right] \quad \text{for} \quad -\infty < l, j < \infty.$$

We now verify that $\mathbf{T}_{22}^{-1}$ is an upper-triangular and bounded operator and that $\mathbf{T}_{12}\mathbf{T}_{22}^{-1}$ is a strictly contractive upper-triangular operator, such that

$$\mathbf{V}(t) = -\mathbf{U}(t)P_{\mathcal{U}}(t)\mathbf{T}_{12}\mathbf{T}_{22}^{-1}/\bigoplus_{j\leq t}\mathcal{V}(j) \quad \text{for all } t \in \mathbb{Z}.$$

For this purpose, note that it follows from the $\mathbf{J}$-losslessness property of $\mathbf{T}$ that

$$\mathbf{T}_{22}\mathbf{T}_{22}^* \geq I, \qquad \mathbf{T}_{22}^*\mathbf{T}_{22} \geq I. \tag{18}$$

Hence, $\mathbf{T}_{22}$ is invertible and $\|\mathbf{T}_{22}^{-1}\| \leq 1$. Define $X(t) = P_{\mathcal{V}}(t)\mathbf{T}_{22}/\oplus_{j\leq t}\mathcal{V}(j)$, then it follows from (18) that $X^*(t)X(t) \geq I$. Define $\mathbf{T}(t) = P_{\mathcal{U}\oplus\mathcal{V}}(t)\mathbf{T}/\oplus_{j\leq t}\mathcal{U}(j)\oplus\mathcal{V}(j)$ and $J_t = \oplus_{j\leq t} J(j)$. It follows from the embedding relation (16) that

$$J_t - \mathbf{T}(t)J_t\mathbf{T}^*(t) = \begin{bmatrix} \dots & F(t)G(t-1) & G(t) \end{bmatrix}^* R^{-1}(t) \begin{bmatrix} \dots & F(t)G(t-1) & G(t) \end{bmatrix} \geq 0.$$

Hence, $X(t)X^*(t) \geq I$ and we conclude that $X(t)$ is invertible for every $t \in \mathbb{Z}$ and that the family $\{X^{-1}(t)\}_{t\in\mathbb{Z}}$ is uniformly bounded by one. Define the following operators (acting on the same space as $\mathbf{T}_{22}$),

$$\tilde{X}(t) = \begin{bmatrix} X(t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Then $\tilde{X}(t+1)$ and $\tilde{X}(t)$ satisfy the following nested property (they differ by just one block column)

$$\tilde{X}(t+1) = \begin{bmatrix} \tilde{X}(t) & ? \\ \mathbf{0} & ? \end{bmatrix}, \tag{19}$$

where ? denotes irrelevant entries. Hence, $\{\tilde{X}(t)\}_{t\in\mathbb{Z}}$ strongly converges to a bounded operator $\tilde{X}$ as $t \to \infty$. It is easily checked that $\tilde{X}$ is upper-triangular and that it actually coincides with $\mathbf{T}_{22}^{-1}$.

The fact that $\mathbf{T}_{12}\mathbf{T}_{22}^{-1} \in \mathcal{L}(\oplus_{t\in\mathbb{Z}}\mathcal{V}(t),\ \oplus_{t\in\mathbb{Z}}\mathcal{U}(t))$, is an upper-triangular strictly contractive operator is a consequence of the $\mathbf{J}$-losslessness of $\mathbf{T}$. We thus conclude that $S = -\mathbf{T}_{12}\mathbf{T}_{22}^{-1} \in \mathcal{L}(\oplus_{t\in\mathbb{Z}}\mathcal{V}(t),\ \oplus_{t\in\mathbb{Z}}\mathcal{U}(t))$ is a strictly contractive upper-triangular operator that satisfies (15). □

Remark. The above argument is based on a recursive construction of $\mathbf{T}$. We can also give a direct (nonrecursive) proof of the same result as follows: First prove the embedding relation (16) as in Theorem 4.1 and the blocking property (17) as in Theorem 4.6. We then conclude the argument as above.

4.3.2. Parametrization of all solutions. We now show how to parametrize all solutions S that satisfy (15).

THEOREM 4.8. *Assuming finite-dimensional spaces $\{\mathcal{R}(t),\mathcal{U}(t),\mathcal{V}(t)\}_{t\in\mathbb{Z}}$ and the nondegeneracy condition $\mathbf{U}(t)\mathbf{U}^*(t) \geq \mu > 0$, for all t, and that the displacement equation (1) has a Pick solution $R(t)$ such that $R(t) > \epsilon I > 0$ for a constant ϵ and for all t. Then all strictly contractive upper-triangular solutions $S \in \mathcal{L}(\oplus_{t\in\mathbb{Z}}\mathcal{V}(t),\ \oplus_{t\in\mathbb{Z}}\mathcal{U}(t))$ are given by*

$$S = -\left[\mathbf{T}_{11}K + \mathbf{T}_{12}\right]\left[\mathbf{T}_{21}K + \mathbf{T}_{22}\right]^{-1},$$

for arbitrary upper-triangular contractive operators

$$K \in \mathcal{L}(\underset{t\in\mathbb{Z}}{\oplus}\mathcal{V}(t),\ \underset{t\in\mathbb{Z}}{\oplus}\mathcal{U}(t)) \quad \text{with} \quad \|\mathcal{K}\| < 1.$$

Proof. One implication is immediate. Consider a K as above. Since $\mathbf{T}_{22}^{-1}$ is a bounded upper-triangular operator, it follows that $S = -\left[\mathbf{T}_{11}K + \mathbf{T}_{12}\right]\left[\mathbf{T}_{21}K + \mathbf{T}_{22}\right]^{-1}$ is bounded upper-triangular and, using the $\mathbf{J}(t)$-losslessness of $\mathbf{T}$, we conclude that $\|S\| < 1$. Let $S_1 = \mathbf{T}_{11}K + \mathbf{T}_{12}$ and $S_2 = \mathbf{T}_{21}K + \mathbf{T}_{22}$. Then,

$$\begin{bmatrix} S_1 \\ S_2 \end{bmatrix} = \mathbf{T}\begin{bmatrix} K \\ I \end{bmatrix},$$

and, because of the blocking property of $\mathbf{T}$, we obtain that S is a solution of (15).

For the converse implication we follow the pattern developed in [6] and adapted for the time-variant Nevanlinna–Pick problem in [5]. Because our framework is more general, we indicate the necessary changes. For an upper-triangular operator $X \in \mathcal{L}(\mathcal{E}, \mathcal{K})$, where $\mathcal{E} = \oplus_{t\in\mathbb{Z}} \mathcal{E}(t)$ and $\mathcal{K} = \oplus_{t\in\mathbb{Z}} \mathcal{K}(t)$ are two families of Hilbert spaces, we define $X(t) = P_{\mathcal{E}}(t)X/\oplus_{j\le t}\mathcal{K}(j)$. We also denote by U_p the set of upper-triangular operators in $\mathcal{L}(\oplus_{t\in\mathbb{Z}}\mathcal{V}(t), \oplus_{t\in\mathbb{Z}}(\mathcal{U}(t)\oplus\mathcal{V}(t)))$. We claim that $\mathbf{T}U_p = \{X \in U_p \,/\mathbf{G}(t)X(t) = \mathbf{0}, \quad t \in \mathbb{Z}\}$, where $\mathbf{G}(t) = [\ldots F(t)F(t-1)G(t-2)F(t)G(t-1)\ G(t)]$. Indeed, take $Y \in U_p$ then $\mathbf{G}(t)(\mathbf{T}Y)(t) = \mathbf{G}(t)\mathbf{T}(t)Y(t) = \mathbf{0}$, by the blocking property of $\mathbf{T}$. Conversely, take $X \in U_p$, $\mathbf{G}(t)X(t) = \mathbf{0}$ for all $t \in \mathbb{Z}$ and define $Y = \mathbf{T}^{-1}X = \mathbf{J}\mathbf{T}^*\mathbf{J}X$. Due to the structure of the Markov parameters of $\mathbf{T}$, it is readily checked that all the entries of Y under the main diagonal are zero. That is, $Y \in U_p$ and the claim is proved. From now on the arguments in Theorem 3.1 of [5] for getting the required representation of the solution of (15) apply directly. □

5. Schur parameters. There are special choices of the parameters $\{h_i(t), k_i(t)\}$ in (11) that would greatly simplify the generator recursion of Theorem 4.2 and lead to the so-called Schur parameters and the corresponding lattice sections. These parameters, which first appeared in the classical paper of Schur [30], have encountered applications in several areas including the study of orthogonal polynomials, inverse scattering, digital filtering, etc. [18]. They were also studied in a general operatorial framework in [3] and [7]. However, we want to emphasize that in our paper the Schur parameters are not the parameters associated with the load (i.e., the upper-triangular contractive operator K in Theorem 4.8), but rather are the parameters associated with the recursive construction of the strictly contractive solution $S = -\mathbf{T}_{12}\mathbf{T}_{22}^{-1}$.

We do not go into the details of the lattice structures here. The reader is referred to [24] and [26] for a detailed derivation. Instead we show how certain so-called time-variant Schur (or reflection) parameters appear in two important special cases.

We continue to require the finite dimensionality assumptions and the nondegeneracy condition $\mathbf{U}(t)\mathbf{U}^*(t) > \mu > 0$, for all t, of the previous section. But we now further assume that $\dim \mathcal{R}_i(t) = 1$ and that there exists $b > 0$ such that

$$b \le |g_i(t)J(t)g_i^*(t)| \quad \text{for all } t \in \mathbb{Z}. \tag{20}$$

We remark that conditions (20) and $\{d_i(t)\}$ *bounded from below* are independent, as can be shown by simple examples. We distinguish between two special cases: $g_i(t)J(t)g_i^*(t) > 0$ or $g_i(t)J(t)g_i^*(t) < 0$. That is, $g_i(t)$ has either positive or negative $J(t)$-norm. We partition $g_i(t)$ accordingly with $J(t)$, viz., $g_i(t) = \begin{bmatrix} u_i(t) & v_i(t) \end{bmatrix}$.

5.1. The positive case. In the positive case, we have

$$g_i(t)J(t)g_i^*(t) = u_i(t)u_i^*(t) - v_i(t)v_i^*(t) > 0,$$

and, by a well-known factorization result (see [21]), it follows that there exists a contraction $\tilde{\gamma}_i(t) : \mathcal{R}(v_i^*(t)) \to \mathcal{R}(u_i^*(t))$ such that $v_i(t) = u_i(t)\tilde{\gamma}_i(t)$, and $\|\tilde{\gamma}_i(t)\| < 1$. We can extend this contraction by zero to another contraction $\gamma_i(t) \in \mathcal{L}(\mathcal{U}(t), \mathcal{V}(t))$ that still satisfies $\|\gamma_i(t)\| < 1$ and $v_i(t) = u_i(t)\gamma_i(t)$. If we now construct the $J(t)$-unitary operator

$$\Theta_i(t) = \begin{bmatrix} I_{\mathcal{U}(t)} & -\gamma_i(t) \\ -\gamma_i^*(t) & I_{\mathcal{V}(t)} \end{bmatrix} \begin{bmatrix} (I-\gamma_i(t)\gamma_i^*(t))^{-1/2} & \mathbf{0} \\ \mathbf{0} & (I-\gamma_i^*(t)\gamma_i(t))^{-1/2} \end{bmatrix},$$

we readily conclude that $\Theta_i(t)$ reduces $g_i(t)$ to the form $g_i(t)\Theta_i(t) = \begin{bmatrix} \delta_i(t) & \mathbf{0} \end{bmatrix}$, for a certain $\delta_i(t)$. Now note that

$$\delta_i(t)\delta_i^*(t) = g_i(t)\Theta_i(t)J(t)\Theta_i^*(t)g_i^*(t) = g_i(t)J(t)g_i^*(t) > b,$$

and consequently, using the boundedness of $\{g_i(t)J(t)g_i^*(t)\}$ from below, there exists a constant $k > 0$ such that $\|\Theta_i^{-1}(t)\| \leq k$ for all t.

5.2. The negative case. In the negative case, we have

$$g_i(t)J(t)g_i^*(t) = u_i(t)u_i^*(t) - v_i(t)v_i^*(t) < 0,$$

and, by the same argument as above, we conclude that there exists a contraction $\gamma_i(t) \in \mathcal{L}(\mathcal{U}(t), \mathcal{V}(t))$ ($\|\gamma_i(t)\| < 1$) such that $u_i(t) = v_i(t)\gamma_i(t)$. If we now define the $J(t)$-unitary operator

$$\Theta_i(t) = \begin{bmatrix} I_{\mathcal{U}(t)} & -\gamma_i^*(t) \\ -\gamma_i(t) & I_{\mathcal{V}(t)} \end{bmatrix} \begin{bmatrix} (I - \gamma_i(t)^*\gamma_i(t))^{-1/2} & \mathbf{0} \\ \mathbf{0} & (I - \gamma_i(t)\gamma_i^*(t))^{-1/2} \end{bmatrix},$$

we readily conclude that $\Theta_i(t)$ reduces $g_i(t)$ to the form $g_i(t)\Theta_i(t) = \begin{bmatrix} \mathbf{0} & \delta_i(t) \end{bmatrix}$, for a certain $\delta_i(t)$. It also follows from

$$-\delta_i(t)\delta_i^*(t) = g_i(t)\Theta_i(t)J(t)\Theta_i^*(t)g_i^*(t) = g_i(t)J(t)g_i^*(t) < -b,$$

that $\|\Theta_i^{-1}(t)\| \leq k$ for some $k > 0$.

The contractions $\{\gamma(t)\}_{t\in\mathbb{Z}}$ are called the *Schur parameters* of the underlying displacement structure (1).

5.3. Strictly lower-triangular F(t). An important special case that often arises is the case of *strictly* lower triangular $F(t)$. That is, $f_i(t) = 0$ for all $t \in \mathbb{Z}$ and $i = 0, 1, \ldots, n-1$ and, consequently, $d_i(t) = g_i(t)J(t)g_i^*(t)$. But $\{d_i(t)\}$ is uniformly bounded from below, viz., $d_i(t) > \epsilon > 0$ for all t. Hence, we now always have

$$u_i(t)u_i^*(t) - v_i(t)v_i^*(t) > \epsilon > 0 \quad \text{for all } t \in \mathbb{Z},$$

and there always exist Schur parameters $\gamma_i(t)$ such that $v_i(t) = u_i(t)\gamma_i(t)$, with the corresponding $J(t)$-unitary operator defined by

$$\Theta_i(t) = \begin{bmatrix} I_{\mathcal{U}(t)} & -\gamma_i(t) \\ -\gamma_i^*(t) & I_{\mathcal{V}(t)} \end{bmatrix} \begin{bmatrix} (I - \gamma_i(t)\gamma_i^*(t))^{-1/2} & \mathbf{0} \\ \mathbf{0} & (I - \gamma_i^*(t)\gamma_i(t))^{-1/2} \end{bmatrix}.$$

The generator recursion in Theorem 4.2 can then be shown to reduce to the compact form (see also [29])

$$\begin{bmatrix} \mathbf{0} \\ G_{i+1}(t) \end{bmatrix} = F_i(t)G_i(t-1)\Theta_i(t-1) \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + G_i(t)\Theta_i(t) \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & I_{(r(t)-1)} \end{bmatrix},$$

which has the the following interpretation: multiply $G_i(t)$ by $\Theta_i(t)$ and keep the last columns; multiply the first column of $G_i(t-1)\Theta_i(t-1)$ by $F_i(t)$; these two steps result in $G_{i+1}(t)$.

6. Concluding remarks. We proved a general result (Theorem 2.2) concerning time-variant displacement equations of the form (1) with Pick operator solutions $R(t)$. We considered several moment, completion, and interpolation problems whose solutions followed as special cases of Theorem 2.2. These problems were stated in a general operator setting, including a time-variant version of the tangential Hermite–Fejér interpolation problem. Under supplementary finite-dimensionality and nondegeneracy conditions, a recursive procedure was derived that led to a recursive construction and parametrization of all solutions of the general result of Theorem 2.2. We also considered special cases where further simplifications were possible.

REFERENCES

[1] N. I. AKHIEZER, *The Classical Moment Problem and Some Related Questions in Analysis*, Hafner Publishing Company, New York, 1965. (In Russian, 1961.)

[2] D. ALPAY, P. DEWILDE, AND H. DYM, *On the existence and construction of solutions to the partial lossless inverse scattering problem with applications to estimation theory*, IEEE Trans. Inform. Theory, 35 (1989), pp. 1184–1205.

[3] G. ARSENE, Z. CEAUSESCU, AND C. FOIAS, *On interwining dilations* VIII, J. Operator Theory, 4 (1980), pp. 55–91.

[4] J. A. BALL AND I. GOHBERG, *A commutant lifting theorem for triangular matrices with diverse applications*, Integral Equations Operator Theory, 8 (1985), pp. 205–267.

[5] J. A. BALL, I. GOHBERG, AND M. A. KAASHOEK, *Nevanlinna–Pick interpolation for time-varying input-output maps: The discrete case*, in Operator Theory: Advances and Applications, Vol. 56, I. Gohberg, ed., Birkhäuser-Verlag, Basel, 1992, pp. 1–51.

[6] J. A. BALL, I. GOHBERG, AND L. RODMAN, *Interpolation of Rational Matrix Functions*, Operator Theory: Advances and Applications, Vol. 45, Birkhäuser-Verlag, Basel, 1990.

[7] Z. CEAUSESCU AND C. FOIAS, *On interwining dilations* V, Acta Sci. Math., 40 (1978), pp. 9–32.

[8] T. CONSTANTINESCU, *Some aspects of nonstationarity* – I, Acta Sci. Math., 54 (1990), pp. 379–389.

[9] T. CONSTANTINESCU, A. H. SAYED, AND T. KAILATH, *Structured matrices and moment problems*, in Proc. Workshop on Challenges of a Generalized System Theory, P. Dewilde, M. Kaashoek, and M. Verhaegen, eds., North Holland, Amsterdam, 1993, pp. 25–43.

[10] P. DEWILDE, *A course on the algebraic Schur and Nevanlinna–Pick interpolation problems*, in Algorithms and Parallel VLSI Architectures, Vol. A: Tutorials, E. F. Deprettere and A. J. van der Veen, eds., Elsevier Science Publications, New York, 1991, pp. 13–69.

[11] P. DEWILDE AND H. DYM, *Interpolation for upper triangular operators*, Operator Theory: Advances and Applications, Vol. 56, I. Gohberg, ed., Birkhäuser-Verlag, Basel, 1992, pp. 153–260.

[12] H. DYM AND I. GOHBERG, *Extensions of band matrices with band inverses*, Linear Algebra Appl., 36 (1981), pp. 1–24.

[13] C. FOIAS AND A. E. FRAZHO, *The Commutant Lifting Approach to Interpolation Problems*, Operator Theory: Advances and Applications, Vol. 44, Birkhäuser-Verlag, Basel, 1990.

[14] Y. GENIN, P. VAN DOOREN, T. KAILATH, J. DELOSME, AND M. MORF, *On Σ-lossless transfer functions and related questions*, Linear Algebra Appl., 50 (1983), pp. 251–275.

[15] I. GOHBERG AND M. A. KAASHOEK, *Time-varying linear systems with boundary conditions and integral operators, I. The transfer operator and its applications*, Integral Equations Operator Theory, 7 (1984), pp. 325–391.

[16] J. W. HELTON, *Operator Theory, Analytic Functions, Matrices and Electrical Engineering*, Conference Board of the Mathematical Sciences, American Mathematical Society, Providence, RI, 1987.

[17] T. KAILATH, *Linear Systems*, Prentice Hall, Englewood Cliffs, NJ, 1980.

[18] ———, *Signal processing applications of some moment problems*, in Moments in Mathematics, Vol. 37, H. Landau, ed., American Mathematical Society, Providence, RI, 1987, pp. 71–109.

[19] H. LEV-ARI AND T. KAILATH, *State-space approach to factorization of lossless transfer functions and structured matrices*, Linear Algebra Appl., 162–164 (1992), pp. 273–295.

[20] A. A. NUDELMAN, *On a generalization of classical interpolation problems*, Dokl. Akad. Nauk SSR, 256 (1981), pp. 790–793; Soviet Math. Dokl., 23 (1981), pp. 125–128.

[21] M. ROSENBLUM AND J. ROVNYAK, *Hardy Classes and Operator Theory*, Oxford University Press, London, 1985.

[22] L. A. SAKHNOVICH, *Interpolation problems, inverse spectral problems and nonlinear equations*, Operator Theory: Advances and Applications, Vol. 59, 1992, pp. 292–304.

[23] D. SARASON, *Generalized interpolation in H^∞*, Trans. Amer. Math. Soc., 127 (1967), pp. 179–203.

[24] A. H. SAYED, *Displacement Structure in Signal Processing and Mathematics*, Ph.D. thesis, Stanford University, Stanford, CA, August 1992.

[25] A. H. SAYED, T. CONSTANTINESCU, AND T. KAILATH, *Square-root algorithms for structured matrices, interpolation, and completion problems*, IMA Vol. Math. Appl., vol. 69, A. Bojanczyk and G. Cybenko, eds., Springer-Verlag, Berlin, to appear.

[26] ———, *Time-variant displacement structure and interpolation problems*, IEEE Trans. Automat. Control, 39 (1994), pp. 960–976.

[27] A. H. SAYED AND T. KAILATH, *A look-ahead block Schur algorithm for Toeplitz-like matrices*, SIAM J. Matrix Anal. Appl., to appear.

[28] A. H. SAYED, T. KAILATH, H. LEV-ARI, AND T. CONSTANTINESCU, *Recursive solutions of rational interpolation problems via fast matrix factorization*, Integral Equations and Operator Theory, to appear.

[29] A. H. SAYED, H. LEV-ARI, AND T. KAILATH, *Time-variant displacement structure and triangular arrays*, IEEE Trans. Signal Process., 42 (1994), pp. 1052–1062.

[30] I. SCHUR, *Über potenzreihen die im Inneren des Einheitskreises beschränkt sind*, J. Reine Angew. Math., 147 (1917), pp. 205–232. (English translation in *Operator Theory: Advances and Applications*, Vol. 18, I. Gohberg, ed., Birkhäuser-Verlag, Boston, 1986, pp. 31–88.)

[31] B. SZ. NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, North Holland, Amsterdam-Budapest, 1970.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 79–92, January 1995

006

A DIVIDE-AND-CONQUER ALGORITHM FOR THE BIDIAGONAL SVD*

MING GU† AND STANLEY C. EISENSTAT‡

Abstract. The authors present a stable and efficient divide-and-conquer algorithm for computing the singular value decomposition (SVD) of a lower bidiagonal matrix. Previous divide-and-conquer algorithms all suffer from a potential loss of orthogonality among the computed singular vectors unless extended precision arithmetic is used. A generalization that computes the SVD of a lower banded matrix is also presented.

Key words. singular value decomposition, divide-and-conquer, bidiagonal matrix

AMS subject classification. 65F15

1. Introduction.

Given an $(N+1) \times N$ lower bidiagonal matrix[1]

$$B = \begin{pmatrix} \alpha_1 & & & \\ \beta_1 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & \beta_{N-1} & \alpha_N \\ & & & \beta_N \end{pmatrix}, \tag{1}$$

its *singular value decomposition* (SVD) is

$$B = X \begin{pmatrix} \Omega \\ 0 \end{pmatrix} Y^T,$$

where X and Y are $(N+1) \times (N+1)$ and $N \times N$ orthogonal matrices, respectively, Ω is an $N \times N$ nonnegative diagonal matrix and 0 is a row of zero elements. The columns of X and Y are the *left singular vectors* and the *right singular vectors* of B, respectively, and the diagonal entries of Ω are the *singular values* of B. This problem arises when one computes the SVD of a general matrix by first reducing it to bidiagonal form [10], [12]. In this paper, we propose a bidiagonal divide-and-conquer algorithm (BDC) for solving this problem.

BDC first partitions B as

$$B = \begin{pmatrix} B_1 & \alpha_k e_k & 0 \\ 0 & \beta_k e_1 & B_2 \end{pmatrix},$$

where B_1 and B_2 are lower bidiagonal matrices, each of which is a submatrix of B. Next it recursively computes the SVDs of B_1 and B_2 and computes orthogonal matrices $(Q,\ q)$ and W such that

$$B = \begin{pmatrix} Q & q \end{pmatrix} \begin{pmatrix} M \\ 0 \end{pmatrix} W^T,$$

* Received by the editors December 31, 1992; accepted for publication (in revised form) by J. R. Bunch, September 1, 1993. This research was supported in part by U. S. Army Research Office contract DAAL03-91-G-0032.

† Department of Mathematics and Lawrence Berkeley Laboratory, University of California, Berkeley, California 94720 (minggu@math.berkeley.edu).

‡ Department of Computer Science, Yale University, Box 208285, New Haven, Connecticut 06520-8285 (eisenstat-stan@cs.yale.edu).

[1] An $N \times N$ lower bidiagonal matrix can be put into the form (1) by appending a zero row; we consider this case since it simplifies the recursion.

where M is an $N \times N$ matrix with nonzero elements only in the first column and on the diagonal. Finally it finds the singular values of B by computing the SVD $M = U\Omega V^T$, where U and V are orthogonal matrices, and then computes the singular vector matrices of B as (QU, q) and WV, respectively.

Since error is associated with computation, a *numerical* SVD of B or M is usually defined as a decomposition of the form

$$B = \hat{X} \begin{pmatrix} \hat{\Omega} \\ 0 \end{pmatrix} \hat{Y}^T + O(\epsilon \|B\|_2) \quad \text{or} \quad M = \hat{U}\hat{\Omega}\hat{V}^T + O(\epsilon \|M\|_2), \tag{2}$$

where ϵ is the machine precision, $\hat{\Omega}$ is diagonal, and $\hat{X}$ and $\hat{Y}$ or $\hat{U}$ and $\hat{V}$ are *numerically orthogonal*. An algorithm that produces such a decomposition is said to be *stable*.

While the singular values of B and M are always well conditioned with respect to perturbations, the singular vectors can be extremely sensitive [13, pp. 419–420]. That is, $\hat{\Omega}$ must be close to Ω, but $\hat{X}$, $\hat{Y}$, $\hat{U}$, and $\hat{V}$ can be very different from X, Y, U, and V, respectively. Thus one is usually content with stable algorithms for computing the SVD of B or M.

Jessup and Sorensen [22] present a divide-and-conquer method that uses basically the same dividing strategy and computes the SVD of M using an algorithm based on the work in [5], [6], and [9]. While it can compute the singular values of M to high absolute accuracy, in the presence of close singular values it can have difficulties in computing numerically orthogonal singular vectors unless extended precision arithmetic is used [22], [23], [27].

In this paper we develop a new algorithm for computing the SVD of M based on the work in [16] and [19]. It uses an approach similar to that of Jessup and Sorensen for computing the singular values, but it uses a completely different approach for computing the singular vectors, one that is stable. The amount of work is roughly the same, yet it does not require the use or simulation of extended precision arithmetic. Since it uses this algorithm, BDC is stable as well. Moreover, BDC uses a new procedure for handling deflation that makes it up to twice as fast asymptotically as the Jessup and Sorensen method.

There are three other divide-and-conquer algorithms for the bidiagonal SVD. Arbenz and Golub [3] follow the Jessup and Sorensen approach, but divide B by removing a column rather than a row. Arbenz [1] and Gragg, Thornton, and Warner [14] (see also Borges and Gragg [4]) each use a divide-and-conquer method for the symmetric tridiagonal eigenproblem to compute a spectral decomposition of a symmetric permutation of the matrix $\left(\begin{smallmatrix} 0 & B^T \\ B & 0 \end{smallmatrix}\right)$ while taking advantage of its special structure. All three algorithms can be unstable as noted above, unless extra precision arithmetic is used. The techniques presented here can be used to stabilize (and speed up deflation in) these algorithms as well.

BDC computes all the singular values in $O(N^2)$ time and all the singular values and singular vectors in $O(N^3)$ time. By using the fast multipole method of Carrier, Greengard, and Rokhlin [7], [15], BDC can be accelerated to compute all the singular values in $O(N \log_2 N)$ time and all the singular values and singular vectors in $O(N^2)$ time (see [16] and [17] for details). These asymptotic times are better than the corresponding worst-case times ($O(N^2)$ and $O(N^3)$) for the Golub–Kahan algorithm [10], [12] and bisection with inverse iteration [20], [21].

Section 2 presents the dividing strategy; §3 presents an algorithm for computing the SVD of M; §4 presents the deflation procedure; and §5 generalizes BDC to

compute the SVD of a lower banded matrix.

We take the usual model of arithmetic[2]

$$\mathrm{fl}(\alpha \circ \beta) = (\alpha \circ \beta)\,(1+\xi),$$

where α and β are floating-point numbers, $\circ$ is one of $+, -, \times$ and $\div$, $\mathrm{fl}(\alpha \circ \beta)$ is the floating-point result of the operation $\circ$, and $|\xi| \leq \epsilon$. We also require that

$$\mathrm{fl}(\sqrt{\alpha}) = \sqrt{\alpha}\,(1+\xi)$$

for any positive floating-point number α. For simplicity, we ignore the possibility of overflow and underflow.

2. "Dividing" the matrix. Given an $(N+1) \times N$ lower bidiagonal matrix B, we divide B into two subproblems as follows (cf. [22]):

$$B = \begin{pmatrix} B_1 & \alpha_k e_k & 0 \\ 0 & \beta_k e_1 & B_2 \end{pmatrix}, \tag{3}$$

where $1 < k < N$, B_1 and B_2 are $k \times (k-1)$ and $(N-k+1) \times (N-k)$ lower bidiagonal matrices, respectively, and e_j is the jth unit vector of appropriate dimension. Usually k is taken to be $\lfloor N/2 \rfloor$.

Let

$$B_i = \begin{pmatrix} Q_i & q_i \end{pmatrix} \begin{pmatrix} D_i \\ 0 \end{pmatrix} W_i^T$$

be the SVD of B_i. Let l_1^T and λ_1 be the last row and last component of Q_1 and q_1, respectively, and let f_2^T and φ_2 be the first row and first component of Q_2 and q_2, respectively. Substituting into (3), we get

$$B = \begin{pmatrix} q_1 & Q_1 & 0 & 0 \\ 0 & 0 & Q_2 & q_2 \end{pmatrix} \begin{pmatrix} \alpha_k\lambda_1 & 0 & 0 \\ \alpha_k l_1 & D_1 & 0 \\ \beta_k f_2 & 0 & D_2 \\ \beta_k\varphi_2 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & W_1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & W_2 \end{pmatrix}^T.$$

There is only one nonzero element in the first and last rows of the middle matrix. Applying a Givens rotation to zero out $\beta_k\varphi_2$, we have

$$\begin{aligned} B &= \left(\begin{pmatrix} c_0 q_1 & Q_1 & 0 \\ s_0 q_2 & 0 & Q_2 \end{pmatrix} \begin{pmatrix} -s_0 q_1 \\ c_0 q_2 \end{pmatrix} \right) \begin{pmatrix} r_0 & 0 & 0 \\ \alpha_k l_1 & D_1 & 0 \\ \beta_k f_2 & 0 & D_2 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & W_1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & W_2 \end{pmatrix}^T \\ &\equiv \begin{pmatrix} Q & q \end{pmatrix} \begin{pmatrix} M \\ 0 \end{pmatrix} W^T, \end{aligned} \tag{4}$$

where

$$r_0 = \sqrt{(\alpha_k\lambda_1)^2 + (\beta_k\varphi_2)^2}, \qquad c_0 = \frac{\alpha_k\lambda_1}{r_0}, \quad \text{and} \quad s_0 = \frac{\beta_k\varphi_2}{r_0}. \tag{5}$$

[2] This model excludes machines such as CRAY and CDC Cyber that do not have a guard digit. BDC can easily be modified for such machines.

Thus we have reduced B to $\binom{M}{0}$ by orthogonal transformations $(Q,\ q)$ and W, and M has nonzero elements only in the first column and on the diagonal.

Let $U\Omega V^T$ be the SVD of M computed using the algorithm described in §3. Substituting into (4) we obtain

$$B = \begin{pmatrix} Q & q \end{pmatrix} \begin{pmatrix} U\Omega V^T \\ 0 \end{pmatrix} W^T \equiv X \begin{pmatrix} \Omega \\ 0 \end{pmatrix} Y^T.$$

The singular values of B are the diagonal elements of Ω, and the singular vectors of B are the columns of X and Y. To compute the SVDs of B_1 and B_2, this process ((3) and (4)) can be recursively applied until the subproblems are sufficiently small. These small subproblems are then solved using the Golub–Kahan algorithm [10], [12]. There can be at most $O(\log_2 N)$ levels of recursion.

Equations (3) and (4) also suggest a recursion for computing only the singular values. Let f_1^T and φ_1 be the first row of Q_1 and the first component of q_1, respectively; let l_2^T and λ_2 be the last row of Q_2 and last component of q_2, respectively; let f^T and φ be the first row of Q and first component of q, respectively; and let l^T and λ be the last row of Q and the last component of q, respectively. Suppose that D_i, f_i, λ_i, l_i, and φ_i are given for $i = 1, 2$. Then we can compute Ω, f, λ, l, and φ by computing r_0, s_0, and c_0 using (5), computing the SVD of M, and computing

$$f^T = \begin{pmatrix} c_0\varphi_1 & f_1^T & 0 \end{pmatrix} U, \quad \lambda = c_0\lambda_2, \quad l^T = \begin{pmatrix} s_0\lambda_2 & 0 & l_2^T \end{pmatrix} U, \quad \text{and} \quad \varphi = -s_0\varphi_1.$$

There is a similar recursion for the divide-and-conquer algorithms in [8], [16],and [18] for the symmetric tridiagonal eigenproblem.

3. Computing the SVD of M. In this section we present a stable and efficient algorithm for finding the SVD of the $n \times n$ matrix

$$M = \begin{pmatrix} z_1 & & & \\ z_2 & d_2 & & \\ \vdots & & \ddots & \\ z_n & & & d_n \end{pmatrix},$$

where $D = \operatorname{diag}(d_1, \ldots, d_n)$, with[3] $0 \equiv d_1 \le d_2 \le \cdots \le d_n$, and $z = (z_1, \ldots, z_n)^T$. We further assume that

$$d_{j+1} - d_j \ge \tau\|M\|_2 \quad \text{and} \quad |z_j| \ge \tau\|M\|_2, \tag{6}$$

where τ is a small multiple of ϵ to be specified later. Any matrix of this form can be reduced to one that satisfies these conditions by using the deflation procedure described in §4.1 and a simple permutation.

The following lemma characterizes the singular values and singular vectors of M.

LEMMA 3.1 (Jessup and Sorensen [22]). *Let $U\Omega V^T$ be the* SVD *of M with*

$$U = (u_1, \ldots, u_n), \quad \Omega = \operatorname{diag}(\omega_1, \ldots, \omega_n), \quad \textit{and} \quad V = (v_1, \ldots, v_n).$$

Then the singular values $\{\omega_i\}_{i=1}^n$ satisfy the interlacing property

$$0 = d_1 < \omega_1 < d_2 < \cdots < d_n < \omega_n < d_n + ||z||_2$$

[3] We introduce d_1 to simplify the presentation.

and the secular equation

$$f(\omega) = 1 + \sum_{k=1}^{n} \frac{z_k^2}{d_k^2 - \omega^2} = 0.$$

The singular vectors are given by

$$u_i = \left(\frac{z_1}{d_1^2 - \omega_i^2}, \ldots, \frac{z_n}{d_n^2 - \omega_i^2} \right)^T \Bigg/ \sqrt{\sum_{k=1}^{n} \frac{z_k^2}{\left(d_k^2 - \omega_i^2\right)^2}}, \tag{7}$$

$$v_i = \left(-1, \frac{d_2 z_2}{d_2^2 - \omega_i^2}, \ldots, \frac{d_n z_n}{d_n^2 - \omega_i^2} \right)^T \Bigg/ \sqrt{1 + \sum_{k=2}^{n} \frac{(d_k z_k)^2}{\left(d_k^2 - \omega_i^2\right)^2}}. \tag{8}$$

On the other hand, given D and all the singular values, we can reconstruct M up to the signs of the z_i (cf. Löwner [25]).

LEMMA 3.2. *Given a diagonal matrix* $D = \operatorname{diag}(d_1, \ldots, d_n)$ *and a set of numbers* $\{\hat{\omega}_i\}_{i=1}^n$ *satisfying the interlacing property*

$$0 = d_1 < \hat{\omega}_1 < d_2 < \cdots < d_n < \hat{\omega}_n, \tag{9}$$

there exists a matrix

$$\hat{M} = \begin{pmatrix} \hat{z}_1 & & & \\ \hat{z}_2 & d_2 & & \\ \vdots & & \ddots & \\ \hat{z}_n & & & d_n \end{pmatrix}$$

whose singular values are $\{\hat{\omega}_i\}_{i=1}^n$. *The vector* $\hat{z} = (\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_n)^T$ *is given by*

$$|\hat{z}_i| = \sqrt{\left(\hat{\omega}_n^2 - d_i^2\right) \prod_{k=1}^{i-1} \frac{\left(\hat{\omega}_k^2 - d_i^2\right)}{\left(d_k^2 - d_i^2\right)} \prod_{k=i}^{n-1} \frac{\left(\hat{\omega}_k^2 - d_i^2\right)}{\left(d_{k+1}^2 - d_i^2\right)}}, \tag{10}$$

where the sign of $\hat{z}_i$ *can be chosen arbitrarily.*

Proof. Assume that $\hat{M}$ (and thus $\hat{z}$) exists. Then

$$\det\left(D^2 + \hat{z}\hat{z}^T - \omega^2 I\right) = \det\left(\hat{M}\hat{M}^T - \omega^2 I\right) = \prod_{k=1}^{n} \left(\hat{\omega}_k^2 - \omega^2\right).$$

On the other hand,

$$\begin{aligned} \det\left(D^2 + \hat{z}\hat{z}^T - \omega^2 I\right) &= \det\left(I + \left(D^2 - \omega^2 I\right)^{-1} \hat{z}\hat{z}^T\right) \det\left(D^2 - \omega^2 I\right) \\ &= \left(1 + \sum_{k=1}^{n} \frac{\hat{z}_k^2}{d_k^2 - \omega^2}\right) \prod_{k=1}^{n} \left(d_k^2 - \omega^2\right). \end{aligned}$$

Combining these two equations,

$$\prod_{k=1}^{n} \left(\hat{\omega}_k^2 - \omega^2\right) = \left(1 + \sum_{k=1}^{n} \frac{\hat{z}_k^2}{d_k^2 - \omega^2}\right) \prod_{k=1}^{n} \left(d_k^2 - \omega^2\right).$$

Setting $\omega = d_i$, we get

$$\hat{z}_i^2 = \prod_{k=1}^{n} (\hat{\omega}_k^2 - d_i^2) \Big/ \prod_{k \neq i} (d_k^2 - d_i^2).$$

Because of the interlacing property (9), the expression on the right-hand side is positive. Taking square roots we get (10). Working backward, if $\hat{z}$ is given by (10), then the singular values of $\hat{M}$ are $\{\hat{\omega}_i\}_{i=1}^n$. □

3.1. Computing the singular vectors. If ω_i were given, then we could compute each difference $d_k^2 - \omega_i^2$ in (7) and (8) to high relative accuracy as $(d_k - \omega_i)(d_k + \omega_i)$. We could also compute each product, each ratio, and each sum to high relative accuracy and thus compute u_i and v_i to componentwise high relative accuracy.

In practice we can only hope to compute an approximation $\hat{\omega}_i$ to ω_i. But problems can arise if we approximate u_i and v_i by

$$\hat{u}_i = \left(\frac{z_1}{d_1^2 - \hat{\omega}_i^2}, \ldots, \frac{z_n}{d_n^2 - \hat{\omega}_i^2} \right)^T \Big/ \sqrt{\sum_{k=1}^{n} \frac{z_k^2}{(d_k^2 - \hat{\omega}_i^2)^2}}$$

and

$$\hat{v}_i = \left(-1, \frac{d_2 z_2}{d_2^2 - \hat{\omega}_i^2}, \ldots, \frac{d_n z_n}{d_n^2 - \hat{\omega}_i^2} \right)^T \Big/ \sqrt{1 + \sum_{k=2}^{n} \frac{(d_k z_k)^2}{(d_k^2 - \hat{\omega}_i^2)^2}}$$

(i.e., replace ω_i by $\hat{\omega}_i$ in (7) and (8) as in [22]). For even if $\hat{\omega}_i$ is close to ω_i, the approximate ratios $z_k/(d_k^2 - \hat{\omega}_i^2)$ and $d_k z_k/(d_k^2 - \hat{\omega}_i^2)$ can still be very different from the exact ratios $z_k/(d_k^2 - \omega_i^2)$ and $d_k z_k/(d_k^2 - \omega_i^2)$, resulting in $\hat{u}_i$ and $\hat{v}_i$ very different from u_i and v_i. And when all the approximate singular values $\{\hat{\omega}_i\}_{i=1}^n$ are computed and all the corresponding singular vectors are approximated in this manner, the resulting singular vector matrices may not be orthogonal.

Lemma 3.2 allows us to overcome this problem. After we have computed all the approximate singular values $\{\hat{\omega}_i\}_{i=1}^n$ of M, we find a *new* matrix $\hat{M}$ whose *exact* singular values are $\{\hat{\omega}_i\}_{i=1}^n$ and then compute the singular vectors of $\hat{M}$ using Lemma 3.1. Note that each difference

$$\hat{\omega}_k^2 - d_i^2 = (\hat{\omega}_k - d_i)(\hat{\omega}_k + d_i) \quad \text{and} \quad d_k^2 - d_i^2 = (d_k - d_i)(d_k + d_i)$$

in (10) can be computed to high relative accuracy, as can each ratio and each product, and we can choose the sign of $\hat{z}_i$ to be the sign of z_i. Thus $\hat{z}_i$ can be computed to high relative accuracy. Substituting the *exact* singular values $\{\hat{\omega}_i\}_{i=1}^n$ and the computed $\hat{z}$ into (7) and (8), each singular vector of $\hat{M}$ can be computed to componentwise high relative accuracy. Consequently, after all the singular vectors are computed, the singular vector matrices of $\hat{M}$ will be numerically orthogonal.

To ensure the existence of $\hat{M}$, we need $\{\hat{\omega}_i\}_{i=1}^n$ to satisfy (9). But since the exact singular values of M satisfy the same interlacing property (see Lemma 3.1), this is only an accuracy requirement on the computed singular values and is not an additional restriction on M.

We can use the SVD of $\hat{M}$ as an approximation to the SVD of M. And since $\|M - \hat{M}\|_2 = \|z - \hat{z}\|_2$, such a substitution is stable (see (2)) as long as $\hat{z}$ is close to z (cf. [16], [19]).

3.2. Computing the singular values. To guarantee that $\hat{z}$ is close to z, we must ensure that the approximations $\{\hat{\omega}_i\}_{i=1}^n$ to the singular values are sufficiently accurate. The key is the stopping criterion for the root-finder, which requires a slight reformulation of the secular equation (cf. [5], [16], [19]).

Consider the singular value $\omega_i \in (d_i, d_{i+1})$, where $1 \le i \le n-1$; we consider the case $i = n$ later.

First assume that[4] $\omega_i \in \left(d_i, \frac{d_i+d_{i+1}}{2}\right)$. Let $\delta_j = d_j - d_i$ and let

$$\psi_i(\mu) \equiv \sum_{j=1}^{i} \frac{z_j^2}{(\delta_j - \mu)(d_j + d_i + \mu)} \quad \text{and} \quad \phi_i(\mu) \equiv \sum_{j=i+1}^{n} \frac{z_j^2}{(\delta_j - \mu)(d_j + d_i + \mu)}.$$

Setting $\omega = d_i + \mu$, we seek the root $\mu_i = \omega_i - d_i \in (0, \delta_{i+1}/2)$ of the reformulated secular equation

$$g_i(\mu) \equiv f(\mu + d_i) = 1 + \psi_i(\mu) + \phi_i(\mu) = 0.$$

Note that we can compute each ratio $z_j^2/((\delta_j - \mu)(d_j + d_i + \mu))$ in $g_i(\mu)$ to high relative accuracy for any $\mu \in (0, \delta_{i+1}/2)$. Indeed, either $\delta_j - \mu$ is a sum of negative terms or $|\mu| \le |\delta_j|/2$, and $d_j + d_i + \mu$ is a sum of positive terms. Thus, since both $\psi_i(\mu)$ and $\phi_i(\mu)$ are sums of terms of the same sign, we can bound the error in computing $g_i(\mu)$ by

$$\eta n(1 + |\psi_i(\mu)| + |\phi_i(\mu)|),$$

where η is a small multiple of ϵ that is independent of n and μ.

Now assume that $\omega_i \in \left[\frac{d_i+d_{i+1}}{2}, d_{i+1}\right)$. Let $\delta_j = d_j - d_{i+1}$ and let

$$\psi_i(\mu) \equiv \sum_{j=1}^{i} \frac{z_j^2}{(\delta_j - \mu)(d_j + d_{i+1} + \mu)} \quad \text{and} \quad \phi_i(\mu) \equiv \sum_{j=i+1}^{n} \frac{z_j^2}{(\delta_j - \mu)(d_j + d_{i+1} + \mu)}.$$

Setting $\omega = d_{i+1} + \mu$, we seek the root $\mu_i = \omega_i - d_{i+1} \in [\delta_i/2, 0)$ of the equation

$$g_i(\mu) \equiv f(\mu + d_{i+1}) = 1 + \psi_i(\mu) + \phi_i(\mu) = 0.$$

For any $\mu \in [\delta_i/2, 0)$, we can compute each ratio $z_j^2/((\delta_j - \mu)(d_j + d_{i+1} + \mu))$ to high relative accuracy (either $\delta_j - \mu$ is a sum of positive terms or $|\mu| \le |\delta_j|/2$; and $d_j + d_i + \mu = d_j + (d_{i+1} + \mu)$, where $|\mu| \le d_{i+1}/2$) and we can bound the error in computing $g_i(\mu)$ as before.

Finally consider the case $i = n$. Let $\delta_j = d_j - d_n$ and let

$$\psi_n(\mu) \equiv \sum_{j=1}^{n} \frac{z_j^2}{(\delta_j - \mu)(d_j + d_n + \mu)} \quad \text{and} \quad \phi_n(\mu) \equiv 0.$$

Setting $\omega = d_n + \mu$, we seek the root $\mu_n = \omega_n - d_n \in (0, \|z\|_2)$ of the equation

$$g_n(\mu) \equiv f(\mu + d_n) = 1 + \psi_n(\mu) + \phi_n(\mu) = 0.$$

[4] This can be checked by computing $f\left(\frac{d_i+d_{i+1}}{2}\right)$. If $f\left(\frac{d_i+d_{i+1}}{2}\right) > 0$, then $\omega_i \in \left(d_i, \frac{d_i+d_{i+1}}{2}\right)$, otherwise $\omega_i \in \left[\frac{d_i+d_{i+1}}{2}, d_{i+1}\right)$.

Again, for any $\mu \in (0, \|z\|_2)$, we can compute each ratio $z_j^2/((\delta_j - \mu)(d_j + d_n + \mu))$ to high relative accuracy, and we can bound the error in computing $g_n(\mu)$ as before.

In practice, the root-finder cannot make any progress at a point μ where it is impossible to determine the sign of $g_i(\mu)$ numerically. Thus we propose the stopping criterion

$$|g_i(\mu)| \leq \eta n \left(1 + |\psi_i(\mu)| + |\phi_i(\mu)|\right), \tag{11}$$

where, as before, the right-hand side is an upper bound on the round-off error in computing $g_i(\mu)$. Note that for each i, there is at least one floating-point number that satisfies this stopping criterion numerically, namely, $\mathrm{fl}(\mu_i)$.

We have not specified the method for finding the root of $g_i(\mu)$. We can use the bisection method or the rational interpolation strategies in [4], [5], [14], [24]. What is most important is the stopping criterion and the fact that, with the reformulation of the secular equation given above, we can find a μ that satisfies it.

3.3. Numerical stability. We now show that $\hat{z}$ is close to z.

THEOREM 3.3. *If* $\tau = 2\eta n^2$ *in* (6) *and each* $\hat{\mu}_i$ *satisfies* (11), *then*

$$|\hat{z}_i - z_i| \leq 4\eta n^2 \|z\|_2. \tag{12}$$

The proof is nearly identical to that of the analogous result in [19]. As argued there, the factor n^2 in τ and (12) is likely to be $O(n)$ in practice.

4. Deflation.

4.1. Deflation for M. Let

$$M = \begin{pmatrix} z_1 & & & \\ z_2 & d_2 & & \\ \vdots & & \ddots & \\ z_n & & & d_n \end{pmatrix},$$

where $D = \mathrm{diag}(d_1, \ldots, d_n)$ with $d_1 \equiv 0$ and $d_i \geq 0$, and $z = (z_1, \ldots, z_n)^T$. We now show that we can stably reduce M to a matrix of the same form that satisfies

$$|d_i - d_j| \geq \tau\|M\|_2 \quad \text{for } i \neq j \quad \text{and} \quad |z_i| \geq \tau\|M\|_2$$

(cf. (6)), where τ is specified in §3.3. We illustrate the reductions for $n = 3$, $i = 3$, and $j = 2$. Similar reductions appear in [22].

Assume that $|z_1| < \tau\|M\|_2$. Changing z_1 to $\tau\|M\|_2$ perturbs M by $O(\tau\|M\|_2)$:

$$M = \begin{pmatrix} z_1 & & \\ z_2 & d_2 & \\ z_3 & & d_3 \end{pmatrix} = \begin{pmatrix} \tau\|M\|_2 & & \\ z_2 & d_2 & \\ z_3 & & d_3 \end{pmatrix} + O(\tau\|M\|_2). \tag{13}$$

The perturbed matrix has the same structure as M and satisfies $|z_1| \geq \tau\|M\|_2$.

Next assume that $|z_i| < \tau\|M\|_2$ for $i \geq 2$. Changing z_i to zero perturbs M by $O(\tau\|M\|_2)$:

$$M = \begin{pmatrix} z_1 & & \\ z_2 & d_2 & \\ z_3 & & d_3 \end{pmatrix} = \begin{pmatrix} z_1 & & \\ z_2 & d_2 & \\ 0 & & d_3 \end{pmatrix} + O(\tau\|M\|_2). \tag{14}$$

In the perturbed matrix, d_i is a singular value and can be deflated, and the $(n-1) \times (n-1)$ leading principle submatrix has the same structure as M.

Now assume that $|d_i - d_1| = |d_i| < \tau\|M\|_2$ for $i \geq 2$. Changing d_i to zero and applying a Givens rotation G to zero out z_i perturbs the matrix GM by $O(\tau\|M\|_2)$:

$$GM = \begin{pmatrix} c & & s \\ & 1 & \\ -s & & c \end{pmatrix} \begin{pmatrix} z_1 & & \\ z_2 & d_2 & \\ z_3 & & 0 \end{pmatrix} + O(\tau\|M\|_2) \tag{15}$$
$$= \begin{pmatrix} r & & \\ z_2 & d_2 & \\ 0 & & 0 \end{pmatrix} + O(\tau\|M\|_2),$$

where $r = \sqrt{z_i^2 + z_1^2}$, $s = z_i/r$, and $c = z_1/r$. In the perturbed matrix, 0 is a singular value and can be deflated, and the $(n-1) \times (n-1)$ leading principle submatrix has the same structure as M.

Finally assume that $|d_i - d_j| < \tau\|M\|_2$ for $i, j \geq 2$. Changing d_j to d_i and symmetrically applying a Givens rotation G to zero out z_i perturbs the matrix GMG^T by $O(\tau\|M\|_2)$:

$$GMG^T = \begin{pmatrix} 1 & & \\ & c & s \\ & -s & c \end{pmatrix} \begin{pmatrix} z_1 & & \\ z_2 & d_3 & \\ z_3 & & d_3 \end{pmatrix} \begin{pmatrix} 1 & & \\ & c & -s \\ & s & c \end{pmatrix} + O(\tau\|M\|_2) \tag{16}$$
$$= \begin{pmatrix} z_1 & & \\ r & d_3 & \\ 0 & & d_3 \end{pmatrix} + O(\tau\|M\|_2),$$

where $r = \sqrt{z_i^2 + z_j^2}$, $s = z_i/r$, and $c = z_j/r$. In the perturbed matrix, d_i is a singular value and can be deflated, and the $(n-1) \times (n-1)$ leading principle submatrix has the same structure as M.

4.2. Local deflation. In the dividing strategy for BDC (see (4)), we write

$$B = \begin{pmatrix} Q & q \end{pmatrix} \begin{pmatrix} M \\ 0 \end{pmatrix} W^T = \begin{pmatrix} QU & q \end{pmatrix} \begin{pmatrix} \Omega \\ 0 \end{pmatrix} (WV)^T, \tag{17}$$

where

$$Q = \begin{pmatrix} c_0 q_1 & Q_1 & 0 \\ s_0 q_2 & 0 & Q_2 \end{pmatrix}, \quad M = \begin{pmatrix} r_0 & 0 & 0 \\ \alpha_k l_1 & D_1 & 0 \\ \beta_k f_2 & 0 & D_2 \end{pmatrix}, \quad \text{and} \quad W = \begin{pmatrix} 0 & W_1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & W_2 \end{pmatrix},$$

l_1^T is the last row of Q_1, f_2^T is first row of Q_2, and $U\Omega V^T$ is the SVD of M.

Note that both Q and W are block matrices with some zero blocks. Since the main cost of BDC is in computing the matrix-matrix products QU and WV, we would like to take advantage of this structure. In this subsection we describe a deflation procedure for BDC that gets a speedup of roughly a factor of two by doing so. This approach is not used in [22].

If $|r_0| < \tau\|M\|_2$, then we apply reduction (13). If the vector $(\alpha_k l_1^T, \beta_k f_2^T)$ has some components with small absolute value, then we apply reduction (14). In both cases the block structure of Q and W is preserved. If D_1 has a small diagonal element, then we apply reduction (15), and if D_1 has two close diagonal elements, then we apply

reduction (16). Again in both cases the block structure is preserved. We do the same when D_2 has a small diagonal element or has two close diagonal elements.

However, when D_1 has a diagonal element that is close to a diagonal element in D_2 and we apply reduction (16), the block structure of Q and W is changed. To illustrate, assume that after applying a permutation, the first diagonal element of D_1 is close to the last diagonal element of D_2. Let

$$Q_1 = (\tilde{q}_1 \;\; \tilde{Q}_1), \quad Q_2 = (\tilde{Q}_2 \;\; \tilde{q}_2), \quad W_1 = (\tilde{w}_1 \;\; \tilde{W}_1), \quad \text{and} \quad W_2 = (\tilde{W}_2 \;\; \tilde{w}_2);$$

and let

$$\alpha_k l_1 = \begin{pmatrix} z_2 \\ \tilde{z}_1 \end{pmatrix}, \quad \beta_k f_2 = \begin{pmatrix} \tilde{z}_2 \\ z_N \end{pmatrix}, \quad D_1 = \operatorname{diag}(d_2, \tilde{D}_1), \quad \text{and} \quad D_2 = \operatorname{diag}(\tilde{D}_2, d_N).$$

Changing d_2 to d_N and applying a Givens rotation G to zero out z_N, we get

$$GMG^T = \begin{pmatrix} r_0 & & & & \\ r & d_N & & & \\ \tilde{z}_1 & & \tilde{D}_1 & & \\ \tilde{z}_2 & & & \tilde{D}_2 & \\ 0 & & & & d_N \end{pmatrix} + O(\tau\|M\|_2) \equiv \begin{pmatrix} \tilde{M}_1 & 0 \\ 0 & d_N \end{pmatrix} + O(\tau\|M\|_2),$$

where $r = \sqrt{z_2^2 + z_N^2}$, $c = z_2/r$, and $s = z_N/r$. Substituting into (17), we have

$$\begin{aligned} B &= (QG^T \;\; q) \begin{pmatrix} GMG^T \\ 0 \end{pmatrix} (WG^T)^T \\ &= (\tilde{X}_1 \;\; \tilde{x} \;\; q) \begin{pmatrix} \tilde{M}_1 & 0 \\ 0 & d_N \\ 0 & 0 \end{pmatrix} (\tilde{Y}_1 \;\; \tilde{y})^T + O(\tau\|M\|_2), \end{aligned}$$

where

$$\tilde{X}_1 = \begin{pmatrix} c_0 q_1 & c\tilde{q}_1 & \tilde{Q}_1 & 0 \\ s_0 q_2 & s\tilde{q}_2 & 0 & \tilde{Q}_2 \end{pmatrix} \quad \text{and} \quad \tilde{x} = \begin{pmatrix} -s\tilde{q}_1 \\ c\tilde{q}_2 \end{pmatrix}$$

and

$$\tilde{Y}_1 = \begin{pmatrix} 0 & c\tilde{w}_1 & \tilde{W}_1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & s\tilde{w}_2 & 0 & \tilde{W}_2 \end{pmatrix} \quad \text{and} \quad \tilde{y} = \begin{pmatrix} -s\tilde{w}_1 \\ 0 \\ c\tilde{w}_2 \end{pmatrix}.$$

d_N is an approximate singular value of B and can be deflated. The corresponding approximate left and right singular vectors are $\tilde{x}$ and $\tilde{y}$, respectively. The matrix $\tilde{M}_1$ has the same structure as M and can be deflated in a similar fashion until no diagonal element of $\tilde{D}_1$ is close to a diagonal element of $\tilde{D}_2$.

Thus, ignoring permutations of the columns of Q_i and W_i and the diagonal entries of D_i, after a series of these interblock deflations B can be written as

$$B = (\tilde{X}_1 \;\; \tilde{X}_2 \;\; q) \begin{pmatrix} \tilde{M}_1 & 0 \\ 0 & \tilde{\Omega}_2 \\ 0 & 0 \end{pmatrix} (\tilde{Y}_1 \;\; \tilde{Y}_2)^T + O(\tau\|B\|_2).$$

$\tilde{\Omega}_2$ is a diagonal matrix whose diagonal elements are the deflated singular values, and the columns of $\tilde{X}_2$ and $\tilde{Y}_2$ are the corresponding approximate left and right singular vectors. $\tilde{M}_1$ is of the form

$$\tilde{M}_1 = \begin{pmatrix} r_0 & & & \\ \tilde{z}_0 & \tilde{D}_0 & & \\ \tilde{z}_1 & & \tilde{D}_1 & \\ \tilde{z}_2 & & & \tilde{D}_2 \end{pmatrix},$$

where the dimension of $\tilde{D}_0$ is the number of deflations, $\tilde{D}_1$ and $\tilde{D}_2$ contain the diagonal elements of D_1 and D_2 not affected by deflation, and $\tilde{z}_0$, $\tilde{z}_1$, and $\tilde{z}_2$ are defined accordingly. $\tilde{X}_1$ and $\tilde{Y}_1$ are of the form

$$\tilde{X}_1 = \begin{pmatrix} c_0 q_1 & \tilde{Q}_{0,1} & \tilde{Q}_1 & 0 \\ s_0 q_2 & \tilde{Q}_{0,2} & 0 & \tilde{Q}_2 \end{pmatrix} \quad \text{and} \quad \tilde{Y}_1 = \begin{pmatrix} 0 & \tilde{W}_{0,1} & \tilde{W}_1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & \tilde{W}_{0,2} & 0 & \tilde{W}_2 \end{pmatrix}, \tag{18}$$

where the column dimension of $\tilde{Q}_{0,1}$, $\tilde{Q}_{0,2}$, $\tilde{W}_{0,1}$, and $\tilde{W}_{0,2}$ is the number of deflations, and the columns of $\tilde{Q}_1$, $\tilde{Q}_2$, $\tilde{W}_1$, and $\tilde{W}_2$ are those of Q_1, Q_2, W_1, and W_2 not affected by deflation.

Let $\tilde{U}_1 \tilde{\Omega}_1 \tilde{V}_1^T$ be the SVD of $\tilde{M}_1$. Then

$$\begin{aligned} B &= \begin{pmatrix} \tilde{X}_1 & \tilde{X}_2 & q \end{pmatrix} \begin{pmatrix} \tilde{U}_1 \tilde{\Omega}_1 \tilde{V}_1^T & 0 \\ 0 & \tilde{\Omega}_2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \tilde{Y}_1 & \tilde{Y}_2 \end{pmatrix}^T + O(\tau \|B\|_2) \\ &= \begin{pmatrix} \tilde{X}_1 \tilde{U}_1 & \tilde{X}_2 & q \end{pmatrix} \begin{pmatrix} \tilde{\Omega}_1 & 0 \\ 0 & \tilde{\Omega}_2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \tilde{Y}_1 \tilde{V}_1 & \tilde{Y}_2 \end{pmatrix}^T + O(\tau \|B\|_2). \end{aligned}$$

Thus $(\tilde{X}_1 \tilde{U}_1, \tilde{X}_2, q)$ and $(\tilde{Y}_1 \tilde{V}_1, \tilde{Y}_2)$ are approximate left and right singular vector matrices of B, respectively. The matrices $\tilde{X}_1 \tilde{U}_1$ and $\tilde{Y}_1 \tilde{V}_1$ can be computed while taking advantage of the block structure of $\tilde{X}_1$ and $\tilde{Y}_1$ in (18).

We refer to these as local deflations since they are associated with individual subproblems.

4.3. Global deflation. To illustrate *global* deflation, we look at two levels of the dividing strategy (see (4)):

$$B = \begin{pmatrix} B_1 & \alpha_{i+j} e_{i+j} & \\ & \beta_{i+j} e_1 & B_2 \end{pmatrix} = \begin{pmatrix} B_{1,1} & \alpha_i e_i & & \\ & \beta_i e_1 & B_{1,2} & \alpha_{i+j} e_j \\ & & & \beta_{i+j} e_1 \quad B_2 \end{pmatrix}, \tag{19}$$

where B_1, B_2, $B_{1,1}$ and $B_{1,2}$ are principle submatrices of B of dimensions $(i+j) \times (i+j-1)$, $(N-i-j+1) \times (N-i-j)$, $i \times (i-1)$, and $j \times (j-1)$, respectively.

Let $X_{1,2} \left(\begin{smallmatrix} D_{1,2} \\ 0 \end{smallmatrix}\right) W_{1,2}^T$ be the SVD of $B_{1,2}$, and let $(f_{1,2}^T, \varphi_{1,2})$ and $(l_{1,2}^T, \lambda_{1,2})$ be the first and last rows of $X_{1,2}$, respectively. Then

$$B = \bar{X} \begin{pmatrix} B_{1,1} & \alpha_i e_i & & \\ & \beta_i f_{1,2} & D_{1,2} & \alpha_{i+j} l_{1,2} \\ & \beta_i \varphi_{1,2} & 0 & \alpha_{i+j} \lambda_{1,2} \\ & & & \beta_{i+j} e_1 \quad B_2 \end{pmatrix} \bar{Y}^T, \tag{20}$$

where $\bar{X} = \mathrm{diag}(I_i,\ X_{1,2},\ I_{N-i-j+1})$ and $\bar{Y} = \mathrm{diag}(I_{i-1},\ 1,\ W_{1,2},\ 1,\ I_{N-i-j})$.

Let $\bar{d}_s$ be the sth diagonal element of $D_{1,2}$, and let $\bar{f}_s$ and $\bar{l}_s$ be the sth components of $f_{1,2}$ and $l_{1,2}$, respectively. Then ignoring all zero components, the $(i+s)$th column and row of the middle matrix in (20) are $(\bar{d}_s)$ and $(\beta_i \bar{f}_s,\ \bar{d}_s,\ \alpha_{i+j}\bar{l}_s)$, respectively. Thus if both $|\beta_i \bar{f}_s|$ and $|\alpha_{i+j}\bar{l}_s|$ are small, then we can perturb them to zero. $\bar{d}_s$ is a singular value of the perturbed matrix and the $(i+s)$th columns of $\bar{X}$ and $\bar{Y}$ are the corresponding left and right singular vectors, respectively. This singular value and its singular vectors can be deflated from all subsequent subproblems. We call this *global* deflation.

Consider the deflation procedure for computing the SVD in §4.2. If $|\beta_i \bar{f}_s|$ is small, then it can be perturbed to zero. This is a local deflation if only $|\beta_i \bar{f}_s|$ is small and a global deflation if $|\alpha_{i+j}\bar{l}_s|$ is also small.

5. Computing the SVD of a banded matrix. We now generalize BDC to compute the SVD of a lower banded matrix. This problem arises when one uses the block Lanczos algorithm to compute the SVD of a sparse matrix [11], [13]. Arbenz [2] has similarly generalized a divide-and-conquer algorithm for the symmetric tridiagonal eigenproblem to solve the symmetric banded eigenproblem.

Let B be an $(N+K)\times N$ lower $(K+1)$-diagonal matrix with $K \ll N$. We divide B into two subproblems as follows:

$$B = \begin{pmatrix} B_{1,1} & B_{1,2} & 0 \\ 0 & B_{2,2} & B_{2,3} \end{pmatrix}, \tag{21}$$

where $1 < k < N$, $B_{1,1}$ and $B_{2,3}$ are $(k+K)\times k$ and $(N-k)\times(N-K-k)$ lower $(K+1)$-diagonal matrices, respectively, $B_{1,2}$ is a $(k+K)\times K$ matrix with nonzero elements only on the lowest K diagonals, and $B_{2,2}$ is an $(N-k)\times K$ matrix with nonzero elements only on the highest K diagonals. Usually k is taken to be $\lfloor (N-K)/2 \rfloor$.

Let

$$B_{1,1} = \begin{pmatrix} Q_1 & S_1 \end{pmatrix} \begin{pmatrix} D_1 \\ 0 \end{pmatrix} W_1^T \quad \text{and} \quad B_{2,3} = \begin{pmatrix} Q_2 & S_2 \end{pmatrix} \begin{pmatrix} D_2 \\ 0 \end{pmatrix} W_2^T$$

be the SVDs of $B_{1,1}$ and $B_{2,3}$, respectively. Substituting into (21), we have

$$B = \begin{pmatrix} S_1 & Q_1 & & \\ & & Q_2 & S_2 \end{pmatrix} \begin{pmatrix} Z_{0,1} & 0 & 0 \\ Z_1 & D_1 & 0 \\ Z_2 & 0 & D_2 \\ Z_{0,2} & 0 & 0 \end{pmatrix} \begin{pmatrix} I_K & & W_1 & \\ & & & W_2 \end{pmatrix}^T, \tag{22}$$

where $Z_{0,1} = S_1^T B_{1,2}$, $Z_1 = Q_1^T B_{1,2}$, $Z_2 = Q_2^T B_{2,2}$, and $Z_{0,2} = S_2^T B_{2,2}$. There exists a $2K \times 2K$ orthogonal matrix

$$\begin{pmatrix} G_{1,1} & G_{1,2} \\ G_{2,1} & G_{2,2} \end{pmatrix}$$

such that

$$\begin{pmatrix} Z_{0,1} \\ Z_{0,2} \end{pmatrix} = \begin{pmatrix} G_{1,1} & G_{1,2} \\ G_{2,1} & G_{2,2} \end{pmatrix} \begin{pmatrix} Z_0 \\ 0 \end{pmatrix},$$

where Z_0 is a $K \times K$ lower triangular matrix. Substituting into (22), we have

$$B = \begin{pmatrix} S_1 G_{1,1} & Q_1 & & S_1 G_{2,1} \\ S_2 G_{1,2} & & Q_2 & S_2 G_{2,2} \end{pmatrix} \begin{pmatrix} Z_0 & 0 & 0 \\ Z_1 & D_1 & 0 \\ Z_2 & 0 & D_2 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} & W_1 & \\ I_K & & \\ & & W_2 \end{pmatrix}^T. \tag{23}$$

The middle matrix in (23) is lower triangular and can have nonzero elements only in the first K columns and on the diagonal. Partition

$$\begin{pmatrix} Z_0 \\ Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} \tilde{Z}_0 & 0 \\ \tilde{Z} & z \end{pmatrix},$$

where $\tilde{Z}_0$ is a $(K-1) \times (K-1)$ lower triangular matrix and $z = (r_0, z_1^T, z_2^T)^T$, with z_i being the last column of Z_i and r_0 being the last diagonal element of Z_0. Let $U\Omega V^T$ be the SVD of

$$M = \begin{pmatrix} r_0 & 0 & 0 \\ z_1 & D_1 & 0 \\ z_2 & 0 & D_2 \end{pmatrix}$$

computed using the algorithm described in §3. Then the middle matrix in (23) can be rewritten as

$$\begin{pmatrix} \tilde{Z}_0 & 0 \\ \tilde{Z} & U\Omega V^T \end{pmatrix} = \begin{pmatrix} I_{K-1} & \\ & U \end{pmatrix} \begin{pmatrix} \tilde{Z}_0 & 0 \\ U^T \tilde{Z} & \Omega \end{pmatrix} \begin{pmatrix} I_{K-1} & \\ & V \end{pmatrix}^T,$$

where the middle matrix is lower triangular and can have nonzero elements only in the first $K-1$ columns and on the diagonal. Thus the SVD of the middle matrix in (23) can be computed by applying this procedure K times.

To compute the SVDs of $B_{1,1}$ and $B_{2,3}$, we can recursively apply (21) and (23) to $B_{1,1}$ and $B_{2,3}$ until the subproblems are sufficiently small. These small subproblems are then solved using the Golub–Kahan algorithm [10], [12]. There can be at most $O(\log_2 N)$ levels of recursion. This algorithm takes $O(KN^3)$ time to compute both the singular values and the singular vectors. Similar to the bidiagonal case, there is an $O(K^2N^2)$ time divide-and-conquer algorithm for computing only the singular values. These times can be reduced to $O(KN^2)$ and $O(K^2N \log_2 N)$, respectively, by using the fast multipole method [16], [17]. These reduced times are better than the corresponding worst-case times ($O(N^3)$ and $O(KN^2)$) for the banded QR algorithm [26, p. 172].

REFERENCES

[1] P. ARBENZ, *Divide-and-conquer algorithms for the computation of the SVD of bidiagonal matrices*, in Vector and Parallel Computing, J. Dongarra, I. Duff, P. Gaffney, and S. McKee, eds., Ellis Horwood, Chichester, 1989, pp. 1–10.

[2] ———, *Divide-and-conquer algorithms for the bandsymmetric eigenvalue problem*, Parallel Comput., 18 (1992), pp. 1105–1128.

[3] P. ARBENZ AND G. H. GOLUB, *On the spectral decomposition of Hermitian matrices modified by low rank perturbations with applications*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 40–58.

[4] C. F. BORGES AND W. B. GRAGG, *A parallel divide and conquer algorithm for the generalized real symmetric definite tridiagonal eigenproblem*, in Numerical Linear Algebra and Scientific Computing, L. Reichel, A. Ruttan, and R. S. Varga, eds., de Gruyter, Berlin, 1993, pp. 10–28.

[5] J. R. BUNCH AND C. P. NIELSEN, *Updating the singular value decomposition*, Numer. Math., 31 (1978), pp. 111–129.
[6] J. R. BUNCH, C. P. NIELSEN, AND D. C. SORENSEN, *Rank-one modification of the symmetric eigenproblem*, Numer. Math., 31 (1978), pp. 31–48.
[7] J. CARRIER, L. GREENGARD, AND V. ROKHLIN, *A fast adaptive multipole algorithm for particle simulations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 669–686.
[8] J. J. M. CUPPEN, *A divide and conquer method for the symmetric tridiagonal eigenproblem*, Numer. Math., 36 (1981), pp. 177–195.
[9] G. H. GOLUB, *Some modified matrix eigenvalue problems*, SIAM Rev., 15 (1973), pp. 318–334.
[10] G. H. GOLUB AND W. KAHAN, *Calculating the singular values and pseudo-inverse of a matrix*, SIAM J. Numer. Anal., 2 (1965), pp. 205–224.
[11] G. H. GOLUB, F. T. LUK, AND M. OVERTON, *A block Lanczos method for computing the singular values and corresponding singular vectors of a matrix*, ACM Trans. Math. Soft., 7 (1981), pp. 149–169.
[12] G. H. GOLUB AND C. REINSCH, *Singular value decomposition and least squares solutions*, Numer. Math., 14 (1970), pp. 403–420.
[13] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1989.
[14] W. B. GRAGG, J. R. THORNTON, AND D. D. WARNER, *Parallel divide and conquer algorithms for the symmetric tridiagonal eigenproblem and bidiagonal singular value problem*, in Modelling and Simulation, W. G. Vogt and M. H. Mickle, eds., Vol. 23, Part 1, University of Pittsburgh School of Engineering, Pittsburgh, 1992, pp. 49–56.
[15] L. GREENGARD AND V. ROKHLIN, *A fast algorithm for particle simulations*, J. Comput. Phys., 73 (1987), pp. 325–348.
[16] M. GU, *Studies in Numerical Linear Algebra*, Ph.D. thesis, Department of Computer Science, Yale University, New Haven, CT, 1993.
[17] M. GU AND S. C. EISENSTAT, *A fast algorithm for updating the singular value decomposition*, manuscript.
[18] ———, *A divide-and-conquer algorithm for the symmetric tridiagonal eigenproblem*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 172–191.
[19] ———, *A stable and efficient algorithm for the rank-one modification of the symmetric eigenproblem*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1266–1276.
[20] E. R. JESSUP, *Parallel Solution of the Symmetric Tridiagonal Eigenproblem*, Ph.D. thesis, Department of Computer Science, Yale University, New Haven, CT, 1989.
[21] E. R. JESSUP AND I. C. F. IPSEN, *Improving the accuracy of inverse iteration*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 550–572.
[22] E. R. JESSUP AND D. C. SORENSEN, *A parallel algorithm for computing the singular value decomposition of a matrix*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 530–548.
[23] W. KAHAN, *Rank-1 perturbed diagonal's eigensystem*, manuscript, July 1989.
[24] R.-C. LI, *Solving secular equations stably and efficiently*, Working Paper, Department of Mathematics, University of California at Berkeley, Oct. 1992.
[25] K. LÖWNER, *Über monotone Matrixfunktionen*, Math. Z., 38 (1934), pp. 177–216.
[26] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice Hall, Englewood Cliffs, NJ, 1980.
[27] D. C. SORENSEN AND P. T. P. TANG, *On the orthogonality of eigenvectors computed by divide-and-conquer techniques*, SIAM J. Numer. Anal., 28 (1991), pp. 1752–1775.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 93–112, January 1995

007

ON THE SENSITIVITY OF SOLUTION COMPONENTS IN LINEAR SYSTEMS OF EQUATIONS*

S. CHANDRASEKARAN† AND I. C. F. IPSEN†

Abstract. Expressions are presented for the errors in *individual* components of the solution to systems of linear equations and linear least squares problems. No assumptions about the structure or distribution of the perturbations are made.

The resulting "componentwise condition numbers" measure the sensitivity of each solution component to perturbations. It is shown that any linear system has at least one solution component whose sensitivity to perturbations is proportional to the condition number of the matrix; but there may exist many components that are much better conditioned. Unless the perturbations are restricted, no norm-based relative error bound can predict the presence of well-conditioned components, so these componentwise condition numbers are essential.

For the class of componentwise perturbations, necessary and sufficient conditions are given under which Skeel's condition numbers are informative, and it is shown that these conditions are similar to conditions where componentwise condition numbers are useful. Numerical experiments not only confirm that these circumstances do occur frequently, they also illustrate that for many classes of matrices the ill conditioning of the matrix is due to a few rows of the inverse only. This means that many of the solution components are computed more accurately than current analyses predict.

Key words. condition number, diagonal scaling, forward error, linear system, least squares, perturbation theory

AMS subject classifications. 65F05, 65F20, 65F35, 15A06, 15A09, 15A12, 15A45, 15A60

1. Introduction.

Certain problems in statistics [33], combustion [26], and molecular conformation [10] require the solution of systems of linear equations whose individual solution components have physical significance; knowledge about the accuracy in the computation of the solution components is important. For the solution of problems involving Markov chains, for instance, it turns out that all solution components exhibit essentially the same sensitivity to perturbations in the data [25]. In [8] it is necessary to analyse individual solution components to demonstrate the convergence of inverse iteration in finite precision.

1.1. Motivation.

Consider the solution of a system of linear equations $Ax = b$ with nonsingular coefficient matrix A. The computed solution $\bar{x}$, which is usually different from the true solution x, can be viewed as the true solution to a perturbed system $(A + F)\bar{x} = b + f$.

So far, little work has dealt with trying to assess the error in individual solution components of a linear system; exceptions are the stability analyses of algorithms for solving particular structured linear systems, e.g., [3], [20], [22], [23]. The conventional way of estimating the error in $\bar{x}$, as compared to the true solution x, is to estimate an upper bound on the norm-based relative error $\|\bar{x} - x\|/\|x\|$. The most commonly used first-order bound is

$$\frac{\|\bar{x} - x\|}{\|x\|} \dot{\leq} \kappa(A)(\rho_A + \rho_b),$$

* Received by the editors May 18, 1992; accepted for publication (in revised form) by J. Bunch, June 29, 1993. The work presented in this paper was supported by Defense Advanced Research Projects Agency contract N00014-88-K-0573 and by National Science Foundation grant CCR-9102853.

† Department of Computer Science, Yale University, New Haven, Connecticut 06520 (chandras@math.ncsu.edu and ipsen@math.ncsu.edu).

where the condition number $\kappa(A) = \|A\| \, \|A^{-1}\| \geq 1$ acts as an amplifier for the relative perturbations in the data $\rho_A = \|F\|/\|A\|$ and $\rho_b = \|f\|/\|b\|$.

In many situations this type of error assessment is just fine unless, however, the individual components of the solution have physical significance. The example of the matrix of order four below, which represents a special case of a class of matrices discussed in §7, illustrates that the condition number $\kappa(A)$ can severely overestimate the error in some components,

$$A = \begin{pmatrix} 0.4919 & 0.1112 & -0.6234 & -0.6228 \\ -0.5050 & -0.6239 & 0.0589 & 0.0595 \\ 0.5728 & -0.0843 & 0.7480 & 0.7483 \\ -0.4181 & 0.7689 & 0.2200 & 0.2204 \end{pmatrix}, \qquad b = \begin{pmatrix} 0.4351 \\ -0.1929 \\ 0.6165 \\ -0.8022 \end{pmatrix}.$$

The first three columns of A are nearly orthogonal while the last two columns are almost identical. Both the two-norm condition number $\kappa_2(A)$ and Skeel's condition number [31] are larger than 10^3 (note that the matrix is not ill scaled). But the "componentwise condition numbers" that we introduce in this paper turn out to be

$$< 1.1, \quad < 1.1, \quad > 10^3, \quad > 10^3 .$$

This means that the first two components of x are well conditioned, regardless of the perturbations, and the remaining two are ill conditioned. To illustrate this, compare the "exact" solution x computed with 16-digit arithmetic with the solution $\bar{x}$ computed with 4-digit arithmetic, which can be viewed as the solution to a perturbed problem,

$$x = \begin{pmatrix} 1.000075414240576 \\ -.5000879795933286 \\ -.0242511388797165 \\ .02624513955005858 \end{pmatrix}, \qquad \bar{x} = \begin{pmatrix} 1.000 \\ -.5003 \\ -.0589 \\ .06090 \end{pmatrix}.$$

As predicted by our componentwise condition numbers, the first two components are accurate to almost four digits, whereas the last two have no accuracy whatsoever. As far as we know no other existing condition numbers can predict the well conditioning of the first two components of this system.

1.2. Overview. Given a linear system $Ax = b$ of full column rank and a perturbed system $(A + F)\bar{x} = b + f$, we derive expressions for the error in individual components of the computed solution $\bar{x}$ (§2). Our work is more general than that of Skeel [31] on componentwise perturbations and that of Stewart [34] on stochastic perturbations because we make no assumptions about the perturbations F and f, their size, structure, or distribution.

We associate with a linear system $Ax = b$ not a single condition number but a set of "componentwise condition numbers," one for each solution component. These condition numbers provide a clear separation of the three factors responsible for the loss of accuracy in the computed solution: relative magnitude of the solution components, matrix condition, and relationship between matrix and right-hand side.

We show that there is at least one component of the solution vector whose sensitivity to relative perturbations is proportional to the condition number of the matrix; but there may exist components that are much better conditioned. Consequently, unless the perturbations are restricted, *no* norm-based relative error bound can ever

predict the presence of well-conditioned components in x. Therefore, our componentwise condition numbers are essential.

Along the way, we comment on the tightness of norm-based error bounds (§3), and we clarify some results of Chan and Foulser [6] regarding the influence of the right-hand side on the sensitivity of the solution to perturbations (§4).

We also provide a geometric interpretation (§5) of our condition numbers, which in turn leads to a geometric interpretation of rank-revealing QR factorisations. Unlike traditional condition numbers, our componentwise condition numbers are able to indicate how linearly dependent individual matrix columns are on other columns. They can therefore be considered a continuation of Stewart's work on collinearity in regression problems [33].

We further show that the relative errors in individual components of a linear system are reduced by column scaling only if column scaling manages to reduce the perturbations (§6). Two simple examples are given where our componentwise condition numbers are significantly more accurate than the norm-based condition numbers (§7). We extend the results for linear systems to the solution of linear least squares problems $\min_y \|Ay - b\|$ of full column rank (§8).

For the class of componentwise perturbations, we give necessary and sufficient conditions under which Skeel's condition numbers are informative, and we show that these conditions are similar to those where componentwise condition numbers are useful (§9). Numerical experiments not only confirm that these circumstances do occur frequently, they also illustrate that for many classes of matrices the ill conditioning of the matrix is due to *a few* rows of the inverse only (§11). This means that many of the solution components are computed more accurately than current analyses would lead us to believe. Finally we demonstrate that a componentwise error bound for componentwise perturbations can be significantly better than the norm-based error bounds.

Existing software can be used to compute or estimate componentwise condition numbers (§10). We also prove that the problem of estimating componentwise condition numbers for triangular matrices by means of the comparison matrix is well conditioned.

2. Condition numbers for linear systems. This section presents expressions for errors in individual solution components of linear systems with full column rank and defines condition numbers for each component.

As for notation, $\|\cdot\|$ represents the two-norm and e_i stands for the ith column of the identity matrix I. Let A be an $n \times m$ matrix A of rank m. Its condition number is $\kappa(A) = \|A\| \, \|A^\dagger\|$ and the rows of its left-inverse $A^\dagger$ are denoted by r_i^T.

Regarding perturbations in the right-hand side, the treatment of linear systems and least squares problems can be combined. Suppose the exact solution $x \neq 0$ solves $\min_y \|Ay - b\|$, while the computed solution $\bar{x}$ solves $\min_y \|Ay - (b + f)\|$. Let β_i be the angle between r_i and b, and ψ_i the angle between r_i and f. If $x_i \neq 0$ and $\epsilon_b = \|f\|/\|b\|$ then

$$\text{(RE1)} \qquad \frac{\bar{x}_i - x_i}{x_i} = \frac{1}{\cos\beta_i} \, \epsilon_b \cos\psi_i = \frac{\|b\|}{\|A\| \, \|x\|} \, \frac{\|x\|}{x_i} \, \|A\| \, \|r_i\| \, \epsilon_b \cos\psi_i.$$

Regarding perturbations in the matrix of a linear system, suppose the exact solution $x \neq 0$ solves $Ax = b$, while the computed solution $\bar{x} \neq 0$ solves $(A + F)\bar{x} = b$.

Denote by ψ_i the angle between r_i and $F\bar{x}$. If $x_i \neq 0$ and $\epsilon_A = \frac{\|F\bar{x}\|}{\|A\|\,\|\bar{x}\|}$ then

$$\text{(RE2)} \qquad \frac{\bar{x}_i - x_i}{x_i} = -\frac{1}{\cos\beta_i}\frac{\|F\bar{x}\|}{\|b\|}\cos\psi_i = -\frac{\|\bar{x}\|}{x_i}\,\|A\|\,\|r_i\|\,\epsilon_A\cos\psi_i.$$

The perturbations in the first expressions for (RE1) and (RE2) are amplified by $1/\cos\beta_i$. Hence the relative error in $\bar{x}_i$ is likely to increase with increasing orthogonality of r_i and b.

The second expressions in (RE1) and (RE2) have two amplification factors in common: the magnitude of x_i relative to $\|x\|$, and the matrix condition $\|A\|\,\|r_i\| \leq \kappa(A)$. The term

$$\frac{\|b\|}{\|A\|\,\|x\|} \geq \frac{1}{\kappa(A)}$$

in (RE1) occurs in the error expressions for all $\bar{x}_i$ and describes the relation between matrix and right-hand side. In the case of linear systems $Ax = b$ it has the upper bound

$$\frac{\|b\|}{\|A\|\,\|x\|} = \frac{\|Ax\|}{\|A\|\,\|x\|} \leq 1.$$

The expressions (RE1) and (RE2) provide a clear separation of the three factors responsible for the loss of accuracy in the computed solution: relative magnitude of the solution components, matrix condition, and relationship between matrix and right-hand side.

Now we determine when the amplification factors are maximal. If $\|r_{\max}\| = \max_k \|r_k\|$ is the row of largest norm in $A^\dagger$ then

$$\text{(CN)} \qquad \|A\|\,\|r_{\max}\| \leq \kappa(A) \leq \sqrt{m}\|A\|\,\|r_{\max}\|.$$

Applying inequalities (CN) to the componentwise relative errors (RE1) and (RE2) shows that there must exist a component $\bar{x}_k$ for which

$$\frac{|\bar{x}_k - x_k|}{|x_k|} \geq \frac{1}{\sqrt{m}}\frac{\|b\|}{\|A\|\,\|x\|}\,\kappa(A)\,\frac{\|x\|}{|x_k|}\,\epsilon_b|\cos\psi_k|$$

and

$$\frac{|\bar{x}_k - x_k|}{|x_k|} \geq \frac{1}{\sqrt{m}}\kappa(A)\,\frac{\|x\|}{|x_k|}\,\epsilon_A|\cos\psi_k|.$$

Therefore, the sensitivity of x_k to matrix perturbations is proportional to the condition number of A, and is proportional to right-hand side perturbations only when the right-hand side has an appropriate direction, that is, whenever $\frac{\|b\|}{\|A\|\,\|x\|}$ is not too small.

DEFINITION 1. *Let $x \neq 0$ solve the linear system $Ax = b$ with $n \times m$ matrix A of rank m, and let $\bar{x} \neq 0$ be the computed solution. If $r_i^T = e_i^T A^\dagger$, then the quantities*

$$\frac{\|\bar{x}\|}{|x_i|}, \quad \|A\|\,\|r_i\|, \quad 1 \leq i \leq m,$$

are called componentwise condition numbers for the linear system *or* condition numbers for x_i.

Support for this kind of definition comes from earlier work of Stewart [33] who introduces the "collinearity indices" $\kappa_i = \|a_i\| \, \|r_i\|$ that represent the scaling-invariant version of $\|A\| \, \|r_i\|$. The main difference between Stewart's condition numbers and ours is that the collinearity indices are designed to reflect the linear dependence of the matrix columns, while our componentwise condition numbers measure the conditioning of the linear system: matrix plus right-hand side.

In 1970 van der Sluis [38], [39] realised the need to distinguish the conditioning of individual components of x and the fact that the conditioning depends on the relative size of a component. He introduced the notion of "ith column condition number of A," $\|A^{-1}\| \, \|a_i\|$, and derived the similar-looking normwise relative error bound (here $f = 0$)

$$\frac{\|\bar{x} - x\|}{\|x\|} \leq \frac{\|F\|}{\|A\|} \sum_i \|A^{-1}\| \, \|a_i\| \frac{|x_i|}{\|x\|}.$$

3. Conventional error bounds. This section argues that for any linear system there exist perturbations for which the norm-based bounds on the relative error are as tight as possible. We also justify our particular representation of the matrix perturbations.

It follows from (RE1) and (CN) that for perturbations of the right-hand side,

$$\frac{1}{\sqrt{m}} \kappa(A) \frac{\|b\|}{\|A\| \, \|x\|} \epsilon_b \mu \leq \frac{\|\bar{x} - x\|}{\|x\|} \leq \sqrt{m} \kappa(A) \frac{\|b\|}{\|A\| \, \|x\|} \epsilon_b,$$

where

$$\epsilon_b = \frac{\|f\|}{\|b\|}, \qquad \mu = \frac{\max_i \{\|r_i\| \, |\cos \psi_i|\}}{\max_k \|r_k\|}.$$

As for perturbations of the matrix,

$$\frac{1}{\sqrt{m}} \kappa(A) \frac{\|\bar{x}\|}{\|x\|} \epsilon_A \mu \leq \frac{\|\bar{x} - x\|}{\|x\|} \leq \sqrt{m} \kappa(A) \frac{\|\bar{x}\|}{\|x\|} \epsilon_A,$$

where $\epsilon_A = \frac{\|F\bar{x}\|}{\|A\| \, \|\bar{x}\|}$.

In the absence of knowledge about the values of $\cos \psi_i$, we must assume the worst case $\mu = 1$, which implies that the norm-based error bounds are tight. Thus the conventional upper bounds are as good as possible given that one has chosen to measure a *norm*-based error. As a consequence, if the normwise bounds give unsatisfying information, it is not because the bounds are loose, but rather because an unsatisfying way of measuring the error was adopted in the first place.

The upper bounds for nonsingular linear systems commonly found in the literature are of the form

$$\frac{\|\bar{x} - x\|}{\|x\|} \leq \frac{\kappa(A)}{1 - \kappa(A)\rho_A} (\rho_A + \epsilon_b), \qquad \|A^{-1}F\| < 1,$$

e.g., §III.2.3 in [35], where the matrix perturbations are represented by $\rho_A = \|F\|/\|A\|$. In contrast, our representation of the matrix perturbations is ϵ_A. This is a sensible measure because ϵ_A represents the smallest possible matrix perturbation, as we now show.

For given $Ax = b$ and $\bar{x}$, let $F_{\min}$ be the perturbation of smallest Frobenius norm among all perturbations F that satisfy $(A+F)\bar{x} = b$ ($F_{\min}$ also has smallest two-norm

among all such perturbations). From Theorem III.2.16 in [35], and also [27], it follows that $F_{\min}$ satisfies

$$(A + F_{\min})\bar{x} = b, \qquad \epsilon_{\min} = \frac{\|F_{\min}\|}{\|A\|} = \frac{\|F\bar{x}\|}{\|A\|\,\|\bar{x}\|},$$

which is exactly the matrix perturbation ϵ_A in the relative error (RE2).

4. Special right-hand sides for linear systems. This section analyses error bounds for linear systems $Ax = b$ whose right-hand side b is a singular vector associated with the smallest singular value σ_m of A. We show that in this case *all* solution components are sensitive to perturbations.

In this case

$$\|A^\dagger b\|/\|b\| = 1/\sigma_m = \|A^\dagger\|$$

and

$$\frac{\|b\|}{\|A\|\,\|x\|} = \frac{1}{\kappa(A)}, \qquad \frac{\|b\|}{\|A\|\,\|x\|}\,\|A\|\,\|r_i\| = \frac{\|r_i\|}{\|A^\dagger\|} \le 1.$$

This implies together with (RE1) that the relative sensitvity of all solution components to right-hand side perturbations is solely determined by their relative magnitude.

According to §2, the norm-based error satisfies

$$\frac{1}{\sqrt{m}}\epsilon_b\mu \le \frac{\|\bar{x} - x\|}{\|x\|} \le \sqrt{m}\epsilon_b.$$

This means the norm-based relative error is about the same magnitude as the perturbation in the right-hand side and does not depend on the condition number of A. This was already observed in [6].

Chan and Foulser [6] try to incorporate a potential relationship between right-hand side and matrix by modifying the conventional bound

$$\frac{\|\bar{x} - x\|}{\|x\|} \le \frac{\kappa(A)}{1 - \kappa(A)\rho_A}(\rho_A + \epsilon_b), \qquad \rho_A = \frac{\|F\|}{\|A\|}.$$

Let

$$A = U\Sigma V^T, \qquad \text{where} \quad U = (\,u_1 \quad \dots \quad u_n\,), \qquad \sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n > 0,$$

be the singular value decomposition (SVD) of a nonsingular matrix A with singular values σ_i and right singular vectors u_i. According to Theorem 1 in [6], if $A\bar{x} = b + f$ and P_k is the orthogonal projection onto the space spanned by $u_{n-k+1}, \dots, u_n$,

$$\frac{\|\bar{x} - x\|}{\|x\|} \le \frac{\sigma_{n-k+1}}{\sigma_n}\left(\frac{\|P_k b\|}{\|b\|}\right)^{-1}\epsilon_b.$$

They conclude that if, for some k, a large fraction of b lies in the space spanned by $u_{n-k+1}, \dots, u_n$, and if $\sigma_{n-k+1} \approx \sigma_n$, then x "is relatively insensitive to perturbations in b." For instance, if $b = u_n$ then $P_1 b = b$,

$$\frac{\|\bar{x} - x\|}{\|x\|} \le \epsilon_b,$$

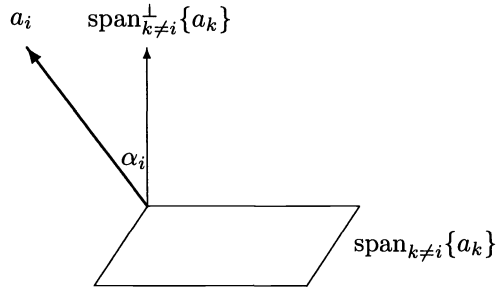


FIG. 1. *Angles associated with columns.*

and we conclude that x is insensitive to perturbations in b.

The interpretation of Theorem 1 given in [6] is valid if f represents the *input* error in the data b. However, we do not agree with the application of Theorem 1 in the case when f represents a backward error chosen to satisfy $A\bar{x} = b + f$ because f depends on the size of $\bar{x}$. Since $F_{\min} = -f\bar{x}^T/\bar{x}^T\bar{x}$ is the perturbation of smallest two-norm and Frobenius norm satisfying $(A + F_{\min})\bar{x} = b$, Theorem III.2.16 in [35], we obtain from the first expression in (RE1)

$$\frac{|\bar{x}_i - x_i|}{|x_i|} \geq \frac{\|A\|\,\|\bar{x}\|}{\|b\|}\,\epsilon_{\min}|\cos\psi_i|.$$

When $b = u_n$, the common term $\|A\|\,\|\bar{x}\|/\|b\|$ is approximately σ_1/σ_n and the sensitivity of *all* solution components is proportional to the condition number. A slightly different argument based on the use of the perturbations

$$\epsilon_{\min} = \frac{\|F_{\min}\|}{\|A\|} = \frac{\|b\|}{\|A\|\,\|\bar{x}\|}\epsilon_b$$

implies that for $b = u_n$ we have $\epsilon_b \approx \kappa(A)\epsilon_{\min}$ and the ill conditioning is merely hidden in the perturbation ϵ_b. Consequently, all components of x are extremely sensitive to perturbations if A is ill conditioned, which disagrees with the interpretation in [6].

5. Geometric interpretation. This section gives a geometric interpretation of the componentwise condition numbers. It is shown that $\|r_i\|$ reflects the linear dependence of column i of A on all other columns. This, in turn, leads to a geometric justification for rank-revealing QR factorisations.

First of all, the size of the $\|r_i\|$ reflects the linear dependence of the ith column of A on all others because

$$\|r_i\| = \frac{1}{\|a_i\|\cos\alpha_i},$$

where α_i is the angle between r_i and a_i. This follows from the expression $1 = r_i^T a_i = \|r_i\|\,\|a_i\|\cos\alpha_i$ for the ith diagonal element of $I = A^\dagger A$, which also implies that $\cos\alpha_i > 0$, so $-\frac{1}{2}\pi < \alpha_i < \frac{1}{2}\pi$. Because $e_i^T = r_i^T A$, r_i is orthogonal to all columns of A except for a_i, see Fig. 1.

To obtain a geometric meaning for r_1, partition $A = (\,a_1 \quad A_1\,)$, where a_1 represents the first column of A and A_1 represents the remaining columns. Let $-\hat{a}_1$ be the residual in the least squares approximation of a_1 by the columns of A_1,

$\|\hat{a}_1\| = \min_y \|A_1 y - a_1\|$ and let $-\check{A}_1$ be the residual in the least squares approximation of the columns of A_1 by a_1, $\|\check{A}_1\| = \min_y \|a_1 y^T - A_1\|$. As in the derivation of the formulae for partial correlation coefficients in [11] one can now show that

$$A^\dagger = (A^T A)^{-1} A^T = \begin{pmatrix} (\hat{a}_1^T a_1)^{-1} \hat{a}_1^T \\ \\ (\check{A}_1^T A_1)^{-1} \check{A}_1^T \end{pmatrix}.$$

It follows that the first row r_1^T of $A^\dagger$ lies in the same direction as the residual $-\hat{a}_1$ in the least squares approximation of column a_1 by the remaining columns. The residual, in turn, is just the projection of a_1 onto the orthogonal complement of the range of A_1. Hence, $\|r_1\| = 1/\|\hat{a}_1\|$, which means that increasing linear dependence of a_1 on the other columns leads to larger $\|r_1\|$. Analogous statements hold for the other rows r_i^T of $A^\dagger$.

Already in [33] Stewart used a different argument to show that

$$\|\hat{a}_i\| = \min_y \|A_i y - a_i\| = 1/\|r_i\|.$$

Here we provide more justification for the choice of r_i as an indicator of sensitivity: because r_i is a multiple of the residual $\hat{a}_i$, the residual is inherent in A and thus represents a most natural choice for sensitivity measure.

Angles between subspaces spanned by different columns of a matrix also occur in the context of nonsymmetric eigenvalue problems [12], [29].

5.1. Application. Our geometric interpretation of the rows of the left-inverse explains certain algorithms for rank-revealing QR factorisations. These factorisations appeared first in [15], [4], [16], [18], and are further analysed and refined in [32], [13], [5], [33], [9]. In the simplest case, the goal of a rank-revealing QR factorisation is to determine the most linearly dependent column of a matrix A.

The idea [9], [32] is based on the existence of a row of $A^\dagger$ that approximates $\|A^\dagger\|$ well. Perform a QR factorisation $AP = QR$, where Q has orthonormal columns, R is upper triangular, and the permutation matrix P is chosen so as to minimise the magnitude of the trailing diagonal element $(R)_{mm}$ of R. Then the inverse of this element, $1/|(R)_{mm}| = \|e_m^T R^{-1}\| = \|r_m^T\|$, is as large as possible, and the residual $1/\|r_m^T\|$ is as small as possible. Therefore the last column of AP is the column that can be best approximated by all other columns and so is the most linearly dependent among all columns.

6. Implications for column scaling. This section shows that the componentwise relative error decreases under column scaling only if column scaling actually reduces the perturbations.

A diagonal column scaling D of the least squares problem $\min_y \|Ay - b\|$ to $\min_z \|(AD)z - b\|$, where D is a nonsingular diagonal matrix, changes only the lengths of the columns but not the angles. In case of a column equilibrated matrix AD, [17, §3.5.2], and [37], [38], where the diagonal matrix D is chosen so that all columns of AD have identical length, the condition number of AD comes from the largest angle of A,

$$\frac{1}{\cos(\max_i \alpha_i)} \leq \|AD\| \, \|(AD)^\dagger\| \leq \frac{\sqrt{m}}{\cos(\max_i \alpha_i)}.$$

This bound already appeared in a different form in [33].

In [37], van der Sluis showed that a column equilibrated matrix A of order n has a condition number that is at most a factor $\sqrt{n}$ away from the lowest condition number among all matrices of the form AD. This would suggest that one could solve only linear systems and least squares problems with column equilibrated matrices so as to minimise the condition number in

$$\frac{\|\bar{x} - x\|}{\|x\|} \leq \sqrt{m}\kappa(A) \frac{\|b\|}{\|A\| \, \|x\|} \epsilon_b.$$

However, note that the condition number occurs in an upper bound.

Based on the expressions for the componentwise errors (RE1) and (RE2) we come to the following conclusions. In contrast to the norm-based condition numbers, the amplification factors $1/\cos\beta_i$ are preserved when the columns of A are multiplied by nonzero scalars. The computed solution $\bar{z}$ of the system $(AD)z = b$, where $z = D^{-1}x$, satisfies a perturbed system $AD\bar{z} = b + g$. Postmultiplication of A by D corresponds to premultiplication of $A^\dagger$ by D^{-1}, which changes only the lengths of the rows r_i^T in $A^\dagger$ but preserves the angles β_i between b and r_i. Hence the amplification factors $1/\cos\beta_i$ remain invariant under column scaling. Therefore the componentwise relative error decreases under column scaling only if column scaling manages to reduce the perturbations.

7. Example. This section contains two examples that illustrate the previous results. The first example represents a generalisation of the example from §1.1 and demonstrates that even a very ill-conditioned matrix may have robust solution components.

Consider a 4×4 orthogonal matrix $A = (\, a_1 \quad a_2 \quad a_3 \quad a_4 \,)$ and define a one-parameter family of matrices by

$$A(\lambda) = \left(a_1 \quad a_2 \quad a_3 \quad \tfrac{1}{\sqrt{1+\lambda^2}}(\lambda a_3 + a_4) \right).$$

Obviously $A(0) = A$ is a well-conditioned matrix and $A(\infty)$ is a singular matrix. For all λ, $\|A(\lambda)\| \leq 2$. When $\lambda < \infty$, the inverse is given by

$$[A(\lambda)]^{-1} = \begin{pmatrix} a_1^T \\ a_2^T \\ a_3^T - \lambda a_4^T \\ \sqrt{1+\lambda^2} a_4^T \end{pmatrix},$$

from which one computes

$$\cos\alpha_1 = \|a_1\| \cos\alpha_1 = \cos\alpha_2 = \|a_2\| \cos\alpha_2 = 1,$$
$$\cos\alpha_3 = \|a_3\| \cos(\alpha_3) = \cos\alpha_4 = \|a_4\| \cos(\alpha_4) = \frac{1}{\sqrt{1+\lambda^2}}.$$

Thus as $\lambda \to \infty$ the matrix $A(\lambda)$ becomes increasingly singular. Its condition number behaves like $O(\lambda)$. Note that the matrix $A(\lambda)$ is column equilibrated (and not necessarily row ill scaled) so the ill conditioning is a result of small angles rather than short columns.

Consider a linear system $A(\lambda)x(\lambda) = b$, where the right-hand side is independent of λ and can be represented as $b = \tau_1 a_1 + \tau_2 a_2 + \tau_3 a_3 + \tau_4 a_4$. Then

$$\cos\beta_1 = \frac{\tau_1}{\|b\|}, \quad \cos\beta_2 = \frac{\tau_2}{\|b\|}, \quad \cos\beta_3 = \frac{\tau_3 - \lambda\tau_4}{\|b\|\sqrt{1+\lambda^2}}, \quad \cos\beta_4 = \frac{\tau_4}{\|b\|}.$$

The solution vector is given by

$$x(\lambda) = (\,\tau_1 \quad \tau_2 \quad \tau_3 - \lambda\tau_4 \quad \sqrt{1+\lambda^2}\tau_4\,)^T .$$

The values of x_1 and x_2 are independent of λ, and so are $\|a_j\| \cos\alpha_j$ and $\cos\beta_j$ for $j = 1, 2$. So the sensitivity of the components x_1 and x_2 depends solely on their size relative to x. If, for instance, $|x_1| \gg |x_i|$ for $i \neq 1$ then, according to (RE), the error in x_1 is not amplified independent of the values of λ and the condition number of $A(\lambda)$.

The second example shows that *all* solution components can be sensitive to perturbations when the choice of right-hand side is unfortunate. In [25] we show that systems with uniformly sensitive solution components also occur in ill-conditioned Markov problems.

The coefficient matrix of the linear system $Ax = b$ is the Hilbert matrix with elements $1/(i+j-1)$ of order 4,

$$A = \begin{pmatrix} 1.00000000000000 & 0.50000000000000 & 0.33333333333333 & 0.25000000000000 \\ 0.50000000000000 & 0.33333333333333 & 0.25000000000000 & 0.20000000000000 \\ 0.33333333333333 & 0.25000000000000 & 0.20000000000000 & 0.16666666666667 \\ 0.25000000000000 & 0.20000000000000 & 0.16666666666667 & 0.14285714285714 \end{pmatrix}$$

while the right-hand side

$$b^T = (-0.02919332316479\ 0.32871205576319\ -0.79141114583313\ 0.51455274999716)$$

is a left singular vector corresponding to the smallest singular value of A. The condition number of A is at least 10^4. If the error matrix F has norm $\|F\| = 10^{-3}\|A\|$, then the solution of the system $(A+F)\bar{x} = b$ contains at least one component that has no accurate digits. We choose the following random matrix with norm $10^{-3}\|A\|$,

$$F = \begin{pmatrix} 0.00057208543036 & 0.00017162562911 & 0.00038139028691 & 0.00038139028691 \\ 0.00019069514345 & 0.00057208543036 & 0.00019069514345 & 0.00057208543036 \\ 0.00019069514345 & 0.00057208543036 & 0.00019069514345 & 0.00057208543036 \\ 0.00038139028691 & 0.00038139028691 & 0.00057208543036 & 0.00005720854304 \end{pmatrix}.$$

Computing x and $\bar{x}$ in 16-digit arithmetic gives

$$x = \begin{pmatrix} -301.88859986174430 \\ 3399.21637943995029 \\ -8183.99472310610599 \\ 5320.99783141589251 \end{pmatrix}, \qquad \bar{x} = \begin{pmatrix} 81.63154025985811 \\ -1310.35333852711346 \\ 3649.17285297454328 \\ -2572.42993839543533 \end{pmatrix}.$$

The components of $\bar{x}$ do not even have the correct sign, let alone any accurate digits. So all solution components of this system are sensitive to perturbations.

8. Condition numbers for least squares problems. This section presents expressions for componentwise errors in the solution of least squares problems of full column rank. The treatment in §2 on perturbations of the right-hand side is now extended to also allow perturbations in the matrix.

Suppose $x \neq 0$ solves

$$\min_y \|Ay - b\|, \qquad \text{where} \qquad r = b - Ax,$$

and $\bar{x} \neq 0$ solves

$$\min_y \|(A+F)y - (b+f)\|, \qquad \text{where} \qquad \bar{r} = b + f - (A+F)\bar{x} \neq 0.$$

Let $q_i^T = e_i^T (A^T A)^{-1}$ and define the following error angles: $\psi_{F,i}$ is the angle between r_i and $F\bar{x}$, $\psi_{f,i}$ is the angle between r_i and f, ω_i is the angle between r_i and $\bar{r}$, and $\omega_{q,i}$ is the angle between q_i and $F^T\bar{r}$. By applying (RE1) to the associated augmented nonsingular system one can show the following.

If $x_i \neq 0$ and $\epsilon_{A,r} = \frac{\|F^T\bar{r}\|}{\|A^T\|\,\|\bar{r}\|}$ then

$$\text{(LS)} \quad \frac{\bar{x}_i - x_i}{x_i} = \text{RE} + \frac{\|\bar{r}\|}{\|b\|}\frac{1}{\cos\beta_i}\cos\omega_i = \text{RE} + \frac{\|\bar{r}\|}{\|A\|\,\|\bar{x}\|}\frac{\|\bar{x}\|}{x_i}\,\|q_i\|\,\|A\|^2\,\epsilon_{A,r}\cos\omega_{q,i},$$

where

$$\text{RE} = -\frac{\|\bar{x}\|}{x_i}\,\|A\|\,\|r_i\|\left[\epsilon_A\cos\psi_{F,i} - \frac{\|b\|}{\|A\|\,\|\bar{x}\|}\epsilon_b\cos\psi_{f,i}\right]$$

is the componentwise relative error in the solution of a linear system solution.

Equations (LS) contain two different expressions that account for the least squares nature. The perturbation in the first expression is amplified by $1/\cos\beta_i$, which reflects how linearly dependent b is on the space spanned by a_k, $k \neq i$; and it is invariant under column scaling.

The relative perturbation $\epsilon_{A,r}\cos\omega_{q,i}$ in the second expression is amplified by three factors. The first factor represents, as in the error for linear system solution, the size of the component x_i relative to $\|\bar{x}\|$. The second factor $\|q_i\|\,\|A\|^2$ has the bounds

$$(\|r_i\|\,\|A\|)^2 \leq \|q_i\|\,\|A\|^2 \leq \|(A^TA)^{-1}\|\,\|A\|^2 = \kappa^2(A),$$

as a result of $\|q_i\| \geq \|r_i\|^2$. Since there exists a row r_k of $A^\dagger$ whose norm approximates $\|A^\dagger\|$ to a factor of $\sqrt{m}$, there must exist at least one component x_k for which

$$\|q_k\|\,\|A\|^2 \geq \frac{1}{m}\kappa^2(A).$$

The third factor $\frac{\|\bar{r}\|}{\|A\|\,\|\bar{x}\|}$ describes the relationship between matrix and right-hand side. If θ is the angle between b and the range of A, then the exact residual r satisfies

$$\frac{1}{\kappa(A)}\tan\theta \leq \frac{\|r\|}{\|A\|\,\|x\|} \leq \tan\theta$$

and for some x_k

$$\frac{1}{m}\kappa(A)\tan\theta \leq \frac{\|r\|}{\|A\|\,\|x\|}\,\|q_k\|\,\|A\|^2 \leq \kappa^2(A)\tan\theta.$$

Consequently, least squares problems are always more sensitive to ill conditioning than linear systems and, depending on the angle between b and the range of A, their sensitivity may be as high as the square of the condition number.

DEFINITION 2. *Let $x \neq 0$ solve the least squares problem $\min_y \|Ay - b\|$ with $n \times m$ matrix A of* rank m, *and let $\bar{x} \neq 0$ be the computed solution with residual $\bar{r} \neq 0$. If $q_i = e_i^T(A^TA)^{-1}$ and $r_i^T = e_i^TA^\dagger$ then the quantities*

$$\frac{\|\bar{x}\|}{|x_i|}, \quad \|A\|\,\|r_i\|, \quad \frac{\|\bar{r}\|}{\|A\|\,\|\bar{x}\|}\,\|A\|^2\|q_i\|, \quad 1 \leq i \leq m,$$

are called componentwise condition numbers for the least squares problem.

The condition numbers for linear systems from [38] and [39] are extended to least squares problems in [14].

9. A special class of perturbations. Unlike the previous sections, which assumed no knowledge about the perturbations, this section analyses the reduction in error bounds brought about by the special structure of perturbations resulting from floating point computations. This issue was first investigated by Skeel in [31] for the case of "componentwise perturbations." We provide necessary and sufficient conditions under which Skeel's condition numbers are useful, and we show that these conditions are similar to those where componentwise condition numbers are useful. The experiments in §11 illustrate that these conditions indeed occur frequently.

For $Ax = b$ and $(A + F)\bar{x} = b + f$ the perturbations F and f are called *componentwise perturbations* if the inequalities

$$|F| \leq \epsilon |A|, \qquad |f| \leq \epsilon |b|$$

hold componentwise for some $\epsilon \geq 0$.

In [31] Skeel defines a condition number that exploits componentwise perturbations. Theorem 2.1 in [31] shows that

$$\frac{\|\bar{x} - x\|_\infty}{\|x\|_\infty} \leq \epsilon \frac{\| \, |A^{-1}| \, |A| \, |x| \, + \, |A^{-1}| \, |b| \, \|_\infty}{(1 - \epsilon \| \, |A^{-1}| \, |A| \, \|_\infty) \, \|x\|_\infty},$$

and Skeel uses

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

as the condition number for the linear system $Ax = b$. He also introduces

$$\mathrm{cond}(A) = \| \, |A^{-1}| \, |A| \, \|_\infty$$

as an upper bound for $\mathrm{cond}(A, x)$. A componentwise version of Skeel's condition number

$$e_i^T (|A^{-1}| \, |A| \, |x| \, + \, |A^{-1}| \, |b|)/|x_i|$$

is advocated in [28]; and [1] introduces condition numbers similar to the one used by Skeel for matrix inversion, least squares problems, and the solution of Vandermonde-like linear systems.

Skeel's condition number is invariant under row-scaling. Therefore, $\mathrm{cond}(A)$ may be much lower than the traditional condition number $\kappa_\infty(A) = \|A^{-1}\|_\infty \|A\|_\infty$ when the rows of A are ill scaled, i.e., when the norms of the rows of A differ widely. But the less known fact is that $\mathrm{cond}(A)$ can be much lower than $\kappa_\infty(A)$ *only* when the rows of A are ill scaled. The reasoning is as follows. Let e be the vector of all ones and D_R a nonsingular diagonal matrix with $D_R|A|e = e$, that is, the diagonal elements of D_R are the inverse row norms of A. Then

$$\frac{\kappa_\infty(A)}{\kappa_\infty(D_R)} \leq \mathrm{cond}(A) \leq \kappa_\infty(A).$$

This means, if $\kappa_\infty(D) \approx 1$ then the rows of A are not ill scaled and $\mathrm{cond}(A) \approx \kappa_\infty(A)$, which limits the applicability of $\mathrm{cond}(A)$.

Remember that our componentwise condition numbers are useful if there are large differences among the $\|r_i\|$. It turns out that something similar holds for $\mathrm{cond}(A,x)$: $\mathrm{cond}(A,x)$ is useful only when the norms of the columns of A^{-1} differ widely in magnitude. Denote the columns of A by a_i and the columns of A^{-1} by p_i. If $j(i)$ is the index of the largest element in column i, $\|a_i\|_\infty = |a_{i,j(i)}|$, then

$$\| \, |A^{-1}| \, |A| \, |x| \, \|_\infty \geq \|p_{j(i)}\|_\infty \|a_i\|_\infty |x_i| = \|A\|_1 \|p_{j(i)}\|_\infty |x_i| \frac{\|a_i\|_\infty}{\|A\|_1}.$$

Choosing i such that $|x_i| = \|x\|_\infty$ and defining D_C as the diagonal matrix that equilibrates the columns of A, $e^T|A|D_C = e^T$, gives

$$\begin{aligned} \mathrm{cond}(A,x) = \frac{\| \, |A^{-1}| \, |A| \, |x| \, \|_\infty}{\|x\|_\infty} &\geq \|p_{j(i)}\|_\infty \|A\|_1 \frac{1}{n\kappa_\infty(D_C)} \\ &\geq \frac{\min_i \|p_i\|_\infty}{\|p_j\|_\infty} \kappa_1(A) \frac{1}{n^2\kappa_\infty(D_C)} \end{aligned}$$

for some column p_j of A^{-1}. This means, $\mathrm{cond}(A,x) \approx \kappa_1(A)$ for all x whenever A^{-1} is not badly column-scaled. Therefore the conditions under which $\mathrm{cond}(A,x)$ is useful are quite similar to those for our componentwise condition numbers.

It is possible to profitably combine Skeel's analysis with our componentwise errors because componentwise perturbations induce upper bounds on the cosines. If $A\bar{x} = b + f$ and $|f| \leq \epsilon|b|$ then

$$|\cos\psi_i| = \frac{|r_i^T f|}{\|r_i\| \, \|f\|} \leq \epsilon \frac{|r_i^T| \, |b|}{\|r_i\| \, \|f\|},$$

implies

$$\frac{|\bar{x}_i - x_i|}{|x_i|} \leq \epsilon \frac{|r_i^T| \, |b|}{|r_i^T b|} \tag{CRE1}$$

for error (RE1). Similarly, if $(A+F)\bar{x} = b$ and $|F| \leq \epsilon|A|$ then the upper bound for (RE2) simplifies to

$$\frac{|\bar{x}_i - x_i|}{|x_i|} \leq \epsilon \frac{|r_i^T| \, |A| \, |\bar{x}|}{|x_i|}. \tag{CRE2}$$

This last inequality illustrates that componentwise perturbations in our error expressions lead to a componentwise version of Skeel's condition number $\mathrm{cond}(A,x)$. Although these expressions already exist implicitly in Skeel's work, it is the observation that we lose a lot by taking norms that is important. Because the rows of the inverse may differ significantly in size, the difference between our bounds and $\mathrm{cond}(A,x)$ may be arbitrarily large as shown in the following example.

Let $\epsilon > 0$ and

$$A = \begin{pmatrix} \epsilon & \frac{1}{\epsilon} \\ \frac{1}{\epsilon} & \end{pmatrix}, \qquad b = \begin{pmatrix} \epsilon^2 + \frac{1}{\epsilon} \\ \frac{1}{\epsilon} \end{pmatrix},$$

so that

$$A^{-1} = \begin{pmatrix} \frac{1}{\epsilon} & -\frac{1}{\epsilon} \\ & \epsilon \end{pmatrix}, \qquad x = \begin{pmatrix} \epsilon \\ 1 \end{pmatrix}.$$

Hence

$$|A^{-1}|\,|A|\,|x| = |A^{-1}|\,|b| = \begin{pmatrix} \epsilon + \frac{2}{\epsilon^2} \\ 1 \end{pmatrix}.$$

Skeel's condition number $\mathrm{cond}(A,x)$ is unbounded as ϵ becomes small. In fact, it has the same order of magnitude $O(1/\epsilon^2)$ as the traditional condition number $\kappa_\infty(A)$, although the columns of A are badly scaled (this is because $|x|$ lies almost in the singular direction corresponding to the largest singular value of $|A^{-1}|\,|A|$). In contrast, the amplifier in our error bound for x_2, which is the largest component of x, equals $|r_2^T|\,|A|\,|x|/|x_2| = 1$.

Because our error expressions represent a componentwise version of Skeel's condition number, we get the same componentwise error bounds as appear in the literature. For instance, when $|A^{-1}|\,|b| = |x|$, as is the case for certain Vandermonde systems [20] and M-matrices with positive right-hand sides, the term amplifying ϵ in (CRE1) equals one. So the individual solution components are insensitive to perturbations in the right-hand side (an algorithm for such systems that gives rise to a small componentwise backward error f is called "weakly stable" in [23]).

For triangular M-matrices A with positive right-hand side b, it is shown in [22] that

$$|A^{-1}|\,|A|\,|x| \leq (2n-1)|x|,$$

which implies

$$|r_i^T|\,|A|\,|\bar{x}| = e_i^T A^{-1}|A|\bar{x} \leq (2n-1)|\bar{x}_i|.$$

Hence the term amplifying ϵ in (CRE2) is essentially bounded above by $2n-1$. This is true in particular if b is the vector of all ones. Thus, estimating the componentwise condition numbers of a triangular matrix by solving a linear system involving the comparison matrix, as in [21] and §10, is a well-conditioned problem.

10. Estimation of componentwise condition numbers. This section shows that componentwise condition numbers can be efficiently estimated with existing software.

For a $n \times m$ matrix A, bounds for $\|A\|$ can be determined in $O(mn)$ operations, and $\|\bar{x}\|/|x_i|$ and $\|\bar{r}\|$ can be estimated a posteriori in $O(mn)$ operations. This leaves the computation of $\|r_i\|$ and $\|q_i\|$. Numerical issues in the computation of the $\|r_i\|$, due to the potential ill conditioning of A, are addressed in [32] and in the context of statistical errors in [33]. If a factorization of A is available, then upper bounds on $\|r_i\|$ can be determined in $O(n^2)$ additional operations and an estimate of $\|q_i\|$ can be obtained by making use of the inequality $\|q_i\| \geq \|r_i\|^2$.

For instance, suppose the QR factorization

$$A = Q \begin{pmatrix} R \\ 0 \end{pmatrix}$$

is available, where Q is a $n \times n$ orthogonal matrix, and R is an $m \times m$ nonsingular upper triangular matrix. To compute $\|r_i\|$ and $\|q_i\|$, it suffices to work with R instead of A. From

$$q_i^T = e_i^T (A^T A)^{-1} = e_i^T R^{-1} R^{-T} = v_i^T R^{-T}, \qquad v_i = R^{-T} e_i,$$

it follows that q_i is the solution of the triangular system $Rq_i = v_i$ and $\|r_i\| = \|v_i\|$.

As for the actual computation of $\|q_i\|$ and $\|r_i\|$, observe that $v_m = R^{-T}e_m = \frac{1}{\rho}e_m$, where ρ is the element of R in position (m,m). Hence $\|r_m\| = 1/|\rho|$ and $Rq_m = \frac{1}{\rho}e_m$. Therefore, if a QR decomposition of A is available, $\|r_m\|$ is available right away and the computation of q_m requires $O(m^2)$ operations. This process can be carried out for all i, and is described in [32] for the computation of $\|r_i\|$ by permuting the columns of A. Gragg and Stewart [18] show how to efficiently "update" the QR factorisation from one permutation to the next in $O(m^2)$ operations; see also [17, §12.6].

Next, we indicate how the condition number estimators for triangular matrices in [21] can be used to compute upper bounds for the $\|r_i\|$ in $O(n^2)$ operations. Since A is triangular, $(A^{-1})_{ii} = 1/a_{ii}$ and $1/|a_{ii}| \leq \|r_i\| \leq \|r_i\|_1$. Replace A by its *comparison matrix* $C(A) = (c_{ij})$ of A [2], which is defined as

$$c_{ij} = \begin{cases} |a_{ii}| & \text{if } i = j, \\ -|a_{ij}| & \text{if } i \neq j, \end{cases}$$

and satisfies the componentwise inequalities

$$C(A)^{-1} \geq 0, \qquad |A^{-1}| \leq C(A)^{-1}$$

because it is an M-matrix [40]. The first inequality implies that the ith element of $C(A)^{-T}e$ equals $\|C(A)^{-T}e_i\|_1$, where e is the vector of all ones, while the second one implies $\|r_i\| \leq \|r_i\|_1 \leq \|C(A)^{-T}e_i\|_1$. Hence all $\|C(A)^{-T}e_i\|_1$ can be computed with a total of $O(n^2)$ operations by solving the system $C(A)^T y = e$. Since $C(A)$ is an M-matrix, so is $C(A)^T$. According to §9, the solution of linear systems with triangular M-matrices and positive right-hand side produces a small componentwise error. Hence, the estimation of componentwise condition numbers from the solution of $C(A)^T y = e$ is a well-conditioned problem.

In [7] we fit the linear-time algorithms in [19] for computing $\|A^{-1}\|_\infty$ for bi or tridiagonal matrices A to the computation of $\|r_i\|$.

We are currently investigating techniques based on appropriate rank-revealing QR decompositions that estimate componentwise condition numbers in $O(n^2)$ operations.

11. Numerical experiments. This section presents numerical experiments that reveal the existence of large classes of matrices for which the componentwise matrix condition numbers vary widely. For these matrices, componentwise condition numbers can therefore predict the sensitivity of individual solution components much more accurately than norm-based or Skeel's condition numbers.

Here we consider only nonsingular linear systems $Ax = b$. The componentwise condition numbers consist of two parts: the relative magnitude $\|\bar{x}\|/x_i$ of the solution component and the associated matrix condition $\|A\|\,\|r_i\|$ where $r_i = e_i^T A^{-1}$. We consider only matrices for which $\|r_i\|$ differ widely in size because they exhibit a large difference between $\|r_i\|$ and $\|A^{-1}\|$, as well as between $|r_i^T|\,|A|$ and $\mathrm{cond}(A)$. According to inequalities (CN), at least one $\|r_k\|$ approximates $\|A^{-1}\|$ to within a factor of $\sqrt{n}$, where n is the order of A. The potential for deviation of other $\|r_i\|$ from $\|A^{-1}\|$ increases, of course, with increasing ill conditioning of A. Below we present examples where some $\|r_i\|$ are orders of magnitude smaller than $\|A^{-1}\|$.

All experiments were performed in CLAM, version 2.00 [30], on a SPARCstation 1. The tests involved more than twenty classes of matrices, most of them from [24], their orders ranging up to $n = 500$. Among these, only the Minij and Pei matrices have r_i that are essentially identical in size. A group of matrices with a little more variation in

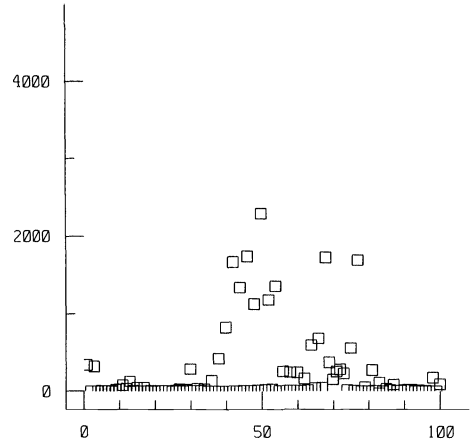


FIG. 2. *Random tridiagonal matrix A with $n = 100$, $\|A^{-1}\| = 4752$, $\min_i \|r_i\| = 1.1$.*

the $\|r_i\|$ are the highly ill-conditioned Pascal, Cauchy, Hilbert, and Lotkin matrices. In the group of matrices that comprises random symmetric and nonsymmetric matrices, random Toeplitz and Vandermonde matrices, at least half of the $\|r_i\|$ differ from $\|A^{-1}\|$ by a small multiple of ten. This means, at least half of the components of x are 1–2 digits more accurate than predicted by $\|A^{-1}\|$ (assuming the components are not too small). The group of matrices with the widest variation in the $\|r_i\|$ includes random tridiagonal matrices, Jordan matrices, Chebyshev–Vandermonde matrices, and triangular comparison matrices.

The surprising outcome of our experiments is that often only a few rows of A^{-1} are responsible for $\|A^{-1}\|$, while most of the remaining rows are small in size. This is more pronounced for ill-conditioned matrices. It also comes out in the plots in Figs. 2–7, where we plot $\|r_i\|$ against i, $1 \leq i \leq n$, for matrices from the last group. In case of high ill conditioning, the difference among the $\|r_i\|$ can be as high as 10^{15} for matrices of order $n = 100$. In addition, preliminary statistical analyses show that for these matrices usually more than half of the $\|r_i\|$ are small. Therefore, although a norm-based error bound would predict a total loss of accuracy, many components could actually be computed to a significant number of correct digits.

Figures 2–4 contain plots of three typical random tridiagonal matrices of order $n = 100$. The differences in the $\|r_i\|$ for each matrix are illustrated in Table 1.

TABLE 1

Figure	$\|A^{-1}\|$	$\min_i \|r_i\|$
2	4752	1.1
3	678	1.2
4	2577	1.1

Figure 5 shows the $\|r_i\|$ for a random Chebyshev–Vandermonde matrix of order $n = 10$, for which $\|A^{-1}\| = 11922$ and $\min \|r_i\| = .72$. Figures 6 and 7 plot the $\|r_i\|$ on a logarithmic scale for a random Jordan and a random unit upper triangular comparison matrix, respectively, both of order $n = 100$. The Jordan matrix has $\|A^{-1}\| = 4 \cdot 10^{14}$ and $\min \|r_i\| = 1.4$, while the triangular matrix has $\|A^{-1}\| = 2 \cdot 10^{17}$ and $\min \|r_i\| = 1.1$. Similar observations about the ill conditioning of random unit-triangular matrices are made in [36].

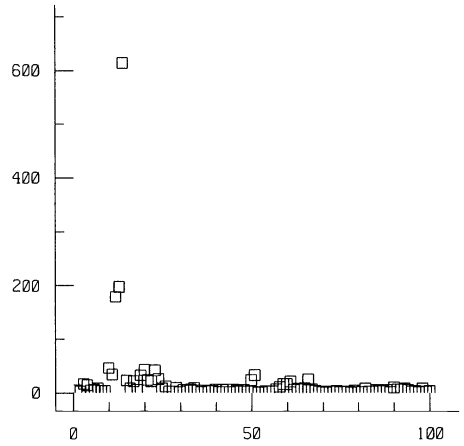


FIG. 3. *Random tridiagonal matrix A with* $n = 100$, $\|A^{-1}\| = 678$, $\min_i \|r_i\| = 1.2$.

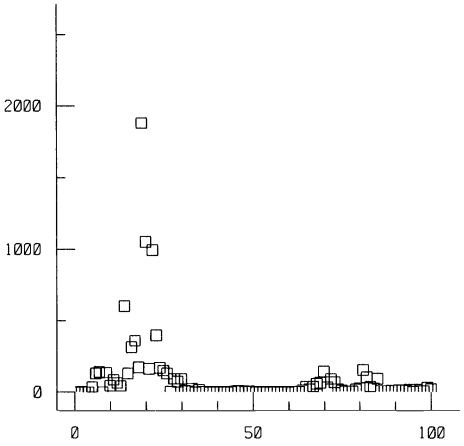


FIG. 4. *Random tridiagonal matrix A with* $n = 100$, $\|A^{-1}\| = 2577$, $\min_i \|r_i\| = 1.1$.

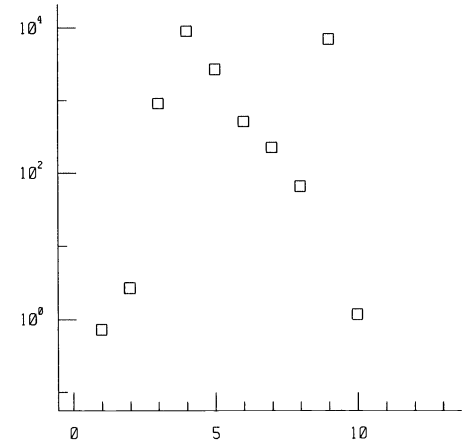


FIG. 5. *Random Chebyshev–Vandermonde matrix A with* $n = 10$, $\|A^{-1}\| = 11922$, $\min_i \|r_i\| = .72$.

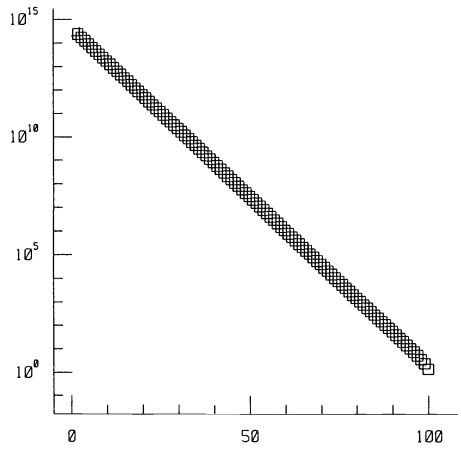


FIG. 6. *Random Jordan matrix A with diagonal element* .7183, $n = 100$, $\|A^{-1}\| = 4 \cdot 10^{14}$, $\min_i \|r_i\| = 1.4$.

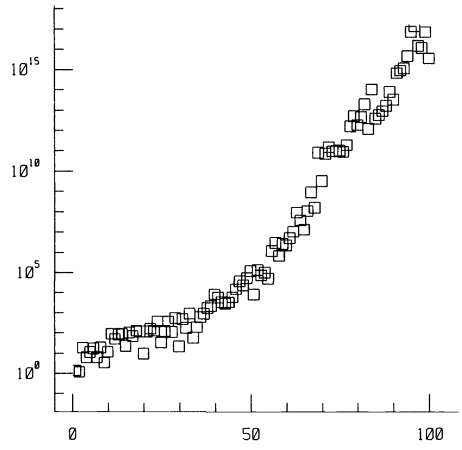


FIG. 7. *Random unit upper triangular comparison matrix A,* $n = 100$, $\|A^{-1}\| = 2 \cdot 10^{17}$, $\min_i \|r_i\| = 1.1$.

Acknowledgments. We would like to thank Stan Eisenstat and Jean-Marc Delosme for helpful discussions, Sham Sao for performing many of the numerical experiments, and Nick Higham for pointing out an error in an earlier version of the manuscript.

REFERENCES

[1] S. BARTELS AND D. HIGHAM, *The structured sensitivity of Vandermonde-like systems*, Numer. Math., 62 (1992), pp. 17–33.

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

[3] Å. BJÖRCK, *Component-wise perturbation analysis and error bounds for linear least squares solutions*, BIT, 31 (1991), pp. 238–244.

[4] P. BUSINGER AND G. GOLUB, *Linear least squares solutions by Householder transformations*, Numer. Math., 7 (1965), pp. 269–276.

[5] T. CHAN, *Rank revealing QR factorizations*, Linear Algebra Appl., 88/89 (1987), pp. 67–82.

[6] T. CHAN AND D. FOULSER, *Effectively well-conditioned linear systems*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 963–969.
[7] S. CHANDRASEKARAN AND I. IPSEN, *Perturbation theory for the solution of linear systems of equations and linear least squares problems*, Research Report 866, Department of Computer Science, Yale University, New Haven, CT, 1991.
[8] ———, *Finite precision analysis of inverse iteration*, manuscript.
[9] ———, *On rank-revealing QR factorisations*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 592–622.
[10] T. COLEMAN, D. SHALLOWAY, AND Z. WU, *Parallel packet annealing for molecular conformation* (II): *Simulated annealing and packet approximation*, Tech. Report, Advanced Computing Research Institute, Cornell University, Ithaca, NY, 1991.
[11] J. DELOSME AND I. IPSEN, *From Bareiss' algorithm to the stable computation of partial correlations*, J. Comput. Appl. Math., 27 (1989), pp. 53–91; *Parallel Algorithms for Numerical Linear Algebra* (*Advances in Parallel Computing*, 1), H. van der Vorst and P. van Dooren, eds., North Holland, Amsterdam, 1990.
[12] J. DEMMEL, *The condition number of equivalence transformations that block diagonalize matrix pencils*, SIAM J. Numer. Anal., 20 (1983), pp. 599–610.
[13] L. FOSTER, *Rank and null space calculations using matrix decomposition without column interchanges*, Linear Algebra Appl., 74 (1986), pp. 47–71.
[14] A. GEURTS, *A contribution to the theory of condition*, Numer. Math., 39 (1982), pp. 85–96.
[15] G. GOLUB, *Numerical methods for solving linear least squares problems*, Numer. Math., 7 (1965), pp. 206–216.
[16] G. GOLUB, V. KLEMA, AND G. STEWART, *Rank degeneracy and least squares problems*, Tech. Report STAN-CS-76-559, Computer Science Department, Stanford University, Stanford, CA, 1976.
[17] G. GOLUB AND C. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1989.
[18] W. GRAGG AND G. STEWART, *A stable variant of the secant method for solving nonlinear equations*, SIAM J. Numer. Anal., 13 (1976), pp. 889–903.
[19] N. HIGHAM, *Efficient algorithms for computing the condition number of a tridiagonal matrix*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 150–165.
[20] ———, *Error analysis of the Björck–Pereyra algorithms for solving Vandermonde systems*, Numer. Math., 50 (1987), pp. 613–632.
[21] ———, *A survey of condition number estimation for triangular matrices*, SIAM Rev., 29 (1987), pp. 575–596.
[22] ———, *The accuracy of solutions to triangular systems*, SIAM J. Numer. Anal., 26 (1989), pp. 1252–1265.
[23] ———, *Stability analysis of algorithms for solving confluent Vandermonde-like systems*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 23–41.
[24] ———, *Algorithm* 694: *A collection of test matrices in MATLAB*, ACM TOMS, 17 (1991), pp. 289–305.
[25] I. IPSEN AND C. MEYER, *Uniform stability of Markov chains*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1061–1074.
[26] D. KEYES AND M. SMOOKE, *Flame sheet starting estimates for counterflow diffusion flame problems*, J. Comput. Physics, 73 (1987), pp. 267–288.
[27] J. RIGAL AND J. GACHES, *On the compatibility of a given solution with the data of a linear system*, JACM, 14 (1967), pp. 543–548.
[28] J. ROHN, *New condition numbers for matrices and linear systems*, Computing, 41 (1989), pp. 167–169.
[29] A. RUHE, *Properties of a matrix with a very ill-conditioned eigenproblem*, Numer. Math, 15 (1970), pp. 57–60.
[30] SCIENTIFIC COMPUTING ASSOCIATES, INC., *CLAM User's Guide, The Computational Linear Algebra Machine*, New Haven, CT, 1990.
[31] R. SKEEL, *Scaling for numerical stability in Gaussian elimination*, JACM, 26 (1979), pp. 494–526.
[32] G. STEWART, *Rank degeneracy*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 403–413.
[33] ———, *Collinearity and least squares regression*, Statist. Sci., 2 (1987), pp. 68–100.
[34] ———, *Stochastic perturbation theory*, SIAM Rev., 32 (1990), pp. 579–610.
[35] G. STEWART AND J. SUN, *Matrix Perturbation Theory*, Academic Press, New York, 1990.
[36] L. TREFETHEN AND R. SCHREIBER, *Average-case stability of Gaussian elimination*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 335–360.

[37] A. VAN DER SLUIS, *Condition numbers and equilibration of matrices*, Numer. Math., 14 (1969), pp. 14–23.
[38] ———, *Condition, equilibration and pivoting in linear algebraic systems*, Numer. Math., 14 (1970), pp. 74–86.
[39] ———, *Stability of solutions of linear algebraic systems*, Numer. Math., 14 (1970), pp. 246–251.
[40] R. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 113–122, January 1995

008

AN INDEX THEOREM FOR MONOTONE MATRIX-VALUED FUNCTIONS*

WERNER KRATZ†

Abstract. The main result of this paper is the following index theorem, which is closely related to oscillation theorems on linear selfadjoint differential systems such as results by M. Morse. Let real $m \times m$-matrices R_1, R_2, X, U be given, which satisfy

$$R_1 R_2^T = R_2 R_1^T, \quad X^T U = U^T X, \quad \text{rank}\,(R_1, R_2) = \text{rank}\,(X^T, U^T) = m.$$

Moreover, assume that $X(t)$, $U(t)$ are real $m \times m$-matrix-valued functions on some interval $\mathcal{J} = [-\varepsilon, \varepsilon]$, $\varepsilon > 0$, such that

$$X^T(t)U(t) = U^T(t)X(t) \quad \text{on} \quad \mathcal{J},$$

$$X(t) \to X \quad \text{and} \quad U(t) \to U \quad \text{as} \quad t \to 0,$$

$$X(t) \quad \text{is invertible for} \quad t \in \mathcal{J}\backslash\{0\}, \quad \text{and such that}$$

$$U(t)X^{-1}(t) \quad \text{is decreasing on} \quad \mathcal{J}\backslash\{0\},$$

and define

$$M(t) \equiv R_1 R_2^T + R_2 U(t) X^{-1}(t) R_2^T, \quad \Lambda(t) \equiv R_1 X(t) + R_2 U(t), \quad \Lambda \equiv R_1 X + R_2 U.$$

Then ind $M(0+)$, ind $M(0-)$, and def $\Lambda(0+)$ exist and

$$\text{ind}\, M(0+) - \text{ind}\, M(0-) = \text{def}\, \Lambda - \text{def}\, \Lambda(0+) - \text{def}\, X,$$

where ind denotes the index (the number of negative eigenvalues) and def denotes the defect (the dimension of the kernel) of a matrix. The basic tool for the proof of this result consists of a theorem on the rank of a certain product of matrices, so that this *rank theorem* is the key result of the present paper.

Key words. rank of products of matrices, monotone matrix-valued functions, index of matrices, oscillation of linear selfadjoint differential systems

AMS subject classifications. 15A03, 15A23, 26A48, 34A30

1. Introduction.

We use the following notation. By ker, Im, rank, def, ind, respectively, we denote the kernel, image, rank, defect (that is, the dimension of the kernel), negative index (that is, the number of negative eigenvalues), respectively, of a matrix; I denotes the identity matrix and Q^T denotes the transpose of Q. Moreover, we say that a (square) matrix-valued function $Q(t)$ is decreasing on some interval $\mathcal{J} \subset \mathbb{R}$, if $Q(t)$ is symmetric for $t \in \mathcal{J}$ and if $Q(t_1) - Q(t_2)$ is nonnegative definite for all $t_1, t_2 \in \mathcal{J}$ with $t_1 \leq t_2$. Of course, increasing for $Q(t)$ is defined similarly. Throughout this paper we deal with real matrices.

We now present the central result of this paper (Theorem 2 in §3). Let real $m \times m$-matrices R_1, R_2, X, U be given, which satisfy

$$R_1 R_2^T = R_2 R_1^T, \quad X^T U = U^T X, \quad \text{rank}\,(R_1, R_2) = \text{rank}\,(X^T, U^T) = m.$$

Moreover, assume that $X(t)$ and $U(t)$ are real $m \times m$-matrix-valued functions on some interval $\mathcal{J} = [-\varepsilon, \varepsilon]$, $\varepsilon > 0$ such that

$$X^T(t)U(t) = U^T(t)X(t) \quad \text{on} \quad \mathcal{J},$$

* Received by the editors November 16, 1992; accepted for publication (in revised form) by R. Horn, November 9, 1993.

† Abteilung Mathematik V, Universität Ulm, D-89069 Ulm, Germany (kratz@dulruu51.bitnet).

$$X(t) \to X \quad \text{and} \quad U(t) \to U \quad \text{as} \quad t \to 0,$$

$$X(t) \quad \text{is invertible (regular) for} \quad t \in \mathcal{J}\backslash\{0\}, \quad \text{and such that}$$

$$U(t)X^{-1}(t) \quad \text{is decreasing on} \mathcal{J}\backslash\{0\},$$

and define

$$M(t) \equiv R_1 R_2^T + R_2 U(t) X^{-1}(t) R_2^T, \quad \Lambda(t) \equiv R_1 X(t) + R_2 U(t), \quad \Lambda \equiv R_1 X + R_2 U.$$

Then, ind $M(0+)$, ind $M(0-)$, and def $\Lambda(0+)$ exist and

$$\text{ind}\, M(0+) - \text{ind}\, M(0-) = \text{def}\, \Lambda - \text{def}\, \Lambda(0+) - \text{def}\, X.$$

The motivation for this result stems from oscillation theorems on linear self-adjoint differential systems. Actually, our index theorem includes (in a certain way) the Oscillation Theorem by Morse [10, Thm. 24.1] (see also [9] and [1, Thm. 1] and [7, Thm. 10]). These oscillation results are intimately related to corresponding quadratic functionals, associated eigenvalue problems, and their Rayleigh principle as discussed in [5]. Our result here is stated in a general setting that does not use the underlying differential systems, but instead uses only a few consequences, particularly the monotonicity (and symmetry) of certain matrix-valued functions. By contrast, the proofs of the corresponding results in [1], [7], [9], and [10] use the differential system time after time.

Finally, we summarize the setup of this paper. The proof of our quoted central result is split into two parts: an *algebraic* part (§2) and an *analytic* part (§3). The content of the algebraic part is a *rank theorem* (Theorem 1) about a certain associated matrix (built up from the matrices R_1, R_2, X, and U above). Actually, this theorem is the key result of this paper. The analytic part in §3 is based mainly on a limit theorem for monotone matrix-valued functions from [6], which in turn uses compactness, monotone convergence, the minimum-maximum principle, and an inequality on symmetric matrices [4, Lem. 1]. This limit theorem and the rank theorem yield the index theorem rather straightforwardly.

2. The rank theorem. In this section we prove a rank theorem on the product of certain matrices. This result is in a sense the algebraic part of an oscillation theorem on linear selfadjoint differential systems [1, Thm. 1] or [7, Thm. 10] as mentioned in the Introduction. Our theorem is stated in a general setting without reference to the differential systems that motivated it. As a consequence, the result is now much clearer, and, correspondingly, the proof is more transparent than its counterpart in [7, p. 132].

The following proposition is contained in [1, Prop. A1] or in [7, Prop. A1], and we cite it here since it is used several times.

PROPOSITION 1. *Given real $m \times m$-matrices Q_1, Q_2 with $Q_1 Q_2^T = Q_2 Q_1^T$ and* $\text{rank}\,(Q_1, Q_2) = m$. *Then* $\text{rank}\,(Q_1^T, Q_2^T) = m$ *and* $\ker Q_1 Q_2^T = \ker Q_1^T \oplus \ker Q_2^T$, *where* $\oplus$ *denotes a direct sum.*

Furthermore, we need a result on the rank of the product of three matrices. While the inequality below is well known (the so-called *Frobenius inequality* [8, (2.17.1)]), the case of equality does not seem to be cited elsewhere. Although a corresponding

statement in [1, Prop. A5] and [7, Prop. A8] is incorrect, this does not affect its application either in [1] or in [7]. So, the following proposition, including its proof, is also a correction of [1, Prop. A5] and [7, Prop. A8].

PROPOSITION 2. *Let real matrices Q_1, Q_2, Q_3 be given and suppose the product $Q_1Q_2Q_3$ is defined. Then*

$$\operatorname{rank} Q_1Q_2Q_3 \geq \operatorname{rank} Q_1Q_2 + \operatorname{rank} Q_2Q_3 - \operatorname{rank} Q_2$$

with equality if and only if

(1) $Q_1Q_2d \in \operatorname{Im} Q_1Q_2Q_3$ *and* $(Q_2Q_3)^TQ_2d = 0$ *always imply* $Q_2d = 0$.

Proof. Consider bases $z_1, \ldots, z_r$, $y_1, \ldots, y_s$, and $x_1, \ldots, x_t$ of $\operatorname{Im} Q_3$, $\operatorname{Im} Q_2Q_3$, and $\operatorname{Im} Q_1Q_2Q_3$, respectively, as in the proof of [7, Prop. A8] (thus $\operatorname{rank} Q_3 = r \geq \operatorname{rank} Q_2Q_3 = s \geq \operatorname{rank} Q_1Q_2Q_3 = t$) with $y_\nu = Q_2z_\nu$ and $x_\mu = Q_1y_\mu = Q_1Q_2z_\mu$ for $\nu = 1, \ldots, s$; $\mu = 1, \ldots, t$. Moreover, supplement these bases to bases $y_1, \ldots, y_s, \tilde{y}_1, \ldots, \tilde{y}_k$ and $x_1, \ldots, x_t, \tilde{x}_1, \ldots, \tilde{x}_\ell$ of $\operatorname{Im} Q_2$, respectively, $\operatorname{Im} Q_1Q_2$ (thus $\operatorname{rank} Q_2 = s + k \geq \operatorname{rank} Q_1Q_2 = t + \ell$) such that

$$y_\nu^T\tilde{y}_\mu = 0, \quad \tilde{x}_i = Q_1\tilde{y}_i \quad \text{for} \quad \nu = 1, \ldots, s, \quad \mu = 1, \ldots, k, \quad \text{and} \quad i = 1, \ldots, \ell.$$

Hence, $\ell \leq k$, i.e.,

$$\ell = \operatorname{rank} Q_1Q_2 - \operatorname{rank} Q_1Q_2Q_3 \leq k = \operatorname{rank} Q_2 - \operatorname{rank} Q_2Q_3,$$

which yields the Frobenius inequality, and we have equality if and only if

(2) $x_1, \ldots, x_t, Q_1\tilde{y}_1, \ldots, Q_1\tilde{y}_k$ are linearly independent.

We show that (1) and (2) are equivalent. First, assume (2) and let $Q_1Q_2d = Q_1Q_2Q_3c \in \operatorname{Im} Q_1Q_2Q_3$, $(Q_2Q_3)^TQ_2d = 0$. Put $y = Q_2d$. Then $Q_1y = \sum_{\nu=1}^t \alpha_\nu x_\nu \in \operatorname{Im} Q_1Q_2Q_3$ and $y_\nu^Ty = 0$ for $\nu = 1, \ldots, s$, so that y is a linear combination of $\tilde{y}_1, \ldots, \tilde{y}_k$; $y = \sum_{\nu=1}^k \beta_\nu\tilde{y}_\nu$ say. Hence,

$$0 = \sum_{\nu=1}^t \alpha_\nu x_\nu - Q_1y = \sum_{\nu=1}^t \alpha_\nu x_\nu - \sum_{\nu=1}^k \beta_\nu Q_1\tilde{y}_\nu,$$

and (2) implies that $\alpha_1 = \cdots = \alpha_t = \beta_1 = \cdots = \beta_k = 0$ so that $y = Q_2d = 0$, which proves (1). Next, assume (1), let

$$0 = \sum_{\nu=1}^t \alpha_\nu x_\nu - \sum_{\nu=1}^k \beta_\nu Q_1\tilde{y}_\nu,$$

and put $y = \sum_{\nu=1}^k \beta_\nu\tilde{y}_\nu$. Then $y = Q_2d \in \operatorname{Im} Q_2$, $y_\nu^Ty = 0$ for $\nu = 1, \ldots, s$, so that $(Q_2Q_3)^TQ_2d = 0$ and $Q_1y = Q_1Q_2d = \sum_{\nu=1}^t \alpha_\nu x_\nu \in \operatorname{Im} Q_1Q_2Q_3$. Thus, by (1), we have $y = Q_2d = \sum_{\nu=1}^k \beta_\nu\tilde{y}_\nu = 0$, and therefore also $\sum_{\nu=1}^t \alpha_\nu x_\nu = 0$. The linear independence of $\tilde{y}_1, \ldots, \tilde{y}_k$ and of $x_1, \ldots, x_t$ implies that $\beta_1 = \cdots = \beta_k = \alpha_1 = \cdots = \alpha_t = 0$, which proves (2). □

Remark 1. It is straightforward to see that if $Q_2 = I$ is the $m \times m$-identity matrix (so that Q_1 and Q_3 must be type $n \times m$, $m \times k$, respectively, for some positive integers n and k), then condition (1) is equivalent to the condition

(1′) $\ker Q_1 \subset \operatorname{Im} Q_3$.

Remark 2. By using the singular value decomposition (SVD) (see [3, Thm. 7.3.5]) of Q_2, it can be shown that the hypothesis of symmetry of $Q_1 Q_2^T$ in Proposition 1 can easily be relaxed to *normality*. This fact was pointed out by the referee. Moreover, the SVD of the matrices Q_1 and Q_3 in Proposition 2 leads to an alternative proof, particularly in the case of equality, i.e., of assertion (1).

The statement and the proof of our main result in this section require some notation and auxiliary lemmas. Let real $m \times m$-matrices R_1, R_2, X, U be given such that

$$\operatorname{rank}(R_1, R_2) = m, \qquad R_1 R_2^T = R_2 R_1^T, \tag{3}$$

and

$$\operatorname{rank}(X^T, U^T) = m, \qquad X^T U = U^T X. \tag{4}$$

The orthogonal decomposition of $\operatorname{Im} R_2^T$ into $\operatorname{Im} X$ (i.e., the column-space of X) and its orthogonal complement leads to a unique matrix $\tilde{S}$ and another matrix S (which is not uniquely determined and which may be any matrix with this property) such that

$$R_2^T = XS + \tilde{S} \quad \text{with } X^T \tilde{S} = 0. \tag{5}$$

Then we consider any matrix T with

$$\operatorname{Im} T = \ker \tilde{S}, \tag{6}$$

which yields $\tilde{S}T = 0$, $T^T \tilde{S}^T = 0$, and $\operatorname{Im} \tilde{S}^T = \ker T^T$.

Finally, we define

$$\Lambda \equiv R_1 X + R_2 U \tag{7}$$

and

$$K \equiv X^T X + U^T U. \tag{8}$$

First, assumptions (3), (4) together with Proposition 1 yield Lemma 1.

LEMMA 1. *It holds that* $\operatorname{rank}(R_1^T, R_2^T) = \operatorname{rank}(X, U) = m$, $\ker(R_1, R_2) = \operatorname{Im} \begin{pmatrix} R_2^T \\ -R_1^T \end{pmatrix}$, $\ker(X^T, U^T) = \operatorname{Im} \begin{pmatrix} U \\ -X \end{pmatrix}$, *and* $K = X^T X + U^T U$ *is invertible.*

Next, we have Lemma 2.

LEMMA 2. *The matrix* $S' \equiv K^{-1} U^T \tilde{S}$ *satisfies*

$$XS' = 0, \quad \tilde{S} = US', \quad \text{and} \quad \operatorname{rank} S' = \operatorname{rank} \tilde{S}. \tag{9}$$

Proof. By (5), we have $0 = X^T \tilde{S} = X^T \tilde{S} - U^T 0$. Hence, by Lemma 1, there exists $\tilde{S}'$ such that $\tilde{S} = U\tilde{S}'$, $0 = X\tilde{S}'$, and therefore

$$U^T \tilde{S} = U^T U \tilde{S}' = \{X^T X + U^T U\} \tilde{S}' = K\tilde{S}'.$$

This implies that $\tilde{S}' = S' = K^{-1} U^T \tilde{S}$ and, together with $\tilde{S} = U\tilde{S}' = US'$, we can conclude that $\operatorname{rank} S' = \operatorname{rank} \tilde{S}$. □

The next lemma is crucial. It corresponds to [7, Lem. 10] and [1, Lem. 4.3] ("crucial result"), but we do not need any of the identities that occur there.

LEMMA 3. *Let* $S' \equiv K^{-1}U^T\tilde{S}$ *as in Lemma* 2. *Then*

$$\ker T^T\Lambda = \ker\Lambda \oplus \operatorname{Im} S' \tag{10}$$

and

$$\operatorname{rank} T^T\Lambda = \operatorname{rank} T + \operatorname{rank}\Lambda - m. \tag{10'}$$

Proof. Because $\operatorname{rank} T = m - \operatorname{rank} S'$ by (6) and (9), the statement (10) implies (10'). First, we show that

$$\ker\Lambda + \operatorname{Im} S' \subset \ker T^T\Lambda.$$

Of course, $\ker\Lambda \subset \ker T^T\Lambda$. So let $d = S'c \in \operatorname{Im} S'$. Then, from (7), (5), (6), (4), and (9), we obtain

$$T^T\Lambda d = T^T\{R_1X + (S^TX^T + \tilde{S}^T)U\}d = T^T\{R_1 + S^TU^T\}Xd = 0,$$

which proves the inclusion above. Next, we show that the sum is direct. Let $d = S'c \in \ker\Lambda \cap \operatorname{Im} S'$. Then, by (7), (5), (4), and (9),

$$0 = \Lambda d = \{(R_1 + S^TU^T)X + \tilde{S}^TU\}S'c = \tilde{S}^T\tilde{S}c.$$

Hence, $d = S'c = K^{-1}U^T\tilde{S}c = 0$ by Lemma 2. Thus, $\ker\Lambda \cap \operatorname{Im} S' = \{0\}$, and, together with the inclusion we have already shown, it follows that

$$\begin{aligned}\operatorname{def} T^T\Lambda &\geq \operatorname{def}\Lambda + \operatorname{rank} S' = m - \{\operatorname{rank}\Lambda + \operatorname{rank} T - m\}\\ &\geq m - \operatorname{rank} T^T\Lambda = \operatorname{def} T^T\Lambda\end{aligned}$$

from (9), (6), and Proposition 2 (with $Q_1 = T^T$, $Q_2 = I$, $Q_3 = \Lambda$). Hence, we must have equality above, which yields (10). □

Remark 3. Observe that, by Remark 1, condition (10') implies

$$\ker T^T \subset \operatorname{Im}\Lambda \quad \text{and} \quad \ker\Lambda^T \subset \operatorname{Im} T.$$

Finally, we need Lemma 4.

LEMMA 4. *We have*

$$\ker T^TR_2 = \ker R_2 \oplus \operatorname{Im}\tilde{S} \tag{11}$$

and

$$\operatorname{rank} T^TR_2 = \operatorname{rank} T + \operatorname{rank} R_2 - m. \tag{11'}$$

Proof. By (6), statement (11) implies (11'). As in the preceding proof we show first that

$$\ker R_2 + \operatorname{Im}\tilde{S} \subset \ker T^TR_2.$$

Of course, $\ker R_2 \subset \ker T^TR_2$. So let $d = \tilde{S}c \in \operatorname{Im}\tilde{S}$. Then $T^TR_2d = T^TS^TX^T\tilde{S}c = 0$ by (5) and (6), which yields the inclusion. To show that the sum is direct, assume that $d = \tilde{S}c \in \ker R_2 \cap \operatorname{Im}\tilde{S}$. Then (5) implies that $0 = R_2d =$

$\{S^TX^T+\tilde{S}^T\}\tilde{S}c=\tilde{S}^T\tilde{S}c$. Hence, $d=\tilde{S}c=0$, so that $\ker R_2\cap \operatorname{Im}\tilde{S}=\{0\}$. It follows that

$$\operatorname{def}T^TR_2 \geq \operatorname{def}R_2+\operatorname{rank}\tilde{S}=m-\{\operatorname{rank}R_2+\operatorname{rank}T-m\}\geq m-\operatorname{rank}T^TR_2=\operatorname{def}T^TR_2$$

by (6) and Proposition 2 (with $Q_1=T^T$, $Q_2=I$, $Q_3=R_2$). Hence, we must have equality above and the assertion (11) follows. □

Remark 4. As in Remark 3, the condition (11′) implies

$$\operatorname{Im}\tilde{S}^T=\ker T^T\subset \operatorname{Im}R_2 \quad \text{and} \quad \ker R_2^T\subset \operatorname{Im}T=\ker\tilde{S}\,.$$

Now we can state the central result.

THEOREM 1 (rank theorem). *Let real $m\times m$-matrices R_1, R_2, X, and U be given, which satisfy* (3) *and* (4), *i.e.,*

$$\operatorname{rank}(R_1,R_2)=m, \qquad R_1R_2^T=R_2R_1^T,$$

and

$$\operatorname{rank}(X^T,U^T)=m, \qquad X^TU=U^TX\,,$$

and suppose that the matrices S, T, and Λ are given by (5), (6), *and* (7), *respectively, i.e.,*

$$R_2^T=XS+\tilde{S} \qquad \textit{with} \qquad X^T\tilde{S}=0,$$

$$\operatorname{Im}T=\ker\tilde{S},$$

and

$$\Lambda=R_1X+R_2U.$$

Then the matrix

$$Q = T^T\Lambda ST \tag{12}$$

is symmetric and it satisfies

$$\operatorname{rank}Q=\operatorname{rank}T^T\Lambda+\operatorname{rank}R_2^TT-\operatorname{rank}X \tag{13}$$

and

$$\operatorname{rank}Q=2\operatorname{rank}T+\operatorname{rank}\Lambda+\operatorname{rank}R_2-\operatorname{rank}X-2m\,. \tag{13′}$$

Proof. First, the conditions (5), (6), (7), (4), and (3) imply tyhe following equations.

$$\begin{gathered} R_2^TT=XST, \qquad T^T\Lambda=T^TR_1'X, \\ Q=T^T\{R_1R_2^T+S^TU^TXS\}T=(T^TR_1')X(ST), \quad \text{where } R_1'=R_1+S^TU^T. \end{gathered} \tag{14}$$

Hence, by (4), Q is symmetric. Moreover, by (10′) and (11′) of Lemmas 3 and 4 above, the statement (13′) follows from (13). Now, we show (13) by applying Proposition 2 with

$$Q_1=T^TR_1', \quad Q_2=X\,, \quad Q_3=ST$$

(as indicated in (14)), such that $Q = Q_1Q_2Q_3$, $Q_1Q_2 = T^T\Lambda$, and $Q_2Q_3 = R_2^TT$ by (14). Hence, we must prove (1). Therefore, assume that

$$T^T\Lambda d = T^T\Lambda STc \in \operatorname{Im} Q_1Q_2Q_3 \quad \text{and} \quad T^TR_2Xd = 0.$$

Then $T^T\Lambda(d - STc) = 0$, and Lemma 3 implies that

$$d - STc = d_1 + d_2 \quad \text{with } \Lambda d_1 = 0, \quad d_2 = S'c' \in \operatorname{Im} S',$$

where S' is as in Lemma 2. By (9), $Xd_2 = 0$ so that

$$Xd = Xd_1 + R_2^TTc \quad \text{with } 0 = \Lambda d_1 = R_1\{Xd_1\} + R_2\{Ud_1\}.$$

By Lemma 1 there exists d' such that $Xd_1 = R_2^Td'$, $Ud_1 = -R_1^Td'$, and therefore $Xd = R_2^T\tilde{c} \in \operatorname{Im} R_2^T$ with $\tilde{c} = d' + Tc$. Moreover, $T^TR_2Xd = 0$ implies by Lemma 4 that $Xd = c_1 + \tilde{S}c_2$ $(= R_2^T\tilde{c})$ with $R_2c_1 = 0$. Altogether, it follows that (use also (5)) $d^TX^TXd = \tilde{c}^TR_2\{c_1 + \tilde{S}c_2\} = \tilde{c}^TR_2\tilde{S}c_2 = d^TX^T\tilde{S}c = 0$. Thus, $Xd = Q_2d = 0$, which establishes (1). □

3. The index theorem. The following index theorem contains, in a rather general setting, the (local) oscillation theorem on linear selfadjoint differential systems stated in [1, Thm. 1] and [7, Thm. 10], and, in particular, it contains the oscillation theorem of Morse [10, Thm. 24.1] (see also the discussion in [1, p. 332]). It turns out that (besides some identities yielding the symmetry of certain matrices, e.g., (4)), we need only the monotonicity of the matrix-valued function under consideration as a consequence of the differential system. The proof below uses (besides the algebraic result of the previous section) as *analytic tools* only limit results on matrix-valued functions (see [1], [4], [6], [7]) and the following proposition that is stated in [7, Prop. A7] and [1, Prop. A4] (and that is based mainly on the minimum-maximum principle for symmetric matrices; see [2] or [11]).

PROPOSITION 3. *Let $Q(t)$ be an $m \times m$-matrix-valued function that is symmetric and increasing on (a, t_0) with $a < t_0 \in \mathbb{R} \cup \{\infty\}$, and suppose T is an $m \times r$-matrix such that*

$$r = \operatorname{rank} T, \quad T^TT = I_{r\times r}, \quad \text{and} \quad \lim_{t\to t_0-} c^TQ(t)c = \infty \quad \text{for all } c \notin \operatorname{Im} T,$$

and is such that $\tilde{Q} = \lim_{t\to t_0-} T^TQ(t)T$ exists. Moreover, let $\mu_i(t)$ and $\tilde{\mu}_i$ denote the eigenvalues of $Q(t)$ and $\tilde{Q}$ with $\mu_1(t) \le \cdots \le \mu_m(t)$, $\tilde{\mu}_1 \le \cdots \le \tilde{\mu}_r$. Then

(i) $\lim_{t\to t_0-} \mu_i(t) = \mu_i$ *for* $i = 1, \ldots, r$, *and*

(ii) $\lim_{t\to t_0-} \mu_i(t) = \infty$ *for* $i = r+1, \ldots, m$.

A similar result holds for right-hand limits. The main result of this paper now follows.

THEOREM 2 (index theorem). *Let real $m \times m$-matrices R_1, R_2, X, U be given, that satisfy* (3) *and* (4), *i.e.,*

$$\operatorname{rank}(R_1, R_2) = \operatorname{rank}(X^T, U^T) = m, \quad R_1R_2^T = R_2R_1^T, \quad X^TU = U^TX,$$

and let real $m \times m$-matrix-valued functions $X(t)$ and $U(t)$ be given on some interval $[-\varepsilon, \varepsilon]$, $\varepsilon > 0$, such that

$$\begin{aligned} &X^T(t)U(t) = U^T(t)X(t) \quad \text{for} \quad t \in [-\varepsilon, \varepsilon] \quad \text{and} \\ &X(t) \to X, \quad U(t) \to U \quad \text{as} \quad t \to 0; \end{aligned} \tag{15}$$

(16) $X(t)$ *is invertible for* $t \in [-\varepsilon, \varepsilon]$, $t \neq 0$ *and*

(17) $U(t)X^{-1}(t)$ *decreases on* $[-\varepsilon, 0)$ *and on* $(0, \varepsilon]$.

Define

(18) $M(t) \equiv R_1 R_2^T + R_2 U(t) X^{-1}(t) R_2^T$, $\Lambda(t) \equiv R_1 X(t) + R_2 U(t)$, *and* $\Lambda \equiv R_1 X + R_2 U$ (*as in* (7)).

Then $\operatorname{ind} M(0+)$, $\operatorname{ind} M(0-)$, *and* $\operatorname{def} \Lambda(0+)$ *exist, and*

(19) $$\operatorname{ind} M(0+) - \operatorname{ind} M(0-) = \operatorname{def} \Lambda - \operatorname{def} \Lambda(0+) - \operatorname{def} X .$$

Proof. Of course, $U(t)X^{-1}(t)$ and $M(t)$ are symmetric by (15) and (3). The limit theorem on monotone matrix-valued functions [6, Thm. 1] implies that

(20) $$\begin{aligned} &X^T U(t) X^{-1}(t) X \to U^T X \text{ as } t \to 0, \text{ and} \\ &c^T U(t) X^{-1}(t) c \to \infty \text{ (respectively, } -\infty) \text{ as} \\ &t \to 0+ \text{ (respectively, } 0-) \text{ for all } c \notin \operatorname{Im} X. \end{aligned}$$

Now suppose that $S, \tilde{S}, T$, and Q are given by (5), (6), and (12) as in the previous section, such that, in particular, $R_2^T c \in \operatorname{Im} X$ if and only if $c \in \operatorname{Im} T$. We may assume that T satisfies the requirement of Proposition 3 that T is of type $m \times r$ with $T^T T = I_{r \times r}$. Then, it follows from (20) (and (18)) that

(20′) $$\begin{aligned} &T^T M(t) T \to Q = T^T \Lambda S T \text{ as } t \to 0, \text{ and} \\ &c^T M(t) c \to \infty \ (-\infty) \text{ as } t \to 0+ \ (0-) \text{ for all } c \notin \operatorname{Im} T. \end{aligned}$$

Proposition 3 can now be applied, and denoting by $\mu_1(t), \ldots, \mu_m(t)$ and $\mu_1, \ldots, \mu_r$ the eigenvalues of $M(t)$ and Q, we get

(i) $\mu_j(t) \to \mu_j$ for $j = 1, \ldots, r$, $\mu_j(t) \to \infty$ for $j = r+1, \ldots, m$ as $t \to 0+$, if $\mu_1(t) \leq \cdots \leq \mu_m(t)$, $\mu_1 \leq \cdots \leq \mu_r$; and

(ii) $\mu_j(t) \to \mu_j$ for $j = 1, \ldots, r$, $\mu_j(t) \to -\infty$ for $j = r+1, \ldots, m$ as $t \to 0-$, if $\mu_1(t) \geq \cdots \geq \mu_m(t)$, $\mu_1 \geq \cdots \geq \mu_r$.

Since $M(t) = \Lambda(t) X^{-1}(t) R_2^T$ (by (18)) is symmetric with

$$\operatorname{rank} (\Lambda(t) X^{-1}(t), R_2) = \operatorname{rank} (R_1 + R_2 U(t) X^{-1}(t), R_2) = \operatorname{rank} (R_1, R_2) = m,$$

Proposition 1 implies that

(21) $$\ker M(t) = \ker \Lambda^T(t) \oplus \ker R_2^T \quad \text{for } t \in [-\varepsilon, \varepsilon], \qquad t \neq 0.$$

The monotonicity of $M(t)$ on $(0, \varepsilon]$ (hypothesis (17)) implies that $\operatorname{def} M(t)$ is constant on $(0, \delta]$ for some $\delta > 0$, and therefore, by (21), $\operatorname{def} M(0+)$ and $\operatorname{def} \Lambda(0+)$ exist with

(21′) $$\operatorname{def} M(0+) = \operatorname{def} \Lambda(0+) + \operatorname{def} R_2.$$

Of course, $\operatorname{ind} M(0+)$ and $\operatorname{ind} M(0-)$ exist as well, and the statements (i) and (ii) above (observe also that the $\mu_j(t)$ are decreasing on $(0,\varepsilon]$ and $[-\varepsilon,0)$ by [7, Prop. A3] or [11, p. 101, 102]) imply that

$$\operatorname{ind} M(0+) = \operatorname{ind} Q + \operatorname{def} Q - \operatorname{def} M(0+), \text{ and } \operatorname{ind} M(0-) = \operatorname{ind} Q + m - \operatorname{rank} T.$$

Since $\operatorname{def} Q = r - \operatorname{rank} Q = \operatorname{rank} T - \operatorname{rank} Q$, we finally obtain from (21′) and (13′) of Theorem 1 the following corollary, which is more specific than assertion (19) and which, of course, yields (19) and completes the proof. □

COROLLARY 1. *Under the assumptions and with the notation of Theorem* 2 *let* S, T, *and* Q *be defined by* (5), (6), (12), *respectively. Then*

$$\begin{aligned} &\operatorname{ind} M(0+) = \operatorname{ind} Q + m - \operatorname{rank} T + \operatorname{def} \Lambda - \operatorname{def} \Lambda(0+) - \operatorname{def} X, \text{ and} \\ &\operatorname{ind} M(0-) = \operatorname{ind} Q + m - \operatorname{rank} T. \end{aligned} \tag{22}$$

Remark 5. Replacing t by $-t$ we obtain corresponding results if $U(t)X^{-1}(t)$ is *increasing* instead of decreasing, while the other assumptions remain unchanged, namely,

$$\operatorname{ind} M(0+) - \operatorname{ind} M(0-) = \operatorname{def} \Lambda(0-) - \operatorname{def} \Lambda + \operatorname{def} X; \tag{19′}$$

and

$$\begin{aligned} &\operatorname{ind} M(0+) = \operatorname{ind} Q + m - \operatorname{rank} T, \quad \text{and} \\ &\operatorname{ind} M(0-) = \operatorname{ind} Q + m - \operatorname{rank} T + \operatorname{def} \Lambda - \operatorname{def} \Lambda(0-) - \operatorname{def} X. \end{aligned} \tag{22′}$$

Finally, in the special case $R_1 = 0$, $R_2 = I$, the condition (3) holds, and then Theorem 2 reduces to the following corollary (which complements [6]).

COROLLARY 2. *Assume that* $X(t)$ *and* $U(t)$ *are real* $m \times m$*-matrix-valued functions such that* (15), (16), (17) *hold, let* X *and* U *be real* $m \times m$*-matrices such that* $\operatorname{rank}(X^T, U^T) = m$, *and let* $Q(t) = U(t)X^{-1}(t)$. *Then* $\operatorname{def} U(0+)$, $\operatorname{ind} Q(0+)$ *and* $\operatorname{ind} Q(0-)$ *exist and*

$$\operatorname{ind} Q(0+) - \operatorname{ind} Q(0-) = \operatorname{def} U - \operatorname{def} U(0+) - \operatorname{def} X. \tag{23}$$

Remark 6. If $A(t)$ is a matrix-valued function defined for all t in a neighborhood of zero and if $\lim_{t\to 0} A(t) = A$, then, by considering subdeterminants of $A(t)$ (or the singular values of $A(t)$ and A), it follows that there is some $\delta > 0$ such that

$$\operatorname{rank} A \leq \operatorname{rank} A(t) \quad \text{for all } 0 < |t| \leq \delta.$$

This implies that we always have

$$\operatorname{def} \Lambda(0+) \leq \operatorname{def} \Lambda, \quad \operatorname{def} \Lambda(0-) \leq \operatorname{def} \Lambda, \quad \operatorname{def} U(0+) \leq \operatorname{def} U$$

in the assertions (19), (19′), (22), (22′), (23), respectively, above.

Remark 7. Since the background of our results is in a sense a *real* theory, we have dealt with real matrices. But the results above (including their proofs) carry over easily to Hermitian matrices (instead of real and symmetric matrices), when the transpose of a matrix is always replaced by the Hermitian adjoint.

Acknowledgment. I am grateful to the anonymous referee for the careful reading of the paper and several useful suggestions.

REFERENCES

[1] G. BAUR AND W. KRATZ, *A general oscillation theorem for self-adjoint differential systems with applications to Sturm–Liouville eigenvalue problems and quadratic functionals*, Rend. Circ. Mat. Palermo, 38 (1989), pp. 329–370.

[2] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics* I, Interscience, New York, 1953.

[3] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

[4] W. KRATZ, *A substitute of l'Hospital's rule for matrices*, Proc. Amer. Math. Soc., 99 (1987), pp. 395–402.

[5] ———, *An oscillation theorem for self-adjoint differential systems and the Rayleigh principle for quadratic functionals*, J. London Math. Soc., to appear.

[6] ———, *A limit theorem for monotone matrix functions*, Linear Algebra Appl., 194 (1993), pp. 205–222.

[7] W. KRATZ AND A. PEYERIMHOFF, *A treatment of Sturm–Liouville eigenvalue problems via Picone's identity*, Analysis, 5 (1985), pp. 97–152.

[8] M. MARCUS AND H. MINC, *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, 1964.

[9] M. MORSE, *The Calculus of Variations in the Large*, AMS Colloquium Publication 18, 1934.

[10] ———, *Variational analysis: Critical Extremals and Sturmian Extensions*, Wiley, New York, 1973.

[11] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. AND APPL.
Vol. 16, No. 1, pp. 123–137, January 1995

009

COMPUTING MOST NEARLY RANK-REDUCING STRUCTURED MATRIX PERTURBATIONS*

M. A. WICKS† AND R. A. DECARLO‡

Abstract. The paper investigates the problem of computing structured matrix perturbations that cause, or most nearly cause, some specified system matrix to fail to have full rank. The paper discusses some theoretical issues concerning the existence of solutions to these problems. It suggests a numerical approach to computing solutions that utilizes some ideas on differentiation of singular values. Finally, an algorithm for finding structured most rank-reducing perturbations and structured most nearly rank-reducing perturbations is developed. The paper demonstrates convergence of the algorithm to a rank-reducing perturbation or to a local minimum for a most nearly rank-reducing perturbation. Numerical examples illustrating the technique are included.

Key words. matrix perturbation theory, rank, rounding errors, singular value decomposition, stability robustness, controllability robustness

AMS subject classifications. 15A03, 65F35, 65G05, 93B35, 93B20, 93D99

1. Introduction.

For a given matrix M, a rank-reducing perturbation R from an allowable perturbation set makes the matrix $M - R$ rank-deficient. Eigenvalue problems can be cast in this context. For example, to find the eigenvalues of a square matrix A, one must compute a "perturbation matrix" of the form λI such that $\det(A - \lambda I) = 0$. Similarly, in the generalized eigenvalue problem, one must find a structured rank-reducing matrix perturbation: given square matrices, A and B, compute a matrix having the form λB for which $\det(A - \lambda B) = 0$. Third, there is the eigentuple problem [1], [2], [15]: given A and B_i, $i = 1, \ldots, n$ find λ_i and $x \neq 0$ for which

$$\left(A - \sum_{i=1}^{n} \lambda_i B_i \right) x = 0.$$

These are special cases of the rank-reducing perturbation (RRP) problem.

DEFINITION 1.1 (The RRP problem). *Given $M \in \mathbb{C}^{n \times m}$ with $n \leq m$ and a vector space $\mathcal{R} \subset \mathbb{C}^{n \times m}$ determine $R \in \mathcal{R}$ (if possible) so that* $\operatorname{rank}(M - R) < n$.

The coefficient field for the vector space $\mathcal{R}$ may be $\mathbb{R}$ or $\mathbb{C}$. The main emphasis of the problem formulation considered in this paper is that $\mathcal{R}$ characterizes the structure of the particular problem under study, such as controllability or stability. However, $\mathcal{R}$ may also represent a space of structured parameter variations. An R satisfying $\operatorname{rank}(M - R) < n$ is called an RRP.

Numerical algorithms solving various special cases of the RRP are well developed, especially for the case when the dimension of $\mathcal{R}$ is one and M is square. Here, the RRP reduces to the familiar generalized eigenvalue problem, for which numerically reliable techniques are available (see, for example, [11]).

---

*Received by the editors March 16, 1992; accepted for publication (in revised form) by C. Van Loan, September 29, 1993.

†Electrical and Computer Engineering Department, GMI Engineering & Management Institute, 1700 West Third Avenue, Flint, Michigan 48504-4898 (mwicks@nova.gmi.edu).

‡School of Electrical Engineering, Purdue University, West Lafayette, Indiana 47907-1285 (decarlo@ecn.purdue.edu).

On the other hand, algorithms for nonsquare M or when $\mathcal{R}$ is multidimensional are not as fully developed. Under certain conditions, the elements of a two-parameter eigentuple problem can be decoupled and the problem can be reduced to two generalized eigenvalue problems [15]. Here, conventional eigenvalue techniques can be used to obtain a solution.

In [1] and [2], iterative gradient methods are explored for the general RRP. The papers remark that these techniques require no large matrix inversions at each iteration. Hence, they may be appropriate for large, sparse systems. However, the authors point out that the rate of convergence makes these techniques rather slow.

The nonlinear eigenvalue problem is similar to the problem stated in Definition 1.1: given $A(\lambda)$ determine λ and $x \neq 0$ such that $A(\lambda)x = 0$. Frequently, the nonlinear eigenvalue problem appears in the form of a polynomial

$$\sum_{i=0}^{r} \lambda^i A_i x = 0.$$

This problem is a multiparameter problem or eigentuple problem with a special structure imposed on the perturbation to A_0.

Several numerical methods for solving the nonlinear eigenvalue problem are reviewed in [14]. It appears possible to generalize these techniques for general multiparameter problems. Indeed, Rayleigh quotient iteration, used for solving the nonlinear eigenvalue problem is similar to the method presented in this paper. Specifics will be presented later in the paper. However, issues arise from the multidimensional and nonsquare nature of the problem given in Definition 1.1 that do not accompany the single parameter, square, generalized, or nonlinear eigenvalue problems. Specifically, the nonexistence of a solution to the general rectangular RRP may be generic, depending on the dimensions involved in the problem. Moreover, in many applications, the existence of a solution is undesirable. In this case, a most nearly rank-reducing perturbation provides information about the property being investigated. For example, consider the usual linear time-invariant state model,

$$\dot{x}(t) = Ax(t) + Bu(t),$$

where $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times r}$. The existence of a $\lambda \in \mathbb{C}$ for which

$$\operatorname{rank}([\,A - \lambda I, \quad B\,]) < n$$

indicates uncontrollability. This is often undesirable. If $\operatorname{rank}([\,A - \lambda I, \quad B\,]) = n$ for all $\lambda \in \mathbb{C}$, one may wish to find the value of λ (there exists one, see [17]) for which the matrix most nearly fails to have full rank, i.e., that value of λ that minimizes $\sigma_n([\,A - \lambda I, \quad B\,])$ as well as the corresponding perturbation to the matrices A and B that would result in uncontrollability ($\sigma_n(M)$ denotes the nth singular value of M, where the singular values are assumed to be ordered as $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$). The problem of determining such a perturbation is called the controllability robustness problem [7], [8], [17]. This desirable nonexistence of solution is also present in the stability robustness problem, where one finds a value of ω that minimizes $\sigma_n([A - j\omega I])$ and a corresponding matrix perturbation which results in instability [4], [18], [12].

The nonexistence of solutions in these problems suggests a more general problem formulation designated as the most nearly rank-reducing perturbation (MNRRP) problem. Definition 1.2 formalizes this problem.

DEFINITION 1.2 (The MNRRP problem). *Given $M \in \mathbb{C}^{n\times m}$ with $n \leq m$ and a vector space $\mathcal{R} \subset \mathbb{C}^{n\times m}$ determine an $R \in \mathcal{R}$ (if one exists) that minimizes $\sigma_n(M-R)$.*

The coefficient field for the vector space $\mathcal{R}$ may be $\mathbb{R}$ or $\mathbb{C}$. The vector space script $\mathcal{R}$ characterizes the structure of the problem under consideration. A minimizing R is called an MNRRP.

Note that any RRP is also an MNRRP. Hence, a method for solving the MNRRP problem will necessarily solve the RRP problem. The converse does not hold. Thus, a single method that solves the MNRRP (and hence the RRP) is desirable.

Numerical methods have been developed for several special cases of the MNRRP. The one-dimensional case where the coefficient field for R is $\mathbb{C}$ is considered in [3]. The method computes $R \in \mathcal{R}$, which minimizes $\|u^H M - u^H R\|$ at each iteration (with u fixed), and then updates u to be a left singular vector associated with $\sigma_n(M-R)$. This method is globally convergent, but not necessarily to a global minimizer. The method easily generalizes to multiparameter problems. The main disadvantage of this technique is the large amount of overhead required per iteration compounded by a slow rate of convergence. Similar techniques are applied to the controllability robustness problem in [17].

As noted above, the stability robustness problem is a special case of the MNRRP where the coefficient field for $\mathcal{R}$ is $\mathbb{R}$ and the dimension of $\mathcal{R}$ is one. A bisection method for finding solutions for the stability robustness problem appears in [4]. This technique converges rapidly to a global minimum. A similar technique appears for structured stability in [12]. However, it is unclear how to generalize the technique to cases where $\mathcal{R}$ is multidimensional or where the coefficient field of $\mathcal{R}$ is $\mathbb{C}$.

Another area of research related to work of this paper focuses on computation of the so-called structured singular value [9], [10], [16]. The structured singular value is used to measure the robustness of feedback systems in the presence of structured uncertainty. Given a square $M \in \mathbb{C}^{n\times n}$, computing the structured singular value involves minimizing the spectral norm of $\|\Delta\|$ subject to $\operatorname{rank}(I + M\Delta) < n$, where Δ has a specified block-diagonal structure.

While similar to the structured singular value (and motivated by similar objectives), the MNRRP differs in several aspects. First, the structured singular value is a frequency response technique measuring robustness at a particular frequency and is recomputed for each desired frequency. The resulting destabilizing perturbation represents a perturbation on a frequency domain model representation. Minimizing $\sigma_n(A - j\omega I)$ finds the minimum destabilizing perturbation for a time domain plant representation. The minimizing value of $j\omega$ is the eigenvalue that would result from the minimum norm, destabilizing parameter variation. The goal is to accomplish the minimization over ω automatically. Second, the MNRRP applies to a broad class of problems having similar structure, including controllability robustness and stability robustness. Being able to determine controllability robustness has numerical implications [13]. Third, the structure imposed by $\mathcal{R}$ in Definition 1.1 reflects the particular robustness problem being investigated. The minimizing singular value reflects the offending parameter variation, which is assumed to be unrestricted in the case of the MNRRP problem. Techniques developed for the MNRRP should prove useful for restricted parameter variation problems, which incorporate a parameter variation structure in addition to the structure represented by $\mathcal{R}$.

The problem of finding minimum RRPs on square matrices where the size of the perturbation is measured using the maximum entry after applying a prespecified scaling is addressed in [6]. Computing an approximation of the solution of this problem to arbitrary accuracy is shown to be NP-complete.

A problem similar to the structured singular value is solved in [5] where a minimum Frobenius, or two-norm perturbation that reduces the rank of the matrix, is constructed. Again, the MNRRP differs from this work as it minimizes $\sigma_n(M - R)$ over a linear space $\mathcal{R}$, where $\mathcal{R}$ characterizes the structure of the problem rather than the structure of the perturbation. A primary motivation for solving the MNRRP is to perform the minimization over $\mathcal{R}$ automatically.

This paper presents a method for numerical solution of the RRP and the MNRRP. The approach views the RRP as a special case of the MNRRP. An algorithm based on this approach attempts to drive the nth singular value of $M - R$ toward zero. From this perspective, singular points are not viewed as isolated points, but as zeros of the surface defined by $\sigma_n(M - R)$ as R varies over $\mathcal{R}$.

Before discussing these algorithms, it is useful to set forth some elementary properties on the existence of solutions to the above problems. Section 2 covers some theoretical issues on the existence of solutions and the asymptotic behavior of $\sigma_n(M - R)$ as $\|R\| \to \infty$. Section 3 discusses differentiation of singular values and develops a Newton method for computing RRP solutions as motivation for an MNRRP algorithm. Such a method will fail near a local minimum of $\sigma_n(M - R)$ as the derivative of $\sigma_n(M - R)$ vanishes. To avoid this difficulty, a modified algorithm is proposed that will always converge to either an RRP or an MNRRP (a local minimum) while retaining the convergence rate of the Newton method to an RRP. Convergence to an MNRRP is nearly as rapid. Section 4 presents a specific algorithmic implementation along with an analysis of its convergence. The paper proves convergence of this algorithm with a suitable stepsize choice under certain conditions. Experimental results are presented in §5 demonstrating the accuracy and efficiency of the algorithm.

2. Existence of solutions. This section discusses some elementary properties relating to the existence of solutions to the problems introduced above. The following property is a restatement of an earlier observation.

PROPERTY 2.1. *For a given matrix, $M \in \mathbb{C}^{n\times m}$, there may be no finite* RRP.

The asymptotic behavior of $\sigma_n(M - R)$ as $\|R\| \to \infty$ helps to understand the global behavior of the singular value surfaces and can provide insight into determining suitable starting values for numerical algorithms. The next set of properties characterize the asymptotic behavior of singular values as functions of matrices as R varies over $\mathcal{R}$.

ASSUMPTION 2.1. *Henceforth, assume that $\mathcal{R} \subset \mathbb{C}^{n\times m}$ is a vector space over the field $\mathbb{R}$.*

ASSUMPTION 2.2. *Singular vector pairs u and v are always chosen to be associated with a singular value σ so that $Rv = \sigma u$, $R^H u = \sigma v$, $\|u\| = 1$, and $\|v\| = 1$.*

PROPERTY 2.2. *Given $M \in \mathbb{C}^{n\times m}$ and $R \in \mathcal{R}$, there exists a pair of left and right unit length singular vectors u and v associated with $\sigma_n(R)$ for which the asymptote of $\sigma_n(M - \alpha R)$ is given by $|-\mathrm{Re}(u^H M v) + \alpha\sigma_n(R)|$.*

PROPERTY 2.3. *If the singular values of R are nonzero for all $R \in \mathcal{R}$, $R \neq 0$, then a finite solution exists to the* MNRRP.

PROPERTY 2.4. *$\lim_{\alpha\to\infty} \sigma_n(M - \alpha R)$ is finite if and only if $\mathrm{rank}(R) < n$.*

Before presenting the next group of properties, it is necessary to distinguish between finite solutions to the RRP and infinite solutions to the RRP.

DEFINITION 2.1. *The* RRP *has a finite solution if and only if some $R \in \mathcal{R}$ satisfies $\sigma_n(M - R) = 0$. The* RRP *is said to have an infinite solution if for some $R \in \mathcal{R}$ $\lim_{\alpha\to\infty} \sigma_n(M - \alpha R) = 0$.*

As mentioned in Property 2.1 there may be no finite solution or infinite solution

to the RRP problem.

DEFINITION 2.2. *The* MNRRP *is said to have an infinite solution if for some* $R_0 \in \mathcal{R}$ $\lim_{\alpha\to\infty} \sigma_n(M - \alpha R_0) = \inf_{R\in\mathcal{R}} \sigma_n(M - R)$.

In contrast to the RRP, the MNRRP has the following property.

PROPERTY 2.5. *The* MNRRP *always has a solution, which may be a finite or an infinite solution.*

The previous property reinforces the desirability of computing MNRRP solutions rather than RRP solutions. The former always exist while the latter may not exist in any sense, finite or infinite. The nonexistence of RRP solutions, finite or infinite, can occur even when the matrices involved are square.

Finally a complete characterization of infinite solutions to the RRP is possible.

PROPOSITION 2.1. *Given* M *and* $\mathcal{R}$ *as before, an infinite solution to the* RRP *exists if and only if for some* $R \in \mathcal{R}$

$$\text{rank}\left(\begin{bmatrix} M & R \\ R & 0 \end{bmatrix}\right) < n + \text{rank}(R), \tag{2.1}$$

where n *is the row dimension of* M *and* R.

Proof. Suppose for some $R \in \mathcal{R}$ (2.1) holds. From (2.1) it follows that $\text{rank}(R) < n$. First, it is necessary to establish the existence of $x, y \in \mathbb{C}^n$, $x \neq 0$ for which $x^H M + y^H R = 0$ and $x^H R = 0$. To do this consider a basis for the left null-space of R, say $\{y_i\}$, $i = 1, \ldots, n - \text{rank}(R)$. The set of vectors $\{\text{col}(\,0, \quad y_i\,)\}$, $i = 1, \ldots, n - \text{rank}(R)$ forms a linearly independent set contained in the left null-space of

$$\begin{bmatrix} M & R \\ R & 0 \end{bmatrix}.$$

But (2.1) implies the dimension of this left null-space is greater than $n - \text{rank}(R)$. This implies the existence of another vector, $\text{col}(\,x, \quad y\,)$, in this left null-space that is linearly independent of the set $\{\text{col}(\,0, \quad y_i\,)\}$, $i = 1, \ldots, n - \text{rank}(R)$. It must be that $x \neq 0$; otherwise one would obtain $y^H R = 0$ with y being linearly independent of the set of $\{y_i\}$, which cannot be.

Using the vectors x and y obtained from $\text{col}(\,x, \quad y\,)$ above, consider the product

$$(x + (1/\alpha)y)^H (M + \alpha R) = (1/\alpha) y^H M.$$

Taking the limit of this product as $\alpha \to \infty$ yields zero, i.e.,

$$\lim_{\alpha\to\infty} (x + (1/\alpha)y)^H (M + \alpha R) = 0.$$

Since $x \neq 0$ and since $n \leq m$, this implies that

$$\lim_{\alpha\to\infty} \sigma_n(M + \alpha R) = 0,$$

satisfying the definition of an infinite RRP solution.

Conversely, suppose for some $R \in \mathcal{R}$

$$\lim_{\alpha\to\infty} \sigma_n(M + \alpha R) = 0.$$

This implies the existence of sequences, α_k and z_k for which $\alpha_k \to \infty$, $\|z_k\| = 1$ and $\|z_k^H (M + \alpha_k R)\| \to 0$. The vector z_k can be written (in a unique way) as $z_k =$

$x_k + (1/\alpha_k)y_k$ for some $x_k \in \ker(R^H)$ and $y_k \in \text{range}(R)$. From this construction it follows that $\|x_k\| \leq 1$ and $\|(1/\alpha_k)y_k\| \leq 1$. Using these vectors, one obtains

$$\|(x_k + (1/\alpha_k)y_k)^H(M + \alpha_k R)\| = \|x_k^H M + y_k^H R + (1/\alpha_k)y_k^H R\| \to 0. \tag{2.2}$$

It follows from (2.2) that $\|y_k\|$ is bounded; otherwise $\|y_k^H R\|$ would be unbounded since $y_k \in \text{range}(R)$ and (2.2) would not converge to zero since all other terms are obviously bounded as described above. Moreover, it follows that $x_k^H M + y_k^H R \to 0$ since boundedness of $\|y_k\|$ implies $(1/\alpha_k)y_k \to 0$.

Also, it follows that $\|x_k\| \to 1$ since $\|y_k\|$ is bounded and $\|z_k\|^2 = \|x_k\|^2 + (1/\alpha)^2\|y_k\|^2 = 1$. Thus, arbitrary cluster points of the sequences, x_k and y_k, say x_* and y_*, must satisfy $x_*^H M + y_*^H R = 0$, $x_*^H R = 0$, and $x_* \neq 0$, i.e., $\text{col}(\, x_* \quad y_* \,)$ is contained in the left null-space of

$$\begin{bmatrix} M & R \\ R & 0 \end{bmatrix}. \tag{2.3}$$

Again, let $\{y_i\}$, $i = 1, \ldots, n - \text{rank}(R)$ be a basis for the left null-space of R. Clearly the vectors, $\{\text{col}(\, 0, \quad y_i \,)\}$ are contained in the left null-space of the matrix in (2.3) and are linearly independent of $\text{col}(\, x_*, \quad y_* \,)$ because $x_* \neq 0$, i.e., there exist $n - \text{rank}(R) + 1$ linearly independent vectors in the left null-space of the matrix given in (2.3). This implies (2.1). □

Depending on the dimension and structure of the matrices involved, the nonexistence of a solution to the RRP may be generic, i.e., even though a solution to the RRP may exist in principle, it may not exist, numerically speaking. Numerically, the RRP problem can be perturbed into an MNRRP problem. This problem is discussed in [17] within the context of the controllability robustness problem mentioned earlier. This provides additional motivation for designing single algorithms that work on either problem. Such an algorithm is developed in the following section.

3. Algorithm development. The Newton approach to the problem is formulated by differentiating the smallest singular value of $M - R$. To avoid differentiation problems associated with singular values, assume that the singular values of $M - R$ are distinct for almost all R in $\mathcal{R}$. Certain problem structures may cause the singular values to be nondistinct throughout $\mathcal{R}$. This paper does not consider such problem structures. The function

$$f(\alpha) \triangleq \sigma_n(M - \alpha R) \tag{3.1}$$

is differentiable with respect to α (α is real) as long as the nth singular value of $M - R$ is distinct and nonzero. The derivative of f is given by $-\text{Re}(u^H R v)$ where u and v form a left and right unit length singular vector pair associated with $\sigma_n(M - \alpha R)$.

When M and all $R \in \mathcal{R}$ are real this differentiation poses no difficulty and there results a simple expression for Newton iteration for finding an RRP. For example, when $\mathcal{R}$ is a one-dimensional space having R as its basis, one obtains the Newton iteration,

$$\begin{aligned} \alpha_{k+1} &= \alpha_k + [u_k^H(M - \alpha_k R)v_k]/(u_k^H R v_k) \\ &= (u_k^H M v_k)/(u_k^H R v_k), \end{aligned} \tag{3.2}$$

where u_k and v_k are left and right unit length singular vectors associated with $\sigma_n(M - \alpha_k R)$. The sequence $\alpha_k \to \alpha_*$ as $k \to \infty$ and the quantity $\alpha_* R$ becomes the RRP.

Here, the similarity of (3.2) and Rayleigh quotient iteration is evident. Rayleigh quotient iteration is used for the generalized eigenvalue problem and the nonlinear eigenvalue problem [11], [14]. When $M-\alpha_k R$ is symmetric, (3.2) is virtually the same as Rayleigh quotient iteration. This follows because $M-\alpha_k R$ being symmetric implies $u_k = \pm v_k$. A notable difference between (3.2) and Rayleigh quotient iteration is the use of different vectors on the left and right in (3.2), obviously necessary if dealing with rectangular matrices. Hence, (3.2) appears to be a generalization of Rayleigh quotient iteration suitable for use with rectangular matrices. The vector used for Rayleigh quotient iteration is usually obtained via inverse iteration. In (3.2), u_k and v_k are obtained from a singular value decomposition. However, one can envision obtaining u_k and v_k via left and right pairs of inverse iterations.

When $\mathcal{R}$ is multidimensional, one must locate $\Delta R_k \in \mathcal{R}$ such that

$$u_k^H(\Delta R_k)v_k = u_k^H(M - R_k)v_k \tag{3.3}$$

and set $R_{k+1} = R_k + \Delta R_k$. This follows because a Newton iteration sets $0 = \sigma_n + \Delta\sigma_n = u^H(M - R)v - u^H(\Delta R)v$ which is equivalent to (3.3).

On the other hand, when M or R have nonzero imaginary parts, the successive approximations of the Newton iteration are determined by locating $\Delta R_k \in \mathcal{R}$ such that

$$\mathrm{Re}\big(u_k^H(\Delta R_k)v_k\big) = u_k^H(M - R_k)v_k \tag{3.4}$$

with $R_{k+1} = R_k + \Delta R_k$. This formulation is more difficult to implement than (3.3) because of the presence of the real part. The real part is present because the derivative of the function in (3.1) is $\mathrm{Re}(u^H R v)$ duly generalized to multiple dimensions in (3.3).

The first main result of this section shows that the quantity $-u^H R v$ may be utilized in place of the actual derivative of $f(\alpha)$ in a Newton iteration when M or R is complex. Strictly speaking, the quantity $-u^H R v$ is not the derivative of $f(\alpha)$ as given in (3.1), but it is the derivative of a complex-valued function having the same modulus as $\sigma_n(M - \alpha R)$. Obviously locating zeros of this function is equivalent to locating zeros of $\sigma_n(M - \alpha R)$. This result is stated below in Proposition 3.1, and justifies using the iteration given by (3.2) even if R and M are complex.

PROPOSITION 3.1. *Let $M, R \in \mathbb{C}^{n\times m}$. Given α_0 for which the first n singular values of $M-\alpha_0 R$ are nonzero and distinct, there exists a locally differentiable function $c(\alpha)$ for which $|c(\alpha)| = \sigma_n(M - \alpha R)$ in a neighborhood of α_0 and which satisfies $c'(\alpha_0) = -u^H R v$, where u and v are a pair of singular vectors associated with $\sigma_n(M - \alpha_0 R)$. Note that α is real.*

Proof. Let $u(\alpha)$ and $v(\alpha)$ be a left and right unit length singular vector pair associated with $\sigma_n(M - \alpha R)$. Consider the solution of the initial value problem

$$c'(\alpha) = -\frac{(u^H(\alpha)Rv(\alpha))}{\sigma_n(M - \alpha R)}c(\alpha)$$

with the initial condition $c(\alpha_0) = \sigma_n(M - \alpha_0 R) \neq 0$. The solution of this problem exists in some neighborhood of α_0. Since α is a real variable

$$\begin{aligned}\frac{d}{d\alpha}c(\alpha)\bar{c}(\alpha) &= -2\frac{c(\alpha)\bar{c}(\alpha)}{\sigma_n(M - \alpha R)}\,\mathrm{Re}\big(u^H(\alpha)Rv(\alpha)\big)\\ &= c(\alpha)\bar{c}(\alpha)\frac{2f'(\alpha)}{f(\alpha)},\end{aligned}$$

where $f(\alpha)$ is defined in (3.1). This implies that $c(\alpha)\bar{c}(\alpha) = [\sigma_n(M - \alpha R)]^2$ and completes the proof. □

The function $c(\alpha)$ may be interpreted as a complex-valued singular value since it satisfies $|c(\alpha)| = \sigma_n(M - \alpha R)$. By associating singular values with the modulus of a differentiable complex valued function, an appealing derivative formulation results. Left and right singular vectors associated with $c(\alpha)$ can also be defined. They must be selected to satisfy $(M - \alpha R)v(\alpha) = c(\alpha)u(\alpha)$ and $(M - \alpha R)^H u(\alpha) = \bar{c}(\alpha)v(\alpha)$.

Clearly $c(\alpha)$ is zero if and only if $\sigma_n(M - \alpha R) = 0$. This suggests applying the Newton method to $c(\alpha)$ rather than applying it to $\sigma_n(M - \alpha R)$. Given any initial guess, say α_0, the derivative of $c(\cdot)$ at α_0 becomes $-u^H(\alpha_0)Rv(\alpha_0)$ since $\sigma_n(M - \alpha_0 R) = c(\alpha_0)$. This amounts to using (3.3) in the complex case as opposed to (3.4).

Iteration based on (3.3) or (3.4) will be unstable near a solution to the MNRRP. This occurs near a positive local minimum of $\sigma_n(M - \alpha R)$ where the derivative vanishes, i.e., $u^H Rv$ approaches zero, making the iteration based on (3.3) unstable. As an illustration of this instability, consider Example 3.1.

Example 3.1. Determine $\alpha \in \mathbb{R}$, which minimizes $\sigma_2(M - \alpha R)$, where

$$M = \begin{bmatrix} 10 & 0 & 10 \\ 0 & 1 & 0 \end{bmatrix}; \qquad R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 100 \end{bmatrix}.$$

A local minimum occurs when α is approximately equal to 2.00975×10^{-4}. Applying (3.3) with $\alpha_0 = 0$ results in $\alpha_1 = 1$, for which $\sigma_n(M - \alpha_1 R)$ is approximately 8.955. The divergence of α_1 from the actual solution occurs because the derivative of $\sigma_n(M - \alpha R)$ is nearly zero at the initial point.

As a method of stabilizing the Newton approach consider determining ΔR_k to minimize the least squares problem,

$$\min_{\Delta R_k \in \mathcal{R}} \| [u]_k^H (M - R_k - \Delta R_k) \, [\, [v_n]_k \quad [v_{n+1}]_k \quad \cdots \quad [v_m]_k \,] \| \tag{3.5}$$

and setting $R_{k+1} = R_k + \Delta R_k$. The vectors $[u]_k$ and $[v_n]_k$ are left and right unit length singular vectors of $M - R_k$ associated with $\sigma_n(M - R_k)$, while the orthonormal vectors $[v_i]_k$ for $i = n+1, \ldots, m$ span the kernel of $M - R_k$. The notation $[\ ]_k$ denotes the kth iterate of the quantity in the brackets. This notation has been introduced to avoid confusion between the ith column vector v_i and the kth iterate of the ith column vector, $[v_i]_k$. Note that the inversion implicit in the solution of (3.5) may be nonsingular even if the quantity $\mathrm{Re}\big([u]_k^H R_k [v_n]_k\big)$ is zero for all $R \in \mathcal{R}$. In the one-dimensional case, the norm of the ΔR_k obtained from (3.5) will be less than or equal to the norm of the ΔR_k obtained from (3.3), thereby adding stability to the algorithm. In the multidimensional case, adding the null-space information has the effect of adding second derivation information. The additional vectors are used to guarantee convergence in Proposition 4.1. Moreover, solution of (3.5) always yields a descent direction for $f(\alpha)$ as given in (3.1), i.e., there will always be some $R_* \in \mathcal{R}$ on the line connecting R_k and $R_k + \Delta R_k$ for which $\sigma_n(M - R_*) < \sigma_n(M - R_k)$. This claim is verified by Lemma 4.2, appearing in §4. By a suitable choice for the stepsize, global convergence is achieved. A specific choice for the stepsize appears in §4. However, it is not immediately apparent that the convergence rate of (3.3) is retained locally after incorporating the null-space information into the iteration formula. The stabilizing effect of using the extra singular vectors can be observed by executing a single iteration of (3.5) on Example 3.1. The resulting value of α_1 is 1.9996×10^{-4} compared with the locally minimizing value 2.0098×10^{-4}.

The remainder of this section serves only to suggest that the convergence rate achieved by (3.5) is nearly the same as that achieved by (3.2) by restricting the discussion to the one-dimensional, real case. A convergence proof appears in §4 for the general, complex case. In the one-dimensional real case, the iteration based on (3.5) is approximately equivalent to the iteration function

$$\begin{aligned}\alpha_{k+1} &= \alpha_k - \frac{f(\alpha_k)}{f'(\alpha_k) + (f(\alpha_k)f''(\alpha_k)/f'(\alpha_k))} \\ &= \alpha_k - \frac{f(\alpha_k)f'(\alpha_k)}{(f'(\alpha_k))^2 + f(\alpha_k)f''(\alpha_k)}.\end{aligned} \tag{3.6}$$

The iteration function of (3.6) converges quadratically in a neighborhood of either a single zero of f or a zero of f'. Near a zero of f' it converges quadratically so long as f and f' do not simultaneously vanish. The iteration specified by (3.6) is an approximation to the iteration of (3.5) in the one-dimensional, real case when $m > n$. Loosely speaking, the approximation becomes better as f becomes smaller so that performance improves as the singular value associated with a solution of the MNRRP diminishes. Performance improves as robustness decreases.

A correspondence between (3.5) and (3.6) becomes clear after examining the closed form solution of (3.5) in the case when M is real, $m > n$, and $\mathcal{R}$ is one-dimensional having a real matrix R as its basis:

$$\alpha_{k+1} = \alpha_k + \frac{([u]_k^H(M - \alpha_k R)[v_n]_k)([u]_k^H R[v_n]_k)}{([u]_k^H R[v_n]_k)^2 + \sum_{i=n+1}^{m}([u]_k^H R[v_i]_k)^2}. \tag{3.7}$$

Most of the quantities in (3.7) correspond to quantities in (3.6), i.e., $f(\alpha_k)$ is given by $[u]_k^H(M - \alpha_k R)[v_n]_k$ and $f'(\alpha_k)$ is given by $-[u]_k^H R[v_n]_k$. The correspondence with (3.6) would be complete if $f''(\alpha_k)$ were equal to the quantity

$$\frac{1}{\sigma_n(M - \alpha_k R)} \sum_{i=n+1}^{m} ([u]_k^H R[v_i]_k)^2. \tag{3.8}$$

The quantity in (3.8) approximates $f''(\alpha_k)$ if $m > n$ and $\sigma_n(M - \alpha_k R) << \sigma_{n-1}(M - \alpha_k R)$. To see this, consider the second derivative of the quantity $\sigma_n(M - \alpha R)$. With the real restriction on M and R, the second derivative of f may be written as

$$\begin{aligned}f''(\alpha) &= -\big(u'(\alpha)^T R v(\alpha) + u(\alpha)^T R v'(\alpha)\big) \\ &= [\,u(\alpha)^T \quad v(\alpha)^T\,] \begin{bmatrix} 0 & -R \\ -R^T & 0 \end{bmatrix} \begin{bmatrix} u'(\alpha) \\ v'(\alpha) \end{bmatrix}.\end{aligned}$$

By differentiation of the equation

$$\begin{bmatrix} -\sigma_n(M - \alpha R)I & (M - \alpha R) \\ (M - \alpha R)^T & -\sigma_n(M - \alpha R)I \end{bmatrix} \begin{bmatrix} u(\alpha) \\ v(\alpha) \end{bmatrix} = 0$$

and recognition that

$$[\,u(\alpha)^T \quad v(\alpha)^T\,] \begin{bmatrix} u'(\alpha) \\ v'(\alpha) \end{bmatrix} = 0,$$

it is possible to show that

$$\begin{bmatrix} u'(\alpha) \\ v'(\alpha) \end{bmatrix} = -D^+(\alpha) \begin{bmatrix} f'(\alpha) & -R \\ -R^T & f'(\alpha) \end{bmatrix} \begin{bmatrix} u(\alpha) \\ v(\alpha) \end{bmatrix}, \tag{3.9}$$

where

$$D(\alpha) = \begin{bmatrix} -\sigma_n(M-\alpha R)I & (M-\alpha R) \\ (M-\alpha R)^T & -\sigma_n(M-\alpha R)I \end{bmatrix}$$

and the quantities $u(\alpha)$ and $v(\alpha)$ are unit length left and right singular vectors associated with $\sigma_n(M-\alpha R)$. Assuming the singular values of $M-\alpha R$ are distinct, the singular values of $D(\alpha)$ in (3.9) can be enumerated as follows: $2n-2$ nonzero singular values of the form $|\sigma_n \pm \sigma_i|$, two zero singular values, and a singular value equal to σ_n having multiplicity $m-n$. The singular value decomposition of the matrix, $D(\alpha)$, (dropping the explicit dependence on α) can be written

$$D(\alpha) = \sum_{i=1}^{n-1}(1/2)(\sigma_i - \sigma_n)\begin{bmatrix} u_i \\ v_i \end{bmatrix}[u_i^T \quad v_i^T] + \sum_{i=1}^{n-1}(1/2)(-\sigma_i - \sigma_n)\begin{bmatrix} u_i \\ -v_i \end{bmatrix}[u_i^T \quad -v_i^T] - \sigma_n \sum_{i=n+1}^{m}\begin{bmatrix} 0 \\ v_i \end{bmatrix}[0 \quad v_i^T],$$

where σ_i is a singular value of $M-\alpha R$ (not of $D(\alpha)$) having associated left and right unit length singular vectors, u_i and v_i, and where v_i, $i = n+1, \ldots, m$ are an orthonormal basis for the kernel of $M-\alpha R$. If σ_n is much smaller than the other nonzero singular values and $m > n$, then an approximation to the pseudoinverse can be obtained by truncating the smallest $2n-2$ nonzero singular values. This results in the approximate equality

$$D^+(\alpha) \approx -\frac{1}{\sigma_n(M-\alpha R)} \sum_{i=n+1}^{m} \begin{bmatrix} 0 & 0 \\ 0 & v_i v_i^T \end{bmatrix}. \tag{3.10}$$

Substituting this approximation into (3.9) results in the approximate equality

$$f''(\alpha_k) \approx \frac{1}{\sigma_n(M-\alpha_k R)} \sum_{i=n+1}^{m} (u_k^H R v_i)^2,$$

where the v_i, $i = n+1, n+2, \ldots, m$ span the null-space of $M-\alpha_k R$. This demonstrates the approximate equivalence of (3.5) and (3.6) when $\sigma_n(M-\alpha_k R) << \sigma_{n-1}(M-\alpha_k R)$ and $m > n$. The convergence rate increases as $\sigma_n(M-\alpha_k R)$ decreases and the approximation of (3.10) improves. Experimental evidence presented in §5 supports this intuition for both the one-dimensional case and multidimensional case. Section 5 presents a multidimensional complex case. Section 4 presents a globally convergent stepsize, demonstrates convergence, and discusses implementation issues.

4. Implementation. This section discusses specific details of an implementation of the algorithm and demonstrates convergence of the algorithm to a local minimum of $\sigma_n(M-\alpha R)$. The first goal of this section is to establish convergence of the following algorithm.

1. Set $k = 0$, select R_0; $R_0 = [0]$ is a convenient choice.
2. Compute the singular value decomposition of the matrix, $M - R_k$, i.e., determine orthonormal sets $\{[u_i]_k\}$, $i = 1, \ldots, n$, and $\{[v_i]_k\}$, $i = 1, \ldots, m$ so that $M - R_k = \sum_{i=1}^{n} [u_i]_k [\sigma_i]_k [v_i]_k^H$.
3. Let $V_k = [\,[v_n]_k \quad [v_{n+1}]_k \quad \ldots \quad [v_m]_k\,]$.

4. Let ΔR_k be the matrix having smallest norm that minimizes the quantity $\|[u_n]_k^H(\Delta R_k)V_k - [u_n]_k^H(M - R_k)V_k\|$.

5. Set $a_k = \|[u_n]_k^H(\Delta R_k)(I - V_kV_k^H)(M - R_k)^+(\Delta R_k)\|$.

6. Set $\gamma_k = \min\left((1/4)\left(\frac{\|[u_n]_k^H(\Delta R_k)V_k\|^2}{a_k[\sigma_n]_k}\right), 1\right)$.

7. Let $R_{k+1} = R_k + \gamma_k \Delta R_k$.

8. For some prespecified ϵ, if $\gamma_k\|\Delta R_k\| < \epsilon\|M - R_k\|$ stop; otherwise let $k = k + 1$ and go to 2.

Identifying some quantities will assist in verifying convergence of the algorithm. Let $c_k^H = [u_n]_k^H(\Delta R_k)V_k$, let the error $e_k^H = [u_n]_k^H(M - R_k)V_k - [u_n]_k^H(\Delta R_k)V_k$, and let $s_k^H = [u_n]_k^H(M - R_k)V_k$. (The quantity e_k represents the error in the least squares solution of Step 4.) Note that e_k is orthogonal to the vector c_k and $e_k + c_k = s_k$, i.e., $\|c_k\|^2 + \|e_k\|^2 = \|s_k\|^2$. Here, orthogonality is not defined with respect to the usual inner product on $\mathbb{C}^{m-n+1}$. Instead, $\mathbb{C}^{m-n+1}$ is viewed as a $2\times(m-n-1)$ dimensional real space having the inner product $< x, y >= \mathrm{Re}(x)^T\,\mathrm{Re}(y) + \mathrm{Im}(x)^T\,\mathrm{Im}(y)$.

Subsequently, a preliminary lemma will prove useful.

LEMMA 4.1. *The orthogonality of c_k and e_k implies that*

$$\left(\|e_k\|^2 + (1-\gamma)^2\|c_k\|^2\right)^{1/2} \le \|s_k\| - \frac{\|c_k\|^2}{2\|s_k\|}\gamma$$

as long as $\gamma \in [0, 1]$.

Proof. Consider that

$$\begin{aligned}\left[\|e_k\|^2 + \|c_k\|^2\right] - \left[\|e_k\|^2 + (1-\gamma)^2\|c_k\|^2\right] &= \gamma(2-\gamma)\|c_k\|^2 \\ &\ge \gamma\|c_k\|^2\end{aligned} \tag{4.1}$$

for $\gamma \in [0, 1]$. Since

$$\left[\|e_k\|^2 + (1-\gamma)^2\|c_k\|^2\right]^{1/2} \le \left[\|e_k\|^2 + \|c_k\|^2\right]^{1/2} = \|s_k\|,$$

it follows that

$$\left[\|e_k\|^2 + \|c_k\|^2\right]^{1/2} + \left[\|e_k\|^2 + (1-\gamma)^2\|c_k\|^2\right]^{1/2} \le 2\|s_k\|. \tag{4.2}$$

Dividing inequality (4.1) by inequality (4.2) provides the stated result. □

The following lemma provides an upper bound for the the nth singular value, $\sigma_n(M - R_k - \gamma\Delta R_k)$ for $\gamma \in [0, 1]$.

LEMMA 4.2. *With the quantities as specified in the algorithm statement above, any* $\gamma \in [0, 1]$ *satisfies*

$$\sigma_n(M - R_k - \gamma\Delta R_k) \le [\sigma_n]_k - \frac{\|c_k\|^2}{2[\sigma_n]_k}\gamma + a_k\gamma^2.$$

Proof. Define $\hat{v}_k^H = [u_n]_k^H\Delta R_k - c_k^H V_k^H$ or equivalently $\hat{v}_k^H = [u_n]_k^H\Delta R_k(I - V_kV_k^H)$ and let $[\hat{u}]_k$ satisfy $[\hat{u}]_k^H(M - R_k) = \hat{v}_k^H$ (a solution is guaranteed because $\hat{v}_k$ is in the row space of $M - R_k$ since $V_k^H\hat{v}_k = 0$). Consider the product

$$\begin{aligned}&([u_n]_k + \gamma\hat{u}_k)^H((M - R_k) - \gamma\Delta R_k) \\ &\quad = s_k^HV_k^H - \gamma(c_k^HV_k^H + \hat{v}_k^H) + \gamma\hat{u}_k^H(M - R_k) - \gamma^2\hat{u}_k^H\Delta R_k \\ &\quad = \left(e_k^H + (1-\gamma)c_k^H\right)V_k^H - \gamma^2\hat{u}_k^H\Delta R_k,\end{aligned} \tag{4.3}$$

where $\hat{u}^H = u_n^H(\Delta R_k)(I - V_k V_k^H)(M - R_k)^+$. Applying Lemma 4.1 to the first term of (4.3) and recognizing that the norm of $[u_n]_k + \gamma\hat{u}_k$ is greater than one results in the statement of the lemma. □

Convergence of the algorithm is readily established in the following proposition. However, the following convergence proof requires a regularity assumption. Specifically, assume that the norm of ΔR_k as computed in Step 4 of the algorithm remains bounded. The norm of ΔR_k will remain bounded as long as the norm of the pseudo-inverse required for the least squares solution of Step 4 remains bounded. As suggested in the previous section, this is related to the assumption that the second derivative of the nth singular value does not vanish at the solution point.

PROPOSITION 4.1. *If $\|\Delta R_k\|$ remains bounded, the algorithm stated above converges to a necessary condition for a solution to an* MNRRP *in the sense that either $\{[\sigma_n]_k\} \to 0$ or $\mathrm{Re}([u_n]_k^H(R)[v_n]_k) \to 0$ for each $R \in \mathcal{R}$.*

Proof. The sequence $\{\sigma_n\}$ is nonincreasing. To see this, consider that the upper bound given by Lemma 4.2 is minimized for $\gamma \in [0, 1]$ by the choice

$$\gamma_k = \min\left[(1/4)\left(\frac{\|c_k\|^2}{a_k[\sigma_n]_k}\right), 1\right].$$

For this choice, one can show that

$$[\sigma_n]_k - [\sigma_n]_{k+1} \geq \min\left[(1/16)\frac{\|c_k\|^4}{[\sigma_n]_k^2 a_k}, (1/4)\frac{\|c_k\|^2}{[\sigma_n]_k}\right].$$

It can be shown that $a_k \leq \|\Delta R_k\|_F^2/[\sigma_{n-1}]_k$. Hence,

$$[\sigma_n]_k - [\sigma_n]_{k+1} \geq \min\left[(1/16)\frac{\|c_k\|^4[\sigma_{n-1}]_k}{[\sigma_n]_k^2\|\Delta R_k\|^2}, (1/4)\frac{\|c_k\|^2}{[\sigma_n]_k}\right].$$

Assume $[\sigma_n]_k$ does not converge to zero. The sequence c_k has 0 as a limit point since the sequence $\{[\sigma_n]_k - [\sigma_n]_{k+1}\} \to 0$, the sequence $\{[\sigma_n]_k\}$ is bounded, and by assumption $\{\|\Delta R_k\|\}$ is bounded. To show that $[u_n]_k^H(R)[v_n]_k \to 0$ for each $R \in \mathcal{R}$, take any $R \in \mathcal{R}$ and any real α and consider that

$$\|\alpha[u_n]_k^H RV_k - [u_n]_k^H(M - R_k)V_k\| \geq \|[u_n]_k^H \Delta R_k V_k - [u_n]_k^H(M - R_k)V_k\|$$

for all k from the least squares solution of Step 4. The sequences, $\{[u_n]_k^H\}$, $\{V_k^H\}$, and $\{s_k\}$ have cluster points, say u_*,V_*, and s_*, which satisfy

$$\|\alpha u_*^H RV_* - s_*\| \geq \|s_*\|.$$

This implies that the row vectors $u_*^H RV_*$ and s_* are orthogonal (with respect to the inner product on the real space described earlier). This can be the case only if $\mathrm{Re}(u_*^H R[v_n]_*) = 0$, or if $s_* = 0$ in which case $\{[\sigma_n]_k\} \to 0$. □

A few comments are in order concerning implementation of the algorithm described above. The least squares solution required in Step 4 may be accomplished with regard to a particular basis for $\mathcal{R}$, say $\{R_i\}$, $i = 1, \ldots, s$. The matrix ΔR_k is expressed in the given basis as $\Delta R_k = \sum_{i=1}^s R_i\xi_i$. Here, assume that the basis is selected so that the norm of ΔR is equal to the norm of the ξ vector. A solution for the real coefficient vector ξ is formulated as follows: define

$$E_k = \begin{bmatrix} V_k^H R_1^H[u_n]_k & V_k^H R_2^H[u_n]_k & \cdots & V_k^H R_s^H[u_n]_k \end{bmatrix}$$

and $F_k = V_k^H (M - R_k)^H [u_n]_k$, then ξ is given by

$$\xi = \begin{bmatrix} \mathrm{Re}(E_k) \\ \mathrm{Im}(E_k) \end{bmatrix}^+ \begin{bmatrix} \mathrm{Re}(F_k) \\ \mathrm{Im}(F_k) \end{bmatrix}.$$

The pseudoinverse of $M - R_k$ needed in Step 5 requires inversion of only the first $n-1$ singular values. The projector $(I - V_k V_k^H)$ projects out the remaining singular space. Hence, the product $(I - V_k V_k^H)(M - R_k)^+$ can be obtained directly from the singular value decomposition of $M - R_k$. Evaluation of the formula directly as written in Step 5 may lead to a poor estimate for a_k.

The algorithm statement presents a coordinate-free version of the algorithm. It may seem that exploitation of a particular coordinate system would make the algorithm more efficient. The algorithm presented requires a full singular value decomposition at each iteration. A coordinate dependent implementation could use the singular value decomposition from each iteration to update the coordinate system so that $M - R_k$ is nearly diagonal at the next iteration, resulting in faster computation of the subsequent decomposition. Unfortunately, an implementation of this idea reveals that this technique requires more time because of the costly full coordinate transformations that must be applied to M and all of the basis matrices for $\mathcal{R}$ at each iteration. For the same reason, it is costly to use approximate singular value/vector information. Because the singular value decomposition is required at the solution point, using approximate singular space information requires transforming the coordinate system at each iteration so as to obtain the final singular value decomposition in the limit. It is not clear if there exists an intelligent choice for the coordinate system at each iteration that may be exploited to increase the efficiency of the technique.

5. Numerical results. This section shows numerical examples illustrating some of the topics discussed in this paper. Data from numerical experiments are presented that evaluate the numerical accuracy and experimental rate of convergence of the algorithms. The numerical experiments exhibit the following topics:

(i) Quadratic convergence to the solution in a neighborhood of an RRP;

(ii) nearly quadratic convergence to the solution of an MNRRP if the final robustness measure is small;

(iii) improvement of the convergence rate as the separation between the last two singular values increases.

Consider the following example, which is a parametrized family of problems having a known solution.

Example 5.1. Let M, P, R_1, R_2, and R_0 be defined as follows:

$$M_0 = \begin{bmatrix} -1-1i & -2+1i & -4-1i & -4-1i & -5-3i \\ -1-1i & -2-1i & -4+1i & -4-3i & -3-1i \\ 1-1i & 2-1i & 4-1i & 4-3i & 5+1i \\ 1-1i & 2+1i & 4+1i & 4-1i & 3+3i \end{bmatrix},$$

$$R_1 = \begin{bmatrix} 0.6 & 3.6 & 0.3 & 0.6 & 0.3 \\ -0.6 & 4.4 & -0.5 & -0.6 & -0.5 \\ 0.5 & -4.9 & 0.5 & -0.2 & -0.2 \\ -0.5 & -3.9 & -0.7 & 0.2 & 0 \end{bmatrix},$$

$$R_2 = \begin{bmatrix} 0.6 & 3.0 & -0.3 & 0.6 & 0.3 \\ -0.6 & 4.6 & -0.3 & -0.6 & -0.5 \\ 0.5 & -6.3 & -0.9 & -0.2 & -0.2 \\ -0.5 & -2.9 & 0.3 & 0.2 & 0 \end{bmatrix},$$

$$R_0 = \begin{bmatrix} 0.18 & 1.02 & 0.03 & 0.18 & 0.09 \\ -0.18 & 1.34 & -0.13 & -0.18 & -0.15 \\ 0.15 & -1.61 & 0.01 & -0.06 & -0.06 \\ -0.15 & -1.07 & -0.11 & 0.06 & 0 \end{bmatrix},$$

$$P = \begin{bmatrix} 0.4+0.8i & -0.1+0.3i & 0.1-0.3i & -0.2-0.4i & -0.1+0.3i \\ -0.4 & -0.1-0.1i & 0.1+0.1i & 0.2 & -0.1-0.1i \\ 1.6+1.2i & 0.1+0.7i & -0.1-0.7i & -0.8-0.6i & 0.1+0.7i \\ -0.8-1.2i & 0.1-0.5i & -0.1+0.5i & 0.4+0.6i & 0.1-0.5i \end{bmatrix}.$$

Let $\mathcal{R} = \{\alpha_1 R_1 + \alpha_2 R_2 : \alpha_1, \alpha_2 \in \mathbb{R}\}$ and let $M = M_0 + sP + R_0$. The matrix R_0 satisfies a necessary condition for being an MNRRP solution. The minimum singular value associated with this perturbation is exactly $s\sqrt{11.44}$. The second to last singular value is approximately 2.32. The algorithm stated in §4 was executed for varying values of s. An implementation of the algorithm was executing using MATLAB 4.0 on a 486. For each value of s, the following quantities are recorded in Table 1: the required number of iterations N, the relative error e, the computed rate of convergence r, the associated coefficient of convergence C, and the floating point operation (flop) count reported by MATLAB. The relative error was computed as the Frobenius norm of the difference between the computed and the actual solution divided by the Frobenius norm of M_0.

TABLE 1
Experimental results for parametrized example.

s	N	e	r	C	Flop count
0	5	2.9×10^{-17}	2.65	2.3284	25,924
0.001	6	4.2×10^{-17}	1.22	0.0065	42,343
0.01	7	9.0×10^{-17}	0.91	0.0004	51,771
0.1	10	1.7×10^{-16}	1.00	0.0232	73,740
0.2	45	2.2×10^{-14}	1.00	0.5307	349,227
0.3	124	9.7×10^{-14}	1.00	0.8409	934,658

This example supports the claim made earlier that performance improves with the separation of the values of σ_n and σ_{n-1} at the solution point.

6. Conclusion. This paper examines a Newton approach to the solution of the structured RRP problem as well as the structured MNRRP problem. The Newton iteration is based on driving the smallest nonzero singular value of the matrix $M-R$ to zero. The paper demonstrates that a direct implementation of the Newton approach will be unstable near an MNRRP that is not also an RRP. The paper proposes a method that stabilizes the Newton iteration, retains quadratic convergence to an RRP, and often has rapid convergence to an MNRRP. Proper choice for the stepsize achieves global convergence to a solution. In addition, the paper provides some insight into the asymptotic behavior of singular value surfaces and into the existence of solutions to the RRP and MNRRP problems. Numerical examples are provided demonstrating the main ideas of the paper.

REFERENCES

[1] E. K. Blum and P. B. Geltner, *Numerical solution of eigentuple-eigenvector problems in Hilbert spaces by a gradient method*, Numer. Math., 31 (1978), pp. 247–263.

[2] E. K. Blum and A. R. Curtis, *A convergent gradient method for matrix eigenvector-eigentuple problems*, Numer. Math., 31 (1978), pp. 231–246.

[3] D. Boley, *Computing rank-deficiency of rectangular matrix pencils*, Syst. Contr. Lett., 9 (1987), pp. 207–214.

[4] R. Byers, *A bisection method for measuring the distance of a stable matrix to the unstable matrices*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 875–881.

[5] J. W. Demmel, *The smallest perturbation of a submatrix which lowers the rank and constrained total least squares problems*, SIAM J. Numer. Anal., 24 (1987), pp. 199–206.

[6] ———, *The componentwise distance to the nearest singular matrix*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 10–19.

[7] R. Eising, *The distance between a system and the set of uncontrollable systems*, in Proc. MTNS, June 1983, Springer-Verlag, Beer-Sheva, 1984, pp. 303–314.

[8] R. Eising, *Between controllable and uncontrollable*, Syst. Contr. Lett., 4 (1984), pp. 263–264.

[9] J. C. Doyle, *Analysis of feedback systems with structured uncertainties*, Proc. IEE-D, 129 (1982), pp. 242–250.

[10] M. K. H. Fan and A. L. Tits, *Characterization and efficient computation of the structured singular value*, IEEE Trans. Automat. Contr., AC-31 (1986), pp. 734–743.

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

[12] D. Hinrichsen, B. Kelb, and A. Linnemann, *An algorithm for the computation of the structured complex stability radius*, Automatica, 25 (1989), pp. 771–775.

[13] C. C. Paige, *Properties of numerical algorithms relating to controllability*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 130–138.

[14] A. Ruhe, *Algorithms for the nonlinear eigenvalue problem*, SIAM J. Numer. Anal., 10 (1973), pp. 674–689.

[15] T. Slivnik and G. Tomšič, *A numerical method for the solution of two-parameter eigenvalue problems*, J. Comput. Appl. Math., 15 (1986), pp. 109–115.

[16] G. A. Watson, *Computing the structured singular value*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1054–1066.

[17] M. A. Wicks and R. A. DeCarlo, *Computing the distance to an uncontrollable system*, IEEE Trans. Automat. Control, AC-36 (1991), pp. 39–49.

[18] C. F. Van Loan, *How near is a stable matrix to an unstable matrix?*, Contemp. Math., 47 (1985), pp. 465–478.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 138–155, January 1995

010

DOWNDATING THE RANK-REVEALING URV DECOMPOSITION*

HAESUN PARK† AND LARS ELDÉN‡

Abstract. An accurate algorithm is presented for downdating a row in the rank-revealing URV decomposition that was recently introduced by Stewart. By downdating the full rank part and the noise part in two separate steps, the new algorithm can produce accurate results even for ill-conditioned problems. Such problems occur, for example, when the rank of the matrix is decreased as a consequence of the downdate. Other possible generalizations of existing QR decomposition downdating algorithms for the rank-revealing URV downdating are discussed. Numerical test results are presented that compare the performance of these new URV decomposition downdating algorithms in the sliding window method.

Key words. downdating, null space, rank-revealing decomposition, sliding window method, two-sided orthogonal decomposition, URV decomposition

AMS subject classifications. 65F20, 65F25

1. Introduction.

The singular value decomposition (SVD) is of great theoretical and practical importance [8]. One of its major merits is that it provides the rank of the matrix and a basis for four important spaces including the null space. However, the SVD has the drawback that it is computationally expensive. Especially when the problem is of recursive nature, the SVD requires $\mathcal{O}(n^3)$ flops for a matrix of order n even for a simple update such as adding a new row. Thus, algorithms that utilize the existing results for incorporating changes in data are desired. Our goals are to perform such modifications with as few operations and as little storage requirement as possible and to compute the new decomposition for rank-deficient matrices to obtain accurate results.

Recently, Stewart [18], [17] introduced two-sided orthogonal decompositions, called the rank-revealing URV and ULV decompositions (RR URVD and RR ULVD) that are effective in exhibiting the rank and the basis for the null space, and can be updated in $\mathcal{O}(n^2)$ flops. They are compromises between the SVD and a QR decomposition with some of the virtues of both.

Given a matrix $X \in \mathbf{R}^{p\times n}$, where $p \geq n$, we say that it has numerical rank r, if its singular values satisfy

$$\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} \geq \cdots \geq \sigma_n,$$

where σ_r is *large* compared to σ_{r+1}. Then there exist orthogonal matrices $U \in \mathbf{R}^{p\times p}$ and $V \in \mathbf{R}^{n\times n}$ such that,

$$U^T X V = \begin{pmatrix} T \\ 0 \end{pmatrix} = \begin{pmatrix} R & F \\ 0 & G \\ 0 & 0 \end{pmatrix}, \tag{1.1}$$

* Received by the editors December 28, 1992; accepted for publication (in revised form) by R. J. Plemmons, October 19, 1993.

† Computer Science Department, University of Minnesota, Minneapolis, Minnesota 55455 (hpark@cs.umn.edu). The work of this author was supported in part by National Science Foundation grant CCR-9209726 and by contract DAAL02-89-C-0038 between the Army Research Office and the University of Minnesota for the Army High Performance Computing Research Center.

‡ Department of Mathematics, Linköping University, S–581 83 Linköping, Sweden (laeld@math.liu.se).

where $T \in \mathbf{R}^{n\times n}$, $R \in \mathbf{R}^{r\times r}$, and $G \in \mathbf{R}^{(n-r)\times(n-r)}$ are upper triangular, and

$$\sigma_{\min}(R) \approx \sigma_r, \qquad \| F \|_F^2 + \| G \|_F^2 \approx \sigma_{r+1}^2 + \cdots + \sigma_n^2,$$

where $||\cdot||_F$ denotes the Frobenius norm. This is the RR URVD. Given an upper triangular matrix $T \in \mathbf{R}^{n\times n}$, let w be the unit right singular vector of T corresponding to the smallest singular value σ_n. If an orthogonal matrix $Q \in \mathbf{R}^{n\times n}$ makes

$$Q^T w = e_n$$

and we have the QR decomposition of the product TQ

$$TQ = Q^{(1)}T^{(1)},$$

where $Q^{(1)} \in \mathbf{R}^{n\times n}$ is orthogonal and $T^{(1)} \in \mathbf{R}^{n\times n}$ is upper triangular, then

$$\sigma_n = \| Tw \|_2 = \| {Q^{(1)}}^T TQQ^T w \|_2 = \| T^{(1)} e_n \|_2$$

($||\cdot||_2$ denotes the Euclidean vector norm). Thus the last column of $T^{(1)}$ is small. This is called a deflation and by applying the deflation repeatedly, the RR URVD for T can be computed [18], [17]. When the right orthogonal matrix $V = (V_1\ V_2)$ is partitioned according to the rank, where $V_1 \in \mathbf{R}^{n\times r}$ and $V_2 \in \mathbf{R}^{n\times(n-r)}$, then V_2 is an orthogonal basis for the null space of X. Note that the SVD is a special case of the RR URVD.

For recursive problems, two common ways of incorporating changes in data are the sliding rectangular window method and the exponential window method. For updating the SVD and the RR URVD after a new data row is attached, see [18], [17]. For phasing out the old data, one or more rows are deleted explicitly from X in the sliding rectangular window method. In the exponential window method, a value β, $0 < \beta < 1$, which is called a *forgetting factor*, is multiplied to existing rows to damp out the effect of the old information. After an update in the exponential window method, the numerical rank can increase, decrease, or stay the same. In particular, since the effect of old data is only gradually phased out, it is likely that situations occur where the numerical rank is not well determined, and thus the rank decisions are difficult. In such situations the choice of the forgetting factor and the tolerance used in the decision of the numerical rank becomes critical, especially since these quantities are closely related to each other.

The sliding window method can track the change in the information statistics more accurately than the forgetting factor method when there is an abrupt change in data such as when signals are turned on and off or outliers are removed [3]. Stewart's original presentation of the URVD was in the context of using an exponential window method in recursive least squares. This paper fills the gap by providing the details for using the URVD in the sliding window setting, making the URVD a complete tool that can handle both types of recursive problems efficiently.

An advantage of the sliding window method is the a priori information on the rank after the modification: mathematically, after adding a row, the rank can only stay the same or increase by one, and, after deleting a row, the rank can only stay the same or decrease by one. Thus, the indefinite steps of deflation in using the forgetting factor that results from not having any similar a priori information on the rank of the modified matrix can be eliminated.

The methods used in this paper are partly based on the algorithms for downdating described in [6], particularly a hybrid between the LINPACK and the CSNE

downdating algorithms. It is necessary to use such accurate algorithms here, since the downdates are often very ill conditioned when the numerical rank decreases.

The rest of this paper is organized as follows. In §2, we discuss the relation between downdating two-sided orthogonal decompositions and the QR decomposition. Then we present a two-step procedure for downdating the QR decomposition. In §3, we discuss the rank-revealing aspect of URVD downdating and present a new algorithm for downdating the RR URVD based on the two-step QR decomposition downdating algorithm together with other generalizations of QR decomposition downdating algorithms to the RR URVD downdating. Section 4 contains numerical test results comparing the accuracy of these algorithms in deciding rank and a basis for the null space.

There is a vast literature on algorithms in the area of subspace-based signal processing. The papers [1], [4], [19] deal with the problem of rank determination in recursive computations. A Lanczos procedure is used in [19] to find the signal subspace. The algorithm in [4] is based on the updating and downdating of a rank-revealing QR decomposition. An application of the exponential window method for the URV decomposition is described in [1].

2. Downdating a two-sided orthogonal decomposition. In this section we first show that the problem of downdating a two-sided orthogonal decomposition is related to that of downdating a QR decomposition. Then we present a two-step procedure for downdating a QR decomposition. This procedure plays an important role in our downdating algorithm for the RR URVD that is presented in the next section. We also introduce a downdating algorithm based on a reduction to a simpler problem.

One important requirement in an algorithm for downdating the rank-revealing URV decomposition is that it must not destroy the "large-small" structure and that it shall reveal the rank after the downdating. In this section, we leave the rank-revealing aspects aside. In the next section, we will show that the two-step procedure combined with condition number estimating methods can fulfill this requirement. The motivation for the two-step procedure is that we separate the downdating of T into two parts to take advantage of the fact that the matrix R is well-conditioned even if T may be ill-conditioned. In a rank-revealing context, since the matrices F and G are considered as "noise," it is undesirable to allow them to destroy the whole downdating procedure due to ill-conditioning.

2.1. A two-step procedure for downdating a QR decomposition. Suppose a two-sided orthogonal decomposition for $X \in \mathbf{R}^{p\times n}$

$$X = U\begin{pmatrix} T \\ 0 \end{pmatrix} V^T = U\begin{pmatrix} R & F \\ 0 & G \\ 0 & 0 \end{pmatrix} V^T \tag{2.1}$$

is given, where U and V are orthogonal, $R \in \mathbf{R}^{r\times r}$ and $G \in \mathbf{R}^{(n-r)\times(n-r)}$ for some $r \leq n$ are upper triangular. We assume that the matrix V is stored, since it is needed in many applications, e.g., subspace-based signal processing and applications, which require least squares solutions. However, if $p >> n$ then the extra storage for U may be prohibitive, thus we assume that U is *not stored.* A QR decomposition of X_v defined as

$$X_v \equiv XV = \begin{pmatrix} z^T \\ \tilde{X} \end{pmatrix} V = \begin{pmatrix} z_v^T \\ \tilde{X}_v \end{pmatrix}, \tag{2.2}$$

is

$$X_v = U \begin{pmatrix} T \\ 0 \end{pmatrix}.$$

The QR decomposition downdating problem for X_v

$$X_v = XV = \begin{pmatrix} z_{v1}^T & z_{v2}^T \\ \tilde{X}V_1 & \tilde{X}V_2 \end{pmatrix} = U \begin{pmatrix} R & F \\ 0 & G \\ 0 & 0 \end{pmatrix}, \qquad V = (V_1 \quad V_2),$$

where $z_v^T = (z_{v1}^T \quad z_{v2}^T)^T$ with $z_{v1} \in \mathbf{R}^{r\times 1}$ and $V_1 \in \mathbf{R}^{n\times r}$ can be formulated as a problem of finding the upper triangular matrix

$$\tilde{T} = \begin{pmatrix} \tilde{R} & \tilde{F} \\ 0 & \tilde{G} \end{pmatrix},$$

given the upper triangular matrix T and the row to be downdated, z_v^T, so that

$$\bar{J}^T \begin{pmatrix} z_{v1}^T & z_{v2}^T \\ \tilde{R} & \tilde{F} \\ 0 & \tilde{G} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ R & F \\ 0 & G \end{pmatrix},$$

is satisfied for some orthogonal matrix $\bar{J}$. Note that R is the upper triangular factor in the QR decomposition of XV_1. Similarly, $\tilde{R}$ is the upper triangular factor in the QR decomposition of $\tilde{X}V_1$.

Now, we show how we can perform the downdating of the QR decomposition of X_v in two steps based on the partitioning (2.1) of T. This is essential in the new downdating algorithm for the rank-revealing URV decomposition presented in the next section. In obtaining $\tilde{T}$ from T, we can first find $\tilde{R}$ such that

$$\bar{J}_1^T \begin{pmatrix} z_{v1}^T \\ \tilde{R} \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ R \\ 0 \end{pmatrix} \tag{2.3}$$

for some orthogonal matrix $\bar{J}_1$. This is a standard downdating problem that can be solved by hyperbolic transformations [2], LINPACK [14], [15], CSNE, or hybrid algorithms [6], for example. With the matrix $\bar{J}_1$ from (2.3), we have the relation

$$\bar{J}_1^T \begin{pmatrix} z_{v1}^T & z_{v2}^T \\ \tilde{R} & \tilde{F} \\ 0 & \tilde{G} \end{pmatrix} = \begin{pmatrix} 0 & h^T \\ R & F \\ 0 & \tilde{G} \end{pmatrix} \tag{2.4}$$

for some vector $h \in \mathbf{R}^{(n-r)\times 1}$. The computation of $\tilde{F}$ and h is discussed at the end of this subsection. Assuming that they are known, we have only to find $\tilde{G}$, such that

$$\bar{J}_2^T \begin{pmatrix} h^T \\ F \\ \tilde{G} \end{pmatrix} = \begin{pmatrix} 0 \\ F \\ G \end{pmatrix} \tag{2.5}$$

for some orthogonal matrix $\bar{J}_2$. Combining (2.4) and (2.5), we can complete the downdating of the QR decomposition since

$$\bar{J}_2^T \bar{J}_1^T \begin{pmatrix} z_{v1}^T & z_{v2}^T \\ \tilde{R} & \tilde{F} \\ 0 & \tilde{G} \end{pmatrix} = \bar{J}_2^T \begin{pmatrix} 0 & h^T \\ R & F \\ 0 & \tilde{G} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ R & F \\ 0 & G \end{pmatrix}. \tag{2.6}$$

The two-step procedure is summarized as follows.

Two-step procedure for downdating a 2×2 block QR decomposition. Given the upper triangular factor $\left(\begin{smallmatrix} R & F \\ 0 & G \end{smallmatrix}\right)$ in the QR decomposition of a matrix, whose first row is $z_v^T = \left(\begin{smallmatrix} z_{v1}^T & z_{v2}^T \end{smallmatrix}\right)$, find the downdated triangular factor $\left(\begin{smallmatrix} \tilde{R} & \tilde{F} \\ 0 & \tilde{G} \end{smallmatrix}\right)$ after deleting z_v^T.

I. Downdate the first part.

(a) Find $\tilde{R}$ such that

$$J_1^T \begin{pmatrix} z_{v1}^T \\ \tilde{R} \end{pmatrix} = \begin{pmatrix} 0 \\ R \end{pmatrix}$$

for some orthogonal matrix J_1.

(b) Determine $\tilde{F}$ and h such that

$$J_1^T \begin{pmatrix} z_{v2}^T \\ \tilde{F} \end{pmatrix} = \begin{pmatrix} h^T \\ F \end{pmatrix}.$$

II. Downdate the second part.

Find $\tilde{G}$ such that

$$J_2^T \begin{pmatrix} h^T \\ \tilde{G} \end{pmatrix} = \begin{pmatrix} 0 \\ G \end{pmatrix}$$

for some orthogonal matrix J_2.

We now discuss Step I(b). After Step I(a), we know $\tilde{R}$ and an orthogonal matrix J_1, such that

$$J_1^T \begin{pmatrix} z_{v1}^T \\ \tilde{R} \end{pmatrix} = \begin{pmatrix} 0 \\ R \end{pmatrix}.$$

Normally J_1 is chosen as a product of plane rotations,

$$J_1^T = P_r \cdots P_2 P_1, \tag{2.7}$$

where P_i is a rotation in the $(1, i+1)$ plane that annihilates the ith element in z_{v1}^T. Specifically, in Step I(b), the first rotation P_1 affects the rows 1 and 2 of

$$\begin{pmatrix} z_{v2}^T \\ \tilde{F} \end{pmatrix},$$

which can be written as

$$P_1(1,2) \begin{pmatrix} z_{v2}^T \\ \tilde{f}_1^T \end{pmatrix} = \begin{pmatrix} h_1^T \\ f_1^T \end{pmatrix}, \qquad P_1(1,2) = \begin{pmatrix} c_1 & s_1 \\ -s_1 & c_1 \end{pmatrix}, \tag{2.8}$$

where $c_1 = \cos(\theta_1)$ and $s_1 = \sin(\theta_1)$ for some angle θ_1, the vectors f_1^T and $\tilde{f}_1^T$ denote the first rows of F and $\tilde{F}$, respectively, and $\tilde{f}_1^T$ and h_1^T are unknown. We use the notation $P_1(i,j)$ for the 2×2 matrix, whose elements are the $(i,i), (i,j), (j,i)$, and (j,j) elements of P_1. Rearranging (2.8), we obtain

$$H_1(1,2) \begin{pmatrix} h_1^T \\ \tilde{f}_1^T \end{pmatrix} = \begin{pmatrix} z_{v2}^T \\ f_1^T \end{pmatrix}, \qquad H_1(1,2) = \begin{pmatrix} 1/c_1 & s_1/c_1 \\ s_1/c_1 & 1/c_1 \end{pmatrix},$$

from which we have

$$\begin{pmatrix} h_1^T \\ \tilde{f}_1^T \end{pmatrix} = H_1(1,2)^{-1} \begin{pmatrix} z_{v2}^T \\ f_1^T \end{pmatrix} = \begin{pmatrix} 1/c_1 & -s_1/c_1 \\ -s_1/c_1 & 1/c_1 \end{pmatrix} \begin{pmatrix} z_{v2}^T \\ f_1^T \end{pmatrix},$$

where $H_1(1,2)$ can be identified as a hyperbolic rotation since $1/c_1 = 1/\cos(\theta_1) = \cosh(\eta_1)$ and $\sin(\theta_1)/\cos(\theta_1) = \sinh(\eta_1)$ for some η_1 [2]. We can continue in a similar way and compute $\tilde{F}$ and h as shown in the following algorithm.

ALGORITHM HYPER
Given $F \in \mathbf{R}^{r\times(n-r)}$, $z_{v2} \in \mathbf{R}^{(n-r)\times 1}$, and rotation parameters in vectors $cc \in \mathbf{R}^{r\times 1}$ and $ss \in \mathbf{R}^{r\times 1}$ from downdating transformations for R, compute $\tilde{F}$ and h.

1. Determine elementary transformations

$$H_i^{-1}(1, i+1) = \begin{pmatrix} 1/cc(i) & -ss(i)/cc(i) \\ -ss(i)/cc(i) & 1/cc(i) \end{pmatrix}.$$

2. Compute

$$\begin{pmatrix} h^T \\ \tilde{F} \end{pmatrix} = H_r^{-1} \cdots H_1^{-1} \begin{pmatrix} z_{v2}^T \\ F \end{pmatrix}.$$

Algorithm HYPER requires about $4r(n-r)$ flops (1 flop $\approx$ 1 multiplication and 1 addition). The matrix V does not change since the transformations are applied from the left side only.

2.2. Reduction to a simpler downdating problem. From §2.1, we know that if we have the downdated QR decomposition of X_v after deleting z_v^T,

$$\tilde{X}_v = \tilde{U} \begin{pmatrix} \tilde{T} \\ 0 \end{pmatrix},$$

where $\tilde{U}$ is orthogonal and $\tilde{T}$ is upper triangular, then we have the downdated two-sided decomposition of X after deleting z^T, since

$$\tilde{X} = \tilde{U} \begin{pmatrix} \tilde{T} \\ 0 \end{pmatrix} V^T.$$

Similarly, if we have a two-sided orthogonal decomposition for $\tilde{X}_v$,

$$\tilde{X}_v = \hat{U} \begin{pmatrix} \hat{T} \\ 0 \end{pmatrix} V_0^T,$$

for some orthogonal matrices $\hat{U}$ and V_0, and a triangular matrix $\hat{T}$, then

$$\tilde{X} = \hat{U} \begin{pmatrix} \hat{T} \\ 0 \end{pmatrix} \hat{V}^T, \qquad \hat{V} = VV_0$$

gives a downdated two-sided orthogonal decomposition of $\tilde{X}$.

In the two-sided orthogonal decomposition, we can transform the downdating problem into a simpler one by reducing z_{v1} to κe_r, where $\kappa = ||z_{v1}||_2$, by a procedure analogous to one in [17]. This can be done without changing the "large-small" structure of the upper triangular matrix, using a sequence of Givens rotations. We can find matrices $V_0' = Y_1 \cdots Y_{r-1}$ and $U_0' = W_1 \cdots W_{r-1}$, where $Y_i \in \mathbf{R}^{r\times r}$ and

$W_i \in \mathbf{R}^{r\times r}$ are rotation matrices, such that $z_{v1}^T V_0' = \kappa e_r^T$, and $R^{(1)} = {U_0'}^T R V_0'$ is upper triangular. Computation of $R^{(1)}$ requires $O(r^2)$ operations.

Defining

$$V^{(1)} = V \begin{pmatrix} V_0' & 0 \\ 0 & I_{n-r} \end{pmatrix}, \qquad U^{(1)} = U \begin{pmatrix} U_0' & 0 \\ 0 & I_{p-r} \end{pmatrix},$$

and $F^{(1)} = {U_0'}^T F$, we now have

$$XV^{(1)} = \begin{pmatrix} \kappa e_r^T & z_{v2}^T \\ \tilde{X}V_1^{(1)} & \tilde{X}V_2^{(1)} \end{pmatrix} = U^{(1)} \begin{pmatrix} R^{(1)} & F^{(1)} \\ 0 & G \\ 0 & 0 \end{pmatrix}, \tag{2.9}$$

where $V^{(1)} = (V_1^{(1)}\, V_2^{(1)})$ and $V_1^{(1)} \in \mathbf{R}^{n\times r}$, and (2.9) is another two-sided orthogonal decomposition of the matrix X. Accordingly, (2.9) gives a QR decomposition of the matrix $XV^{(1)}$ for which the first row to be downdated has a simpler form than that in (2.2). Furthermore, the norm of each submatrix of T has been maintained, i.e., $||R||_F = ||R^{(1)}||_F$ and $||F||_F = ||F^{(1)}||_F$.

In Step I(a) of the two-step procedure of §2, we now want to find an upper triangular $\tilde{R}$, such that

$$J_1^T \begin{pmatrix} \kappa e_r^T \\ \tilde{R} \end{pmatrix} = \begin{pmatrix} 0 \\ R^{(1)} \end{pmatrix}, \tag{2.10}$$

for some orthogonal matrix J_1. Writing the matrix $R^{(1)}$ in the form

$$R^{(1)} = \begin{pmatrix} R_1 & s \\ 0 & \rho \end{pmatrix},$$

where $R_1 \in \mathbf{R}^{(r-1)\times(r-1)}$, and $s \in \mathbf{R}^{(r-1)\times 1}$, we see that the downdated matrix $\tilde{R}$ differs from $R^{(1)}$ only in the (r,r) element, and it is given by

$$\tilde{R} = \begin{pmatrix} R_1 & s \\ 0 & \tilde{\rho} \end{pmatrix}, \qquad \tilde{\rho} = \sqrt{\rho^2 - \kappa^2}. \tag{2.11}$$

Now Step I(a) in the two-step procedure consists of computing $\tilde{\rho}$ as in (2.11). Since the row to be downdated has already been transformed to κe_r^T, we only need to multiply by one hyperbolic rotation in Step I(b) to obtain h^T and the last row of $\tilde{F}^{(1)}$.

This procedure can be used also when the downdating fails due to $\rho^2 - \kappa^2 < 0$ in floating point arithmetic. In this case we can simply put $\tilde{\rho}$ equal to zero. In the next section we also use it for downdating G in the two-step algorithm. We summarize the algorithm as follows.

ALGORITHM REDUCTION

Given $R \in \mathbf{R}^{r\times r}$ and $v \in \mathbf{R}^{r\times 1}$, this algorithm computes the downdated matrix $\tilde{R} \in \mathbf{R}^{r\times r}$ by using two-sided transformations on R to reduce the vector v into a simpler form keeping the triangular structure of R.

1. Determine plane rotations Y_i in the plane $(i, i+1)$, $1 \le i \le r-1$, such that

$$\kappa e_r^T = v^T Y_1 \cdots Y_{r-1},$$

where $\kappa = ||v||_2$.

2. Determine plane rotations W_i in the plane $(i, i+1)$, $1 \le i \le r-1$, such that

$$\tilde{R} := W_{r-1}^T \cdots W_1^T R Y_1 \cdots Y_{r-1}$$

is upper triangular.
Compute $\tilde{\alpha} := \tilde{R}(r,r)^2 - \kappa^2$
If $\tilde{\alpha} > 0$ then
    $\tilde{R}(r,r) = \sqrt{\tilde{\alpha}}$
else
    $\tilde{R}(r,r) = 0$
end if
Save the information for Y_i and W_i if necessary.

This algorithm requires $4r^2$ flops. When Algorithm REDUCTION is used in Step I(a) of the two-step downdating, the matrix F must also be modified with W_i's ($4r(n-r)$ flops). Also the matrix V has to be modified due to the right side transformations $Y_1 \cdots Y_{r-1}$ ($4nr$ flops). Altogether, this adds up to $8nr$ flops. Similar operation counts are obtained when this algorithm is used in Step II. As a result of reducing the vector v into κe_r^T, the actual downdating occurs only on the last column of the new triangular factor. The information from the plane rotations are saved if necessary.

3. Downdating the rank-revealing URV decomposition. We now extend the results from the previous section and consider downdating the rank-revealing URV decomposition of X

$$X = U \begin{pmatrix} R & F \\ 0 & G \\ 0 & 0 \end{pmatrix} V^T, \tag{3.1}$$

where $R \in \mathbf{R}^{r \times r}$ and G are upper triangular, and the matrices F and G are assumed to be small in the sense that they satisfy

$$\nu = \sqrt{||F||_F^2 + ||G||_F^2} \le \text{tol},$$

for a given value tol. Thus X has numerical rank r.

We have seen in Section 2 that downdating the QR decomposition of XV is closely related to downdating the two-sided orthogonal decomposition of X. In this section, we incorporate the aspects of numerical rank decisions, in particular. The problem of downdating the RR URVD (3.1) can be formulated as follows.

PROBLEM. Downdating of RR URVD. Given $\left(\begin{smallmatrix} R & F \\ 0 & G \end{smallmatrix}\right)$, V, and $r = \text{rank}(R)$ in the RR URVD of

$$X = \begin{pmatrix} z^T \\ \tilde{X} \end{pmatrix} = U \begin{pmatrix} R & F \\ 0 & G \\ 0 & 0 \end{pmatrix} V^T,$$

find the downdated matrices $\bar{V}$, $\bar{R}$, $\bar{F}$, and $\bar{G}$, and $\bar{r} = \text{rank}(\bar{R})$ such that

$$\tilde{X} = \bar{U} \begin{pmatrix} \bar{R} & \bar{F} \\ 0 & \bar{G} \\ 0 & 0 \end{pmatrix} \bar{V}^T$$

for some orthogonal matrix $\bar{U}$.

When the numerical rank of $\bar{T}$ is reduced by one as a result of downdating, then $\bar{r}$ is $r-1$, and otherwise it is equal to r. Thus, after downdating, $\bar{T}$ not only is upper triangular but also reveals the new rank, i.e.,

$$\bar{\nu} = \sqrt{||\bar{F}||_F^2 + ||\bar{G}||_F^2} \leq \text{tol}.$$

Note the difference in notations from those in §2: $\tilde{}$ is used to denote the downdated matrices before the rank decision is made and $\bar{}$ is used to denote those after the rank decision is made on downdated matrices. Thus the order of $\bar{R}$ is one less than that of $\tilde{R}$ when the rank of R is reduced by one as a result of downdating.

3.1. Algorithms for two-step downdating. The downdating algorithm based on hyperbolic transformations [2], the LINPACK [14], [15], hybrid, and CSNE [6] downdating algorithms can be used for the two-step downdating. In fact, when hyperbolic transformations are used, it is not necessary to divide the downdating of X_v into two steps, as the same procedure can be applied throughout. Below we summarize the LINPACK and LINPACK/CSNE hybrid algorithms. For more details, see [6], [14], [15].

When the LINPACK algorithm is used for Step I(a) in the two-step downdating, we first solve the triangular system $R^T q_1 = z_{v1}$ and compute the plane rotations $P_1, \ldots, P_r$ so that

$$P_1^T \cdots P_r^T \begin{pmatrix} \gamma \\ q_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{3.2}$$

where $\gamma = \sqrt{1 - ||q_1||^2}$, which also gives

$$P_1^T \cdots P_r^T \begin{pmatrix} 0 \\ R \end{pmatrix} = \begin{pmatrix} z_{v1}^T \\ \tilde{R} \end{pmatrix}. \tag{3.3}$$

In (3.2), each P_i^T is a rotation in the $(1, i+1)$ plane and (c_i, s_i) from P_i can be saved to obtain H_i to perform Step I(b).

Equation (3.3) and the discussion of Step I(b) in §2.1 show how the LINPACK algorithm and the algorithm based on hyperbolic transformations are related: equivalent transformations are applied, in the first case as Givens rotations, and in the second as hyperbolic transformations (cf. also [11]). The LINPACK algorithm is summarized as follows.

ALGORITHM LINPACK

Given $R \in \mathbf{R}^{r \times r}$ and $v \in \mathbf{R}^{r \times 1}$, this algorithm computes the downdated matrix $\tilde{R}$ and saves the rotation parameters from downdating transformation in $cc \in \mathbf{R}^{r \times 1}$ and $ss \in \mathbf{R}^{r \times 1}$.

1. Compute q_1 and γ from

$$R^T q_1 = v, \qquad \gamma := (1 - ||q_1||^2)^{1/2}.$$

2. Determine plane rotations P_i in the $(1, i+1)$ plane, $1 \leq i \leq r$, such that

$$\begin{pmatrix} 1 & v^T \\ 0 & \tilde{R} \end{pmatrix} := P_1^T \cdots P_r^T \begin{pmatrix} \gamma & 0 \\ q_1 & R \end{pmatrix}$$

and save

$$cc(i) = c_i, \qquad ss(i) = s_i$$

where (c_i, s_i) is from $P_i(1, i+1) = \begin{pmatrix} c_i & s_i \\ -s_i & c_i \end{pmatrix}$, if necessary.

The above algorithm requires about $2.5r^2$ flops. There is no change in the matrix V in the URVD.

The downdating problem in Step I(a) in the two-step procedure is ill conditioned in two cases: (a) *when a row containing an outlier is removed*, and (b) *when there is a decrease in rank* (cf. [15], [6]). The ill conditioning reveals itself in that the quantity $1 - ||q_1||_2^2$ computed in the LINPACK algorithm is small or even negative in floating point arithmetic. In [6] it was shown that the method of corrected seminormal equations (CSNE) can be used to recover q_1 and γ more accurately in ill-conditioned cases. Accordingly, the P_i's in (3.2) are also determined more accurately. The accurate results from the CSNE method are obtained due to the fact that the original data matrix is used in the refinement of q_1 and γ. The hybrid method [6] that applies the refinement only when necessary is an excellent compromise when both computational complexity and accuracy are concerns.

We refer to the LINPACK, hybrid, and CSNE algorithms as LINPACK-type algorithms since the ways the downdating transformations are computed in these methods are similar. The essential difference among LINPACK-type algorithms is in the computation of q_1 and γ. After q_1 and γ are determined, the rest of the algorithm is the same. Thus, each P_i is computed to satisfy (3.2) and the Step I(b) can be performed using H_i's obtained from P_i's. Note that in the first step of downdating, these LINPACK-type algorithms will generally do well in computing q_1 since R is relatively well conditioned according to the definition of the RR URVD. The hybrid LINPACK/CSNE algorithm is summarized below.

ALGORITHM LINPACK/CSNE
Given $X \in \mathbf{R}^{p\times n}$, the first r columns of V, which is $V_1 \in \mathbf{R}^{n\times r}$, $R \in \mathbf{R}^{r\times r}$, $v \in \mathbf{R}^{r\times 1}$, and a tolerance tol1, this algorithm computes the downdated matrix $\tilde{R}$ and saves the rotation parameters from downdating transformation in $cc \in \mathbf{R}^{r\times 1}$ and $ss \in \mathbf{R}^{r\times 1}$.

1. Compute q_1 and α from

$$R^T q_1 = v, \qquad \alpha := 1 - ||q_1||_2^2.$$

2. If $\alpha >$ tol1 then (LINPACK)
   $\gamma := \sqrt{\alpha}$
   else (CSNE)
   Compute y and t from

$$Ry = q_1, \qquad t := e_1 - XV_1 y.$$

   Improve q_1, y and compute γ

$$R^T \delta q_1 = V_1^T X^T t, \qquad q_1 := q_1 + \delta q_1,$$
$$R\delta y = \delta q_1, \qquad t := t - XV_1 \delta y, \qquad \gamma := \|t\|_2.$$

   end if

3. Determine plane rotations P_i in the $(1, i+1)$ plane, $1 \le i \le r$ such that

$$\begin{pmatrix} 1 & v^T \\ 0 & \tilde{R} \end{pmatrix} := P_1^T \cdots P_r^T \begin{pmatrix} \gamma & 0 \\ q_1 & R \end{pmatrix},$$

and save

$$cc(i) = c_i, \qquad ss(i) = s_i,$$

where (c_i, s_i) is from $P_i(1, i+1) = \begin{pmatrix} c_i & s_i \\ -s_i & c_i \end{pmatrix}$, if necessary.

This algorithm requires $3np + 3nr + 4r^2$ flops if the CSNE branch is taken and $2.5r^2$ flops otherwise. There is no change in V in the URVD. Another alternative for Step I(a) downdating is the reduction algorithm described in §2.2.

When G is ill conditioned or numerically singular, then downdating in Step II is difficult and the LINPACK-type algorithms suffer since the downdating transformation relies on the solution of the triangular system where G is the coefficient matrix. The downdating method based on hyperbolic transformations has problems similar to the LINPACK algorithm. If we use the reduction algorithm in Step II, then we first reduce the vector h^T obtained from Step I(b) into τe_{n-r}^T where $\tau = ||h||_2$ while restoring the triangular structure of G as shown in §2.2. Even when G is ill conditioned or singular, the reduction step does not break down.

3.2. Deflation. In downdating, the numerical rank cannot increase, but can decrease. Therefore, we need to test $\tilde{R}$ obtained from Step I(a) for numerical rank deficiency and deflate it, if necessary. Rank deficiency of a matrix can be detected by estimating its smallest singular value [8, p. 128], [9]. In downdating, there are two cases to consider. After downdating and before deciding the new rank, the new value for ν is

$$\tilde{\nu}^2 = \nu^2 - ||z_{v2}||_2^2, \tag{3.4}$$

which is from

$$\nu^2 = ||F||_F^2 + ||G||_F^2 = ||\tilde{F}||_F^2 + ||\tilde{G}||_F^2 + ||z_{v2}||_2^2 = \tilde{\nu}^2 + ||z_{v2}||_2^2.$$

Suppose

$$\tilde{R} = \begin{pmatrix} \tilde{R}_1 & \tilde{s} \\ 0 & \tilde{\rho} \end{pmatrix}, \tag{3.5}$$

where $\tilde{R}_1 \in \mathbf{R}^{(r-1)\times(r-1)}$. If

$$\sqrt{\tilde{\nu}^2 + ||\tilde{s}||^2 + \tilde{\rho}^2} \le \text{tol}, \tag{3.6}$$

then the numerical rank is decreased as a result of downdating and $\bar{R} = \tilde{R}_1 \in \mathbf{R}^{(r-1)\times(r-1)}$.

If (3.6) is not satisfied, then the rank may decrease or stay the same. To check this, the smallest singular value of $\tilde{R}$ is estimated and the deflation step [17] mentioned in the introduction is applied to transform $\tilde{R}$ to

$$W^T \tilde{R} Y = \begin{pmatrix} \tilde{R}_d & \tilde{s}_d \\ 0 & \tilde{\rho}_d \end{pmatrix}, \tag{3.7}$$

where W and Y are orthogonal, $\tilde{R}_d \in \mathbf{R}^{(r-1)\times(r-1)}$ is upper triangular, and $\hat{\sigma}_d = \sqrt{\tilde{\rho}_d^2 + ||\tilde{s}_d||_2^2}$ is a good approximation of the smallest singular value of $\tilde{R}$. If, after transforming $\tilde{R}$ as in (3.7),

$$\sqrt{\tilde{\nu}^2 + \hat{\sigma}_d^2} \leq \text{tol},$$

then the numerical rank of $\tilde{R}$ is $r-1$, and we can modify $\tilde{F}$ and $\tilde{G}$ accordingly. In this case, the matrices F and V must be updated by W and Y, respectively. Otherwise, the rank after the downdating is the same as before, so $\bar{r} = r$, and $\bar{R} = \tilde{R}$.

We outline the deflation procedure as follows.

ALGORITHM DEFLATE
Given the downdated matrices $\tilde{R} \in \mathbf{R}^{r\times r}$, $\tilde{F} \in \mathbf{R}^{r\times(n-r)}$, $\tilde{G} \in \mathbf{R}^{(n-r)\times(n-r)}$, and the orthogonal $\tilde{V} \in \mathbf{R}^{n\times n}$, test for deflation and perform deflation if needed.

1. Compute $\tilde{\nu}^2 = \nu^2 - ||z_{v2}||^2$.

2. $\bar{R} = \tilde{R}$, $\bar{F} = \tilde{F}$, $\bar{G} = \tilde{G}$, $\bar{V} = \tilde{V}$.
   If $\sqrt{\tilde{\nu}^2 + ||\bar{R}(:,r)||_2^2} \leq$ tol then
   $\bar{r} := r-1$, $\bar{R} = \bar{R}(1:r-1, 1:r-1)$, $\bar{\nu} := \sqrt{\tilde{\nu}^2 + ||\bar{R}(:,r)||_2^2}$,
   change dimensions of $\bar{F}$ and $\bar{G}$ accordingly.
   else
   Estimate the smallest singular value, $\hat{\sigma}$ of $\bar{R}$.
   If $\sqrt{\tilde{\nu}^2 + \hat{\sigma}^2} <$ tol,
   Transform $\bar{R}$ as in (3.7), $\bar{R} := W^T\bar{R}Y$, and modify $\bar{F}$ and $\bar{V}$.
   $\bar{r} := r-1$, $\bar{R} = \bar{R}(1:r-1, 1:r-1)$, $\bar{\nu} := \sqrt{\tilde{\nu}^2 + \hat{\sigma}^2}$.
   Modify $\bar{F}$ and $\bar{G}$ accordingly.
   end if
   end if

The estimation of the smallest singular value requires αr^2 flops, where α is a small constant [9]. The transformation (3.7) can be done in $8nr$ flops (including the modification of F and V).

3.3. New algorithms. Incorporating the rank decision with the downdating algorithms described in the previous sections, we present three RR URVD downdating algorithms. Various algorithms can be constructed by combining the algorithms presented for Step I(a) and Step II downdating. Our goal is to obtain accurate solutions with less computational work. The first obvious choice is applying the LINPACK algorithm for both Steps I and II. However, LINPACK can fail in Step II (in fact, it can fail also in Step I(a), see the numerical tests in §4). Therefore, if the downdating in Step II is very ill conditioned, or, if it breaks down completely (this can be seen in the LINPACK algorithm in that the quantity, $1 - ||q_1||_2^2$, becomes negative, which, in theory, is positive), then we use REDUCTION algorithm, in which we simply set the last diagonal element of $\tilde{G}$ equal to zero when the downdating breaks down.

Another choice is to use the REDUCTION algorithm in Steps I(a) and II. In this case, only one hyperbolic rotation is needed in Step I(b).

We want our algorithm to be robust when the downdating is ill conditioned in Step I(a) or Step II and we want it to be fast at the same time. The LINPACK algorithm is considerably less expensive than the CSNE algorithm, while the latter has much

better accuracy than the former [6]. Thus we can use the hybrid LINPACK/CSNE algorithm in Step I(a). In Step II, we prefer the LINPACK algorithm in case the downdating is well conditioned, and the REDUCTION algorithm otherwise, since the latter is less expensive than the CSNE algorithm, and can also deal with the case when G is exactly singular. This is the third algorithm.

The three algorithms are summarized in Table 1 and the total computational complexity is shown in Table 2. In combining the different algorithms described above into a two-step procedure and in computing complexity, the following rules should be applied.

1. If R is modified from the left, then the same transformation must be applied to F.
2. If G is modified from the right, then the same transformation must be applied to F.
3. If any part is modified from the right, then V must be modified accordingly.

TABLE 3.1
Three algorithms.

Step	Algorithm A	Algorithm B	Algorithm C
I(a)	LINPACK	REDUCTION	LINPACK/CSNE
I(b)	HYPER	HYPER	HYPER
II	LINPACK/REDUCTION	REDUCTION	LINPACK/REDUCTION
III	DEFLATE	DEFLATE	DEFLATE

TABLE 3.2
Computational complexity (flops).

Step	Algorithm A	Algorithm B	Algorithm C
I(a)	$2.5r^2$	$8nr$	$2.5r^2$ or $3np + 3nr + 4r^2$
I(b)	$4r(n-r)$	$4(n-r)$	$4r(n-r)$
II	$2.5(n-r)^2$ or $4(n-r)(2n-r)$	$4(n-r)(2n-r)$	$2.5(n-r)^2$ or $4(n-r)(2n-r)$
III	αr^2 $(+8nr)$	αr^2 $(+8nr)$	αr^2 $(+8nr)$

4. Numerical experiments. Numerical tests have been performed in Pro-MATLAB with IEEE double precision floating point arithmetic to compare the accuracy of the RR URVD downdating algorithms that have been presented. In a sliding window method, the data matrix consists of the p latest rows of an observation matrix. In each step, a new row of observations is updated into the RR URVD and an existing row of the data matrix is downdated from the decomposition on a first in-first out basis. The QR decomposition or the SVD of the window matrix was used as a reference in checking the accuracy.

In the plots where signal and noise space errors are shown, the solid, dashed, and dashed-dotted lines denote results from Algorithms A, B, and C, respectively. Errors are plotted against the step number in the sliding window method.

The error in the signal part was taken to be the 2-norm of the difference between the recursively computed matrix R and the R factor from the QR decomposition of the window matrix multiplied by the first r columns of V, where r is the computed rank from the corresponding algorithm.

The basis of the numerical null-space given by the last $n - r$ columns of V was compared to the null-space obtained from the SVD of the window matrix. The largest principal angle between these two null space bases was computed as $\arccos(\nu_{\min})$,

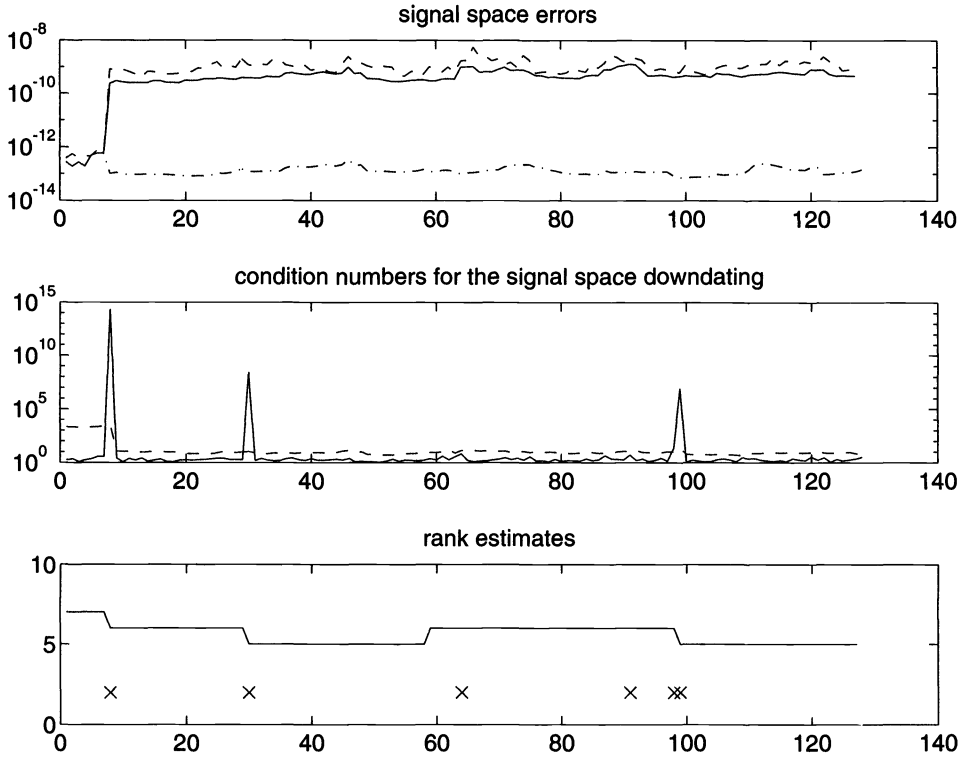


FIG. 1. *Test* I. *Signal space errors, downdating conditioning, and rank estimates for* 12×8 *window matrices.*

where $\nu_{\min}$ is the smallest singular value of $V_0^T Z_0$, and V_0 and Z_0 are the orthonormal null space bases from the computed V and the SVD, respectively, see [5] and [8, p. 585]. From [10], the imprecision in the null-space due to the nonzero block F in the RR URVD is essentially $||F||_F^2/\sigma_r(R)$. This level of uncertainty depends on the details of the deflation procedure and thus it may vary for different algorithms.

According to the perturbation analysis result of Pan [12] (see also [15], [7]), the downdating condition number is $\kappa_{\text{down}} = \kappa^2(R)/(1 - ||q_1||_2^2)$, where $\kappa(R) = ||R||_2 \, ||R^{-1}||_2$ is the condition number of the matrix to be downdated. In each plot where the condition numbers for downdating are shown, the solid line and the dashed line represents $1/(1-||q_1||_2^2)$ and $\kappa(R)$, respectively. When $1-||q_1||_2 < 0$ numerically, then $1/(1-||q_1||_2^2)$ is set to be 10^{20} in the plot. In Algorithm C, Step I(a), the CSNE branch was taken when $\kappa(R) > 10^4$ or $1/(1 - ||q_1||_2^2) > 5$. In Algorithms A and C, Step II, the Linpack branch was taken if $\kappa(G) < 6 \cdot 10^6$, and $1/(1 - ||q_1||_2^2) \leq 5$. It is possible to estimate the condition numbers of the triangular matrices after updating/downdating in $O(n)$ operations with the adaptive methods such as ACE [13].

In the graph where rank estimates are presented, the $\times$ signs indicate where the CSNE branch was used in Algorithm C and $+$ signs indicate where the downdating in the REDUCTION algorithm in Step I(a) was made by setting the diagonal element equal to zero.

Test I. A random matrix $A \in \mathbf{R}^{140 \times 8}$ was constructed with elements taken from a uniform distribution in $(0, 1)$. To vary the numerical ranks of the window matrices,

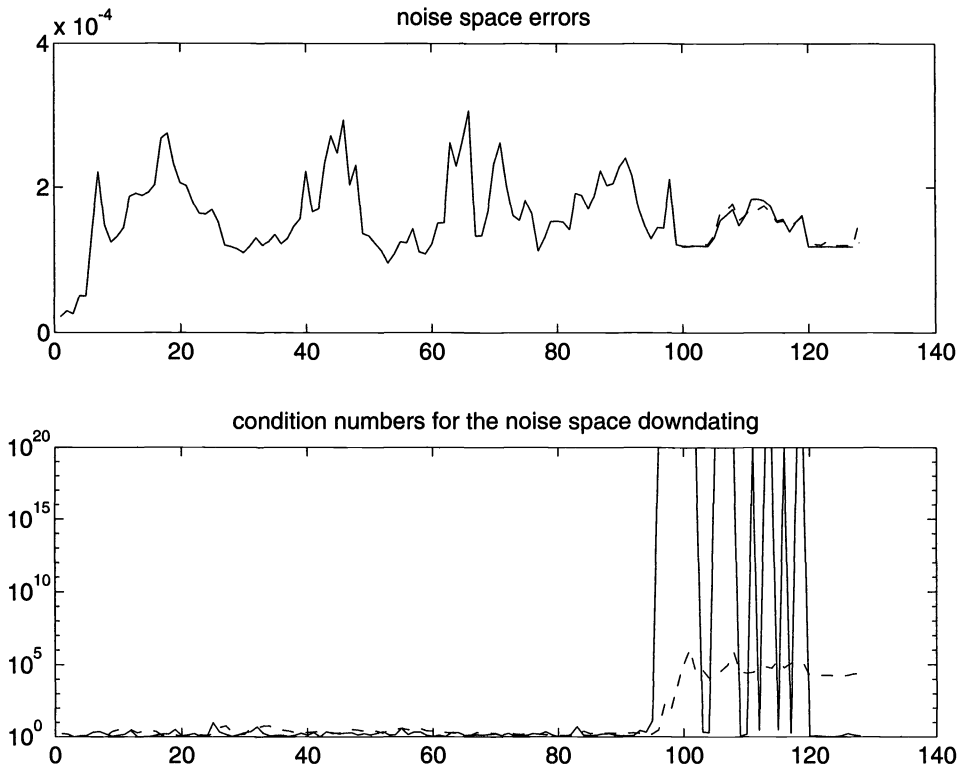


FIG. 2. *Test* I. *Noise space error and downdating conditioning.*

certain parts of the matrix were multiplied by $\delta = 10^{-4}$. Then the matrix was multiplied by a random, orthogonal, 8×8 matrix from the right. Finally, an outlier equal to 10^3 was added in position (8,4). The window size p was 12. The tolerance value for determining the numerical rank was 10δ. The results are shown in Figs. 1 and 2.

Algorithms A and B lost accuracy significantly in signal space estimation after the outlier was deleted. Then the accuracy was not recovered even after the downdating problems became well conditioned. This phenomenon is consistent with what was observed in the LINPACK algorithm for QR decomposition downdating [6].

All three algorithms gave the correct estimate of numerical ranks. Note that the first rank reduction happened when the outlier was removed. The noise space error was below the level of uncertainty given by the data for all three algorithms. They produced almost identical results. In noise space downdating, the numerical value for $1 - ||q_1||_2^2$ was often negative toward the last steps.

Test II. A random matrix $A \in \mathbf{R}^{74\times 9}$ was constructed with ill-conditioned diagonal blocks and zero blocks. Then the matrix was multiplied by random orthogonal matrices from the left and right sides. The window size p was 13. The tolerance value for determining the numerical rank was 30δ. The results are shown in Figs. 3 and 4.

Algorithm A failed when κ_{down} became negative in Step I(a). Algorithms B and C produced comparable results although Algorithm C was consistently better. The

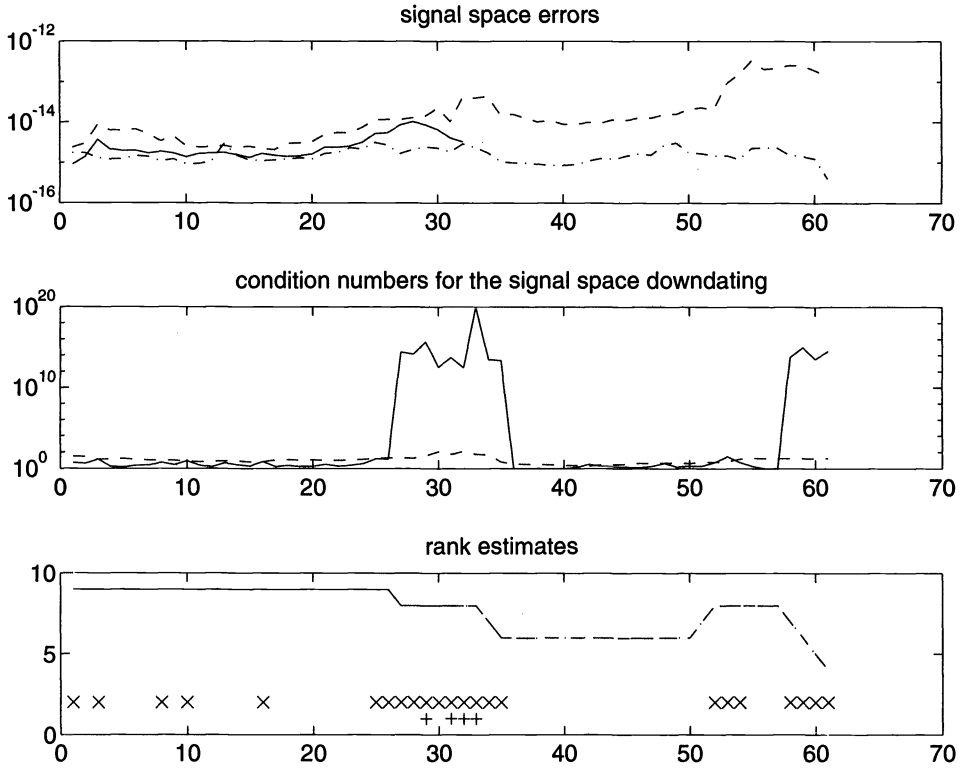


FIG. 3. *Test* II. *Signal space errors, downdating conditioning, and rank estimates for* 13 × 9 *window matrices.*

error in the noise subspace was again of the magnitude that one would expect from the noise level in the data and the tolerance. For a sequence of steps, the plot of the noise error shows an error equal to zero since the window matrix had full column rank, and thus there was no nontrivial null space.

All three algorithms gave the correct estimate of numerical ranks. Note that the CSNE branch was heavily used when the rank kept decreasing in Algorithm C.

We ran several other tests. The results reported here are typical. However, there were a few cases with singular G when the performance of the three algorithms was less satisfactory. For such problems, we developed a method that is based on the observation that G is the R factor of the QR decomposition of the residual matrix for the least squares problem $\min_W \|XV_1W - XV_2\|_F$, where $V = (V_1\ V_2)$, $V_1 \in \mathbf{R}^{n\times r}$. The solution of this problem is obtained from the linear system $RW = F$ and can be refined using CSNE. Preliminary tests have shown promising results, and, although this method is considerably more costly than Algorithm REDUCTION, it may be an alternative for handling extreme cases with almost singular G. Also Algorithms A and B seem to be more sensitive to the choice of the tolerance value for the rank decision. The actual applications should provide helpful information for the tolerances. Further work is needed to make the algorithms more robust.

5. Conclusion. In this paper we introduced a two-step procedure for downdating the rank-revealing URV decomposition, where the downdatings of the signal and

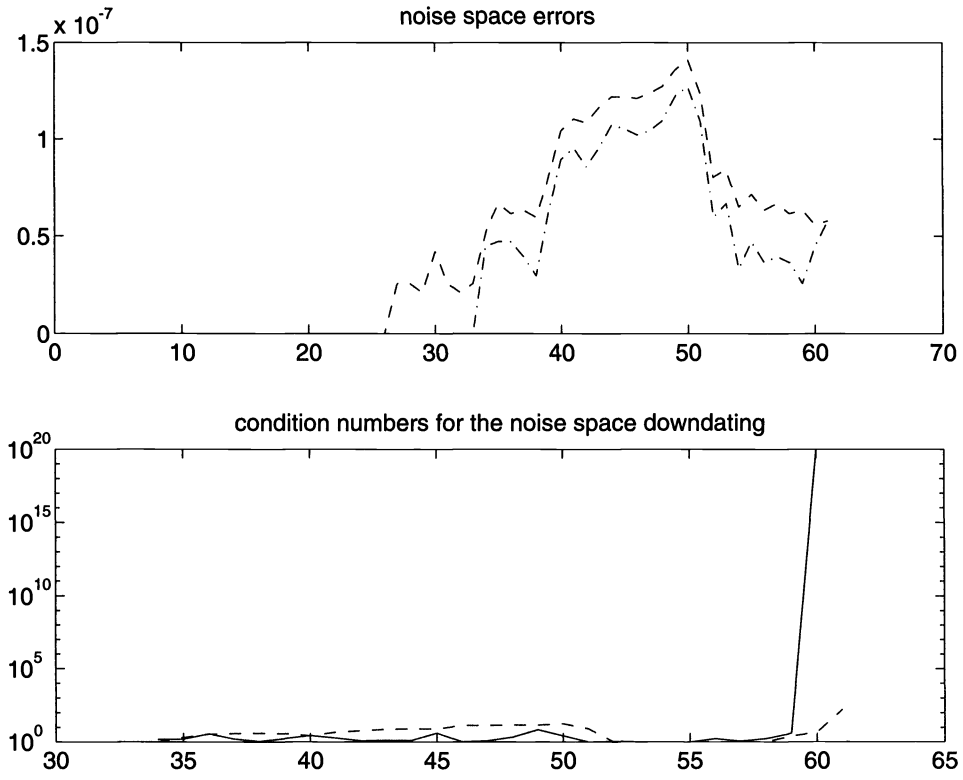


FIG. 4. *Test* II. *Noise space error and downdating conditioning.*

noise parts are performed separately. This enables us to obtain accurate results using the LINPACK/CSNE hybrid algorithm for the ill-conditioned downdates that occur when the numerical rank is decreased.

Three different algorithms based on the two-step procedure have been described. The numerical tests indicate that it is necessary to use the more sophisticated variants to handle ill-conditioned downdates.

In signal processing applications, where signals are turned on and off (thus the numerical ranks of the window matrices vary), the sliding window method was reported to be highly sensitive to round-off errors [3], and therefore the exponential windowing method is preferred. However, our preliminary tests indicate that the new algorithms, and Algorithm C in particular, can give accurate results for the ill-conditioned downdates that occur when rank decreases. We expect that the new algorithms can be used successfully for such applications. Further research is needed to investigate their applicability to real problems in signal processing and to study the choice of the tolerance for rank decision (see [16]).

REFERENCES

[1] G. ADAMS, M. F. GRIFFIN, AND G. W. STEWART, *Direction-of-arrival estimation using the rank-revealing* URV *decomposition*, Tech. Report CS-TR-2640, Dept. of Computer Science,

University of Maryland, College Park, March 1991.

[2] 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.

[3] M. G. BELLANGER, *The family of fast least squares algorithms for adaptive filtering*, in Mathematics in Signal Processing, J. McWhirter, ed., Clarendon Press, Oxford, 1990, pp. 415–434.

[4] C. H. BISCHOF AND G. M. SCHROFF, *On updating signal subspaces*, IEEE Trans. Signal Proc., 40 (1992), pp. 96–105.

[5] A. BJÖRCK AND G. H. GOLUB, *Numerical methods for computing angles between subspaces*, Math. Comp., 27 (1973), pp. 579–594.

[6] Å. BJÖRCK, H. PARK, AND L. ELDÉN, *Accurate downdating of least squares solutions*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 549–568.

[7] L. ELDÉN AND H. PARK, *Perturbation analysis for block downdating of a Cholesky decomposition*, Numer. Math., to appear, 1994.

[8] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

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

[10] R. MATHIAS AND G. W. STEWART, *A block QR algorithm and the singular value decomposition*, Linear Algebra Appl., 182 (1993), pp. 91–100.

[11] C.-T. PAN, *A modification to the Linpack downdating algorithm*, BIT, 30 (1990), pp. 707–722.

[12] ———, *A perturbation analysis of the problem of downdating a Cholesky factorization*, Linear Algebra Appl., 183 (1993), pp. 103–116.

[13] D. J. PIERCE AND R. J. PLEMMONS, *Fast adaptive condition estimation*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 274–291.

[14] M. A. SAUNDERS, *Large-scale linear programming using the Cholesky factorization*, Tech. Report CS252, Computer Science Department, Stanford University, Stanford, CA, 1972.

[15] G. W. STEWART, *The effects of rounding error on an algorithm for downdating a Cholesky factorization*, J. Inst. Math. Appl., 23 (1979), pp. 203–213.

[16] ———, *Determining rank in the presence of error*, Tech. Report CS-TR-2972, Dept. of Computer Science, University of Maryland, College Park, 1992.

[17] ———, *An updating algorithm for subspace tracking*, IEEE Trans. Signal Proc., 40 (1992), pp. 1535–1541.

[18] ———, *Updating a rank-revealing* ULV *decomposition*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 494–499.

[19] G. XU, H. ZHA, G. H. GOLUB, AND T. KAILATH, *Fast and robust algorithms for updating signal subspaces*, Tech. Report, Information Systems Laboratory, Stanford University, Stanford, CA, 1991.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 156–171, January 1995

011

PRECONDITIONED KRYLOV SUBSPACE METHODS FOR LYAPUNOV MATRIX EQUATIONS*

MARLIS HOCHBRUCK† AND GERHARD STARKE‡

Abstract. The authors study the iterative solution of Lyapunov matrix equations

$$AX + XA^T = -D^T D$$

by preconditioned Krylov subspace methods. These solution techniques are of interest for problems leading to large and sparse matrices A as those arising from certain applications in large space structure control theory. We show how conjugate gradient (CG)-type methods for nonsymmetric linear systems can be applied to this type of equation utilizing the special structure when computing matrix-vector and inner products. In contrast to recently developed methods for such matrix equations based on Krylov subspaces associated with A, the authors implicitly work with the equivalent system of linear equations involving the Kronecker sum $M = A \otimes I_N + I_N \otimes A$. Motivation for this new approach comes from the observation that it allows the straightforward incorporation of preconditioners. Several preconditioners for such problems are presented and analyzed. In particular, since the solution matrix X is known to be symmetric, it is of interest to know which of the methods produces symmetric iterates in each step. It is proven that this is the case for alternating direction implicit (ADI)-type and (point) symmetric successive overrelaxation (SSOR) preconditioning in association with the quasiminimal residual (QMR) method. As numerical results show, it is essential to use preconditioning in association with Krylov subspace methods. Several preconditioners for such problems are presented and analyzed. Finally numerical examples are presented where the different preconditioners are compared.

Key words. Lyapunov matrix equations, iterative methods, Krylov subspace methods, QMR, preconditioning, ADI preconditioning, SSOR preconditioning, non-Hermitian matrices

AMS subject classifications. 65F10, 65N22, 93B40

1. Introduction.

In the past ten years the interest in *iterative methods* for the solution of *Lyapunov matrix equations*

$$AX + XA^T = -D^T D \tag{1}$$

has been growing substantially. This is due to the fact that applications have been found that lead to such matrix equations where A is large and sparse. For example, for constructing a near-optimal reduced-order model for a dynamical system

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ y &= Cx \end{aligned}$$

with state $x \in \mathbb{R}^N$, input $u \in \mathbb{R}^p$ and output $y \in \mathbb{R}^q$ ($A \in \mathbb{R}^{N\times N}, B \in \mathbb{R}^{N\times p}, C \in \mathbb{R}^{N\times q}$), one must solve the Lyapunov equations

$$\begin{aligned} AX + XA^T &= -BB^T, \\ A^T Y + YA &= -C^T C \end{aligned}$$

(cf. Moore [23], Hodel [15]). In large space structure control theory, it is often the case that the system may be described as a continuum by a system of partial differential

* Received by the editors October 19,1992; accepted for publication (in revised form) by R. Freund, November 4, 1993.

† Mathematisches Institut, Universität Tübingen, 72076 Tübingen, Germany (marlis@na.-uni-tuebingen.de).

‡ Institut für Praktische Mathematik, Universität Karlsruhe, Englerstrasse 2, 76128 Karlsruhe, Germany (starke@ipmsun1.mathematik.uni-karlsruhe.de).

equations (cf. Balas [1]). Discretizing this partial differential operator (in the space variables) using finite elements or finite differences leads to a matrix A, which is large and sparse. Another application of Lyapunov equations comes from the use of Newton's method for Riccati matrix equations arising in optimal control problems (cf. Mehrmann [22, §11], Hodel [15]).

The Lyapunov matrix equation has a unique solution if and only if $\sigma(A) \cap \sigma(-A) = \emptyset$ (see [17, Cor. 4.4.7]). Moreover, for symmetric right-hand side, as in (1), this solution is also symmetric (see [17, Cor. 4.4.10]).

The Lyapunov matrix equation is a special case of the more general Sylvester matrix equation $AX - XB = C$. There are also applications of Sylvester equations with $B \neq -A^T$ in control theory (Datta and Datta [5] and references therein). Although most of the theory presented below can be easily generalized to Sylvester matrix equations, in this paper we restrict ourselves to the Lyapunov case, i.e., $B = -A^T$. For the general case, we refer the reader to [14, Chap. 6].

Direct methods for the solution of Lyapunov matrix equations, such as those proposed by Bartels and Stewart [3] and Hammarling [13], are attractive if the matrix A is of moderate size. Parallel versions of Hammarling's algorithm and an iterative procedure designed for larger order Lyapunov matrix equations are studied by Hodel and Poolla [16]. However, since these direct methods are based on the Schur decomposition of A, their complexity is $O(N^3)$, which restricts their use to problems of relatively small size. One of the most popular iterative methods for solving (1) was introduced more than twenty years ago by Smith [27] and Barnett and Storey [2]. In [34], Wachspress showed that Smith's method can be regarded as the natural application of the well-known ADI method to such matrix equations (see, also, Lu and Wachspress [21]). If we collect the coefficients of the unknown matrix $X \in \mathbb{R}^{N \times N}$ column by column in the vector

$$\mathbf{x} = (x_{11}, \ldots, x_{N1}, \ldots, x_{1N}, \ldots, x_{NN})^T,$$

then we can rewrite the Lyapunov matrix equation as a linear system of equations with the coefficient matrix

$$M := A \otimes I_N + I_N \otimes A \in \mathbb{R}^{N^2 \times N^2}, \tag{2}$$

where $\otimes$ denotes the standard *Kronecker product* (cf. Horn and Johnson [17, p. 243]).

One possibility for deriving iterative methods for the solution of (1) is to take any of the well-known iterative schemes for the solution of the large system (2) with coefficient matrix $M \in \mathbb{R}^{N^2 \times N^2}$ and reformulate it in terms of (1). In this manner, the ADI method with respect to the splitting of the linear system into $A \otimes I_N$ and $I_N \otimes A$ leads to Smith's method

$$\begin{aligned} (A - \varphi_j I_N)\tilde{X}_{j-1} &= -[X_{j-1}(A + \varphi_j I_N)^T + D^T D], \\ X_j(A - \varphi_j I_N)^T &= -[(A + \varphi_j I_N)\tilde{X}_{j-1} + D^T D] \end{aligned} \tag{3}$$

with real parameters $\varphi_j, j = 1, 2, \ldots$. In what follows, we refer to this method as the ADI method (for Lyapunov matrix equations). The interpretation of the block (or line) successive overrelaxation (SOR) method as an iterative technique for solving Sylvester matrix equations was studied by Starke and Niethammer in [31], a combination of the alternating direction idea with SOR in [29]. It should be noted that, in contrast to SOR, ADI is always (in theory) convergent if all the eigenvalues of A are contained in the left half-plane, an assumption that is always fulfilled in our applications since it follows from the prerequisite that the underlying dynamical system is

stable. To be precise, we can always find parameters $\varphi_j > 0$ (it is actually sufficient to use the same parameter φ throughout the iteration) such that the scheme (3) is convergent.

All of the iterative techniques mentioned above have in common that the solution of the Lyapunov matrix equation (1) is reduced to an iterative series of sets of systems of linear equations with the matrix A, which can be solved in parallel independently for each column or each row of X, respectively. It is, of course, also possible to apply the CG method or, if A is nonsymmetric, any of the CG-like methods such as Bi-CG, generalized minimal residual (GMRES), and QMR (see Freund, Golub, and Nachtigal [9] for an overview of Krylov subspace methods for linear systems) to the equivalent large linear system (2) and utilize the special structure of the Kronecker sum $M = A \otimes I_N + I_N \otimes A$ when computing matrix-vector and inner products. Different, more sophisticated, types of Krylov subspace methods for Lyapunov matrix equations have recently been proposed by Saad [25], Hu and Reichel [18], and Jaimoukha and Kasenally [19]. Their methods are based on Krylov subspaces associated with the matrix A itself rather than M. We exclude the latter class of methods from our study since it is not clear if and how preconditioning can be incorporated into these algorithms (see §3).

Our purpose in this paper is to study Krylov subspace methods based on the Kronecker sum formulation $M = A \otimes I_N + I_N \otimes A$ and to present and analyze several preconditioners for this approach. Our numerical results in §5 show that it is *essential* when using Krylov subspace methods for solving Lyapunov matrix equations to use *preconditioning*.

Section 2 gives a brief review of the Lanczos process and the QMR method in terms of Lyapunov matrix equations and discusses implementational issues that arise in this connection. In §3 we deal with preconditioning these methods and present several approaches to this problem. Since the solution $X \in \mathbb{R}^{N \times N}$ of (1) is a symmetric matrix, it is of particular interest to have symmetric iterates throughout the iteration. We will prove that this is the case for the QMR method without preconditioning and when using (point) SSOR or ADI-type preconditioners.

The use of ADI (Smith's method) as a preconditioner for Lyapunov matrix equations is the topic of §4. In [4], Chin, Manteuffel, and de Pillis used the stationary ADI method, i.e., $\varphi_j = \varphi, j = 1, 2, \ldots$, to precondition the Chebyshev iteration applied to discretized elliptic boundary value problems of convection-diffusion type. Translated into the language of Lyapunov matrix equations, this approach takes the form

$$\begin{aligned}(A - \varphi I_N)^{-1} A X (A - \varphi I_N)^{-T} + (A - \varphi I_N)^{-1} X A^T (A - \varphi I_N)^{-T} \\ = -(A - \varphi I_N)^{-1} D^T D (A - \varphi I_N)^{-T}\end{aligned} \tag{4}$$

(corresponding to left preconditioning for the corresponding linear system) or

$$\begin{aligned}A(A - \varphi I_N)^{-1} Y (A - \varphi I_N)^{-T} + (A - \varphi I_N)^{-1} Y (A - \varphi I_N)^{-T} A^T = -D^T D, \\ (A - \varphi I_N)^{-1} Y (A - \varphi I_N)^{-T} = X\end{aligned} \tag{5}$$

(corresponding to right preconditioning). Since the two operations "multiply a matrix from the left with A" and "multiply a matrix from the right with A^T" obviously commute, the two types of preconditioning (4) and (5) are mathematically equivalent here. Of course, the need for an appropriate choice of the parameter φ (or parameter sets for higher order ADI preconditioning) implies that some information about the location of the eigenvalues of the matrix A must be known. However, since the matrix A is relatively small compared to the size of the problem, it pays to compute (or at

least estimate) its eigenvalues and compute optimal parameters with respect to this information.

We also include a brief review of the approximation problem associated with the optimal choice of the ADI parameters as it was developed in [33] for the symmetric and in [28] for the nonsymmetric case. ADI preconditioning for elliptic boundary value problems is also studied in [30]. Finally, §5 contains a collection of numerical experiments carried out for dynamical systems modeled by partial differential equations.

2. Krylov subspace methods for matrix equations. Krylov subspace methods, especially when combined with preconditioning, are known as powerful methods for the solution of linear systems. Krylov subspace methods for Lyapunov and Sylvester matrix equations were developed by Saad [25], Hu and Reichel [18], and Jaimoukha and Kasenally [19]. All of these methods are based on Krylov subspaces associated with the matrix A.

The main idea for our approach is to write the matrix equation (1) as a linear system with the coefficient matrix M defined in (2) and then apply a Krylov subspace method to this large linear system of order N^2. However, we do not want to work with this linear system explicitly, but instead rewrite the algorithms in terms of the original matrix equation. One advantage is that we can then implement the algorithm with BLAS 3 instead of BLAS 2 (cf., e.g., [6]) subroutines. Another important issue is due to the question: In the case of a Lyapunov equation with a symmetric right-hand side does the algorithm also compute symmetric iterates? Recall from the introduction that for $\sigma(A) \cap \sigma(-A) = \emptyset$ there exists a unique solution of (1) and that this solution is symmetric.

In this paper, we have chosen to use a slightly different approach and rewrite the Krylov subspace methods in an operator formulation instead of a Kronecker product formulation. This allows us to work completely in the space $\mathbb{R}^{N\times N}$ with the inner product in $\mathbb{R}^{N^2}$:

$$\langle X, Y\rangle := \operatorname{trace}(X^T Y). \tag{6}$$

This inner product induces the Frobenius norm, which is denoted by $\|X\|_F = \sqrt{\langle X, X\rangle}$. To this end, we define the Lyapunov operator $\mathbf{L}$ as

$$\mathbf{L}: \begin{cases} \mathbb{R}^{N\times N} & \to & \mathbb{R}^{N\times N}, \\ X & \mapsto & AX + XA^T. \end{cases} \tag{7}$$

From $\langle X, \mathbf{L}Y\rangle = \langle X, AY + YA^T\rangle = \operatorname{trace}(X^T(AY + YA^T)) = \langle A^T X + XA, Y\rangle$ it follows that

$$\mathbf{L}^\star: \begin{cases} \mathbb{R}^{N\times N} & \to & \mathbb{R}^{N\times N}, \\ X & \mapsto & A^T X + XA, \end{cases} \tag{8}$$

is the adjoint operator of $\mathbf{L}$.

To keep this paper self-contained and because we feel that this can be useful also in other applications, we now explain Krylov subspace methods for general matrix equations of the form

$$\mathbf{H}X = F, \quad \mathbf{H}: \mathbb{R}^{N\times N} \to \mathbb{R}^{N\times N}, \quad X, F \in \mathbb{R}^{N\times N} \tag{9}$$

in more detail. First, we call

$$\mathcal{K}_n(\mathbf{H}, B) = \operatorname{span}\{B, \mathbf{H}B, \ldots, \mathbf{H}^{n-1}B\} \tag{10}$$

the nth Krylov subspace with respect to the operator $\mathbf{H}$ and the matrix B. Here, and in the sequel, $\mathbf{H}^k B$ is defined recursively as $\mathbf{H}^k B := \mathbf{H}(\mathbf{H}^{k-1} B)$ for $k \geq 1$ and $\mathbf{H}^0 B := B$. We denote by $\mathbf{H}^\star$ the adjoint operator of $\mathbf{H}$ with respect to the inner product (6).

Now, what remains to be done for applying a Krylov subspace method to the matrix equation (9) is to take ones favorite Krylov subspace method (for the linear system $M\mathbf{x} = \mathbf{f}$, where $\mathbf{f}$ is the vector containing the columns of the right-hand side matrix F) and replace each matrix-vector multiplication $M\mathbf{x}$ by $\mathbf{H}X$, i.e., by applying the operator $\mathbf{H}$ to the matrix X. If the method additionally requires the transpose M^T, one needs to replace $M^T\mathbf{x}$ by $\mathbf{H}^\star X$, i.e., by applying the adjoint operator $\mathbf{H}^\star$ to the matrix X. The inner products are as defined in (6).

For example, for the extension of the classical Lanczos algorithm [20] to matrix equations, it is easily verified that it constructs two sequences of matrices $V_1, V_2, \ldots, V_{n+1}$ and $W_1, W_2, \ldots, W_{n+1}$ that span the nth Krylov subspace with respect to the operator $\mathbf{H}$ and R_0 and the adjoint operator $\mathbf{H}^\star$ and S_0, respectively:

$$\begin{aligned} \operatorname{span}\{V_1, \ldots, V_n\} &= \mathcal{K}_n(\mathbf{H}, R_0), \\ \operatorname{span}\{W_1, \ldots, W_n\} &= \mathcal{K}_n(\mathbf{H}^\star, S_0). \end{aligned}$$

The so-called left and right Lanczos matrices W_n and V_n fulfill the biorthogonality condition

$$\langle W_i, V_j \rangle = 0 \quad \text{for } i \neq j, \qquad i, j \leq L. \tag{11}$$

Here, L denotes the termination index of the classical Lanczos algorithm. Moreover, the following three-term recurrence relations hold true:

$$\mathbf{H}V_j = \gamma_{j+1} V_{j+1} + \alpha_j V_j + \beta_j V_{j-1} \quad \text{for } j = 1, 2, \ldots, n. \tag{12}$$

The classical Lanczos algorithm breaks down prematurely whenever it terminates with $L < \min\{\dim \mathcal{K}_n(\mathbf{H}, R_0),\ \dim \mathcal{K}_n(\mathbf{H}^\star, S_0)\}$. However, except in very special cases, the breakdowns can be cured using look-ahead [10]. We do not want to present the details here since, in our examples, no look-ahead steps have been necessary. We refer the reader to [14] for a look-ahead Lanczos algorithm for general Sylvester matrix equations.

The nth iterate of a Krylov subspace method is of the form

$$X_n \in X_0 + \mathcal{K}_n(\mathbf{H}, R_0), \tag{13}$$

where $R_0 = F - \mathbf{H}X_0$ is the initial residual of (9). Thus, we have

$$X_n = X_0 + [\, V_1 \ \cdots \ V_n \,] \, (z_n \otimes I_N). \tag{14}$$

The expression $[\, V_1 \ \cdots \ V_n \,] \, (z_n \otimes I_N)$ simply represents a linear combination of the matrices $V_1, \ldots, V_n$.

Since our main interest in this paper is in preconditioning, from now on we concentrate on a particular scheme, namely, the QMR algorithm proposed by Freund and Nachtigal [11]. However, the ideas presented for this method can clearly be applied to any other Krylov subspace method. For simplicity, we consider only a version of QMR without look-ahead. Let us write the recurrence coefficients from (12) into a tridiagonal matrix $T_n^{(e)} = \operatorname{tridiag}(\gamma_j, \alpha_j, \beta_j) \in \mathbb{R}^{(n+1)\times n}$, where γ_j, α_j, and β_j are the

entries of the jth row of $T_n^{(e)}$. Then, for the QMR algorithm, the coefficient vector z_n in (14) is determined by means of a quasi-minimization property

$$\left\| \omega_1 \|R_0\|_F \, e_1 - \Omega_{n+1} T_n^{(e)} z_n \right\| = \min_{z \in \mathbb{R}^n} \left\| \omega_1 \|R_0\|_F \, e_1 - \Omega_{n+1} T_n^{(e)} z \right\|. \tag{15}$$

Here, $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^{n+1}$, $e_1 = [1\ 0\ \cdots\ 0]^T \in \mathbb{R}^{n+1}$, and $\Omega_{n+1} = \operatorname{diag}(\omega_1, \ldots, \omega_{n+1})$ is a diagonal scaling matrix. The natural choice for the weights is $\omega_j = \|V_j\|_F$. In this case, all the matrices in the representation (14) are treated equally. We always assume the weights to be chosen in this fashion.

Note, that the minimization problem (15) is exactly the same as for the QMR algorithm for linear systems. Thus we refer to [11] for a detailed description of the solution of this problem.

Recently, Freund and Nachtigal [12] developed a different implementation of the QMR algorithm which is based on coupled two-term recurrences that seem to be more stable, instead of the three-term recurrences used in the Lanczos algorithm.

For the consideration of the symmetry of QMR iterates, we have the following theorem.

THEOREM 2.1. *If X_0, F, and W_1 are symmetric $N \times N$ matrices, then the* QMR *algorithm computes symmetric iterates $X_n, n = 1, 2, \ldots$, if for any symmetric matrix X holds*

$$\mathbf{H}X = (\mathbf{H}X)^T \quad \text{and} \quad \mathbf{H}^\star X = (\mathbf{H}^\star X)^T. \tag{16}$$

Proof. Since X_0 is symmetric and because of (16), the initial residual R_0, and hence V_1, is symmetric. Going through the algorithm, it is easily verified by induction, that all matrices occurring are symmetric for $n = 1, 2, \ldots$. □

COROLLARY 2.2. *If X_0 and W_1 are symmetric, then the* QMR *algorithm applied to the Lyapunov matrix equation* (1) *computes symmetric iterates X_n.*

Proof. Using Theorem 2.1, it only remains to show that (16) holds for the Lyapunov operator $\mathbf{L}$ and its adjoint $\mathbf{L}^\star$. Let $X = X^T \in \mathbb{R}^{N \times N}$ be arbitrary. Then

$$\begin{aligned} (\mathbf{L}X)^T &= (AX + XA^T)^T = XA^T + AX = \mathbf{L}X, \\ (\mathbf{L}^\star X)^T &= (A^T X + XA)^T = XA + A^T X = \mathbf{L}^\star X. \end{aligned}$$ □

3. Preconditioning Lyapunov matrix equations. When looking for preconditioners for Lyapunov matrix equations, one of the first things that comes to mind is to multiply (1) by nonsingular matrices $Q_1 \in \mathbb{R}^{N \times N}$ (from the left) and $Q_2 \in \mathbb{R}^{N \times N}$ (from the right). This leads to the equivalent formulation of (1),

$$(Q_1 A Q_1^{-1}) Q_1 X Q_2 + Q_1 X Q_2 (Q_2^{-1} A^T Q_2) = -Q_1 D^T D Q_2. \tag{17}$$

This is now a Sylvester matrix equation in $\tilde{X} = Q_1 X Q_2$:

$$\tilde{A}\tilde{X} + \tilde{X}\tilde{B} = -Q_1 D^T D Q_2,$$

where $\tilde{A} = Q_1 A Q_1^{-1}$ and $\tilde{B} = Q_2^{-1} A^T Q_2$. To make this a Lyapunov matrix equation we must choose $Q_2 = Q_1^T$. This leads to

$$\tilde{A}\tilde{X} + \tilde{X}\tilde{A}^T = -\tilde{D}^T \tilde{D} \tag{18}$$

with $\tilde{X} = Q_1 X Q_1^T, \tilde{A} = Q_1 A Q_1^T$, and $\tilde{D} = D Q_1^T$. However, since the matrices A and $\tilde{A}$ have the same eigenvalues, this is also true for the corresponding Lyapunov

operators $\mathbf{L}$ and $\tilde{\mathbf{L}}$ (cf. (7)). This means that, at least with respect to eigenvalues, (18) cannot be easier to solve than the original equation (1). Another explanation why this approach is not useful, is as follows.

Associated with the Lyapunov operator $\mathbf{L}$ and $\tilde{\mathbf{L}}$ are the Krylov subspaces $\mathcal{K}_n(\mathbf{L}, R_0)$ and $\mathcal{K}_n(\tilde{\mathbf{L}}, \tilde{R}_0)$. If we denote by R_0 the initial residual of the original equation (1) and by $\tilde{R}_0$ the initial residual of the preconditioned equation (18), then, from $\tilde{X}_n = Q_1 X_n Q_1^T$, we obtain

$$\tilde{R}_0 = -\tilde{D}^T \tilde{D} - \tilde{\mathbf{L}}\tilde{X}_0 = Q_1 R_0 Q_1^T. \tag{19}$$

LEMMA 3.1. *The Krylov subspaces $\mathcal{K}_n(\mathbf{L}, R_0)$ and $\mathcal{K}_n(\tilde{\mathbf{L}}, \tilde{R}_0)$ fulfill*

$$\mathcal{K}_n(\mathbf{L}, R_0) = Q_1^{-1} \mathcal{K}_n(\tilde{\mathbf{L}}, \tilde{R}_0) Q_1^{-T}.$$

Proof. From the definition (10) of the Krylov subspaces $\mathcal{K}_n(\mathbf{L}, R_0)$ and $\mathcal{K}_n(\tilde{\mathbf{L}}, \tilde{R}_0)$, we conclude that it is sufficient to show that

$$Q_1(\mathbf{L}^k R_0) Q_1^T = \tilde{\mathbf{L}}^k \tilde{R}_0$$

holds for $k = 0, 1, \ldots, n-1$. For $k = 0$ this is obviously true. Then, by induction, we get

$$\begin{aligned} Q_1(\mathbf{L}^{k+1} R_0) Q_1^T &= Q_1(A(\mathbf{L}^k R_0) + (\mathbf{L}^k R_0) A^T) Q_1^T \\ &= Q_1(A Q_1^{-1}(\tilde{\mathbf{L}}^k \tilde{R}_0) Q_1^{-T} + Q_1^{-1}(\tilde{\mathbf{L}}^k \tilde{R}_0) Q_1^{-T} A^T) Q_1^T \\ &= \tilde{A}(\tilde{\mathbf{L}}^k \tilde{R}_0) + (\tilde{\mathbf{L}}^k \tilde{R}_0)\tilde{A}^T \\ &= \mathbf{L}^{k+1} \tilde{R}_0. \qquad \square \end{aligned}$$

The iterates of a Krylov subspace method, applied to the original equation, are contained in $X_0 + \mathcal{K}_n(\mathbf{L}, R_0)$, and those of the preconditioned system in $\tilde{X}_0 + \mathcal{K}_n(\tilde{\mathbf{L}}, \tilde{R}_0)$. Now Lemma 3.1 tells us that if we transform the iterates $\tilde{X}_n$ of the preconditioned system back by $X_n = Q_1 \tilde{X}_n Q_1^T$, then we find ourselves in the Krylov subspace associated with the original equation. Thus, we do not change the Krylov subspace by this preconditioning strategy. Of course, this does not mean that the iterates are the same in both cases. More precisely, the QMR iterates are contained in the same Krylov subspace but constructed by quasiminimization in different norms. However, this can generally not be expected to lead to a faster convergent iteration.

We would like to remark at this point that a result similar to Lemma 3.1 can be proved along the same lines for the more general Sylvester matrix equations $AX - XB = C$ if we do not restrict ourselves to $Q_2 = Q_1^T$ in the above similarity transformation. The result is the same: the iterates are still contained in the same Krylov subspace (cf. [14, §6.6]).

The first specific example of a preconditioner for Lyapunov matrix equations that we want to consider here is the well-known SSOR preconditioning of the corresponding system (2). If we split the matrix A according to $A = D_A - L_A - U_A$ into its diagonal, (strictly) lower triangular and (strictly) upper triangular part, then the corresponding decomposition of $M = A \otimes I_N + I_N \otimes A$ is given by

$$\begin{aligned} M &= D_M - L_M - U_M \\ &= (D_A \otimes I_N + I_N \otimes D_A) - (L_A \otimes I_N + I_N \otimes L_A) \\ &\quad -(U_A \otimes I_N + I_N \otimes U_A). \end{aligned} \tag{20}$$

The SSOR preconditioning matrix is given by

$$S_{\mathrm{SSOR}} = \frac{1}{\omega(2-\omega)}(D_M - \omega L_M)D_M^{-1}(D_M - \omega U_M)$$

(see Young [35, Chap. 15]). In the course of evaluating the SSOR preconditioner, we must solve the linear systems

$$\begin{aligned}(D_M - \omega L_M)\mathbf{y} = \mathbf{x} &\iff (D_A - \omega L_A)Y + Y(D_A - \omega L_A^T) = X,\\ (D_M - \omega U_M)\mathbf{y} = \mathbf{x} &\iff (D_A - \omega U_A)Y + Y(D_A - \omega U_A^T) = X.\end{aligned} \tag{21}$$

Of course, it goes without saying that $\mathbf{x}$ and $\mathbf{y}$ again denote the vectors created by columnwise storing of X and Y, respectively. In analogy to the SOR method for linear systems, the equations in (21) can be solved without inverting matrices. For the first equation, the matrix Y can be computed columnwise from left to right and from top to bottom in each column. The second equation can be solved in exactly the opposite order. For the Lanczos process one also needs to evaluate the transpose operator. This leads to

$$\begin{aligned}(D_M - \omega L_M)^T\mathbf{y} = \mathbf{x} &\iff (D_A - \omega L_A^T)Y + Y(D_A - \omega L_A) = X,\\ (D_M - \omega U_M)^T\mathbf{y} = \mathbf{x} &\iff (D_A - \omega U_A^T)Y + Y(D_A - \omega U_A) = X,\end{aligned} \tag{22}$$

which can obviously be solved in a similar fashion.

If, for any $\alpha \in \mathbb{R}$, the matrix $A + \alpha I_N$ can be easily inverted (in the sense of solving linear systems with this matrix), we can use the block SSOR method instead of the point SSOR method as preconditioner. The associated preconditioning matrix is

$$S_{\mathrm{SSOR}}^b = \frac{1}{\omega(2-\omega)}(D_M^b - \omega L_M^b)(D_M^b)^{-1}(D_M^b - \omega U_M^b),$$

where

$$\begin{aligned}D_M^b &:= I_N \otimes A + D_A \otimes I_N,\\ L_M^b &:= L_A \otimes I_N, .\\ U_M^b &:= U_A \otimes I_N.\end{aligned}$$

In this case we must solve the linear systems as follows:

$$\begin{aligned}(D_M^b - \omega L_M^b)\mathbf{y} = \mathbf{x} &\iff AY + Y(D_A - \omega L_A^T) = X;\\ (D_M^b - \omega U_M^b)\mathbf{y} = \mathbf{x} &\iff AY + Y(D_A - \omega U_A^T) = X;\\ (D_M^b - \omega L_M^b)^T\mathbf{y} = \mathbf{x} &\iff A^TY + Y(D_A - \omega L_A) = X;\\ (D_M^b - \omega U_M^b)^T\mathbf{y} = \mathbf{x} &\iff A^TY + Y(D_A - \omega U_A) = X.\end{aligned} \tag{23}$$

To do this, we compute the columns of Y from left to right for the first and the last equation and from right to left for the second and third equation. Now, for each column of Y, we must solve a linear system with the coefficient matrix $A + a_{ii}I_N, i = 1, \ldots, N$.

We previously showed in Theorem 2.1 that if we start with symmetric matrices, then the QMR method produces symmetric iterates. Is the same true if we use SSOR preconditioning? If we look back at Theorem 2.1, we see that the basic ingredient was

the fact that (16) holds for any symmetric matrix X. Similarly, we must show now that the preconditioned operator maps symmetric matrices to symmetric matrices. For point SSOR with $\omega \in \mathbb{R}$ this can easily be seen: each of the equations in (21) and (22) is a Lyapunov equation with symmetric right-hand side and therefore the solution is also symmetric. We state this result in the following theorem.

THEOREM 3.2. *If we start with symmetric matrices $X_0, S_0 \in \mathbb{R}^{N\times N}$, the* QMR *method using (point)* SSOR *preconditioning* (21) *with $\omega \in \mathbb{R}$ produces symmetric iterates X_n.*

For block SSOR preconditioning, the corresponding result is generally not true since the corresponding matrix equations in (23) may have nonsymmetric solutions.

4. ADI preconditioning. In this section we are concerned with the construction of effective preconditioners based on the ADI splitting of (1). Let us first consider the stationary case $l = 1$ as introduced in (5). (We restrict ourselves to right preconditioning here.)

For the ADI method, it is crucial to achieve good convergence results that the parameter φ is properly chosen. We now study this problem for ADI preconditioning. When using one of the CG-like methods for nonsymmetric systems (like Bi-CG, GMRES, or QMR), the goal is to get the preconditioned problem "as close as possible" to the identity operator. This is motivated by the fact that convergence bounds for QMR methods involve the quantity

$$\inf\{\max_{\lambda\in\Lambda} |p_m(\lambda)| : p_m \in \Pi_m, p_m(0) = 1\}, \tag{24}$$

where Λ denotes the spectrum of the underlying (preconditioned) operator (see Saad and Schultz [26] for GMRES and Freund and Nachtigal [11] for QMR). With this, our goal is to choose the preconditioner in such a way that the quantity in (24) is minimized. Since it is usually too difficult to achieve this task, the easiest method is to get the spectrum of the preconditioned operator into a disk Ω, excluding the origin, such that

$$\inf\{\max_{z\in\Omega} |p_m(z)| : p_m \in \Pi_m, p_m(0) = 1\} \tag{25}$$

becomes as small as possible. For

$$\Omega = \{z \in \mathbb{C} : |z - \alpha| \le \rho\}$$

with $0 < \rho < |\alpha|$, the polynomial for which (25) attains its minimum, is obviously given by $p_m(z) = (1 - z/\alpha)^m$ and, therefore,

$$\min\{\max_{z\in\Omega} |p_m(z)| : p_m \in \Pi_m, p_m(0) = 1\} = (\rho/\alpha)^m .$$

Note that this quantity only depends on the ratio ρ/α so that without loss of generality we can set $\alpha = 1$.

Specifically, for ADI preconditioning, our aim is to choose the parameter φ in such a way that the spectrum of the operator $\mathbf{S}_1$ defined by

$$\mathbf{S}_1 : \begin{cases} \mathbb{R}^{N\times N} & \to & \mathbb{R}^{N\times N}, \\ X & \mapsto & -2\varphi[A(A-\varphi I_N)^{-1}X(A-\varphi I_N)^{-T} \\ & & \quad + (A-\varphi I_N)^{-1}X(A-\varphi I_N)^{-T}A^T] \end{cases} \tag{26}$$

is contained in a small disk around one. Here, we normalized $\mathbf{S}_1$ by multiplying (5) with the constant -2φ (note that this has no effect on the preconditioned iterative method) to center the spectrum of $\mathbf{S}_1$ at one. From (3), the ADI iteration operator $\mathbf{T}_1$ is given by

$$\mathbf{T}_1 : \begin{cases} \mathbb{R}^{N\times N} & \to & \mathbb{R}^{N\times N}, \\ X & \mapsto & (A+\varphi I_N)(A-\varphi I_N)^{-1}X(A+\varphi I_N)^T(A-\varphi I_N)^{-T}. \end{cases} \tag{27}$$

Comparing (26) and (27), we conclude that $\mathbf{T}_1 = \mathbf{I} - \mathbf{S}_1$ where $\mathbf{I}$ denotes the identity operator on $\mathbb{R}^{N\times N}$. Note that this is similar to the situation for systems of linear equations $M\mathbf{x} = \mathbf{b}$. There, if we have a preconditioned linear system with the coefficient matrix $S = G^{-1}M$, then $T = I - S = I - G^{-1}M = G^{-1}(G - M)$ is the iteration matrix of the classical iterative method associated with the splitting $M = G - (G - M)$.

A straightforward generalization of Theorem 4.2.12 in [17] shows that

$$\sigma(\mathbf{T}_1) = \{\frac{(\lambda+\varphi)(\mu+\varphi)}{(\lambda-\varphi)(\mu-\varphi)} : \lambda, \mu \in \sigma(A)\},$$

which leads to the following result:

$$\max\{|1-\tau| : \tau \in \sigma(\mathbf{S}_1)\} = \max_{\lambda\in\sigma(A)} \left|\frac{\lambda+\varphi}{\lambda-\varphi}\right|^2. \tag{28}$$

This means that the search for an optimal ADI preconditioner in the sense of (5) leads to the well-known ADI parameter problem of choosing φ in such a way that (28) is minimized (see [33], [28]). We will now derive a similar approach to ADI preconditioning of higher degree.

Let the polynomial $q_l(z) = (z-\varphi_1)\cdots(z-\varphi_l)$ with $\operatorname{Re}\varphi_j > 0, j = 1,\ldots,l$ be given. The corresponding ADI iteration operator $\mathbf{T}_l$ (using the parameters $\varphi_1,\ldots,\varphi_l$ in a cyclic fashion) is, in analogy to (27), defined by

$$\mathbf{T}_l : \begin{cases} \mathbb{R}^{N\times N} & \to & \mathbb{R}^{N\times N}, \\ X & \mapsto & q_l(-A)[q_l(A)]^{-1}X[q_l(A)]^{-T}q_l(-A)^T. \end{cases} \tag{29}$$

This gives rise to define the operator $\mathbf{S}_l$ as

$$\mathbf{S}_l : \begin{cases} \mathbb{R}^{N\times N} & \to & \mathbb{R}^{N\times N}, \\ X & \mapsto & X - q_l(-A)[q_l(A)]^{-1}X[q_l(A)]^{-T}q_l(-A)^T \end{cases} \tag{30}$$

and view this as preconditioned operator.

To choose the parameters $\varphi_1,\ldots,\varphi_l$ in such a way that the preconditioned operator (30) is as close as possible to the identity operator, we must minimize $\rho(\mathbf{T}_l)$. Again, this leads to the ADI parameter problem

$$\min_{q_l\in\Pi_l} \max_{\lambda\in\sigma(A)} \left|\frac{q_l(-\lambda)}{q_l(\lambda)}\right|^2 \tag{31}$$

(see [33], [28]).

Let us consider the case $l = 2$ to illustrate that (30) can indeed be interpreted as a preconditioned version of (1). With

$$q_2(z) = (z-\varphi_1)(z-\varphi_2) =: z^2 - \tau_1 z + \tau_0,$$

we obtain

$$X - q_2(-A)[q_2(A)]^{-1}X[q_2(A)]^{-T}q_2(-A)^T$$
$$= -2\tau_1[q_2(A)]^{-1}[A(AXA^T + \tau_0 X) + (AXA^T + \tau_0 X)A^T][q_2(A)]^{-T}.$$

Thus, (right) ADI preconditioning of degree 2 has the form

$$\begin{aligned} &A[q_2(A)]^{-1}(AYA^T + \tau_0 Y)[q_2(A)]^{-T} \\ &\quad +[q_2(A)]^{-1}(AYA^T + \tau_0 Y)[q_2(A)]^{-T}A^T = D^T D, \\ &X = [q_2(A)]^{-1}[AYA^T + \tau_0 Y][q_2(A)]^{-T}. \end{aligned} \tag{32}$$

In this form, evaluating the ADI preconditioner of degree 2 consists of the following four steps:

$$\begin{array}{ll} \text{compute} & \hat{Y} = AYA^T + \tau_0 Y; \\ \text{solve} & q_2(A)\tilde{Y} = \hat{Y}; \\ \text{solve} & \check{Y}q_2(A)^T = \tilde{Y}; \\ \text{compute} & Z = A\check{Y} + \check{Y}A^T. \end{array}$$

Thus, one iteration with this ADI preconditioner involves four matrix-matrix multiplications of size N (two with A from the left and two with A^T from the right) and the solution of two matrix equations of the form $q_2(A)Y = Z$. Since q_2 can have (conjugate) complex roots, it is preferable, to stay in real arithmetic, to actually form the matrix $q_2(A)$ instead of using it in factored form. From (30) we see that an alternative way for evaluating this preconditioner consists of carrying out one step of the ADI iteration and computing the difference between the two consecutive iterates. Obviously, the cost of this implementation is exactly the same as for the one based on (32). A similar observation can be made for ADI preconditioning of degree 1 using (26) and (27).

Let us now turn to the problem of choosing the parameters $\varphi_1, \ldots, \varphi_l$ to fulfill (31). With respect to compact sets containing $\sigma(A)$, this minimization problem is studied in [33] for the symmetric and in [28] for the nonsymmetric case. Explicit formulas for the optimal parameters for $l = 1$ and $l = 2$ are known for special regions, e.g., rectangles, enclosing $\sigma(A)$ (see again [28]). Since the dimension of the matrix A is small compared to the complexity of the overall problem here, we may assume that the eigenvalues of A (or good approximations to them) are known. Then, for small l, the parameter problem (31) can be solved using minimization procedures. We restrict ourselves to $l = 1$ and $l = 2$ in our computations. The corresponding parameter problems are given by

$$\min_{\varphi\in\mathbb{R}} \max_{\lambda\in\sigma(A)} \left|\frac{\lambda+\varphi}{\lambda-\varphi}\right| \quad \text{and} \quad \min_{\tau_0,\tau_1\in\mathbb{R}} \max_{\lambda\in\sigma(A)} \left|\frac{\lambda^2+\tau_1\lambda+\tau_0}{\lambda^2-\tau_1\lambda+\tau_0}\right|. \tag{33}$$

The restriction to real parameters φ and τ_0, τ_1, respectively, is justified by the fact that, for real matrices A, $\sigma(A)$ is symmetric with respect to the real axis.

For Lanczos-based methods (like QMR) each iteration also involves the transposed (preconditioned) operator. It is easy to see from the Kronecker product representation that the adjoint operator $\mathbf{S}_l^\star$ of $\mathbf{S}_l$ from (30) is given by

$$\mathbf{S}_l^\star : \quad \begin{cases} \mathbb{R}^{N\times N} & \to & \mathbb{R}^{N\times N}, \\ X & \mapsto & X - q_l(-A)^T[q_l(A)]^{-T}X[q_l(A)]^{-1}q_l(-A). \end{cases}$$

The following theorem states that, under similar conditions as before (cf. Cor. 2.2), ADI preconditioning also has the desirable property that, starting with a symmetric matrix, the QMR iterates are all symmetric.

THEOREM 4.1. *Let the polynomial q_l have real coefficients. Then, starting with symmetric matrices $X_0, S_0 \in \mathbb{R}^{N\times N}$, the* QMR *method using* ADI *preconditioning* (30) *produces symmetric iterates X_n.*

Proof. From Theorem 2.1 we know that it is sufficient to show (16), i.e., that the operators $\mathbf{S}_l$ and $\mathbf{S}_l^\star$ map symmetric matrices to symmetric matrices. Since q_l is a real polynomial we have, with $X = X^T$ and $Y := \mathbf{S}_l X$,

$$\begin{aligned} Y^T &= X^T - q_l(-A)[q_l(A)]^{-1}X^T[q_l(A)]^{-T}q_l(-A)^T, \\ &= X - q_l(-A)[q_l(A)]^{-1}X[q_l(A)]^{-T}q_l(-A)^T = Y. \end{aligned}$$

The analogous result for $\mathbf{S}_l^\star$ follows along the same lines. □

5. Computational results. Recall from the introduction that one important application area for large order Lyapunov matrix equations comes from model reduction for dynamical systems described by finite element or finite difference discretizations of elliptic differential operators. As a test problem we consider a dynamical system that is derived from the discretization of an elliptic boundary value problem of the form

$$\dot{u} = u'' + \tau(x)u' + f(x)g(t) \tag{34}$$

for $x \in (0,1)$ with $u(0) = u(1) = 0$ (cf. Saad [24], [25]). Here τ is a function that we assume to be continuous on $(0,1)$. Discretizing (34) by central differences with grid sizes $h = 1/(N+1)$ results in a tridiagonal matrix $A \in \mathbb{R}^{N\times N}$ which is itself sparse and of special block tridiagonal structure.

In our computational experiments we compared different preconditioners for the solution of such problems. We restricted ourselves to ADI and (point) SSOR preconditioning since, as we have shown in the previous sections, starting with a symmetric matrix, all the iterates are then also symmetric. As pointed out in §4, ADI preconditioning is based on the solution of linear systems with the coefficient matrix A. In the two- or three-dimensional case, the solution of these linear systems would be done by methods that exploit the sparsity and, if possible, special structure of A. It might even be attractive to solve linear systems with the discretized operator A iteratively leading to an inner-outer iteration for the overall problem.

As a basic iterative method in our experiments, we use QMR since it combines short recurrences with a weakened residual minimization property. Short recurrences are especially important for these problems since, for large enough N, we cannot afford to store too many consecutive iterates $X_n \in \mathbb{R}^{N\times N}$. On the other hand, we would also prefer a method with a provable convergence bound depending on the spectrum of the operator since we based the construction of our ADI preconditioner on that property.

In our examples, we consider

$$\dot{u} = u'' + \gamma x u' + f(x)g(t) \tag{35}$$

where $\gamma \in \mathbb{R}$ is still to be chosen. We discretized this problem in $(0,1)$ using central differences on a grid with $N = 127$. Note that the reformulation of (1) using the Kronecker sum representation (2) would already lead to a linear system of dimension 16129 here. We also emphasize once more that the computational work for direct

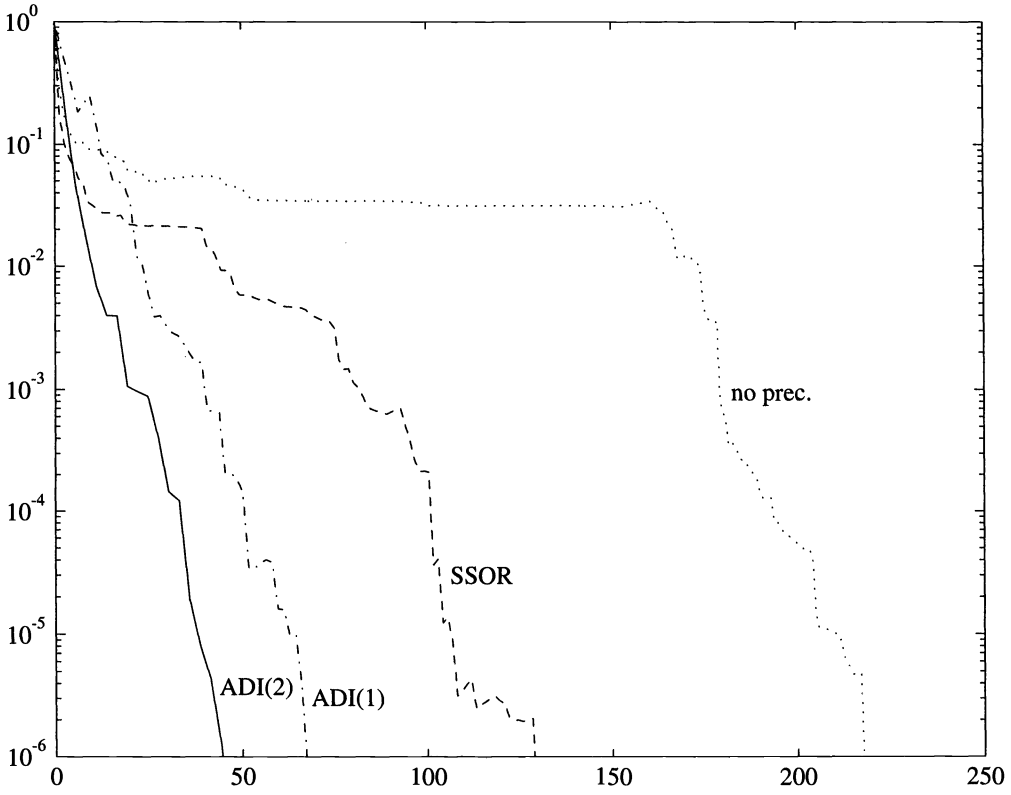


FIG. 1. QMR *convergence curves for Example* 1.

methods grows like $O(N^3)$ here. Following [25], we did our experiments for (1) with the special right-hand side $-\mathbf{b}\mathbf{b}^T$, where we chose $\mathbf{b}$ to be a random vector with entries uniformly distributed in $[-1, 1]$.

Example 1. The choice of $\gamma = 100$ leads to a Lyapunov matrix equation (1) where A is nonsymmetric but has real spectrum.

Example 2. Now we choose $\gamma = 500$ leading to a nonsymmetric A with complex spectrum.

Figures 1 and 2 show the norm reduction of the relative residual norm

$$\frac{\|R_n\|_F}{\|R_0\|_F}$$

versus the number of operations (Mflops).

Clearly, the computation of the Lyapunov operator $\mathbf{L}X = AX + XA^T$ requires two matrix-matrix multiplications of size N (where the structure of A might be utilized). For SSOR preconditioning, the computation can be arranged in such a way that the number of operations is exactly the same as for the evaluation of the Lyapunov operator (see Eisenstat [7]). However, it should be noted that SSOR preconditioning

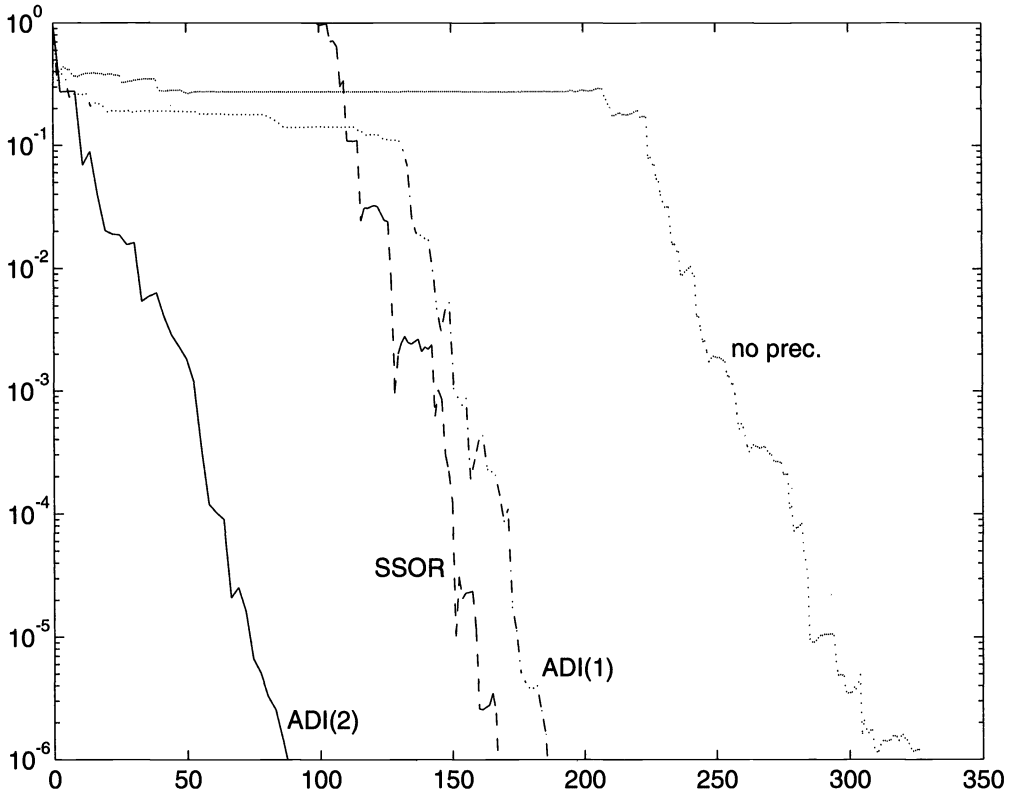


FIG. 2. QMR *convergence curves for Example* 2.

is a purely sequential task, so, in contrast to the unpreconditioned problem and also in contrast to ADI, higher level BLAS routines cannot be utilized here. It is well known that the use of Level 3 BLAS, for example, can significantly speed up computations, particularly on vector and parallel machines (see [6, §5.1]). Figures 1 and 2 show that the use of SSOR preconditioning does not lead to significantly improved convergence here.

Figures 1 and 2 show that all the preconditioners lead to significantly improved convergence in our examples. Clearly the ADI(2) preconditioner outperforms all the other methods in terms of flops. We expect the gain from using ADI preconditioning to become more dramatic in terms of computing time on parallel computers and/or when higher level BLAS is used. Finally, we would like to mention that the difference in the convergence rate between the unpreconditioned equation, ADI(1) and ADI(2) preconditioning becomes more significant as h gets smaller. From the analysis in [30], it can be shown that the number of iterations without preconditioning grows like $O(N)$, while with ADI(1) preconditioning this is reduced to $O(N^{1/2})$ and with ADI(2) preconditioning even to $O(N^{1/4})$.

Acknowledgments. The first author would like to thank Roland Freund for

bringing the application of the QMR method to matrix equations to her attention and for his helpful advice. We are also grateful to Michael Eiermann and Martin Hanke for their careful reading of this manuscript.

REFERENCES

[1] M. J. BALAS, *Trends in large space structure control theory: Fondest hopes, wildest dreams*, IEEE Trans. Automat. Control, 27 (1982), pp. 522–535.

[2] S. BARNETT AND C. STOREY, *Some application of the Lyapunov matrix equation*, J. Inst. Math. Appl., 4 (1968), pp. 33–42.

[3] R. BARTELS AND G. W. STEWART, *Algorithm* 432*: Solution of the matrix equation* $AX + XB = C$, Comm. ACM, 15 (1972), pp. 820–826.

[4] R. C. Y. CHIN, T. A. MANTEUFFEL, AND J. DE PILLIS, ADI *as a preconditioning for solving the convection-diffusion equation*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 281–299.

[5] B. N. DATTA AND K. DATTA, *Theoretical and computational aspects of some linear algebra problems in control theory*, in Computational and Combinatorial Methods in Systems Theory, C. I. Byrnes and A. Lindquist, eds., Elsevier, Amsterdam, 1986, pp. 201–212.

[6] J. J. DONGARRA, I. S. DUFF, D. C. SORENSEN, AND H. A. VAN DER VORST, *Solving Linear Systems on Vector and Shared Memory Computers*, Society for Industrial and Applied Mathematics, Philadelphia, 1991.

[7] S. C. EISENSTAT, *Efficient implementation of a class of preconditioned conjugate gradient methods*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 1–4.

[8] D. FISCHER, G. GOLUB, O. HALD, C. LEIVA, AND O. WIDLUND, *On Fourier–Toeplitz methods for separable elliptic problems*, Math. Comp., 28 (1974), pp. 349–368.

[9] R. W. FREUND, G. H. GOLUB, AND N. M. NACHTIGAL, *Iterative solution of linear systems*, *Acta Numerica*, 1 (1992), pp. 57–100.

[10] R. W. FREUND, M. H. GUTKNECHT, AND N. M. NACHTIGAL, *An implementation of the look-ahead Lanczos algorithm for non-Hermitian matrices*, SIAM J. Sci. Comput., 14 (1993), pp. 137–158.

[11] R. W. FREUND AND N. M. NACHTIGAL, QMR: *a quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.

[12] ———, *An implementation of the* QMR *method based on coupled two-term recurrences*, Tech. Report 92.15, RIACS, NASA-Ames Research Center, Moffatt Field, CA, June 1992.

[13] S. J. HAMMARLING, *Numerical solution of the stable, non-negative definite Lyapunov equation*, IMA J. Numer. Anal., 2 (1982), pp. 303–323.

[14] M. HOCHBRUCK, *Lanczos- und Krylov-Verfahren für nicht-Hermitesche lineare Systeme*, Ph.D. thesis, Universität Karlsruhe, Germany, 1992.

[15] A. S. HODEL, *Recent applications of the Lyapunov equation in control theory*, in Proc. IMACS Internat. Symp. Iterative Methods in Linear Algebra, North-Holland, Amsterdam, 1992, pp. 217–227.

[16] A. S. HODEL AND K. POOLLA, *Parallel solution of large Lyapunov equations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1189–1203.

[17] R. A. HORN AND C. R. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1991.

[18] D. Y. HU AND L. REICHEL, *Krylov subspace methods for the Sylvester equation*, Linear Algebra Appl., 172 (1992), pp. 283–313.

[19] I. M. JAIMOUKHA AND E. M. KASENALLY, *Krylov subspace methods for solving large Lyapunov equations*, SIAM J. Numer. Anal., 31 (1994), pp. 227–251.

[20] C. LANCZOS, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. Natl. Bur. Stand., 45 (1950), pp. 255–282.

[21] A. LU AND E. L. WACHSPRESS, *Solution of Lyapunov equations by* ADI *iteration*, Comput. Math. Appl., 21 (1991), pp. 43–58.

[22] V. MEHRMANN, *The Autonomous Linear Quadratic Control Problem*, Lecture Notes in Control and Information Sciences, Vol. 163, Springer-Verlag, Berlin, Heidelberg, 1991.

[23] B. C. MOORE, *Principal component analysis in linear systems: Controllability, observability and model reduction*, IEEE Trans. Automat. Control, 26 (1981), pp. 17–32.

[24] Y. SAAD, *Projection and deflation methods for partial pole assignment in linear state feedback*, IEEE Trans. Automat. Control, 33 (1988), pp. 290–297.

[25] ———, *Numerical solution of large Lyapunov equations*, in Signal Processing, Scattering, Operator Theory and Numerical Methods, M. A. Kaashoek, J. H. van Schuppen, and A. C. M. Ran, eds., Birkhäuser, Boston, 1990, pp. 503–511.

[26] Y. SAAD AND M. H. SCHULTZ, GMRES: *A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[27] R. A. SMITH, *Matrix equation* $XA + BX = C$, SIAM J. Appl. Math., 16 (1968), pp. 198–201.
[28] G. STARKE, *Optimal* ADI *parameters for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1431–1445.
[29] G. STARKE, SOR-*like methods for Lyapunov matrix equations*, in Proc. IMACS Internat. Symp. Iterative Methods in Linear Algebra, North-Holland, Amsterdam, 1992, pp. 233–240.
[30] ———, *Alternating direction preconditioning for nonsymmetric systems of linear equations*, SIAM J. Sci. Statist. Comput., 15 (1994), pp. 369–384.
[31] G. STARKE AND W. NIETHAMMER, SOR *for* $AX - XB = C$, Linear Algebra Appl., 154–156 (1991), pp. 355–375.
[32] P. N. SWARZTRAUBER, *A direct method for the discrete solution of separable elliptic equations*, SIAM J. Numer. Anal., 11 (1974), pp. 1136–1150.
[33] E. L. WACHSPRESS, *Extended application of alternating direction implicit iteration model problem theory*, J. Soc. Indust. Appl. Math., 11 (1963), pp. 994–1016.
[34] ———, *Iterative solution of the Lyapunov matrix equation*, Appl. Math. Lett., 1 (1988), pp. 87–90.
[35] D. M. YOUNG, *Iterative Solution of Large Linear Systems*, Academic Press, New York, London, 1971.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 172–191, January 1995

012

A DIVIDE-AND-CONQUER ALGORITHM FOR THE SYMMETRIC TRIDIAGONAL EIGENPROBLEM*

MING GU† AND STANLEY C. EISENSTAT‡

Abstract. The authors present a stable and efficient divide-and-conquer algorithm for computing the spectral decomposition of an $N \times N$ symmetric tridiagonal matrix. The key elements are a new, stable method for finding the spectral decomposition of a symmetric arrowhead matrix and a new implementation of deflation. Numerical results show that this algorithm is competitive with bisection with inverse iteration, Cuppen's divide-and-conquer algorithm, and the QR algorithm for solving the symmetric tridiagonal eigenproblem.

Key words. symmetric tridiagonal eigenproblem, divide-and-conquer, arrowhead matrix

AMS subject classification. 65F15

1. Introduction.

Given an $N \times N$ symmetric tridiagonal matrix

$$T = \begin{pmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{N-1} & \alpha_{N-1} & \beta_N \\ & & & \beta_N & \alpha_N \end{pmatrix},$$

the *symmetric tridiagonal eigenproblem* is to find the spectral decomposition $T = X\Lambda X^T$, where Λ is diagonal and X is orthogonal. The diagonal elements of Λ are the eigenvalues of T, and the columns of X are the corresponding eigenvectors. In this paper we propose an *arrowhead divide-and-conquer* algorithm (ADC) for solving this problem.

ADC divides[1] T into two smaller symmetric tridiagonal matrices T_1 and T_2, each of which is a principle submatrix of T. It then recursively computes the spectral decompositions of T_1 and T_2 and constructs an orthogonal matrix Q such that $T = QHQ^T$, where

$$H = \begin{pmatrix} \alpha & z^T \\ z & D \end{pmatrix},$$

with D a diagonal matrix and z a vector, is a *symmetric arrowhead matrix*. Finally it finds the eigenvalues of T by computing the spectral decomposition $H = U\Lambda U^T$, where U is an orthogonal matrix, and computes the eigenvector matrix of T as QU.

Since error is associated with computation, a *numerical spectral decomposition* of T or H is usually defined as a decomposition of the form

$$T = \hat{X}\hat{\Lambda}\hat{X}^T + O(\epsilon\,\|T\|_2) \quad \text{or} \quad H = \hat{U}\hat{\Lambda}\hat{U}^T + O(\epsilon\,\|H\|_2), \tag{1}$$

* Received by the editors December 7, 1992; accepted for publication (in revised form) by J. R. Bunch, September 1, 1993. This research was supported in part by U. S. Army Research Office contract DAAL03-91-G-0032.

† Department of Mathematics and Lawrence Berkeley Laboratory, University of California, Berkeley, California 94720 (minggu@math.berkeley.edu).

‡ Department of Computer Science, Yale University, Box 208285, New Haven, Connecticut 06520-8285 (eisenstat-stan@cs.yale.edu).

[1] This strategy has previously appeared in [1], [3], [13], [15], [19], and [22].

where ϵ is the machine precision, $\hat{\Lambda}$ is diagonal, and $\hat{X}$ or $\hat{U}$ is *numerically orthogonal*. An algorithm that produces such a decomposition is said to be *stable*.

While the eigenvalues of T and H are always well conditioned with respect to a symmetric perturbation, the eigenvectors can be extremely sensitive to such perturbations [14, pp. 413–414]. That is, $\hat{\Lambda}$ must be close to Λ, but $\hat{X}$ and $\hat{U}$ can be very different from X and U, respectively. Thus one is usually content with stable algorithms for computing the spectral decompositions of T and H.

Finding the spectral decomposition of a symmetric arrowhead matrix is an interesting problem in its own right (see [3], [4], [26]–[28] and references therein). Several methods for solving this problem have been proposed [3], [15], [26], [28]. While they can compute the eigenvalues to high absolute accuracy, in the presence of close eigenvalues they can have difficulties in computing numerically orthogonal eigenvectors, unless extended precision arithmetic is used [24], [29]. In this paper we present a new algorithm for computing the spectral decomposition of a symmetric arrowhead matrix. It is similar to previous methods for finding the eigenvalues, but it uses a completely different approach to finding the eigenvectors, one that is stable. The amount of work is roughly the same as for previous methods, yet it does not require the use or simulation of extended precision arithmetic. Since it uses this algorithm, ADC is stable as well.

ADC computes all the eigenvalues of T in $O(N^2)$ time and both the eigenvalues and eigenvectors of T in $O(N^3)$ time. We show that by using the fast multipole method of Carrier, Greengard, and Rokhlin [10], [16], ADC can be accelerated to compute all the eigenvalues in $O(N \log_2 N)$ time and both the eigenvalues and eigenvectors in $O(N^2)$ time. These asymptotic time requirements are better than the corresponding worst-case times ($O(N^2)$ and $O(N^3)$) for bisection with inverse iteration [21], [23] and the QR algorithm [8]. Our algorithm for finding all the eigenvalues of H takes $O(N^2)$ time as do previous methods [3], [15], [26], [28]. By using the fast multipole method, it can be accelerated to compute all the eigenvalues in $O(N)$ time.

Cuppen's divide-and-conquer algorithm (CDC) [11], [12] uses a similar dividing strategy, but it reduces T to a symmetric rank-one modification to a diagonal matrix rather than to a symmetric arrowhead matrix. However, in the presence of close eigenvalues it can have difficulties in computing numerically orthogonal eigenvectors [11], [12], unless extended precision arithmetic is used [5], [24], [29]. In contrast, ADC is stable and is roughly twice as fast as existing implementations of CDC (e.g., `TREEQL` [12]) for large matrices due to the differences in how deflation is implemented.[2] ADC is also very competitive with bisection with inverse iteration [21], [23] and the QR algorithm [8].

Section 2 presents the dividing strategy used in ADC; §3 develops an efficient algorithm for the spectral decomposition of a symmetric arrowhead matrix and shows that it is stable; §4 discusses the deflation procedure used in ADC; §5 discusses the application of the fast multipole method to speed up ADC; and §6 presents some numerical results.

[2] Our techniques [17], [20] can be used to stabilize CDC without the need for extended precision arithmetic; our deflation procedure can be adapted to CDC, as can the fast multipole method.

We take the usual model of arithmetic[3]

$$\mathrm{fl}(x \circ y) = (x \circ y)\,(1+\xi),$$

where x and y are floating-point numbers; $\circ$ is one of $+, -, \times$, and $\div$; $\mathrm{fl}(x \circ y)$ is the floating-point result of the operation $\circ$; and $|\xi| \leq \epsilon$. We also require that

$$\mathrm{fl}(\sqrt{x}) = \sqrt{x}\,(1+\xi)$$

for any positive floating-point number x. For simplicity we ignore the possibility of overflow and underflow.

2. "Dividing" the matrix. Given an $N \times N$ symmetric tridiagonal matrix T, ADC divides T into two subproblems as follows:

$$T = \begin{pmatrix} T_1 & \beta_{k+1}e_k & 0 \\ \beta_{k+1}e_k^T & \alpha_{k+1} & \beta_{k+2}e_1^T \\ 0 & \beta_{k+2}e_1 & T_2 \end{pmatrix}, \tag{2}$$

where $1 < k < n$, T_1 and T_2 are $k \times k$ and $(N-k-1) \times (N-k-1)$ principle submatrices of T, respectively, and e_j is the jth unit vector of appropriate dimension. Usually k is taken to be $\lfloor N/2 \rfloor$.

Let $Q_i D_i {Q_i}^T$ be a spectral decomposition of T_i. Substituting into (2), we get

$$\begin{aligned} T &= \begin{pmatrix} Q_1 D_1 {Q_1}^T & \beta_{k+1}e_k & 0 \\ \beta_{k+1}e_k^T & \alpha_{k+1} & \beta_{k+2}e_1^T \\ 0 & \beta_{k+2}e_1 & Q_2 D_2 {Q_2}^T \end{pmatrix} \\ &= \begin{pmatrix} 0 & Q_1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & Q_2 \end{pmatrix} \begin{pmatrix} \alpha_{k+1} & \beta_{k+1}l_1^T & \beta_{k+2}f_2^T \\ \beta_{k+1}l_1 & D_1 & 0 \\ \beta_{k+2}f_2 & 0 & D_2 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ {Q_1}^T & 0 & 0 \\ 0 & 0 & {Q_2}^T \end{pmatrix} \\ &\equiv QHQ^T, \end{aligned} \tag{3}$$

where l_1^T is the last row of Q_1 and f_2^T is the first row of Q_2. Thus T is reduced to the symmetric arrowhead matrix H by the orthogonal similarity transformation Q.

ADC computes the spectral decomposition $H = U\Lambda U^T$ using the algorithm described in §3. The eigenvalues of T are the diagonal elements of Λ, and the eigenvector matrix of T is obtained by computing the matrix-matrix product $X = QU$. To compute the spectral decompositions of T_1 and T_2, this process ((2) and (3)) can be recursively applied until the subproblems are sufficiently small. These small subproblems are solved using the QR algorithm. There can be at most $O(\log_2 N)$ levels of recursion.

Equations (2) and (3) also suggest a recursion for computing only the eigenvalues. Let f_1^T be the first row of Q_1 and let l_2^T be the last row of Q_2. Suppose that D_i, f_i, and l_i are given for $i = 1, 2$. Then after finding the spectral decomposition of H, the first row of X can be computed as $(0,\ f_1^T,\ 0)\ U$ and the last row of X can be computed as $(0,\ 0,\ l_2^T)\ U$. There is a similar recursion for CDC [11].

[3] This model excludes machines like CRAY and CDC Cyber that do not have a guard digit. ADC can easily be modified for such machines.

3. Computing the spectral decomposition of a symmetric arrowhead matrix. In this section we develop a stable and efficient method for finding the spectral decomposition of an $n \times n$ symmetric arrowhead matrix

$$H = \begin{pmatrix} \alpha & z^T \\ z & D \end{pmatrix},$$

where $D = \operatorname{diag}(d_2, \ldots, d_n)$ is an $(n-1) \times (n-1)$ matrix with $d_2 \leq d_3 \leq \cdots \leq d_n$, $z = (z_2, \ldots, z_n)^T$ is a vector of length $n-1$, and α is a scalar. The development closely parallels that in [17] and [20] for finding the spectral decomposition of a symmetric rank-one modification to a diagonal matrix.

We further assume that

$$d_{j+1} - d_j \geq \tau \|H\|_2 \quad \text{and} \quad |z_i| \geq \tau \|H\|_2, \tag{4}$$

where τ is a small multiple of ϵ to be specified later. Any symmetric arrowhead matrix can be reduced to one that satisfies these conditions by using the deflation procedure described in §4.1 and a simple permutation.

The following lemma characterizes the eigenvalues and eigenvectors of symmetric arrowhead matrices.

LEMMA 3.1 (Wilkinson [30, pp. 95–96], O'Leary and Stewart [26]). *The eigenvalues* $\{\lambda_i\}_{i=1}^n$ *of* H *satisfy the* interlacing property

$$\lambda_1 < d_2 < \lambda_2 < \cdots < d_n < \lambda_n$$

and the secular equation

$$f(\lambda) = \lambda - \alpha + \sum_{j=2}^{n} \frac{z_j^2}{d_j - \lambda} = 0.$$

For each eigenvalue λ_i *of* H, *the corresponding eigenvector is given by*

$$u_i = \left(-1, \frac{z_1}{d_2 - \lambda_i}, \ldots, \frac{z_n}{d_n - \lambda_i}\right)^T \Bigg/ \sqrt{1 + \sum_{j=2}^{n} \frac{z_j^2}{(d_j - \lambda_i)^2}}\,. \tag{5}$$

The following lemma allows us to construct a symmetric arrowhead matrix from its eigenvalues and its shaft.

LEMMA 3.2 (Boley and Golub [6]). *Given a set of numbers* $\{\hat{\lambda}_i\}_{i=1}^n$ *and a diagonal matrix* $D = \operatorname{diag}(d_2, \ldots, d_n)$ *satisfying the interlacing property*

$$\hat{\lambda}_1 < d_2 < \hat{\lambda}_2 < \cdots < d_n < \hat{\lambda}_n, \tag{6}$$

there exists a symmetric arrowhead matrix

$$\hat{H} = \begin{pmatrix} \hat{\alpha} & \hat{z}^T \\ \hat{z} & D \end{pmatrix}$$

whose eigenvalues are $\{\hat{\lambda}_i\}_{i=1}^n$. *The vector* $\hat{z} = (\hat{z}_2, \ldots, \hat{z}_n)^T$ *and the scalar* $\hat{\alpha}$ *are given by*

$$|\hat{z}_i| = \sqrt{(d_i - \hat{\lambda}_1)\,(\hat{\lambda}_n - d_i) \prod_{j=2}^{i-1} \frac{(\hat{\lambda}_j - d_i)}{(d_j - d_i)} \prod_{j=i}^{n-1} \frac{(\hat{\lambda}_j - d_i)}{(d_{j+1} - d_i)}} \tag{7}$$

$$\hat{\alpha} = \hat{\lambda}_1 + \sum_{j=2}^{n} \left(\hat{\lambda}_j - d_j\right), \tag{8}$$

where the sign of $\hat{z}_i$ *can be chosen arbitrarily.*

3.1. Computing the eigenvectors. If λ_i were given *exactly*, then we could compute each difference, each ratio, each product, and each sum in (5) to high relative accuracy, and thus compute u_i to componentwise high relative accuracy. In practice we can only hope to compute an approximation $\hat{\lambda}_i$ to λ_i. But problems can arise if we approximate u_i by

$$\hat{u}_i = \left(-1, \frac{z_1}{d_2 - \hat{\lambda}_i}, \ldots, \frac{z_n}{d_n - \hat{\lambda}_i}\right)^T \Bigg/ \sqrt{1 + \sum_{j=2}^{n} \frac{z_j^2}{(d_j - \hat{\lambda}_i)^2}}$$

(i.e., replace λ_i by $\hat{\lambda}_i$ in (5), as in [3], [15], and [26]). For even if $\hat{\lambda}_i$ is close to λ_i, the approximate ratio $z_j/(d_j - \hat{\lambda}_i)$ can be very different from the exact ratio $z_j/(d_j - \lambda_i)$, resulting in a $\hat{u}_i$ very different from u_i. And when all the approximate eigenvalues $\{\hat{\lambda}_i\}_{i=1}^n$ are computed and all the corresponding eigenvectors are approximated in this manner, the resulting eigenvector matrix may not be numerically orthogonal.

Lemma 3.2 allows us to overcome this problem. After we have computed all the approximate eigenvalues $\{\hat{\lambda}_i\}_{i=1}^n$ of H, we can find a *new* matrix $\hat{H}$ whose *exact* eigenvalues are $\{\hat{\lambda}_i\}_{i=1}^n$, and then compute the eigenvectors of $\hat{H}$ using Lemma 3.1. Note that each difference, each product, and each ratio in (7) can be computed to high relative accuracy, and the sign of $\hat{z}_i$ can be taken to be the sign of z_i. Thus $\hat{z}_i$ can be computed to componentwise high relative accuracy. Substituting the *exact* eigenvalues $\{\hat{\lambda}_i\}_{i=1}^n$ and the computed $\hat{z}$ into (5), each eigenvector of $\hat{H}$ can be computed to componentwise high relative accuracy. And, after all the eigenvectors are computed, the computed eigenvector matrix of $\hat{H}$ will be numerically orthogonal.

To ensure the existence of $\hat{H}$, the approximations $\{\hat{\lambda}_i\}_{i=1}^n$ must satisfy the interlacing property (6). But since the exact eigenvalues of H satisfy the same interlacing property (see Lemma 3.1), this is only an accuracy requirement on the computed eigenvalues and is not an additional restriction on H.

We can use the spectral decomposition of $\hat{H}$ as an approximation to that of H. Since

$$H = \begin{pmatrix} \alpha & z^T \\ z & D \end{pmatrix} = \hat{H} + \begin{pmatrix} \alpha - \hat{\alpha} & (z - \hat{z})^T \\ z - \hat{z} & 0 \end{pmatrix},$$

we have

$$\|\hat{H} - H\|_2 \leq |\alpha - \hat{\alpha}| + \|z - \hat{z}\|_2.$$

Thus such a substitution is stable (see (1)) as long as $\hat{\alpha}$ and $\hat{z}$ are close to α and z, respectively (cf. [17], [20]).

3.2. Computing the eigenvalues. To guarantee that $\hat{z}$ is close to z and $\hat{\alpha}$ is close to α, we must ensure that $\{\hat{\lambda}_i\}_{i=1}^n$ are sufficiently accurate approximations to the eigenvalues. The key is the stopping criterion for the root-finder, which requires a slight reformulation of the secular equation (cf. [9], [17], [20]).

Consider the eigenvalue $\lambda_i \in (d_i, d_{i+1})$, where $2 \le i \le n-1$; we consider the cases $i = 1$ and $i = n$ later. λ_i is a root of the secular equation

$$f(\lambda) = \lambda - \alpha + \sum_{j=2}^{n} \frac{z_j^2}{d_j - \lambda} = 0.$$

We first assume that[4] $\lambda_i \in \left(d_i, \frac{d_i+d_{i+1}}{2}\right)$. Let $\alpha_i = d_i - \alpha$ and $\delta_j = d_j - d_i$, and let

$$\psi_i(\mu) \equiv \sum_{j=2}^{i} \frac{z_j^2}{\delta_j - \mu} \quad \text{and} \quad \phi_i(\mu) \equiv \sum_{j=i+1}^{n} \frac{z_j^2}{\delta_j - \mu}.$$

Since

$$f(\mu + d_i) = \mu + \alpha_i + \psi_i(\mu) + \phi_i(\mu) \equiv g_i(\mu),$$

we seek the root $\mu_i = \lambda_i - d_i \in (0, \delta_{i+1}/2)$ of $g_i(\mu) = 0$. Let $\hat{\mu}_i$ be the computed root so that $\hat{\lambda}_i = d_i + \hat{\mu}_i$ is the computed eigenvalue.

An important property of $g_i(\mu)$ is that each difference $\delta_j - \mu$ can be evaluated to high relative accuracy for any $\mu \in (0, \delta_{i+1}/2)$. Indeed, since $\delta_i = 0$, we have $\mathrm{fl}(\delta_i - \mu) = -\mathrm{fl}(\mu)$. Since $\mathrm{fl}(\delta_{i+1}) = \mathrm{fl}(d_{i+1} - d_i)$ and $0 < \mu < (d_{i+1} - d_i)/2$, we can compute $\mathrm{fl}(\delta_{i+1} - \mu)$ as $\mathrm{fl}(\mathrm{fl}(d_{i+1} - d_i) - \mathrm{fl}(\mu))$. In a similar fashion, we can compute $\delta_j - \mu$ to high relative accuracy for any $j \neq i, i+1$.

Because of this property, each ratio $z_j^2/(\delta_j - \mu)$ in $g_i(\mu)$ can be evaluated to high relative accuracy for any $\mu \in (0, \delta_{i+1}/2)$. Moreover, α_i can be computed to high relative accuracy. Thus, since both $\psi_i(\mu)$ and $\phi_i(\mu)$ are sums of terms of the same sign, we can bound the error in computing $g_i(\mu)$ by

$$\eta n(|\mu| + |\alpha_i| + |\psi_i(\mu)| + |\phi_i(\mu)|),$$

where η is a small multiple of ϵ that is independent of n and μ.

We now assume that $\lambda_i \in \left[(d_i + d_{i+1})/2, d_{i+1}\right)$. Let $\alpha_i = d_{i+1} - \alpha$ and $\delta_j = d_j - d_{i+1}$, and let

$$\psi_i(\mu) \equiv \sum_{j=2}^{i} \frac{z_j^2}{\delta_j - \mu} \quad \text{and} \quad \phi_i(\mu) \equiv \sum_{j=i+1}^{n} \frac{z_j^2}{\delta_j - \mu}.$$

We seek the root $\mu_i = \lambda_i - d_{i+1} \in [\delta_i/2, 0)$ of the equation

$$g_i(\mu) \equiv f(\mu + d_{i+1}) = \mu + \alpha_i + \psi_i(\mu) + \phi_i(\mu) = 0.$$

[4] This can be checked by computing $f\left(\frac{d_i+d_{i+1}}{2}\right)$. If $f\left(\frac{d_i+d_{i+1}}{2}\right) > 0$, then $\lambda_i \in \left(d_i, \frac{d_i+d_{i+1}}{2}\right)$, otherwise $\lambda_i \in \left[\frac{d_i+d_{i+1}}{2}, d_{i+1}\right)$.

Let $\hat{\mu}_i$ be the computed root so that $\hat{\lambda}_i = d_{i+1} + \hat{\mu}_i$. For any $\mu \in [\delta_i/2, 0)$, each difference $\delta_j - \mu$ can again be computed to high relative accuracy, as can each ratio $z_j^2/(\delta_j - \mu)$ and the scalar α_i, and we can bound the error in computing $g_i(\mu)$ as before.

Next we consider the case $i = 1$. Let $\alpha_1 = d_2 - \alpha$ and $\delta_j = d_j - d_2$, and let

$$\psi_1(\mu) \equiv 0 \quad \text{and} \quad \phi_1(\mu) \equiv \sum_{j=2}^{n} \frac{z_j^2}{\delta_j - \mu}.$$

We seek the root $\mu_1 = \lambda_1 - d_2 \in (-\|z\|_2 - |\alpha_1|, 0)$ of the equation

$$g_1(\mu) \equiv f(\mu + d_2) = \mu + \alpha_1 + \psi_1(\mu) + \phi_1(\mu) = 0.$$

Let $\hat{\mu}_1$ be the computed root so that $\hat{\lambda}_1 = d_2 + \hat{\mu}_1$. For any $\mu \in (-\|z\|_2 - |\alpha_1|, 0)$, each ratio $z_j^2/(\delta_j - \mu)$ can be computed to high relative accuracy, as can α_1, and we can bound the error in computing $g_1(\mu)$ as before.

Finally we consider the case $i = n$. Let $\alpha_n = d_n - \alpha$ and $\delta_j = d_j - d_n$, and let

$$\psi_n(\mu) \equiv \sum_{j=2}^{n} \frac{z_j^2}{\delta_j - \mu} \quad \text{and} \quad \phi_n(\mu) \equiv 0.$$

We seek the root $\mu_n = \lambda_n - d_n \in (0, \|z\|_2 + |\alpha_n|)$ of the equation

$$g_n(\mu) \equiv f(\mu + d_n) = \mu + \alpha_n + \psi_n(\mu) + \phi_n(\mu) = 0.$$

Let $\hat{\mu}_n$ be the computed root so that $\hat{\lambda}_n = d_n + \hat{\mu}_n$. For any $\mu \in (0, \|z\|_2 + |\alpha_n|)$, each ratio $z_j^2/(\delta_j - \mu)$ can be computed to high relative accuracy, as can α_n, and we can bound the error in computing $g_n(\mu)$ as before.

In practice the root-finder cannot make any progress at a point μ where it is impossible to determine the sign of $g_i(\mu)$ numerically. Thus we propose the stopping criterion

$$|g_i(\mu)| \leq \eta n \left(|\mu| + |\alpha_i| + |\psi_i(\mu)| + |\phi_i(\mu)| \right), \tag{9}$$

where, as before, the right-hand side is an upper bound on the round-off error in computing $g_i(\mu)$. Note that for each i, there is at least one floating-point number that satisfies this stopping criterion numerically, namely, $\mathrm{fl}(\mu_i)$.

We have not specified the method used to find the root of $g_i(\mu)$. We used a modified version of the rational interpolation strategy in [9] for the numerical experiments, but bisection and its variations [26], [28] or the improved rational interpolation strategies in [15], [25] would also work. What is most important is the stopping criterion and the fact that, with the reformulation of the secular equation given above, we can find a μ that satisfies it.

3.3. Numerical stability. In this subsection we show that $\hat{\alpha}$ and $\hat{z}$ are indeed close to α and z, respectively, as long as the root-finder guarantees that each $\hat{\mu}_i$ satisfies the stopping criterion (9).

Since $f(\lambda_i) = 0$, we have

$$|\alpha_i| = \left| -\mu_i - \sum_{j=2}^{n} \frac{z_j^2}{d_j - \lambda_i} \right| \leq |\mu_i| + \sum_{j=2}^{n} \frac{z_j^2}{|d_j - \lambda_i|}$$

and

$$f(\hat{\lambda}_i) = f(\hat{\lambda}_i) - f(\lambda_i) = (\hat{\lambda}_i - \lambda_i)\left(1 + \sum_{j=2}^{n} \frac{z_j^2}{(d_j - \hat{\lambda}_i)(d_j - \lambda_i)}\right).$$

Since the computed eigenvalue $\hat{\lambda}_i$ satisfies (9), we have

$$|f(\hat{\lambda}_i)| \le \eta n \left(|\mu_i| + |\hat{\mu}_i| + \sum_{j=2}^{n} \frac{z_j^2}{|d_j - \lambda_i|} + \sum_{j=2}^{n} \frac{z_j^2}{|d_j - \hat{\lambda}_i|}\right),$$

so that

$$|\hat{\lambda}_i - \lambda_i| \left(1 + \sum_{j=2}^{n} \frac{z_j^2}{|(d_j - \hat{\lambda}_i)(d_j - \lambda_i)|}\right) \le \eta n \left(|\mu_i| + |\hat{\mu}_i| + \sum_{j=2}^{n} \frac{z_j^2}{|d_j - \hat{\lambda}_i|} + \sum_{j=2}^{n} \frac{z_j^2}{|d_j - \lambda_i|}\right). \tag{10}$$

Note that for any i and j,

$$|\mu_i| + |\hat{\mu}_i| \le 4\|H\|_2 + |\hat{\lambda}_i - \lambda_i| \quad \text{and} \quad |d_j - \hat{\lambda}_i| + |d_j - \lambda_i| \le 4\|H\|_2 + |\hat{\lambda}_i - \lambda_i|.$$

Substituting these relations into (10), we get

$$|\hat{\lambda}_i - \lambda_i| \left(1 + \sum_{j=2}^{n} \frac{z_j^2}{|(d_j - \hat{\lambda}_i)(d_j - \lambda_i)|}\right) \le \eta n \left(4\|H\|_2 + |\hat{\lambda}_i - \lambda_i|\right) \left(1 + \sum_{j=2}^{n} \frac{z_j^2}{|d_j - \hat{\lambda}_i||d_j - \lambda_i|}\right),$$

or

$$|\hat{\lambda}_i - \lambda_i| \le \frac{4\eta n \|H\|_2}{1 - \eta n},$$

i.e., all the eigenvalues are computed to high absolute accuracy. Applying (8) in Lemma 3.2 to both H and $\hat{H}$, we have

$$\alpha = \lambda_1 + \sum_{j=2}^{n} (\lambda_j - d_j) \quad \text{and} \quad \hat{\alpha} = \hat{\lambda}_1 + \sum_{j=2}^{n} (\hat{\lambda}_j - d_j),$$

and therefore

$$|\alpha - \hat{\alpha}| = \left|\sum_{j=1}^{n} (\lambda_j - \hat{\lambda}_j)\right| \le \sum_{j=1}^{n} \left|\lambda_j - \hat{\lambda}_j\right| \le \frac{4\eta n^2 \|H\|_2}{1 - \eta n}. \tag{11}$$

To show that $\hat{z}$ is close to z, we further note that for any i and j, we have

$$|\hat{\mu}_i| \leq |\mu_i| + |\hat{\lambda}_i - \lambda_i|$$

and

$$\frac{1}{|d_j - \hat{\lambda}_i|} + \frac{1}{|d_j - \lambda_i|} \leq \frac{2}{|(d_j - \hat{\lambda}_i)(d_j - \lambda_i)|^{\frac{1}{2}}} + \frac{|\hat{\lambda}_i - \lambda_i|}{|(d_j - \hat{\lambda}_i)(d_j - \lambda_i)|}.$$

Substituting these relations into (10) and using the Cauchy–Schwartz inequality, we get

$$\begin{aligned}
|\hat{\lambda}_i - \lambda_i| & \left(1 + \sum_{j=2}^{n} \frac{z_j^2}{|(d_j - \hat{\lambda}_i)(d_j - \lambda_i)|}\right) \\
& \leq \frac{2\eta n}{1 - \eta n} \left(|\mu_i| + \sum_{j=2}^{n} \frac{z_j^2}{|(d_j - \hat{\lambda}_i)(d_j - \lambda_i)|^{1/2}}\right) \\
& \leq \frac{2\eta n}{1 - \eta n} \sqrt{|\mu_i|^2 + \|z\|_2^2} \sqrt{1 + \sum_{j=2}^{n} \frac{z_j^2}{|(d_j - \hat{\lambda}_i)(d_j - \lambda_i)|}}.
\end{aligned}$$

Since $|\mu_i|^2 + \|z\|_2^2 \leq 5\|H\|_2^2$, we have

$$\begin{aligned}
|\hat{\lambda}_i - \lambda_i| & \leq \frac{2\eta n}{1 - \eta n} \sqrt{|\mu_i|^2 + \|z\|_2^2} \Bigg/ \sqrt{1 + \sum_{j=2}^{n} \frac{z_j^2}{|(d_j - \hat{\lambda}_i)(d_j - \lambda_i)|}} \\
& \leq \frac{2\sqrt{5}\eta n \|H\|_2}{(1 - \eta n)|z_j|} \sqrt{|(d_j - \hat{\lambda}_i)(d_j - \lambda_i)|} \\
& \leq \frac{2\sqrt{5}\eta n \|H\|_2}{(1 - \eta n)|z_j|} \left(|d_j - \lambda_i| + \frac{1}{2}|\hat{\lambda}_i - \lambda_i|\right).
\end{aligned}$$

Letting $\beta_j = 2\sqrt{5}\eta n \|H\|_2 / ((1 - \eta n)|z_j|)$, this implies that

$$|\hat{\lambda}_i - \lambda_i| \leq \frac{\beta_j}{1 - \frac{1}{2}\beta_j} |d_j - \lambda_i| \tag{12}$$

for every $2 \leq j \leq n$, provided that $\beta_j < 2$.

Let $\hat{\lambda}_i - \lambda_i = \alpha_{ij}(d_j - \lambda_i)/z_j$ for all i and j. Suppose that we pick $\tau = 6\eta n^2$ in (4). Then $|z_j| \geq 6\eta n^2 \|H\|_2$. Assume further that $\eta n < 1/100$. Then $\beta_j \leq 2/5$, and (12) implies that $|\alpha_{ij}| \leq \alpha \equiv 6\eta n \|H\|_2$ for all i and j. Thus

$$|\hat{z}_i| = \sqrt{\frac{\prod_{j=1}^{n} (\hat{\lambda}_j - d_i)}{\prod_{j=2, j\neq i}^{n} (d_j - d_i)}} = \sqrt{\frac{\prod_{j=1}^{n} (\lambda_j - d_i)\left(1 + \frac{\alpha_{ji}}{z_i}\right)}{\prod_{j=2, j\neq i}^{n} (d_j - d_i)}} = |z_i| \sqrt{\prod_{j=1}^{n} \left(1 + \frac{\alpha_{ji}}{z_i}\right)}$$

and, since $\hat{z}_i$ and z_i have the same sign,

$$|\hat{z}_i - z_i| = |z_i| \left|\sqrt{\prod_{j=1}^{n} \left(1 + \frac{\alpha_{ji}}{z_i}\right)} - 1\right| \leq |z_i| \left(\left(1 + \frac{\alpha}{|z_i|}\right)^{\frac{n}{2}} - 1\right)$$

$$
\begin{aligned}
(13) \qquad & \leq |z_i| \left(\exp\left(\frac{\alpha n}{2|z_i|} \right) - 1 \right) \leq (e-1)\, \alpha n/2 \\
& \leq 6\eta n^2 \|H\|_2,
\end{aligned}
$$

where we have used the fact that $\alpha n/(2|z_i|) \leq 1$ and that $(e^x - 1)/x \leq e - 1$ for $0 < x \leq 1$.

One factor of n in τ and the bounds (11) and (13) comes from the stopping criterion (9). This is quite conservative and could be reduced to $\log_2 n$ by using a binary tree structure for summing the terms in $\psi_i(\mu)$ and $\phi_i(\mu)$. The other factor of n comes from the upper bound for $\sum_{j=1}^{n} (\lambda_j - \hat{\lambda}_j)$ in (11) and $\prod_{j=1}^{n}(1 + \alpha_{ji}/z_i)$ in (13). This also seems quite conservative. Thus we might expect the factor of n^2 in τ and the bounds (11) and (13) to be more like $O(n)$ in practice.

4. Deflation.

4.1. Deflation for symmetric arrowhead matrices.

Let

$$
H = \begin{pmatrix} \alpha & z^T \\ z & D \end{pmatrix},
$$

where $D = \operatorname{diag}(d_2, \ldots, d_n)$ and $z = (z_2, \ldots, z_n)^T$. We now show that we can reduce[5] H to a symmetric arrowhead matrix that further satisfies

$$
|d_i - d_j| \geq \tau \|H\|_2, \quad \text{for} \quad i \neq j \quad \text{and} \quad |z_i| \geq \tau \|H\|_2
$$

(cf. (4)), where τ is specified in §3.3. We illustrate the reductions for $n = 3$, $i = 3$, and $j = 2$.

Assume that $|z_i| < \tau \|H\|_2$. Then

$$
(14) \qquad H = \begin{pmatrix} \alpha & z_2 & z_3 \\ z_2 & d_2 & \\ z_3 & & d_3 \end{pmatrix} = \begin{pmatrix} \alpha & z_2 & 0 \\ z_2 & d_2 & \\ 0 & & d_3 \end{pmatrix} + O(\tau \|H\|_2).
$$

We perturb z_i to zero. Then H is perturbed by $O(\tau \|H\|_2)$. d_i is an eigenvalue of the perturbed matrix and is deflated. The $(n-1) \times (n-1)$ leading principle submatrix of the perturbed matrix is another symmetric arrowhead matrix with smaller dimensions. This deflation rule is stable (see (1)).

Now assume that $|d_i - d_j| < \tau \|H\|_2$. Apply a Givens rotation G to H to zero out z_i:

$$
\begin{aligned}
GHG^T &= \begin{pmatrix} 1 & & \\ & c & s \\ & -s & c \end{pmatrix} \begin{pmatrix} \alpha & z_2 & z_3 \\ z_2 & d_2 & \\ z_3 & & d_3 \end{pmatrix} \begin{pmatrix} 1 & & \\ & c & -s \\ & s & c \end{pmatrix} \\
(15) \qquad &= \begin{pmatrix} \alpha & r & 0 \\ r & d_2c^2 + d_3s^2 & cs(d_3 - d_2) \\ 0 & cs(d_3 - d_2) & d_2s^2 + d_3c^2 \end{pmatrix} \\
&= \begin{pmatrix} \alpha & r & 0 \\ r & d_2c^2 + d_3s^2 & \\ 0 & & d_2s^2 + d_3c^2 \end{pmatrix} + O(\tau \|H\|_2),
\end{aligned}
$$

[5] These rules have previously appeared in [15] and [19].

where $r = \sqrt{z_i^2 + z_j^2}$, $c = z_j/r$, and $s = z_i/r$. We perturb $cs(d_i - d_j)$ to zero. Then GHG^T is perturbed by $O(\tau\|H\|_2)$. $d_j s^2 + d_i c^2$ is an eigenvalue of the perturbed matrix and can be deflated. The $(n-1) \times (n-1)$ leading principle submatrix of the perturbed matrix is another symmetric arrowhead matrix with smaller dimensions. This deflation rule is also stable (see (1)).

4.2. Local deflation. In the dividing strategy for ADC (see (3)), we write

$$T = \begin{pmatrix} 0 & Q_1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & Q_2 \end{pmatrix} \begin{pmatrix} \alpha_{k+1} & \beta_{k+1} l_1^T & \beta_{k+2} f_2^T \\ \beta_{k+1} l_1 & D_1 & 0 \\ \beta_{k+2} f_2 & 0 & D_2 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ {Q_1}^T & 0 & 0 \\ 0 & 0 & {Q_2}^T \end{pmatrix} \tag{16}$$
$$= (QU)\Lambda(QU)^T,$$

where Q is the first matrix in (16) and $U\Lambda U^T$ is the spectral decomposition of the middle matrix.

Note that Q is a block matrix with some zero blocks. When we compute the matrix-matrix product QU, we would like to take advantage of this structure. Since the main cost of ADC is in computing such products, we get a speedup of close to a factor of two by doing so. This is not done in any current implementation of CDC.

If the vector $(\beta_{k+1} l_1^T,\ \beta_{k+2} f_2^T)$ has components with small absolute value, then we can apply reduction (14). The block structure of Q is preserved. If D_1 has two close diagonal elements, then we can apply reduction (15). The block structure of Q is again preserved. We can do the same when D_2 has two close diagonal elements.

However, when D_1 has a diagonal element that is close to a diagonal element in D_2 and we apply reduction (15), the block structure of Q is changed. To illustrate, assume that after applying a permutation the first diagonal element of D_1 is close to the last diagonal element of D_2. Let $Q_1 = (q_1, \tilde{Q}_1)$ and $Q_2 = (\tilde{Q}_2, q_2)$; let $D_1 = \mathrm{diag}(d_2, \tilde{D}_1)$ and $D_2 = \mathrm{diag}(\tilde{D}_2, d_N)$; and let $\beta_{k+1} l_1^T = (z_2, \tilde{z}_1^T)$ and $\beta_{k+2} f_2^T = (\tilde{z}_2^T, z_N)$. By assumption, d_2 and d_N are close. When we apply the Givens rotation

$$G = \begin{pmatrix} 1 & & & & \\ & c & & & s \\ & & & I_{N-3} & \\ & & & & \\ & -s & & & c \end{pmatrix}$$

to the middle matrix in (16) to zero out z_N, we create some nonzero elements in the second and Nth columns of Q:

$$T = \begin{pmatrix} 0 & q_1 & \tilde{Q}_1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \tilde{Q}_2 & q_2 \end{pmatrix} G^T G \begin{pmatrix} \alpha_{k+1} & z_2 & \tilde{z}_1^T & \tilde{z}_2^T & z_N \\ z_2 & d_2 & & & \\ \tilde{z}_1 & & \tilde{D}_1 & & \\ \tilde{z}_2 & & & \tilde{D}_2 & \\ z_N & & & & d_N \end{pmatrix} G^T G \begin{pmatrix} 0 & 1 & 0 \\ q_1^T & 0 & 0 \\ \tilde{Q}_1^T & 0 & 0 \\ 0 & 0 & \tilde{Q}_2^T \\ 0 & 0 & q_2^T \end{pmatrix}$$

$$= \begin{pmatrix} 0 & cq_1 & \tilde{Q}_1 & 0 & -sq_1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & sq_2 & 0 & \tilde{Q}_2 & cq_2 \end{pmatrix} \begin{pmatrix} \alpha_{k+1} & r & \tilde{z}_1^T & \tilde{z}_2^T & 0 \\ r & \tilde{d}_2 & & & \\ \tilde{z}_1 & & \tilde{D}_1 & & \\ \tilde{z}_2 & & & \tilde{D}_2 & \\ 0 & & & & \tilde{d}_N \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ cq_1^T & 0 & sq_2^T \\ \tilde{Q}_1^T & 0 & 0 \\ 0 & 0 & \tilde{Q}_2^T \\ -sq_2^T & 0 & cq_2^T \end{pmatrix}$$

$$+ O(\tau\|T\|_2),$$

where $\tilde{d}_2 = d_2c^2 + d_Ns^2$ and $\tilde{d}_N = d_2s^2 + d_Nc^2$.

Note that $\tilde{d}_N$ is an approximate eigenvalue of T and can be deflated. The corresponding approximate eigenvector is the last column of the first matrix. The leading $(N-1) \times (N-1)$ principle submatrix of the middle matrix is again a symmetric arrowhead matrix and can be deflated in a similar fashion until no diagonal element of $\tilde{D}_1$ is close to a diagonal element of $\tilde{D}_2$.

Thus, ignoring permutations of the columns of Q_i and the diagonal elements of D_i, after a series of these interblock deflations T can be written as

$$T = \begin{pmatrix} \tilde{X}_1 & \tilde{X}_2 \end{pmatrix} \begin{pmatrix} \tilde{H}_1 & \\ & \tilde{\Lambda}_2 \end{pmatrix} \begin{pmatrix} \tilde{X}_1 & \tilde{X}_2 \end{pmatrix}^T + O(\tau\|T\|_2). \tag{17}$$

$\tilde{\Lambda}_2$ is a diagonal matrix whose diagonal elements are the deflated eigenvalues and the columns of $\tilde{X}_2$ are the corresponding approximate eigenvectors. $\tilde{H}_1$ is the symmetric arrowhead matrix

$$\tilde{H}_1 = \begin{pmatrix} \alpha_{k+1} & \tilde{z}_0^T & \tilde{z}_1^T & \tilde{z}_2^T \\ \tilde{z}_0 & \tilde{D}_0 & & \\ \tilde{z}_1 & & \tilde{D}_1 & \\ \tilde{z}_2 & & & \tilde{D}_2 \end{pmatrix},$$

where the dimension of $\tilde{D}_0$ is the number of deflations, $\tilde{D}_1$ and $\tilde{D}_2$ contain the undeflated diagonal elements of D_1 and D_2, and $\tilde{z}_0$, $\tilde{z}_1$, and $\tilde{z}_2$ are defined accordingly. $\tilde{X}_1$ is of the form

$$\tilde{X}_1 = \begin{pmatrix} 0 & \tilde{Q}_{0,1} & \tilde{Q}_1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & \tilde{Q}_{0,2} & 0 & \tilde{Q}_2 \end{pmatrix}, \tag{18}$$

where the column dimension of both $\tilde{Q}_{0,1}$ and $\tilde{Q}_{0,2}$ is the number of deflations and the columns of $\tilde{Q}_1$ and $\tilde{Q}_2$ are those columns of Q_1 and Q_2 not affected by deflation.

If some diagonal element of $\tilde{D}_0$ is close to a diagonal element of either $\tilde{D}_1$ or $\tilde{D}_2$, then we can use reduction (15) to deflate without changing the structure of $\tilde{X}_1$. In the following we assume that no further such deflation is possible.

Let $\tilde{U}_1\tilde{\Lambda}_1\tilde{U}_1^T$ be the spectral decomposition of $\tilde{H}_1$. Then

$$\begin{aligned} T &= \begin{pmatrix} \tilde{X}_1 & \tilde{X}_2 \end{pmatrix} \begin{pmatrix} \tilde{U}_1\tilde{\Lambda}_1\tilde{U}_1^T & \\ & \tilde{\Lambda}_2 \end{pmatrix} \begin{pmatrix} \tilde{X}_1 & \tilde{X}_2 \end{pmatrix}^T + O(\tau\|T\|_2) \\ &= \begin{pmatrix} \tilde{X}_1\tilde{U}_1 & \tilde{X}_2 \end{pmatrix} \begin{pmatrix} \tilde{\Lambda}_1 & \\ & \tilde{\Lambda}_2 \end{pmatrix} \begin{pmatrix} \tilde{X}_1\tilde{U}_1 & \tilde{X}_2 \end{pmatrix}^T + O(\tau\|T\|_2). \end{aligned}$$

Thus $(\tilde{X}_1\tilde{U}_1, \tilde{X}_2)$ is an approximate eigenvector matrix of T. The matrix $\tilde{X}_1\tilde{U}_1$ can be computed while taking advantage of the block structure of $\tilde{X}_1$.

We refer to these deflations as *local* deflations since they are associated with individual subproblems of ADC.

4.3. Global deflation. To illustrate *global* deflation, we look at two levels of the dividing strategy (see (2)); for simplicity, we denote unimportant entries of T by x:

$$T = \begin{pmatrix} T_1 & \beta_{i+j+2}e_{i+j+1} & \\ \beta_{i+j+2}e_{i+j+1}^T & x & xe_1^T \\ & xe_1 & T_2 \end{pmatrix}$$
$$= \begin{pmatrix} T_{1,1} & xe_i & & & \\ xe_i^T & x & \beta_{i+2}e_1^T & & \\ & \beta_{i+2}e_1 & T_{1,2} & \beta_{i+j+2}e_j & \\ & & \beta_{i+j+2}e_j^T & x & xe_1^T \\ & & & xe_1 & T_2 \end{pmatrix},$$

where T_1, T_2, $T_{1,1}$, and $T_{1,2}$ are principle submatrices of T of dimensions $(i+j+1)\times(i+j+1)$, $(N-i-j-2)\times(N-i-j-2)$, $i\times i$, and $j\times j$, respectively.

Let $Q_{1,2}D_{1,2}Q_{1,2}^T$ be the spectral decomposition of $T_{1,2}$, and let $f_{1,2}^T$ and $l_{1,2}^T$ be the first and last rows of $Q_{1,2}$, respectively. Then

$$T = \begin{pmatrix} T_{1,1} & xe_i & & & \\ xe_i^T & x & \beta_{i+2}e_1^T & & \\ & \beta_{i+2}e_1 & Q_{1,2}D_{1,2}Q_{1,2}^T & \beta_{i+j+2}e_j & \\ & & \beta_{i+j+2}e_j^T & x & xe_1^T \\ & & & xe_1 & T_2 \end{pmatrix}$$
$$= Y \begin{pmatrix} T_{1,1} & xe_i & & & \\ xe_i^T & x & \beta_{i+2}f_{1,2}^T & & \\ & \beta_{i+2}f_{1,2} & D_{1,2} & \beta_{i+j+2}l_{1,2} & \\ & & \beta_{i+j+2}l_{1,2}^T & x & xe_1^T \\ & & & xe_1 & T_2 \end{pmatrix} Y^T, \tag{19}$$

where $Y = \operatorname{diag}(I_i,\ 1,\ Q_{1,2},\ 1,\ I_{N-i-j-2})$.

Let $\bar{d}_s$ be the sth diagonal element of $D_{1,2}$, and let $\bar{f}_s$ and $\bar{l}_s$ be the sth components of $f_{1,2}$ and $l_{1,2}$, respectively. Then, ignoring all zero components, the $(i+s+1)$st row of the middle matrix in (19) is $(\beta_{i+2}\bar{f}_s,\ \bar{d}_s,\ \beta_{i+j+2}\bar{l}_s)$. Thus if both $|\beta_{i+2}\bar{f}_s|$ and $|\beta_{i+j+2}\bar{l}_s|$ are small, then we can perturb them both to zero. $\bar{d}_s$ is an approximate eigenvalue of T and the $(i+s+1)$st column of Y is the corresponding approximate eigenvector. This eigenvalue and its eigenvector can be deflated from all subsequent subproblems. We call this *global* deflation.

Consider the deflation procedure for computing the spectral decomposition of T_1 in §4.2. If $|\beta_{i+2}\bar{f}_s|$ is small, then it can be perturbed to zero. This is a local deflation if only $|\beta_{i+2}\bar{f}_s|$ is small, and a global deflation if $|\beta_{i+j+2}\bar{l}_s|$ is also small.

5. Acceleration by the fast multipole method. Suppose that we want to evaluate the complex function

$$\Phi(\zeta) = \sum_{j=1}^{n} c_j \varphi(\zeta - \zeta_j) \tag{20}$$

at m points in the complex plane, where $\{c_j\}_{j=1}^n$ are constants and $\varphi(\zeta)$ is one of $\log(\zeta)$, $1/\zeta$, and $1/\zeta^2$. The direct computation takes $O(nm)$ time. But the *fast*

multipole method (FMM) of Carrier, Greengard, and Rokhlin [10], [16] takes only $O(n+m)$ time to approximate $\Phi(\zeta)$ at these points to a precision specified by the user.[6] In this section we briefly describe how FMM can be used to accelerate ADC. A more detailed description appears in [17] and [18] in the context of updating the singular value decomposition.

Let

$$H = \begin{pmatrix} \alpha & z^T \\ z & D \end{pmatrix},$$

where $D = \operatorname{diag}(d_2, \ldots, d_n)$ is an $(n-1) \times (n-1)$ matrix with $d_2 \le d_3 \le \cdots \le d_n$, $z = (z_2, \ldots, z_n)^T$ is a vector of length $n-1$, and α is a scalar. Let $U \Lambda U^T$ denote the spectral decomposition of H, with $U = (u_1, \ldots, u_n)$ and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$.

Consider computing $U^T q$ for a vector $q = (q_1, \ldots, q_n)^T$. By (5) in Lemma 3.1, the ith component $u_i^T q$ of $U^T q$ can be written as

$$u_i^T q = \frac{-q_1 + \Phi_1(\lambda_i)}{\sqrt{1 + \Phi_2(\lambda_i)}},$$

where

$$\Phi_1(\lambda) = \sum_{k=2}^{n} \frac{z_k q_k}{d_k - \lambda} \quad \text{and} \quad \Phi_2(\lambda) = \sum_{k=2}^{n} \frac{z_k^2}{(d_k - \lambda)^2}.$$

Thus we can compute $U^T q$ by evaluating $\Phi_1(\lambda)$ and $\Phi_2(\lambda)$ at n points. Since these functions are of the form (20), we can do this in $O(n)$ time using FMM. To achieve better efficiency, we modify FMM to take advantage of the fact that all the computations are real (see [17]–[19]).

Let T be an $N \times N$ symmetric tridiagonal matrix. When ADC is used to compute all the eigenvalues and eigenvectors, the main cost for each subproblem is in forming $\tilde{X}_1 U$ (see (17)), where $\tilde{X}_1$ is a column orthogonal matrix.[7] Each row of $\tilde{X}_1 U$ is of the form $q^T U = (U^T q)^T$ and there are $O(n)$ rows. Thus the cost of computing $\tilde{X}_1 U$ is $O(n^2)$ using FMM. There are $\log_2 N$ levels of recursion and 2^{k-1} subproblems at the kth level, each of size $O(N/2^k)$. Thus the cost at the kth level is $O(N^2/2^k)$ and the total time is $O(N^2)$.

We may also have to apply the eigenvector matrix of T to an orthogonal matrix Y, e.g., when T is obtained by reducing a dense matrix to tridiagonal form [14, pp. 419–420]. For each subproblem, we can apply the eigenvector matrix of the corresponding symmetric arrowhead matrix directly to Y. The cost for each subproblem is $O(Nn)$ using FMM, and there are $O(N/n)$ subproblems at each level. Thus the cost at each level is $O(N^2)$ and the total time is $O(N^2 \log_2 N)$.

When ADC is used to compute only the eigenvalues, the main cost for each subproblem is computing two vectors of the form $q^T U$, finding all the roots of the reformulated secular equation, and computing $\hat{z}$. We now show how to find all the eigenvalues of H and all the components of $\hat{z}$ in $O(n)$ time.

[6] The constant hidden in the $O(\cdot)$ notation depends on the logarithm of the precision.

[7] $\tilde{X}_1$ is also a block-structured matrix (see (18)). Here we view it as a dense matrix to simplify the presentation, even though FMM is more efficient when it exploits this structure.

A root-finder computes successive approximations to each eigenvalue λ_i. The main cost is in evaluating the function[8]

$$f(\lambda) = \lambda - \alpha + \sum_{j=2}^{n} \frac{z_j^2}{d_j - \lambda}.$$

To compute new approximations to all the eigenvalues simultaneously, we must evaluate $f(\lambda)$ at n points. Since this function is similar to the form (20), we can do this in $O(n)$ time using FMM. Thus, assuming that the number of approximations to each eigenvalue is bounded, all the eigenvalues of H can be computed in $O(n)$ time.

To compute $\hat{z}$, note that (7) can be rewritten as

$$|\hat{z}_i| = \sqrt{(d_i - \hat{\lambda}_1)\,(\hat{\lambda}_n - d_i)}\ \exp(\Phi_3(d_i))$$

where

$$\Phi_3(d_i) = \frac{1}{2}\left(\sum_{j=2}^{n-1} \log(\hat{\lambda}_j - d_i) - \sum_{j=2}^{i-1} \log(d_j - d_i) - \sum_{j=i}^{n-1} \log(d_{j+1} - d_i)\right).$$

Thus we can compute all the components of $\hat{z}$ in $O(n)$ time using FMM.

We have shown that when computing all the eigenvalues of T using ADC, we can solve each subproblem in $O(n)$ time. Since there are $O(N/n)$ subproblems at each level, the cost at each level is $O(N)$ and thus the total time is $O(N \log_2 N)$.

6. Numerical results. In this section we compare ADC with three other algorithms for solving the symmetric tridiagonal eigenproblem.

- B/II: Bisection with inverse iteration [21], [23] (subroutines `DSTEBZ` and `DSTEIN` from LAPACK [2]).
- CDC: Cuppen's divide-and-conquer algorithm [11], [12] (subroutine `TREEQL` from `netlib`).
- QR: The QR algorithm [8] (subroutine `DSTEQR` from LAPACK [2]).

ADC solves subproblems of size $N \leq 6$ using the QR algorithm. Since none of the test matrices is particularly large, FMM was not used.

All codes are written in FORTRAN and were compiled with optimization enabled. All computations were done on a SPARCstation/1 in double precision. The machine precision is $\epsilon = 1.1 \times 10^{-16}$.

Let $[\beta, \alpha_i, \beta]$ denote the $N \times N$ symmetric tridiagonal matrix with β on the off-diagonals and $\alpha_1, \ldots, \alpha_N$ on the diagonal. We use the following test matrices, most of which are taken from [21]:

- a random matrix, where the diagonal and off-diagonal elements are uniformly distributed in $[-1, 1]$;
- the Wilkinson matrix $W_N^+ = [1, w_i, 1]$, where $w_i = |(N+1)/2 - i|$;
- a glued Wilkinson matrix W_g^+: a 25×25 block matrix, where each diagonal block is the Wilkinson matrix W_k^+ and the off-diagonal elements $\beta_{i \times k+1} = g$, for $i = 1, \ldots, 24$;

[8] For simplicity we consider the original secular equation. See [17] and [18] for a version of FMM that can compute each $g_i(\mu)$ (and $\psi_i(\mu)$ and $\phi_i(\mu)$ and their derivatives) at a different point in $O(n)$ time. This is needed for the root-finders in [9], [15], [25] and to check the stopping criterion (9).

TABLE 1
Execution time.

Matrix type	Order N	Execution time (seconds) ADC	B/II	CDC	QR
Random	128	3.12	8.50	3.90	11.63
	256	10.43	33.35	14.88	85.86
	512	20.89	133.61	34.31	654.52
W_N^+	129	1.44	6.54	1.46	9.87
	257	3.43	25.00	3.74	66.86
	513	8.26	97.57	14.76	497.55
$W_{10^{-14}}^+$	125	0.63	5.88	*	5.12
	275	2.22	28.83	*	47.35
	525	8.23	121.84	*	353.41
$[1, 2, 1]$	128	3.91	8.49	3.72	10.21
	256	21.89	33.68	22.77	72.40
	512	138.79	144.43	213.01	545.05
$[1, \gamma_i, 1]$	128	4.48	8.54	6.66	10.17
	256	24.20	33.64	43.02	72.14
	512	148.95	135.48	302.06	544.65
$[1/100, 1+\gamma_i, 1/100]$	128	4.57	16.93	6.86	9.83
	256	24.45	102.81	43.01	70.65
	512	149.50	692.64	301.58	539.48

TABLE 2
Residual.

Matrix type	Order N	$\frac{\max_i \|T\hat{x}_i - \hat{\lambda}_i\hat{x}_i\|_2}{N\epsilon\|T\|_2}$ ADC	B/II	CDC	QR
Random	128	0.49×10^{-1}	0.11×10^{-1}	0.10×10^{1}	0.16×10^{0}
	256	0.43×10^{-1}	0.47×10^{-2}	0.74×10^{0}	0.82×10^{-1}
	512	0.23×10^{-1}	0.28×10^{-2}	0.13×10^{1}	0.69×10^{-1}
W_N^+	129	0.67×10^{-1}	0.86×10^{-2}	0.59×10^{0}	0.61×10^{-1}
	257	0.17×10^{-1}	0.39×10^{-2}	0.15×10^{0}	0.35×10^{-1}
	513	0.44×10^{-2}	0.21×10^{-2}	0.67×10^{0}	0.21×10^{-1}
$W_{10^{-14}}^+$	125	0.11×10^{0}	0.16×10^{0}	*	0.22×10^{0}
	275	0.27×10^{-1}	0.36×10^{-1}	*	0.11×10^{0}
	525	0.15×10^{-1}	0.66×10^{-1}	*	0.14×10^{0}
$[1, 2, 1]$	128	0.41×10^{-1}	0.70×10^{-2}	0.31×10^{-1}	0.52×10^{-1}
	256	0.22×10^{-1}	0.12×10^{-1}	0.25×10^{-1}	0.35×10^{-1}
	512	0.12×10^{-1}	0.35×10^{-2}	0.20×10^{-1}	0.25×10^{-1}
$[1, \gamma_i, 1]$	128	0.46×10^{-1}	0.16×10^{-1}	0.67×10^{-1}	0.90×10^{-1}
	256	0.23×10^{-1}	0.11×10^{-1}	0.47×10^{-1}	0.64×10^{-1}
	512	0.12×10^{-1}	0.79×10^{-2}	0.36×10^{-1}	0.47×10^{-1}
$[1/100, 1+\gamma_i, 1/100]$	128	0.22×10^{-1}	0.79×10^{-2}	0.11×10^{-1}	0.60×10^{-1}
	256	0.12×10^{-1}	0.42×10^{-2}	0.11×10^{-1}	0.38×10^{-1}
	512	0.59×10^{-2}	0.21×10^{-2}	0.64×10^{-2}	0.28×10^{-1}

TABLE 3
Orthogonality.

Matrix type	Order N	$\frac{\max_i \|X^T\hat{x}_i - e_i\|_2}{N\epsilon}$ ADC	B/II	CDC	QR
Random	128	0.94×10^{-1}	0.30×10^{0}	0.54×10^{-1}	0.59×10^{0}
	256	0.66×10^{-1}	0.86×10^{-1}	0.17×10^{0}	0.54×10^{0}
	512	0.35×10^{-1}	0.72×10^{-1}	0.30×10^{0}	0.47×10^{0}
W_N^+	129	0.78×10^{-1}	0.35×10^{-1}	0.54×10^{-1}	0.80×10^{0}
	257	0.39×10^{-1}	0.19×10^{-1}	0.89×10^{-1}	0.13×10^{1}
	513	0.19×10^{-1}	0.12×10^{-1}	0.72×10^{-1}	0.13×10^{1}
$W_{10^{-14}}^+$	125	0.64×10^{-1}	0.56×10^{-1}	*	0.38×10^{0}
	275	0.33×10^{-1}	0.16×10^{0}	*	0.31×10^{0}
	525	0.20×10^{-1}	0.34×10^{-1}	*	0.32×10^{0}
$[1, 2, 1]$	128	0.70×10^{-1}	0.78×10^{0}	0.36×10^{0}	0.13×10^{0}
	256	0.47×10^{-1}	0.35×10^{0}	0.18×10^{0}	0.70×10^{-1}
	512	0.39×10^{-1}	0.21×10^{0}	0.14×10^{0}	0.44×10^{-1}
$[1, \gamma_i, 1]$	128	0.62×10^{-1}	0.92×10^{0}	0.17×10^{0}	0.12×10^{0}
	256	0.49×10^{-1}	0.12×10^{1}	0.21×10^{0}	0.62×10^{-1}
	512	0.35×10^{-1}	0.55×10^{0}	0.76×10^{0}	0.48×10^{-1}
$[1/100, 1 + \gamma_i, 1/100]$	128	0.78×10^{-1}	0.35×10^{-1}	0.91×10^{-1}	0.12×10^{0}
	256	0.62×10^{-1}	0.23×10^{-1}	0.11×10^{0}	0.78×10^{-1}
	512	0.61×10^{-1}	0.21×10^{-1}	0.93×10^{-1}	0.40×10^{-1}

- the Toeplitz matrix $[1, 2, 1]$;
- the matrix $[1, \gamma_i, 1]$, where $\gamma_i = i \times 10^{-6}$;
- the matrix $[1/100, 1 + \gamma_i, 1/100]$, where $\gamma_i = i \times 10^{-6}$;
- the test matrices of types 8–21 in the LAPACK test suite.[9]

W_N^+ has pairs of close eigenvalues, W_g^+ has clusters of 50 close eigenvalues, $[1, 2, 1]$ has no close eigenvalues, $[1, \alpha_i, 1]$ and $[1/100, 1 + \alpha_i, 1/100]$ do not deflate, and $[1/100, 1 + \alpha_i, 1/100]$ forces B/II to reorthogonalize all of the eigenvectors.

The numerical results are presented in Tables 1–4. An asterisk means that the algorithm failed. Since the numerical results in Tables 1–3 suggest that CDC and QR are not as competitive, we only compare ADC with B/II for the LAPACK test matrices (see Table 4).

The residual and orthogonality measures for ADC are always comparable with those for QR and B/II, and ADC is roughly twice as fast as CDC for large matrices, due to the differences in how deflation is implemented (see §4.2). In most cases ADC is faster than the others by a considerable margin and in many cases is more than 5–10 times faster. When ADC is slower than B/II (by at most 10%), the matrix size is large ($N \approx 512$) and there are few deflations. These are cases where FMM would make ADC significantly faster.

Acknowledgment. The results in §3 were first announced in a preprint of [20]. Using the ideas there, Borges and Gragg [7] independently derived similar results.

[9] Types 1–7 are all diagonal matrices.

TABLE 4
LAPACK test matrices.

Matrix type	Order N	Execution time		$\frac{\max_i \|T\hat{x}_i - \hat{\lambda}_i\hat{x}_i\|_2}{N\epsilon\|T\|_2}$		$\frac{\max_i \|X^T\hat{x}_i - e_i\|_2}{N\epsilon}$	
		ADC	B/II	ADC	B/II	ADC	B/II
8	128	4.64	8.68	0.47×10^{-1}	0.82×10^{-2}	0.86×10^{-1}	0.16×10^{0}
	256	24.27	33.63	0.27×10^{-1}	0.54×10^{-2}	0.66×10^{-1}	0.25×10^{0}
	512	140.04	133.58	0.17×10^{-1}	0.30×10^{-2}	0.47×10^{-1}	0.21×10^{0}
9	128	2.32	12.66	0.13×10^{0}	0.14×10^{-1}	0.12×10^{0}	0.20×10^{-1}
	256	9.37	76.99	0.28×10^{-1}	0.30×10^{-2}	0.53×10^{-1}	0.27×10^{-1}
	512	44.62	517.45	0.12×10^{-1}	0.92×10^{-1}	0.33×10^{-1}	0.84×10^{-1}
10	128	0.01	12.36	0.11×10^{-1}	0.10×10^{-1}	0.14×10^{-1}	0.70×10^{0}
	256	0.04	84.54	0.45×10^{-2}	0.70×10^{-2}	0.78×10^{-2}	0.52×10^{-1}
	512	0.17	613.86	0.38×10^{-2}	0.21×10^{-2}	0.78×10^{-2}	0.21×10^{-1}
11	128	5.24	8.64	0.45×10^{-1}	0.11×10^{-1}	0.62×10^{-1}	0.22×10^{0}
	256	25.88	33.52	0.28×10^{-1}	0.53×10^{-2}	0.55×10^{-1}	0.17×10^{0}
	512	144.37	132.32	0.17×10^{-1}	0.29×10^{-2}	0.41×10^{-1}	0.20×10^{0}
12	128	4.54	8.75	0.46×10^{-1}	0.85×10^{-2}	0.86×10^{-1}	0.19×10^{0}
	256	24.44	33.94	0.25×10^{-1}	0.54×10^{-2}	0.51×10^{-1}	0.30×10^{0}
	512	141.57	133.83	0.16×10^{-1}	0.31×10^{-2}	0.47×10^{-1}	0.20×10^{0}
13	128	4.61	8.66	0.43×10^{-1}	0.69×10^{-2}	0.62×10^{-1}	0.33×10^{0}
	256	23.49	33.53	0.20×10^{-1}	0.53×10^{-2}	0.70×10^{-1}	0.15×10^{0}
	512	131.57	132.03	0.13×10^{-1}	0.24×10^{-2}	0.41×10^{-1}	0.25×10^{0}
14	128	5.16	8.61	0.46×10^{-1}	0.79×10^{-2}	0.12×10^{0}	0.28×10^{0}
	256	24.88	33.55	0.24×10^{-1}	0.50×10^{-2}	0.51×10^{-1}	0.39×10^{0}
	512	134.86	131.96	0.12×10^{-1}	0.24×10^{-2}	0.43×10^{-1}	0.29×10^{0}
15	128	4.50	8.71	0.49×10^{-1}	0.73×10^{-2}	0.78×10^{-1}	0.12×10^{0}
	256	23.44	33.99	0.24×10^{-1}	0.45×10^{-2}	0.41×10^{-1}	0.17×10^{0}
	512	131.71	133.53	0.13×10^{-1}	0.26×10^{-2}	0.41×10^{-1}	0.13×10^{0}
16	128	4.54	8.68	0.22×10^{-1}	0.87×10^{-2}	0.11×10^{0}	0.11×10^{0}
	256	24.18	33.73	0.13×10^{-1}	0.41×10^{-2}	0.64×10^{-1}	0.93×10^{-1}
	512	139.29	132.47	0.14×10^{-1}	0.19×10^{-2}	0.35×10^{-1}	0.15×10^{0}
17	128	2.67	12.88	0.30×10^{-1}	0.30×10^{-2}	0.11×10^{0}	0.58×10^{-1}
	256	11.78	76.55	0.38×10^{-1}	0.31×10^{-2}	0.51×10^{-1}	0.82×10^{-1}
	512	63.23	521.70	0.16×10^{-1}	0.17×10^{-2}	0.33×10^{-1}	0.29×10^{-1}
18	128	0.01	12.34	0.13×10^{-1}	0.78×10^{-2}	0.16×10^{-1}	0.39×10^{-1}
	256	0.04	83.95	0.98×10^{-2}	0.39×10^{-2}	0.59×10^{-2}	0.29×10^{-1}
	512	0.17	614.26	0.44×10^{-2}	0.19×10^{-2}	0.20×10^{-2}	0.16×10^{0}
19	128	5.08	8.58	0.25×10^{-1}	0.79×10^{-2}	0.90×10^{-1}	0.92×10^{-1}
	256	25.47	33.32	0.14×10^{-1}	0.40×10^{-2}	0.70×10^{-1}	0.16×10^{0}
	512	142.16	131.26	0.13×10^{-1}	0.20×10^{-2}	0.31×10^{-1}	0.97×10^{-1}
20	128	4.46	8.68	0.20×10^{-1}	0.75×10^{-2}	0.62×10^{-1}	0.10×10^{0}
	256	24.12	33.72	0.13×10^{-1}	0.47×10^{-2}	0.51×10^{-1}	0.17×10^{0}
	512	139.29	132.75	0.13×10^{-1}	0.21×10^{-2}	0.33×10^{-1}	0.12×10^{0}
21	128	1.85	12.86	0.45×10^{-1}	0.34×10^{-2}	0.70×10^{-1}	0.45×10^{-1}
	256	6.26	75.81	0.38×10^{-1}	0.27×10^{-2}	0.35×10^{-1}	0.16×10^{-1}
	512	21.07	517.17	0.15×10^{-1}	0.12×10^{-2}	0.31×10^{-1}	0.20×10^{-1}

REFERENCES

[1] L. ADAMS AND P. ARBENZ, *Towards a divide and conquer algorithm for the real nonsymmetric eigenvalue problem*, Tech. Report No. 91-8, Department of Applied Mathematics, University of Washington, Seattle, Aug. 1991.

[2] E. ANDERSON, Z. BAI, C. BISCHOF, J. DEMMEL, J. DONGARRA, J. DU CROZ, A. GREENBAUM, S. HAMMARLING, A. MCKENNEY, S. OSTROUCHOV, AND D. SORENSEN, *LAPACK Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.

[3] P. ARBENZ, *Divide-and-conquer algorithms for the bandsymmetric eigenvalue problem*, Parallel Comput., 18 (1992), pp. 1105–1128.

[4] P. ARBENZ AND G. H. GOLUB, *QR-like algorithms for symmetric arrow matrices*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 655–658.

[5] J. L. BARLOW, *Error analysis of update methods for the symmetric eigenvalue problem*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 598–618.

[6] D. BOLEY AND G. H. GOLUB, *Inverse eigenvalue problems for band matrices*, in Numerical Analysis, Proceedings, Biennial Conference, Dundee 1977, G. A. Watson, ed., Vol. 630, Lecture Notes in Mathematics, Springer-Verlag, New York, 1977, pp. 23–31.

[7] C. F. BORGES AND W. B. GRAGG, *A parallel divide and conquer algorithm for the generalized real symmetric definite tridiagonal eigenproblem*, in Numerical Linear Algebra and Scientific Computing, L. Reichel, A. Ruttan, and R. S. Varga, eds., de Gruyter, Berlin, 1993, pp. 10–28.

[8] H. BOWDLER, R. S. MARTIN, C. REINSCH, AND J. WILKINSON, *The QR and QL algorithms for symmetric matrices*, Numer. Math., 11 (1968), pp. 293–306.

[9] J. R. BUNCH, C. P. NIELSEN, AND D. C. SORENSEN, *Rank-one modification of the symmetric eigenproblem*, Numer. Math., 31 (1978), pp. 31–48.

[10] J. CARRIER, L. GREENGARD, AND V. ROKHLIN, *A fast adaptive multipole algorithm for particle simulations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 669–686.

[11] J. J. M. CUPPEN, *A divide and conquer method for the symmetric tridiagonal eigenproblem*, Numer. Math., 36 (1981), pp. 177–195.

[12] J. J. DONGARRA AND D. C. SORENSEN, *A fully parallel algorithm for the symmetric eigenvalue problem*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. s139–s154.

[13] K. GATES, *Divide and Conquer Methods for the Symmetric Tridiagonal Eigenproblem*, Ph.D. thesis, University of Washington, Seattle, 1991.

[14] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, second ed., 1989.

[15] W. B. GRAGG, J. R. THORNTON, AND D. D. WARNER, *Parallel divide and conquer algorithms for the symmetric tridiagonal eigenproblem and bidiagonal singular value problem*, in Modelling and Simulation, W. G. Vogt and M. H. Mickle, eds., Vol. 23, Part 1, University of Pittsburgh School of Engineering, Pittsburgh, 1992, pp. 49–56.

[16] L. GREENGARD AND V. ROKHLIN, *A fast algorithm for particle simulations*, J. Comput. Phys., 73 (1987), pp. 325–348.

[17] M. GU, *Studies in Numerical Linear Algebra*, Ph.D. thesis, Department of Computer Science, Yale University, New Haven, CT, 1993.

[18] M. GU AND S. C. EISENSTAT, *A fast algorithm for updating the singular value decomposition*, manuscript.

[19] ———, *A fast divide-and-conquer method for the symmetric tridiagonal eigenproblem.* Presented at the Fourth SIAM Conference on Applied Linear Algebra, Minneapolis, MN, Sept. 1991.

[20] ———, *A stable and efficient algorithm for the rank-one modification of the symmetric eigenproblem*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1266–1276.

[21] E. R. JESSUP, *Parallel Solution of the Symmetric Tridiagonal Eigenproblem*, Ph.D. thesis, Department of Computer Science, Yale University, New Haven, CT, 1989.

[22] ———, *A case against a divide and conquer approach to the nonsymmetric eigenvalue problem*, Applied Numerical Mathematics, 12 (1993), pp. 403–420.

[23] E. R. JESSUP AND I. C. F. IPSEN, *Improving the accuracy of inverse iteration*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 550–572.

[24] W. KAHAN, *Rank-1 perturbed diagonal's eigensystem*, manuscript, 1989.

[25] R.-C. LI, *Solving secular equations stably and efficiently*, Working paper, Department of Mathematics, University of California at Berkeley, Oct. 1992.

[26] D. P. O'Leary and G. W. Stewart, *Computing the eigenvalues and eigenvectors of symmetric arrowhead matrices*, J. Comput. Phys., 90 (1990), pp. 497–505.

[27] B. N. Parlett and B. Nour-Omid, *The use of a refined error bound when updating eigenvalues of tridiagonals*, Linear Algebra Appl., 68 (1985), pp. 179–219.

[28] W. E. Shreve and M. R. Stabnow, *An eigenvalue algorithm for symmetric bordered diagonal matrices*, in Current Trends in Matrix Theory, F. Uhlig and R. Grone, eds., Elsevier Science Publishing Co., Inc., New York, 1987, pp. 339–346.

[29] D. C. Sorensen and P. T. P. Tang, *On the orthogonality of eigenvectors computed by divide-and-conquer techniques*, SIAM J. Numer. Anal., 28 (1991), pp. 1752–1775.

[30] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 192–206, January 1995

013

ALGEBRAIC ANALYSIS OF THE HIERARCHICAL BASIS PRECONDITIONER*

HOWARD C. ELMAN† AND XUEJUN ZHANG‡

Abstract. The use of the hierarchical basis in finite element discretizations of two-dimensional elliptic partial differential equations produces matrices with condition numbers of order $O((\log h^{-1})^2)$. Standard proofs of this result are functional analytic in style. In this paper, it is shown that for uniform grids the result can be obtained using a purely linear algebraic argument.

Key words. hierarchical basis, preconditioning, condition number, elliptic problems

AMS subject classifications. primary 65F10, 65N20; secondary 15A06

1. Introduction.

Consider the two-dimensional elliptic partial differential equation

$$-\Delta u = f \text{ on } \Omega = (0,1) \times (0,1), \qquad u = 0 \text{ on } \partial\Omega. \tag{1}$$

When this problem is discretized using standard finite element techniques, e.g., the Galerkin method with piecewise linear nodal basis functions on a uniform mesh of triangles of width h, the result is a linear system of equations

$$By = c, \tag{2}$$

where the condition number of B is of order $O(h^{-2})$ [1], [6], [7].

An alternative discretization strategy is to use a "hierarchical basis," in which the basis functions are defined in a hierarchical manner. Given a basis for a discretization on a (coarse) grid of width $2h$, the basis for the grid of width h is determined by augmenting the coarse grid basis with functions centered at nodes in the new grid and having support on the fine grid elements. In the resulting linear system

$$Ax = b, \tag{3}$$

the stiffness matrix A has condition number of order $O((\log h^{-1})^2)$ [2], [8].

In this paper we present a purely algebraic analysis of the condition number of the matrix A, for problems derived from uniform meshes. We derive an upper bound on the largest eigenvalue of A that is independent of the mesh size by explicitly examining the nonzero structure of A. For a lower bound on the smallest eigenvalue, we use the fact that $A = S^T B S$ where S represents a change of basis from hierarchical basis to nodal basis. The smallest eigenvalue is the inverse of the maximum of the Rayleigh quotient $(v, Qv)/(v, Bv)$, where $Q = (SS^T)^{-1}$ and (v, w) denotes the Euclidean inner product. A loose statement of the derivation of the bound is as follows. Suppose $j =$

* Received by the editors September 14, 1992; accepted for publication (in revised form) by A. Greenbaum, August 17, 1993.

† Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, Maryland 20742 (elman@cs.umd.edu). The work of this author was supported by U. S. Army Research Office grant DAAL-0392-G-0016 and by National Science Foundation grants ASC-8958544 and CCR-8818340.

‡ Department of Computer Science, University of Maryland, College Park, Maryland 20742 (xzhang@cs.umd.edu). The work of this author was supported by National Science Foundation grant ASC-8958544.

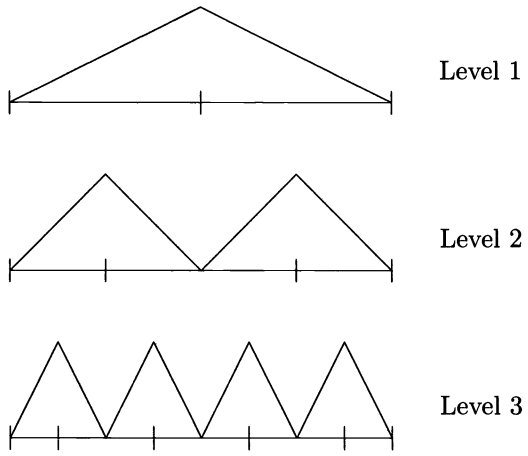


FIG. 1. *Hierarchical basis functions for* $h = 1/8$.

$\log h^{-1}$, so that a sequence of j hierarchical refinements produces the discretization mesh. We show that the operator Q satisfies

$$Q \approx B + B_1 + \cdots + B_{j-1}, \tag{4}$$

where B_i can be thought of as a discrete Laplacian operator restricted to the grid at refinement level i. The bound is obtained by showing that $(v, B_i v)/(v, Bv) \leq c(j-i)$. A related analysis for multigrid methods is given in [4].

An outline of the paper is as follows. In §2, we present a condition number analysis for the coefficient matrix arising from one-dimensional problems discretized using a hierarchical basis. The two-dimensional analysis is presented in §3. Unless otherwise specified, the natural nodal ordering is used for both A and B, i.e., degrees of freedom are ordered using the natural ordering of the underlying fine grid.

2. Analysis of the one-dimensional problem. Consider the one-dimensional problem

$$-u'' = f \text{ on } (0,1), \qquad u(0) = u(1) = 0. \tag{5}$$

Examples of the hierarchical basis functions for a mesh of width $h = 1/8$ are shown in Fig. 1. It is straightforward to show [9] that these functions are orthogonal with respect to the energy inner product

$$a(u,v) = \int_0^1 u'v'.$$

Consequently, the global stiffness matrix A is a diagonal matrix, and it can be transformed via a symmetric scaling TAT into the identity matrix, where T is also a diagonal matrix. Equivalently, the use of appropriately scaled hierarchical basis functions produces a perfectly conditioned global stiffness matrix.

To motivate the two-dimensional analysis, we present an algebraic derivation of this result. Let B denote the coefficient matrix for the standard nodal basis on a

uniform mesh of width $h = 1/2^j$. B is a tridiagonal matrix of the form

$$\begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}, \tag{6}$$

of order $n = 2^j - 1$. It is shown in [8] that the hierarchical basis matrix is

$$A = S^T B S, \tag{7}$$

where

$$S = S_{j-1} \cdots S_2 S_1,$$

and the computation $\beta \leftarrow S_i \alpha$ represents a change of basis in which nodal basis functions on the mesh of width $1/2^i$ are replaced by nodal basis functions on the mesh of width $1/2^{i+1}$. These matrices have the form

$$S_i = I + R_i,$$

where for $1 \le i \le j-1$, R_i is defined by

$$[R_i]_{rs} = \begin{cases} .5 & \text{if } s \text{ is divisible by } 2^{j-i} \quad \text{and} \quad r = s \pm \tfrac{1}{2} 2^{j-i}, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Examples of $\{S_i\}$ in the case $j = 4$ ($n = 15$) are shown below.

$$S_1 = \begin{pmatrix} 1 & & & & & & & & & & & & & & \\ & 1 & & & & & & & & & & & & & \\ & & 1 & & & & & & & & & & & & \\ & & & 1 & & & .5 & & & & & & & & \\ & & & & 1 & & & & & & & & & & \\ & & & & & 1 & & & & & & & & & \\ & & & & & & 1 & & & & & & & & \\ & & & & & & & 1 & & & & & & & \\ & & & & & & & & 1 & & & & & & \\ & & & & & & .5 & & & 1 & & & & & \\ & & & & & & & & & & 1 & & & & \\ & & & & & & & & & & & 1 & & & \\ & & & & & & & & & & & & 1 & & \\ & & & & & & & & & & & & & 1 & \\ & & & & & & & & & & & & & & 1 \end{pmatrix},$$

$$S_2 = \begin{pmatrix} 1 & & & & & & & & & & & & & & \\ & 1 & & .5 & & & & & & & & & & & \\ & & 1 & & & & & & & & & & & & \\ & & & 1 & & & & & & & & & & & \\ & & & .5 & 1 & & .5 & & & & & & & & \\ & & & & & 1 & & & & & & & & & \\ & & & & & & 1 & & & & & & & & \\ & & & & & & .5 & 1 & & .5 & & & & & \\ & & & & & & & & 1 & & & & & & \\ & & & & & & & & & 1 & & & & & \\ & & & & & & & & & .5 & 1 & & .5 & & \\ & & & & & & & & & & & 1 & & & \\ & & & & & & & & & & & & 1 & & \\ & & & & & & & & & & & .5 & & 1 & \\ & & & & & & & & & & & & & & 1 \end{pmatrix},$$

$$S_3 = \begin{pmatrix} 1 & .5 & & & & & & & & & & & & & \\ & 1 & & & & & & & & & & & & & \\ & .5 & 1 & .5 & & & & & & & & & & & \\ & & & 1 & & & & & & & & & & & \\ & & & .5 & 1 & .5 & & & & & & & & & \\ & & & & & 1 & & & & & & & & & \\ & & & & & .5 & 1 & .5 & & & & & & & \\ & & & & & & & 1 & & & & & & & \\ & & & & & & & .5 & 1 & .5 & & & & & \\ & & & & & & & & & 1 & & & & & \\ & & & & & & & & & .5 & 1 & .5 & & & \\ & & & & & & & & & & & 1 & & & \\ & & & & & & & & & & & .5 & 1 & .5 & \\ & & & & & & & & & & & & & 1 & \\ & & & & & & & & & & & & & .5 & 1 \end{pmatrix}.$$

From (7), the matrix $Q = (SS^T)^{-1}$ can be viewed as a preconditioner for B, and bounds on the condition number of A can be obtained from bounds on the Rayleigh quotient

$$\frac{(v, Bv)}{(v, Qv)} = \frac{(v, Bv)}{(v, (SS^T)^{-1} v)}. \tag{9}$$

We establish such bounds as follows.

LEMMA 2.1. *The off-diagonal parts of* $\{S_i\}$ *satisfy*

$$R_{i_1}R_{i_2} = 0 \text{ for } i_1 \le i_2, \qquad R_{i_1}^T R_{i_2} = 0 \text{ for } i_1 \ne i_2.$$

Proof. By (8), any column index s for which R_{i_1} contains a nonzero entry has the form $s = k \times 2^{j-i_1}$, and any row index for which R_{i_2} contains a nonzero entry has the form $r = \left(l \pm \frac{1}{2}\right) \times 2^{j-i_2}$. Therefore,

$$\frac{s}{r} = \frac{k \times 2^{i_2 - i_1}}{l \pm \frac{1}{2}}.$$

For $i_1 \le i_2$, the numerator of this expression is an integer and the denominator is not, which implies $r \ne s$. Consequently, there cannot be any nonzero entries in $R_{i_1}R_{i_2}$.

Similarly, if r_1 is a column index for which $R_{i_1}^T$ contains a nonzero and r_2 is a row index for which R_{i_2} contains a nonzero, then

$$\frac{r_1}{r_2} = \frac{(k \pm \frac{1}{2}) \times 2^{j-i_1}}{(l \pm \frac{1}{2}) \times 2^{j-i_2}} = \left(\frac{k \pm \frac{1}{2}}{l \pm \frac{1}{2}}\right) \times 2^{i_2 - i_1}.$$

If $i_1 \ne i_2$, then this expression cannot equal 1, so that $R_{i_1}^T R_{i_2} = 0$. □

LEMMA 2.2. *The change of basis operator* S *satisfies*

$$S^{-1} = I - (R_1 + \cdots + R_{j-1}), \tag{10}$$

and the preconditioning operator satisfies

$$(SS^T)^{-1} = I - [(R_1 + R_1^T) + \cdots + (R_{j-1} + R_{j-1}^T)] + R_1^T R_1 + \cdots + R_{j-1}^T R_{j-1}. \tag{11}$$

Proof. It follows from the first equality of Lemma 2.1 that $S_i^{-1} = I - R_i$. Therefore, using this equality again, we have

$$S_1^{-1}S_2^{-1} = (I - R_1)(I - R_2) = I - (R_1 + R_2).$$

Assertion (10), for $S^{-1} = S_1^{-1}S_2^{-1}\cdots S_{j-1}^{-1}$, then follows using a straightforward inductive argument. Assertion (11) follows immediately from (10) and the second equality of Lemma 2.1. □

For $1 \le i \le j$, let D_i denote the identity matrix restricted to level i; that is, $[D_i]_{rr} = 1$ if r is divisible by 2^{j-i}, and $[D_i]_{rs} = 0$ otherwise. Let C_i be defined by

$$[C_i]_{rs} = \begin{cases} 1 & \text{if } r \text{ is divisible by } 2^{j-i} \quad \text{and} \quad s = r \pm 2^{j-i}, \\ 0 & \text{otherwise.} \end{cases}$$

Note that

$$R_i + R_i^T = \frac{1}{2}C_{i+1} \quad \text{and} \quad R_i^T R_i = \frac{1}{4}(2D_i + C_i). \tag{12}$$

Let $B_i = 2D_i - C_i$. This matrix has the form of a discrete Laplacian operator restricted to the grid of level i, and in particular, $B_j = B$ of (6). Combining (11) and (12) gives

$$(SS^T)^{-1} = \frac{1}{4}(B_1 + \cdots + B_{j-1}) + \frac{1}{2}B_j. \tag{13}$$

That is, the preconditioner has the form $Q = \frac{1}{2}B + R$ where $R = \frac{1}{4}\sum_{i=1}^{j-1} B_i$. R is positive semidefinite, so that an upper bound for (9) is obtained from

$$\frac{(v,Bv)}{(v,Qv)} = 2\left(\frac{(v,Qv)}{(v,Qv)} - \frac{(v,Rv)}{(v,Qv)}\right) \leq 2. \tag{14}$$

For the lower bound, note that

$$\frac{(v,Qv)}{(v,Bv)} = \frac{1}{2}\frac{(v,Bv)}{(v,Bv)} + \frac{1}{4}\sum_{i=1}^{j-1}\frac{(v,B_iv)}{(v,Bv)}. \tag{15}$$

The following result determines an upper bound for $(v,Qv)/(v,Bv)$.

LEMMA 2.3. *For $1 \leq i \leq j$, the generalized eigenvalue problem $B_iv = \lambda Bv$ has eigenvalues $\lambda = 2^{j-i}$ of multiplicity $2^i - 1$ and $\lambda = 0$ of multiplicity $2^j - 2^i$.*

Proof. For each i, B_i has order $2^j - 1$ and rank $2^i - 1$, so that zero is an eigenvalue of multiplicity $2^j - 2^i$. Now consider the nonzero eigenvalues. The case $i = j$ is trivial. For $i = j - 1$, the generalized eigenvalue problem can be stated as

$$\begin{aligned} -v_{s-2} + 2v_s - v_{s+2} &= \lambda(-v_{s-1} + 2v_s - v_{s+1}) \quad \text{for } s \text{ even}, \\ 0 &= \lambda(-v_{s-1} + 2v_s - v_{s+1}) \quad \text{for } s \text{ odd}, \end{aligned}$$

where $1 \leq s \leq n$. Therefore, for all even s,

$$v_{s-1} = \frac{v_{s-2} + v_s}{2}, \qquad v_{s+1} = \frac{v_s + v_{s+2}}{2},$$

and substitution of these expressions into the equation centered at v_s gives

$$-v_{s-2} + 2v_s - v_{s+2} = \frac{\lambda}{2}(-v_{s-2} + 2v_s - v_{s+2}).$$

Consequently, the only nonzero generalized eigenvalue for $i = j - 1$ is $\lambda = 2$. An identical argument shows that for $j = i - 2$, when s is divisible by four,

$$-v_{s-4} + 2v_s - v_{s+4} = \frac{\lambda}{4}(-v_{s-4} + 2v_s - v_{s+4}),$$

giving $\lambda = 4$ as the only nonzero eigenvalue. More generally, for $i = j - k$, when s is divisible by 2^k, we have

$$-v_{s-2^k} + 2v_s - v_{s+2^k} = \frac{\lambda}{2^k}(-v_{s-2^k} + 2v_s - v_{s+2^k}),$$

and $\lambda = 2^k$. □

This result implies that

$$\frac{(v,Qv)}{(v,Bv)} \leq \frac{1}{2} + \frac{1}{4}\sum_{i=1}^{j-1} 2^{j-i} = 2^{j-2}.$$

Combining this with (14) and (15) gives

$$\frac{1}{2^{j-2}} \leq \frac{(v,Bv)}{(v,Qv)} \leq 2.$$

Consequently, we have the following condition number bound for the unscaled one-dimensional hierarchical basis.

THEOREM 2.4. *The condition number of the stiffness matrix A derived from the hierarchical basis is bounded by $(n+1)/2$.*

Now let A represent the coefficient matrix derived from a scaled hierarchical basis,

$$A = (ST)^T B(ST),$$

where $T = T_j \cdots T_2 T_1$, and T_i is a diagonal matrix associated with a scaling of the basis functions of level i,

$$[T_i]_{rr} = \begin{cases} \tau_i & \text{if } r/2^{j-i} \text{ is an odd integer,} \\ 1 & \text{otherwise.} \end{cases}$$

As in the derivation of (13), it can be shown that

$$Q = [(ST)(ST)^T]^{-1} = \sum_{i=1}^{j-1} \frac{1}{2}\left(\tau_i^{-2} - \frac{1}{2}\tau_{i+1}^{-2}\right) B_i + \frac{\tau_j^{-2}}{2} B_j. \tag{16}$$

When $\tau_i^2 = 2\tau_{i+1}^2$, the first $j-1$ terms of (16) are zero, giving the following result.

THEOREM 2.5. *If the scalings of the hierarchical basis satisfy $\tau_i^2 = 2\tau_{i+1}^2$, $\tau_j = 1/\sqrt{2}$, then $[(ST)(ST)^T]^{-1} = B$ and the coefficient matrix A derived from the hierarchical basis is the identity matrix.*

3. Analysis of the two-dimensional problem. We will consider a sequence of two-dimensional hierarchical meshes each consisting of a set of triangles of width $h_i = 1/2^i$, with nodes $x_{kl}^{(i)} = (kh_i, lh_i)$, $1 \le k, l \le n_i = 2^i - 1$. At level i, $1 \le i \le j$, there is a set of piecewise linear basis functions $\{\phi_{kl}^{(i)}\}$ whose support is defined by

$$\phi_{kl}^{(i)}(x) = \begin{cases} 1 & \text{if } x = x_{kl}^{(i)}, \\ 0 & \text{if } x = x_{k\pm1,l}^{(i)} \text{ or } x = x_{k,l\pm1}^{(i)} \text{ or } x = x_{k-1,l-1}^{(i)} \text{ or } x = x_{k+1,l+1}^{(i)}. \end{cases}$$

Examples of the supports of such functions are shown in Fig. 2. The hierarchical basis of level i consists of the hierarchical basis of level $i-1$, together with those basis functions $\phi_{kl}^{(i)}$ associated with mesh points $x_{kl}^{(i)}$ not in the mesh of level $i-1$.

The main result of the paper is as follows.

THEOREM 3.1. *The condition number of the coefficient matrix derived from the hierarchical basis of level j is of order $O((\log h_j^{-1})^2)$.*

The proof consists of deriving upper and lower bounds on the eigenvalues of A.

3.1. Upper bound. The following results will be used in the derivation of the upper bound. For the moment, let A denote an arbitrary matrix.

LEMMA 3.2. *Suppose A is blocked into submatrices as $A = [A_{rs}]$, and let $A^{(b)} = [\|A_{rs}\|_2]$, the matrix whose entries are the norms of the blocks of A. Then $\|A\|_2 \le \|A^{(b)}\|_2$.*

Proof. Let

$$v = \begin{pmatrix} v^{(1)} \\ \vdots \\ v^{(r_{\max})} \end{pmatrix}, \qquad w = \begin{pmatrix} w^{(1)} \\ \vdots \\ w^{(s_{\max})} \end{pmatrix},$$

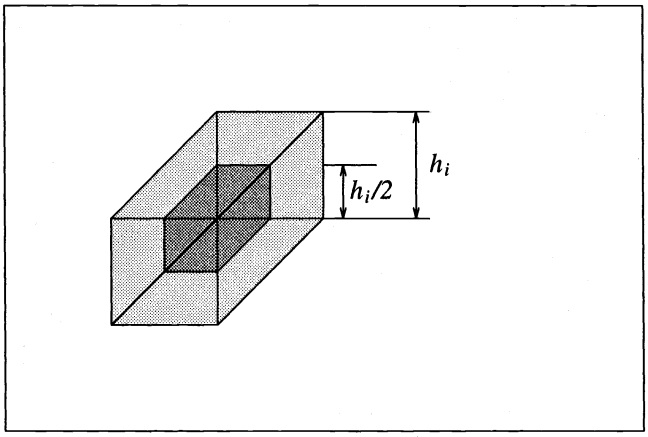


FIG. 2. *Supports of basis functions for meshes of width* h_i *(light and dark shading) and* $h_{i+1} = h_i/2$ *(dark shading).*

where the sizes of the component vectors $v^{(r)}$ and $w^{(s)}$ are the same as the numbers of rows and columns, respectively, of A_{rs}. Using

$$\|A\|_2 = \max_{v,w\neq 0} \frac{|(v, Aw)|}{\|v\|_2\|w\|_2},$$

we have

$$|(v, Aw)| = \left|\sum_{r,s}(v^{(r)}, A_{rs}w^{(s)})\right| \leq \sum_{r,s}\|A_{rs}\|_2\,\|v^{(r)}\|_2\,\|w^{(s)}\|_2 \leq \|A^{(b)}\|_2\,\|v\|_2\,\|w\|_2\,.$$

Thus, $\|A\|_2 \leq \|A^{(b)}\|_2$. □

For $A = [a_{rs}]$, let $\|A_{r,*}\|_2$ denote the Euclidean norm of the vector consisting of the rth row of A.

LEMMA 3.3. *If* $A = [a_{rs}]$ *has at most* M_{col} *nonzero elements in any column, then*

$$\|A\|_2 \leq \sqrt{M_{\text{col}}}\,\max_r \|A_{r,*}\|_2.$$

Proof. Let $\varepsilon_{rs} = |\text{sign}(a_{rs})|$, i.e., $\varepsilon_{rs} = 1$ if $a_{rs} \neq 0$ and 0 otherwise. Then the number of nonzeros in column s is $\sum_r \varepsilon_{rs}$, and

$$\sum_r \varepsilon_{rs} \leq M_{\text{col}}\,, \qquad \varepsilon_{rs}^2 = \varepsilon_{rs}\,. \tag{17}$$

Using the Cauchy–Schwarz inequality and (17), we have

$$\begin{aligned}\|Av\|_2^2 &= \sum_r \left(\sum_s a_{rs}\varepsilon_{rs}v_s\right)^2 \leq \sum_r \left(\sum_s a_{rs}^2\right)\left(\sum_s \varepsilon_{rs}^2 v_s^2\right)\\ &\leq \max_r \|A_{r,*}\|_2^2 \sum_r\sum_s \varepsilon_{rs}v_s^2 = \max_r \|A_{r,*}\|_2^2 \sum_s \left(\sum_r \varepsilon_{rs}\right) v_s^2\\ &\leq M_{\text{col}}\,\max_r \|A_{r,*}\|_2^2\,\|v\|_2^2. \qquad \square\end{aligned}$$

Now, let A denote the coefficient matrix for the hierarchical basis with a hierarchical ordering of rows and columns. That is, A has the block structure $[A_{rs}]$ where $A_{rs} = [a(\phi_{kl}^{(r)}, \phi_{k'l'}^{(s)})]$, k, l and k', l' range over all indices for levels r and s, respectively, and $a(u, v) = \int_\Omega u_x v_x + u_y v_y$.

THEOREM 3.4. *The maximum eigenvalue of A is bounded by a constant independent of h_j.*

Proof. We use Lemma 3.3 to bound $\|A_{rs}\|_2$. Let $\Omega_{kl}^{(i)}$ denote the support of $\phi_{kl}^{(i)}$. If the interiors of $\Omega_{kl}^{(r)}$ and $\Omega_{k'l'}^{(s)}$ are disjoint, then the entry of A_{rs} derived from $a(\phi_{kl}^{(r)}, \phi_{k'l'}^{(s)})$ is zero. Moreover, if $\Omega_{k'l'}^{(s)}$ is wholly contained in one of the triangles determining $\Omega_{kl}^{(r)}$, then this entry of A_{rs} is also zero, because $\phi_{k'l'}^{(s)}$ is harmonic in $\Omega_{k'l'}^{(s)}$. Therefore, for $s \geq r$, we need only consider the case where the interior of $\Omega_{k'l'}^{(s)}$ intersects an edge of one of the triangles defining $\Omega_{kl}^{(r)}$. For fixed k, l, there are at most $O(h_r/h_s)$ examples of such $\Omega_{k'l'}^{(s)}$, so that the row of A_{rs} corresponding to mesh index (k, l) (call it row t) contains $O(h_r/h_s)$ nonzero entries. Any such entry satisfies

$$\begin{aligned} a(\phi_{kl}^{(r)}, \phi_{k'l'}^{(s)}) &\leq \int_\Omega |\nabla \phi_{kl}^{(r)}|\,|\nabla \phi_{k'l'}^{(s)}| = \int_{\Omega_{kl}^{(r)} \cap \Omega_{k'l'}^{(s)}} |\nabla \phi_{kl}^{(r)}|\,|\nabla \phi_{k'l'}^{(s)}| \\ &\leq c \int_{\Omega_{kl}^{(r)} \cap \Omega_{k'l'}^{(s)}} h_r^{-1} h_s^{-1} \leq c h_s^2 \cdot h_r^{-1} h_s^{-1} = c \frac{h_s}{h_r}. \end{aligned}$$

Consequently,

$$\| [A_{rs}]_{t,*} \|_2^2 \leq c \frac{h_r}{h_s} \left(\frac{h_s}{h_r} \right)^2 = c \frac{h_s}{h_r}, \tag{18}$$

where here and in the sequel c represents a generic constant. The number of nonzero entries in a column of A_{rs} corresponding, say, to mesh index (k', l'), is bounded by the number of subdomains $\Omega_{kl}^{(r)}$ whose interiors intersect $\Omega_{k'l'}^{(s)}$; this is at most $c_T =$ (one plus the number of edges of vertex $x_{k'l'}$) ≤ 7. It follows from (18) and Lemma 3.3 that

$$\|A_{rs}\|_2^2 \leq c_T \max_t \| [A_{rs}]_{t,*} \|_2^2 = c \frac{h_s}{h_r} = c \frac{1}{2^{s-r}}. \tag{19}$$

Then, by Lemma 3.2 and Gerschgorin's theorem,

$$\|A\|_2 \leq \|A^{(b)}\|_2 \leq \|A^{(b)}\|_\infty \leq c \left(1 + \sum_{i=1}^{\lfloor j/2 \rfloor} \frac{1}{\sqrt{2}^{\,i}} \right) < c \frac{\sqrt{2}+1}{\sqrt{2}-1}. \qquad \square$$

Remark. Inequality (19) is analogous to Lemma 2.7 in [8]. Similar bounds can be derived for general conforming finite element discretizations and problems in higher dimensions.

3.2. Lower bound. Now, let A again denote the hierarchical basis matrix with the nodal ordering of rows and columns, i.e., degrees of freedom are ordered using the natural ordering of the underlying fine grid. As in the one-dimensional case, this matrix satisfies $A = S^T B S$ where B is the coefficient matrix derived from the nodal

basis,

$$B = \begin{pmatrix} T & -I & & \\ -I & T & -I & \\ & \ddots & \ddots & -I \\ & & -I & T \end{pmatrix}, \qquad T = \begin{pmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{pmatrix},$$

and $S = S_{j-1} \cdots S_2 S_1$. A and B are of order n^2 where $n = 2^j - 1$. The change of basis matrix S_i represents the effect of replacing basis functions associated with the mesh of width $1/2^i$ (referred to here as the *coarse mesh*) with those for the mesh of width $1/2^{i+1}$ (the *fine mesh*, whose nodes include the coarse mesh nodes). The finite element solution on the fine mesh can be represented as a linear combination of coarse mesh functions and fine mesh functions,

$$u = \sum_{x_{kl} \in \text{coarse}} \alpha_{kl} \phi_{kl}^{(i)} + \sum_{x_{kl} \in \text{fine-coarse}} \alpha_{kl} \phi_{kl}^{(i+1)}.$$

The representation using only fine mesh functions is

$$u = \sum_{x_{kl} \in \text{coarse}} \beta_{kl} \phi_{kl}^{(i+1)} + \sum_{x_{kl} \in \text{fine-coarse}} \beta_{kl} \phi_{kl}^{(i+1)},$$

where $\beta_{kl} = \alpha_{kl}$ for coarse mesh indices, $\beta_{kl} = \alpha_{kl} + \frac{1}{2}(\alpha_{k_1 l_1} + \alpha_{k_2 l_2})$ for fine mesh indices, and (k_1, l_1) and (k_2, l_2) are indices of the neighboring coarse grid points that are affected by the change of basis. The coefficients are related by $\beta = S_i \alpha$.

Examples of the matrices S_1 and S_2 for a 7×7 grid are as follows:

$$S_1 = \begin{pmatrix} I & & & & & & \\ & I & & \Upsilon_1 & & & \\ & & I & & & & \\ & & & I + \Theta_1 & & & \\ & & & & I & & \\ & & & \Lambda_1 & & I & \\ & & & & & & I \end{pmatrix}, \qquad S_2 = \begin{pmatrix} I & & \Upsilon_2 & & & & \\ & I + \Theta_2 & & & & & \\ & \Lambda_2 & I & \Upsilon_2 & & & \\ & & & I + \Theta_2 & & & \\ & & & \Lambda_2 & I & \Upsilon_2 & \\ & & & & & I + \Theta_2 & \\ & & & & & \Lambda_2 & I \end{pmatrix},$$

where

$$\Theta_1 = \begin{pmatrix} 0 & & & & & & \\ & 0 & .5 & & & & \\ & 0 & & & & & \\ & & 0 & & & & \\ & & & 0 & & & \\ & & .5 & 0 & & & \\ & & & & 0 & & \end{pmatrix}, \quad \Lambda_1 = \begin{pmatrix} 0 & & & & & & \\ & 0 & & & & & \\ & & 0 & & & & \\ & & & .5 & & & \\ & & & & 0 & & \\ & & & .5 & 0 & & \\ & & & & & 0 & \end{pmatrix}, \quad \Upsilon_1 = \begin{pmatrix} 0 & & & & & & \\ & 0 & .5 & & & & \\ & 0 & & & & & \\ & & .5 & & & & \\ & & & 0 & & & \\ & & & & 0 & & \\ & & & & & 0 & \end{pmatrix},$$

$$\Theta_2 = \begin{pmatrix} 0 & .5 & & & & & \\ & 0 & & & & & \\ & .5 & 0 & .5 & & & \\ & & & 0 & & & \\ & & & .5 & 0 & .5 & \\ & & & & & 0 & \\ & & & & & .5 & 0 \end{pmatrix}, \quad \Lambda_2 = \begin{pmatrix} 0 & & & & & & \\ & .5 & & & & & \\ & .5 & 0 & & & & \\ & & & .5 & & & \\ & & & .5 & 0 & & \\ & & & & & .5 & \\ & & & & & .5 & 0 \end{pmatrix}, \quad \Upsilon_2 = \begin{pmatrix} 0 & .5 & & & & & \\ & .5 & & & & & \\ & & 0 & .5 & & & \\ & & & .5 & & & \\ & & & & 0 & .5 & \\ & & & & & .5 & \\ & & & & & & 0 \end{pmatrix}.$$

$S_i = I + R_i$ is a block matrix of block order n, where each block is itself a matrix of order n. Considered as a block matrix, R_i has a nonzero structure similar to that of the corresponding matrix in the one-dimensional case:

$$(20)\quad [R_i]_{rs} \neq 0 \text{ if and only if } \begin{cases} s \text{ is divisible by } 2^{j-i} \quad \text{and} \quad r = s \pm \frac{1}{2}2^{j-i}, \\ \quad \text{or} \\ r = s \text{ is divisible by } 2^{j-i}. \end{cases}$$

There are three types of nonzero blocks in R_i: Λ_i below the block diagonal, Υ_i above the block diagonal, and Θ_i on the block diagonal. The off-diagonal blocks correspond to the effects of the change of basis in the vertical direction, and the diagonal blocks to the effects in the horizontal direction. For the nonzero structure of these blocks, we have

$$[\Theta_i]_{rs} = \begin{cases} 1/2 & \text{if } s \text{ is divisible by } 2^{j-i} \quad \text{and} \quad r = s \pm \frac{1}{2}2^{j-i}, \\ 0 & \text{otherwise,} \end{cases}$$

$$(21)\quad [\Lambda_i]_{rs} = \begin{cases} 1/2 & \text{if } s \text{ is divisible by } 2^{j-i} \quad \text{and} \quad (r = s \text{ or } r = s + \frac{1}{2}2^{j-i}), \\ 0 & \text{otherwise,} \end{cases}$$

$$[\Upsilon_i]_{rs} = \begin{cases} 1/2 & \text{if } s \text{ is divisible by } 2^{j-i} \quad \text{and} \quad (r = s \text{ or } r = s - \frac{1}{2}2^{j-i}), \\ 0 & \text{otherwise.} \end{cases}$$

Note that Θ_i is identical to R_i of (8).

Lemma 2.1 for the two-dimensional case follows from (20) and (21). The proof is essentially identical to the proof from §2. For example, to show that $R_{i_1}R_{i_2} \neq 0$ for $i_1 \neq i_2$, the argument in the proof of §2 shows that there can be no nonzero contributions of the form $\Lambda_{i_1}\Upsilon_{i_2}$ or $\Upsilon_{i_1}\Lambda_{i_2}$. An identical argument shows that there cannot be any contributions of the form $\Theta_{i_1}\Lambda_{i_2}$ or $\Theta_{i_1}\Upsilon_{i_2}$, i.e., there are no indices s such that R_{i_1} contains $\Theta_{i_1} \neq 0$ in position s of the block diagonal and R_{i_2} contains either Λ_{i_2} or Υ_{i_2} in row s. The same reasoning shows that $\Lambda_{i_1}\Theta_{i_2} = \Upsilon_{i_1}\Theta_{i_2} = 0$ whenever these figure in the product.

Lemma 2.2 for two dimensions follows immediately from Lemma 2.1. We can then specify the structure of $(SS^T)^{-1}$ using the expansion (11). To facilitate the discussion, we introduce the notation $\mathcal{T}_i(X, Y, Z)$, to represent a (block) tridiagonal matrix associated with the grid at level i. That is,

$$[\mathcal{T}_i(X, Y, Z)]_{rs} = \begin{cases} X & \text{if } r \text{ is divisible by } 2^{j-i} \quad \text{and} \quad s = r - 2^{j-i}, \\ Y & \text{if } r \text{ is divisible by } 2^{j-i} \quad \text{and} \quad s = r, \\ Z & \text{if } r \text{ is divisible by } 2^{j-i} \quad \text{and} \quad s = r + 2^{j-i}, \\ 0 & \text{otherwise.} \end{cases}$$

If the arguments used in place of X, Y, Z contain upper case characters, then $\mathcal{T}_i$ represents a block matrix of order n, each of whose entries is itself a matrix of order n; otherwise $\mathcal{T}_i$ represents an ordinary matrix of order n. Then, we have

$$R_i^T R_i = \mathcal{T}_i(\Upsilon_i^T\Lambda_i, \Upsilon_i^T\Upsilon_i + \Theta_i^T\Theta_i + \Lambda_i^T\Lambda_i, \Lambda_i^T\Upsilon_i),$$

where

$$\begin{aligned} \Upsilon_i^T\Lambda_i &= \mathcal{T}_i\left(\tfrac{1}{4}, \tfrac{1}{4}, 0\right), \\ \Upsilon_i^T\Upsilon_i + \Theta_i^T\Theta_i + \Lambda_i^T\Lambda_i &= \mathcal{T}_i\left(\tfrac{1}{4}, \tfrac{3}{2}, \tfrac{1}{4}\right), \\ \Lambda_i^T\Upsilon_i &= \mathcal{T}_i\left(0, \tfrac{1}{4}, \tfrac{1}{4}\right). \end{aligned}$$

That is, $R_i^T R_i$ has a regular structure. By analogy with the one-dimensional case, it has the form

$$R_i^T R_i = \frac{3}{2} D_i + \frac{1}{4} C_i, \tag{22}$$

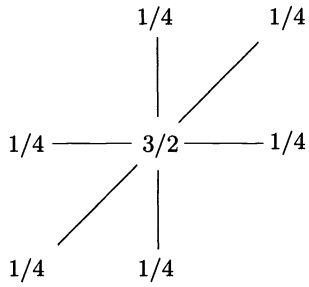
where D_i is the identity operator restricted to level i, and C_i corresponds to the off-diagonal terms: $C_i = 2\mathcal{T}_i(\widehat{\Lambda}_i, \widehat{\Theta}_i, \widehat{\Upsilon}_i)$ with

$$\widehat{\Theta}_i = \Theta_{i-1} + \Theta_{i-1}^T,$$

$$[\widehat{\Lambda}_i]_{rs} = \begin{cases} 1/2 & \text{if } s \text{ is divisible by } 2^{j-i} \quad \text{and} \quad (r = s \text{ or } r = s + 2^{j-i}), \\ 0 & \text{otherwise,} \end{cases} \tag{23}$$

$$[\widehat{\Upsilon}_i]_{rs} = \begin{cases} 1/2 & \text{if } s \text{ is divisible by } 2^{j-i} \quad \text{and} \quad (r = s \text{ or } r = s - 2^{j-i}), \\ 0 & \text{otherwise.} \end{cases}$$

$R_i^T R_i$ can be represented in terms of a computational molecule as follows:



The terms $\{R_i + R_i^T\}$ of (11) have less regular structure, in that there is no computational molecule for every grid point. For example, for the 7×7 grid,

$$R_1 + R_1^T = \begin{pmatrix} 0 & & & & & & \\ & 0 & & \Upsilon_1 & & & \\ & & 0 & & & & \\ & \Upsilon_1^T & & \Theta_1 + \Theta_1^T & & \Lambda_1^T & \\ & & & & 0 & & \\ & & & \Lambda_1 & & 0 & \\ & & & & & & 0 \end{pmatrix} = \begin{pmatrix} 0 & & & & & & \\ & \widehat{\Theta}_2 & & \widehat{\Upsilon}_2 & & & \\ & & 0 & & & & \\ & \widehat{\Upsilon}_2^T & & \widehat{\Theta}_2 & & \widehat{\Lambda}_2^T & \\ & & & & 0 & & \\ & & & \widehat{\Lambda}_2 & & \widehat{\Theta}_2 & \\ & & & & & & 0 \end{pmatrix} - \begin{pmatrix} 0 & & & & & & \\ & \widehat{\Theta}_2 & & \widetilde{\Upsilon}_1 & & & \\ & & 0 & & & & \\ & \widetilde{\Upsilon}_1^T & & 0 & & \widetilde{\Lambda}_1^T & \\ & & & & 0 & & \\ & & & \widetilde{\Lambda}_1 & & \widehat{\Theta}_2 & \\ & & & & & & 0 \end{pmatrix}, \tag{24}$$

where, for $1 \le i \le j-1$,

$$\widetilde{\Lambda}_i = \widehat{\Lambda}_{i+1} - \Lambda_i, \qquad \widetilde{\Upsilon}_i = \widehat{\Upsilon}_{i+1} - \Upsilon_i.$$

The matrices $\widetilde{\Lambda}_i$ and $\widetilde{\Upsilon}_i$ can be thought of as complements to Λ_i and Υ_i that produce the regular operators $\widehat{\Lambda}_{i+1}$ and $\widehat{\Upsilon}_{i+1}$; compare (21) and (23). Note that $\widehat{\Lambda}_i = \widehat{\Upsilon}_i^T$, so that the right side of (24) is a representation of $R_1 + R_1^T$ in the form $\frac{1}{2}C_2 - E_2$. More generally, we have

$$R_i + R_i^T = \frac{1}{2} C_{i+1} - E_{i+1},$$

where E_{i+1} contains the values $\widetilde{\Lambda}_i$ and $\widetilde{\Upsilon}_i$ in all entries where Λ_i or Υ_i appear in $R_i + R_i^T$, and $\hat{\Theta}_{i+1}$ in all diagonal entries (r,r) where r is divisible by $\frac{1}{2}2^{j-i}$ but not 2^{j-i}.

Combining this observation with (22), we have for $i < j$,

$$R_i^T R_i - (R_{i-1} + R_{i-1}^T) = \frac{3}{2}D_i - \frac{1}{4}C_i + E_i,$$

where $R_0 = 0$; and for $i = j$,

$$I - (R_{j-1} + R_{j-1}^T) = \left(3D_j - \frac{1}{2}C_j\right) - (2I - E_j),$$

where D_j is the identity matrix. Let $F_i = 3D_i - \frac{1}{2}C_i$ and $G_j = \left(2I - \sum_{i=1}^{j} E_i\right)$. Then, for the two-dimensional problem, (11) can be written as

$$Q = (SS^T)^{-1} = \frac{1}{2}(F_1 + \cdots + F_{j-1}) + F_j - G_j. \tag{25}$$

LEMMA 3.5. *The matrix G_j is symmetric positive semidefinite.*

Proof. By definition, E_i is symmetric for each i, so that G_j is symmetric. E_i contains at most four nonzeros in any row, with value $\frac{1}{2}$, and for $i_1 \neq i_2$, E_{i_1} and E_{i_2} contain no common nonzero indices. Therefore, G_j is diagonally dominant, whence positive semidefinite. □

Let $\mathcal{V}_i$ denote the space of vectors v with value zero in all entries associated with grid points of level $i' > i$. We will represent members of $\mathcal{V}_i$ as grid functions v_{kl} defined on the two-dimensional grid at level i with index values derived from the finest grid. That is, k and l are divisible by 2^{j-i}, and $1 \leq k/2^{j-i} \leq n_i$, $1 \leq l/2^{j-i} \leq n_i$, with $n_i = 2^i - 1$. Let $r_i = 2^{j-i}$, the offset associated with the grid at level i, and let B_i denote the five-point discrete Laplacian on level i, scaled by h_i^2:

$$[B_i]_{kl} = 4v_{kl} - v_{k-r_i,l} - v_{k+r_i,l} - v_{k,l-r_i} - v_{k,l+r_i}.$$

The following result shows that the operators B_i and F_i are spectrally equivalent.

LEMMA 3.6. *For any nonzero $v \in \mathcal{V}_i$, B_i and F_i satisfy*

$$\frac{1}{2} \leq \frac{(v, F_i v)}{(v, B_i v)} \leq \frac{3}{2}. \tag{26}$$

Proof. Note that $F_i = \frac{1}{2}B_i + \hat{F}_i$ where

$$[\hat{F}_i v]_{kl} = v_{kl} - \frac{1}{2}v_{k-r_i,l-r_i} - \frac{1}{2}v_{k+r_i,l+r_i}.$$

By ordering the grid by diagonals, we find that $\hat{F}_i$ is similar to a block diagonal matrix, each of whose diagonal blocks is irreducibly diagonally dominant. Thus, $\hat{F}_i$ is symmetric positive-definite, and the lower bound of (26) follows. For the upper bound, note that summation by parts [5, p. 213] gives

$$(v, B_i v) = \|\delta_x^{(i)} v\|_2^2 + \|\delta_y^{(i)} v\|_2^2, \tag{27}$$

where

$$[\delta_x^{(i)} v]_{kl} = v_{kl} - v_{k-r_i,l}, \quad [\delta_y^{(i)} v]_{kl} = v_{kl} - v_{k,l-r_i}, \quad 1 \leq k/r_i, l/r_i \leq n_i,$$

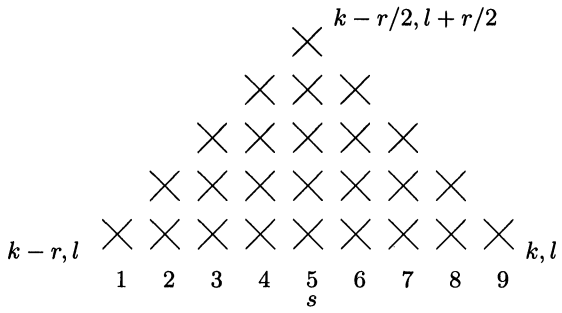


FIG. 3. *Grid triangle* $\Delta_r(k,l)$.

and $v_{k0} = v_{k,2^j} = v_{0l} = v_{2^j,l} = 0$. Similarly,

$$(v, \hat{F}_i v) \le \|\delta_x^{(i)} v\|_2^2 + \|\delta_y^{(i)} v\|_2^2 = (v, B_i v),$$

which leads to $(v, F_i v) \le \frac{3}{2}(v, B_i v)$. □

THEOREM 3.7. *The minimum eigenvalue of A is bounded below by* c/j^2 *where c is independent of j.*

Proof. We will show that $(v,Qv)/(v,Bv) \le O(j^2)$. To begin, note that (25) and Lemma 3.5 imply

$$\frac{(v,Qv)}{(v,Bv)} \le \frac{(v,F_j v)}{(v,Bv)} + \frac{1}{2}\sum_{i=1}^{j-1} \frac{(v,F_i v)}{(v,Bv)}.$$

Application of Lemma 3.6 gives

$$\begin{aligned} \max_{v\neq 0} \frac{(v,Qv)}{(v,Bv)} &\le \max_{v\neq 0} \frac{(v,F_j v)}{(v,Bv)} + \frac{1}{2}\sum_{i=1}^{j-1} \left(\max_{\substack{v\in\mathcal{V}_i \\ v\neq 0}} \frac{(v,F_i v)}{(v,B_i v)} \max_{v\neq 0} \frac{(v,B_i v)}{(v,Bv)} \right) \\ &\le \frac{3}{2} + \frac{3}{4}\sum_{i=1}^{j-1} \max_{v\neq 0} \frac{(v,B_i v)}{(v,Bv)}. \end{aligned} \tag{28}$$

Thus, we seek an upper bound on $(v,B_i v)/(v,Bv)$, where $B = B_j$.

To simplify notation, we drop the subscript i from $r_i = 2^{j-i}$. Let $\Delta_r = \Delta_r(k,l)$ denote the set of grid points contained in the triangle determined by the indices (k,l), $(k-r,l)$ and $(k-r/2, l+r/2)$. An example for the case $r = 8$ is shown in Fig. 3. Using (27), we will relate $(v,B_i v)$ to (v,Bv) by bounding $[\delta_x^{(i)} v]_{kl}^2$ and $[\delta_y^{(i)} v]_{kl}^2$ in terms of a partial sum from Δ_r contributing to $\|\delta_x^{(j)} v\|_2^2$ and $\|\delta_y^{(j)} v\|_2^2$.

Let $v^{(s)}$ denote the average of the entries of v in the sth grid column of Δ_r, i.e.,

$$v^{(s)} = \frac{1}{\eta_s}\sum_{t=0}^{\eta_s - 1} v_{k-r+s-1,l+t},$$

where $\eta_s = \min(s, r+2-s)$ is the number of grid points in this column. (See Fig. 3.) Then

$$[\delta_x^{(i)} v]_{kl} = v_{kl} - v_{k-r,l} = v^{(r+1)} - v^{(1)} = \sum_{s=1}^{r} (v^{(s+1)} - v^{(s)}).$$

Multiplying and dividing the sth term by $(\eta_s+1)^{1/2}$ and applying the Cauchy–Schwarz inequality gives

$$[\delta_x^{(i)}v]_{kl}^2 \le \left(\sum_{s=1}^{r}\frac{1}{\eta_s+1}\right)\left(\sum_{s=1}^{r}(\eta_s+1)(v^{(s+1)}-v^{(s)})^2\right). \tag{29}$$

The first sum on the right side of this inequality is bounded by

$$\sum_{s=1}^{r}\frac{1}{\eta_s+1} \le 2\log(r/2+2) \le c\log r. \tag{30}$$

To obtain a bound on the second term, we simplify notation by letting $m = m_s = k-r+s$, the index of the sth grid column of Δ_r. Consider the case where we are in the left side of Δ_i, i.e., $s \le \frac{r}{2}$, so that $\eta_s = s$ and $\eta_{s+1} = s+1$. Then

$$\begin{aligned} v^{(s+1)} &= \frac{1}{s+1}\sum_{t=0}^{s} v_{m,l+t} = \frac{1}{s+1}\sum_{t=0}^{s-1} v_{m,l+t} + \frac{1}{s(s+1)}\sum_{t=0}^{s-1} v_{m-1,l+s} \\ &\qquad + \frac{1}{s+1}\left(v_{m,l+s} - v_{m-1,l+s}\right), \\ v^{(s)} &= \frac{1}{s}\sum_{t=0}^{s-1} v_{m-1,l+t} = \frac{1}{s+1}\sum_{t=0}^{s-1} v_{m-1,l+t} + \frac{1}{s(s+1)}\sum_{t=0}^{s-1} v_{m-1,l+t}. \end{aligned}$$

Thus

$$v^{(s+1)} - v^{(s)} = \frac{1}{s+1}\sum_{t=0}^{s}(v_{m,l+t} - v_{m-1,l+t}) + \frac{1}{s(s+1)}\sum_{t=0}^{s-1}(v_{m-1,l+s} - v_{m-1,l+t}).$$

The second term of this expression is bounded in absolute value by

$$\frac{1}{s(s+1)}\left|\sum_{t=0}^{s-1}\sum_{q=t}^{s-1}(v_{m-1,l+q+1} - v_{m-1,l+q})\right| \le \frac{1}{s+1}\sum_{q=0}^{s-1}|v_{m-1,l+q+1} - v_{m-1,l+q}|.$$

Consequently,

$$\left|v^{(s+1)} - v^{(s)}\right| \le \frac{1}{s+1}\left(\sum_{t=0}^{s}\left|[\delta_x^{(j)}v]_{m,l+t}\right| + \sum_{t=0}^{s-1}\left|[\delta_y^{(j)}v]_{m-1,l+t+1}\right|\right),$$

which implies

$$\begin{aligned} (\eta_s+1)(v^{(s+1)} - v^{(s)})^2 &\le \frac{1}{s+1}\left(\sum_{t=0}^{s}\left|[\delta_x^{(j)}v]_{m,l+t}\right| + \sum_{t=0}^{s-1}\left|[\delta_y^{(j)}v]_{m-1,l+t+1}\right|\right)^2 \\ &\le \frac{2s+1}{s+1}\left(\sum_{t=0}^{s}([\delta_x^{(j)}v]_{m,l+t})^2 + \sum_{t=0}^{s-1}([\delta_y^{(j)}v]_{m-1,l+t+1})^2\right) \\ &< 2\left(\sum_{t=0}^{s}([\delta_x^{(j)}v]_{m,l+t})^2 + \sum_{t=0}^{s-1}([\delta_y^{(j)}v]_{m-1,l+t+1})^2\right). \end{aligned}$$

Essentially the same argument applies for the case $s > \frac{r}{2}$, so that

$$\sum_{s=1}^{r}(\eta_s + 1)(v^{(s+1)} - v^{(s)})^2 < 2\left(\sum_{\Delta_r}[\delta_x^{(j)}v]_{st}^2 + \sum_{\Delta_r}[\delta_y^{(j)}v]_{st}^2\right), \tag{31}$$

where the sum for the horizontal differences $[\delta_x^{(j)}v]_{st}$ is over all indices $(s,t) \in \Delta_r$, and the sum for the vertical differences $[\delta_y^{(j)}v]_{st}$ is over indices such that $(s,t-1) \in \Delta_r$. Combining (29), (30), and (31) gives

$$[\delta_x^{(i)}]_{kl}^2 \le c \log r \left(\sum_{\Delta_r}[\delta_x^{(j)}v]_{st}^2 + \sum_{\Delta_r}[\delta_y^{(j)}v]_{st}^2\right). \tag{32}$$

An analogous bound holds for $[\delta_y^{(i)}]_{kl}^2$. Hence, summing over all k,l, recalling that $r = 2^{j-i}$, and using (27) with $i = j$, we have

$$(v, B_i v) \le c\,(j-i)\,(v, B_j v).$$

The assertion then follows from (28). □

Remark. Inequality (32) is a variant of the discrete Sobolev inequality

$$\|u_h\|_{L^\infty} \le c(\log h^{-1})^{1/2}\,\|u_h\|_{H^1}$$

for finite element functions in two dimensions; see, e.g., [3, Lem. 3.2] and [8, Lem. 2.1].

REFERENCES

[1] O. Axelsson and V. A. Barker, *Finite Element Solution of Boundary Value Problems: Theory and Computation*, Academic Press, Orlando, FL, 1984.

[2] R. E. Bank, T. F. Dupont, and H. Yserentant, *The hierarchical basis multigrid method*, Numer. Math., 52 (1988), pp. 427–458.

[3] J. H. Bramble, *A second order finite difference analogue of the first biharmonic boundary value problem*, Numer. Math., 9 (1966), pp. 236–249.

[4] M. Griebel, *Multilevel algorithms considered as iterative methods on indefinite systems*, SIAM J. Sci. Comp., 15 (1994), pp. 547–565.

[5] E. Isaacson and H. B. Keller, *Analysis of Numerical Methods*, John Wiley & Sons, New York, 1966.

[6] C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge University Press, New York, 1987.

[7] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[8] H. Yserentant, *On the multi-level splitting of finite element spaces*, Numer. Math., 49 (1986), pp. 379–412.

[9] O. C. Zienkiewicz, D. W. Kelly, J. Gago, and I. Babuška, *Hierarchical finite element approaches, error estimates and adaptive refinement*, in The Mathematics of Finite Elements and Applications IV, J. R. Whiteman, ed., Academic Press, London, 1982.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 207–217, January 1995

014

CONSTRUCTING A HERMITIAN MATRIX FROM ITS DIAGONAL ENTRIES AND EIGENVALUES*

MOODY T. CHU†

Abstract. Given two vectors $a, \lambda \in R^n$, the Schur–Horn theorem states that a majorizes λ if and only if there exists a Hermitian matrix H with eigenvalues λ and diagonal entries a. While the theory is regarded as classical by now, the known proof is not constructive. To construct a Hermitian matrix from its diagonal entries and eigenvalues therefore becomes an interesting and challenging inverse eigenvalue problem. Two algorithms for determining the matrix numerically are proposed in this paper. The lift and projection method is an iterative method that involves an interesting application of the Wielandt–Hoffman theorem. The projected gradient method is a continuous method that, besides its easy implementation, offers a new proof of existence because of its global convergence property.

Key words. Schur–Horn theorem, majorization, inverse eigenvalue problem, lift and projection, projected gradient

AMS subject classifications. 65F15, 65H15

1. Introduction.

The well-known Schur–Horn theorem [14] deals with the relationships between the main diagonal entries and eigenvalues of a Hermitian matrix. For the reference, we restate the theorem in the following form [15, Thms. 4.3.26, 4.3.32, and 4.3.33].

THEOREM 1.1 (Schur–Horn Theorem). 1. *Let H be a Hermitian matrix. Let $\lambda = [\lambda_i] \in R^n$ and $a = [a_i] \in R^n$ denote the vectors of eigenvalues and diagonal entries of H, respectively. If the entries are arranged in increasing order $a_{j_1} \le \cdots \le a_{j_n}$, $\lambda_{m_1} \le \cdots \le \lambda_{m_n}$, then*

$$\sum_{i=1}^{k} a_{j_i} \ge \sum_{i=1}^{k} \lambda_{m_i}, \tag{1}$$

for all $k = 1, 2, \ldots, n$ with equality for $k = n$.

2. *Given any $a, \lambda \in R^n$ satisfying (1), there exists a Hermitian matrix H with eigenvalues λ and diagonal entries a.*

The notion of (1) is also known as *a majorizing* λ, which has arisen as the precise relationship between two sets of numbers in many areas of disciplines, including matrix theory and statistics. There are extensive research results on this subject. See, for example, [2], [16], and the references contained therein. The Schur–Horn theorem itself has many important applications. For instance, through the Schur–Horn theorem the total least squares problem can be seen to be equivalent to a linear programming problem [3]. Some other applications can be found, for example, in [6] and [7].

The second part of the Schur–Horn Theorem gives rise to an interesting inverse eigenvalue problem, namely, to construct such a Hermitian matrix from the given eigenvalues and diagonal entries. For convenience, we shall refer to this problem as (SHIEP). The proof of existence is usually known as the harder part of the Schur–Horn Theorem. One point worthy of note is that there are far more variables in the

* Received by the editors January 19, 1993; accepted for publication (in revised form) by N. J. Higham, November 9, 1993.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205 (chu@math.ncsu.edu). This research was supported in part by National Science Foundation grants DMS-9006135 and DMS-9123448.

(SHIEP) than constraints, which presumably implies that the solution is far from unique. It turns out that most of the proofs in the literature are not practicable for the (SHIEP) in that a construction by mathematical induction, if possible at all, would be overwhelmingly complicated. See, for example, [14], [15], [17]. In this paper we propose numerical algorithms that are different from the classical ways of authenticating the existence.

Henceforth, we shall denote the diagonal matrix whose main diagonal entries are the same as those of the matrix M as $\mathrm{diag}(M)$ and the diagonal matrix whose diagonal entries are formed from the vector v as $\mathrm{diag}(v)$. This notation will prove to be convenient in the discussion. Any ambiguity can be clarified from the context. Also, we shall define

$$\mathcal{T}(a) := \{T \in R^{n\times n} | \mathrm{diag}(T) = \mathrm{diag}(a)\} \tag{2}$$

and

$$\mathcal{M}(\Lambda) := \{Q^T \Lambda Q | Q \in \mathcal{O}(n)\}, \tag{3}$$

where $\Lambda := \mathrm{diag}(\lambda)$ and $\mathcal{O}(n)$ is the group of all orthogonal matrices in $R^{n\times n}$.

Our algorithms are based on the idea of finding the shortest distance between $\mathcal{T}(a)$ and $\mathcal{M}(\Lambda)$, i.e., we want to solve

$$\min_{T\in\mathcal{T}(a), Z\in\mathcal{M}(\Lambda)} \|T - Z\|_F, \tag{4}$$

where $\|\cdot\|_F$ means the Frobenius norm. Clearly, the Schur–Horn theorem attests that $\mathcal{T}(a)$ and $\mathcal{M}(\Lambda)$ intersect and hence the minimal value of (4) should be zero. Our goal is, starting with an arbitrary point on either $\mathcal{T}(a)$ or $\mathcal{M}(\Lambda)$, to find the intersection point.

The (SHIEP) is fundamentally different from a classical inverse eigenvalue problem (CIEP) that has been discussed, for example, in [12]. Given symmetric matrices $A_0, A_1, \ldots, A_n$, the (CIEP) is to find a vector $c \in R^n$ such that the matrix $A(c)$ where

$$A(c) := A_0 + c_1 A_1 + \cdots + c_n A_n \tag{5}$$

has the prescribed spectrum λ. In relating the (CIEP) to the (SHIEP), one must specify each A_i of the basis matrices. To characterize the matrices $A_1, \ldots, A_n$ a priori so that a solution to the (SHIEP) may be written in the form of (5), however, is by no means easy. One may select, for example, $A_0 = \mathrm{diag}(a)$ and all other A_i, $1 \le i \le n$, such that $\mathrm{diag}(A_i) = 0$. Even so, the off-diagonal entries of A_i are still completely free. Picking the remaining part of A_i arbitrarily would be an absurd thing to do since the resulting (CIEP) may very well not have a solution at all. In fact, one can easily construct a 2×2 example to demonstrate a *near miss* case where A_i's are such that a certain combination $A(c)$ from a special c is very near a solution of the (SHIEP), yet the entire affine subspace of $A(c)$ from every possible c does not intersect $\mathcal{M}(\Lambda)$. One may then wonder why not to exploit the freedom in the (SHIEP) by further restricting the structure of the matrix. For example, it seems to be a sensible requirement that the matrix being constructed should be a Jacobi matrix, since there are really $2n-1$ given data elements (both a and λ have the same sum.) Again, one can easily check that there are no real numbers b_1, b_2 so that the 3×3 matrix

$$\begin{bmatrix} 1 & b_1 & 0 \\ b_1 & 2 & b_2 \\ 0 & b_2 & 3 \end{bmatrix}$$

will have eigenvalues $\{-5, -4, 15\}$. That is, the (SHIEP) for structured matrices is not necessarily solvable. It is an interesting question to study what additional conditions must be imposed so that a more specified problem has a solution; however, in this paper attention is paid only to the (SHIEP) that arises from the Schur–Horn Theorem. Consequently, until $A_1, \ldots, A_n$ are properly selected, any numerical techniques proposed for (CIEP) are not directly applicable for the (SHIEP).

In contrast, a much easier iterative method that alternates points between $\mathcal{T}$ and $\mathcal{M}(\Lambda)$ is possible. This procedure, called *lift-and-projection* for the reason that will become clear from the geometry, is discussed in §2. The lift and projection method is essentially the same as the so-called alternating projection method [4], [8], [11], [13] except that the latter requires the underlying sets to be convex. The set $\mathcal{M}(\Lambda)$ is not convex. Nevertheless, we shall see for our problem that the so-called *proximity map* can still be formulated. In particular, the projection is almost free and the Wielandt–Hoffman theorem makes the action of lifting possible at the cost of a spectral decomposition per step. We think this connection is worth mentioning even though the rate of convergence is expected to be linear only.

Our main contribution in this paper is in the construction of a gradient flow on the surface $\mathcal{M}(\Lambda)$ that moves toward the desired intersection point. No spectral decomposition is needed. Our approach is similar to that developed in [9] with slight modifications, but the application to the Schur–Horn theorem is apparently new. The gradient flow is derived in §3. We should emphasize that our goal in this paper is not to redo the proof of the Schur–Horn theorem, but rather to develop an algorithm that can compute the results promised by the theorem. On the other hand, if we can show that our algorithm always finds a solution, then in return we have indeed offered a different proof for the Schur–Horn theorem. Numerical examples are demonstrated in §4.

2. Lift and projection. In [10] we introduced a notion that interprets numerical methods proposed in [12] for the (CIEP) as a *coordinate-free* Newton method. For the (SHIEP), however, one quickly discovers that the same idea does not work. The search for a $\mathcal{T}$-intercept of a tangent array from $\mathcal{M}(\Lambda)$ amounts to a nonlinear system of $n(n+1)/2$ equations in $n(n-1)$ unknowns. When $n > 3$, this is an underdetermined system. Unlike those methods discussed in [12], there is no clear strategy on how the this system could be solved. In this section, we replace the concept of "tangent" by that of "projection" and propose an analogous but easier iteration method called lift and projection.

The main idea is to alternate between $\mathcal{T}$ and $\mathcal{M}(\Lambda)$ in the following way: From a given $T^{(k)} \in \mathcal{T}$, first we find the point $Z^{(k)} \in \mathcal{M}(\Lambda)$ such that $\|T^{(k)} - Z^{(k)}\|_F = \operatorname{dist}(T^{(k)}, \mathcal{M}(\Lambda))$. Then we find $T^{(k+1)} \in \mathcal{T}$ such that $\|T^{(k+1)} - Z^{(k)}\|_F = \operatorname{dist}(\mathcal{T}, Z^{(k)})$. Here, as usual, the distance is measured in the Frobenius norm. A schematic diagram of the iteration is illustrated in Fig. 1, even though the topology of $\mathcal{M}(\Lambda)$ is much more complicated. We call $Z^{(k)}$ a *lift* of $T^{(k)}$ onto $\mathcal{M}(\Lambda)$ and $T^{(k+1)} \in \mathcal{T}$ a *projection* of $Z^{(k)}$ onto $\mathcal{T}$.

The projection is easy to formulate. In fact, the projection $T = [t_{ij}]$ of any $Z = [z_{ij}] \in \mathcal{M}(\Lambda)$ onto $\mathcal{T}$ must be given by

$$t_{ij} := \begin{cases} z_{ij}, & \text{if } i \neq j \\ a_i, & \text{if } i = j. \end{cases} \tag{6}$$

The Wielandt–Hoffman theorem [15, Thm. 6.3.5], on the other hand, furnishes a mechanism for lifting. For demonstration purposes, we assume that both Λ and the

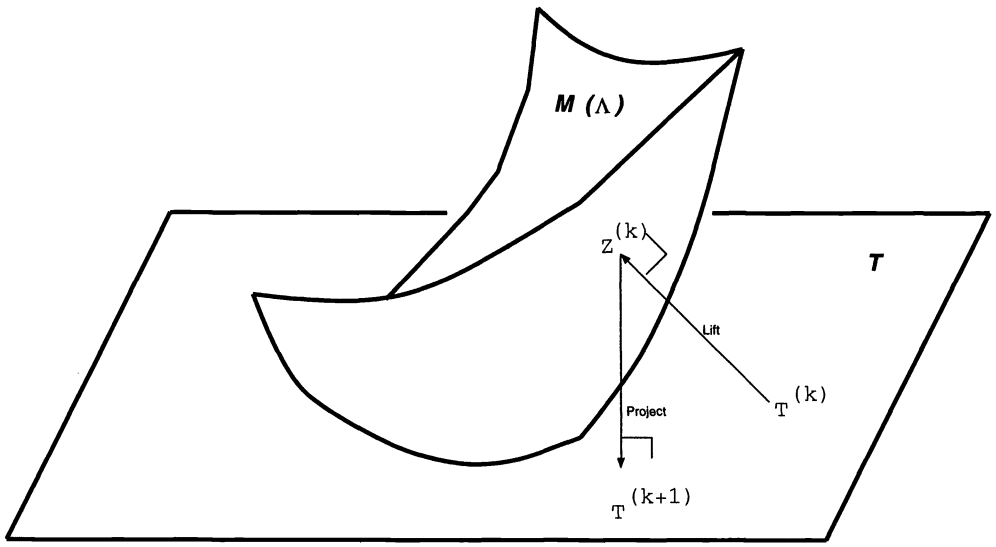


FIG. 1. *Geometric sketch of lifting and projection.*

given $T \in \mathcal{T}$ have simple spectrum. (For the case of multiple eigenvalues, the result only needs a slight modification.) Suppose $T = Q^T DQ$ is a spectral decomposition of T where D is a diagonal matrix of eigenvalues. Then the shortest distance between T and $\mathcal{M}(\Lambda)$ is attained at the point (i.e., the lift of T onto $\mathcal{M}(\Lambda)$)

$$Z := Q^T \mathrm{diag}(\lambda_{\pi_1}, \ldots, \lambda_{\pi_n}) Q, \tag{7}$$

where π is a permutation so that $\lambda_{\pi_1}, \ldots, \lambda_{\pi_n}$ are in the same algebraic ordering as the diagonal entries in D. The justification on why this assertion is correct has already been proved in [7] and [9].

Since in either step of lifting or projection we are minimizing the distance between a point and a set, we have

$$\|T^{(k+1)} - Z^{(k+1)}\|_F^2 \leq \|T^{(k+1)} - Z^{(k)}\|_F^2 \leq \|T^{(k)} - Z^{(k)}\|_F^2. \tag{8}$$

The lift and projection clearly is a descent method. The sequence $\{(T^{(k)}, Z^{(k)})\}$ will converge to a stationary point for the problem (4).

Because $\mathcal{M}(\Lambda)$ is not a convex set, a stationary point for (4) is not necessarily an intersection point of $\mathcal{T}$ and $\mathcal{M}(\Lambda)$. This is a major difference between our method and the alternating projection method [4], [8], [11], [13]. On the other hand, the application of the Wielandt–Hoffman theorem to formulate the proximity map despite of the non-convexity is remarkably simple and quite interesting. The rate of convergence being expected to be linear only, this method might not be very efficient.

3. Gradient flow. We now consider a continuous approach for the (SHIEP) that does not need any spectral decomposition. Similar to (4), our idea now is to

solve the problem

$$\min_{Q\in\mathcal{O}(n)} F(Q) := \frac{1}{2}\|\mathrm{diag}(Q^T\Lambda Q) - \mathrm{diag}(a)\|_F^2. \tag{9}$$

The Schur–Horn theorem guarantees that there exists a Q at which F vanishes. We now explain how to improve the matrix Q when $F(Q)$ is not minimal.

In terms of the Frobenius inner product

$$\langle A, B\rangle := \sum_{i,j} a_{ij}b_{ij}, \tag{10}$$

the Fréchet derivative of F at Q acting on an arbitrary matrix $U \in R^{n\times n}$ can be calculated as follows:

$$\begin{aligned} F'(Q)U &= 2\langle \mathrm{diag}(Q^T\Lambda Q) - \mathrm{diag}(a), \mathrm{diag}(Q^T\Lambda U)\rangle \\ &= 2\langle \mathrm{diag}(Q^T\Lambda Q) - \mathrm{diag}(a), Q^T\Lambda U\rangle \\ &= 2\langle \Lambda Q(\mathrm{diag}(Q^T\Lambda Q) - \mathrm{diag}(a), U\rangle. \end{aligned} \tag{11}$$

The second equality in (11) follows from the observation that the first entry in the inner product is a diagonal matrix that results in the same inner product with either $\mathrm{diag}(Q^T\Lambda U)$ or $Q^T\Lambda U$. The third equality follows from the adjoint property of the inner product. The gradient ∇F at Q can now be represented as

$$\nabla F(Q) = 2\Lambda Q\beta(Q) \tag{12}$$

with $\beta(Q) := \mathrm{diag}(Q^T\Lambda Q) - \mathrm{diag}(a)$.

Once we have (12), the entire framework developed in [9] for projected gradient method can be applied. In particular, the projected gradient is readily available.

THEOREM 3.1. *The projection $g(Q)$ of $\nabla F(Q)$ onto $\mathcal{O}(n)$ is given by*

$$g(Q) = Q[Q^T\Lambda Q, \beta(Q)], \tag{13}$$

where $[A, B] := AB - BA$ is the Lie bracket.

Proof. See [9, formulas (20), (21), and (22)]. □

We may also calculate the projected Hessian as follows.

THEOREM 3.2. *Let $Q \in \mathcal{O}(n)$ be a stationary point of* (9). *Any tangent vector H of the manifold $\mathcal{O}(n)$ at Q is of the form $H = QK$ for some skew symmetric matrix K. The projected Hessian $g'(Q)$ acting on H is given by*

$$\langle g'(Q)QK, QK\rangle = \langle \mathrm{diag}[\mathrm{Q^T\Lambda Q, K}] - [\beta(\mathrm{Q}), \mathrm{K}], [\mathrm{Q^T\Lambda Q, K}]\rangle. \tag{14}$$

Proof. See [9, formulas (27), (28), and (29)]. □

The vector field

$$\dot{Q} = -g(Q) \tag{15}$$

defines a steepest descent flow on the manifold $\mathcal{O}(n)$ for the function $F(Q)$. Let $X := Q^T\Lambda Q$ and $\alpha(X) := \beta(Q) = \mathrm{diag}(X) - \mathrm{diag}(a)$, then correspondingly

$$\dot{X} = [X, [\alpha(X), X]] \tag{16}$$

defines an isospectral flow on $\mathcal{M}(\Lambda)$ that moves to decrease the distance between $\text{diag}(X)$ and $\text{diag}(a)$.

The algorithm for solving the (SHIEP) is then simply to integrate the differential equation (16) from a starting point $X_0 \in \mathcal{M}(\Lambda)$.

We now take a closer look at where the flow of (16) will converge. By the way we construct (16), a natural Lyapunov function

$$G(t) := \frac{1}{2}\|\text{diag}(X(t)) - \text{diag}(a)\|_F^2 \tag{17}$$

exists. Lyapunov's second method [5, Thm. 5.5] implies that a limit point of (16) must satisfy $[\alpha(X), X] = 0$. For simplicity, we consider only the generic case where all $\lambda_1, \ldots, \lambda_n$ are distinct. The case with some eigenvalues equal to each other is a little bit more complicated to analyze, although our numerical experiences seem to indicate that the convergence behavior should be similar. Under our assumption, a stationary point Q of (9) necessarily corresponds to an equilibrium point $X = Q^T \Lambda Q$ of (16) and vice versa.

Recall that $\beta(Q) = \text{diag}(Q^T \Lambda Q) - \text{diag}(a)$. Obviously if $\beta(Q) = 0$ at a stationary point Q, then the corresponding $X = Q^T \Lambda Q$ is a solution to the (SHIEP). Indeed, we have the following observation which shows that the stationary point Q in this case satisfies the second order *necessary* condition for being a minimum of (9).

THEOREM 3.3. *If $\beta(Q) = 0$ for some $Q \in \mathcal{O}(n)$, then for all skew symmetric matrices K it is true that $< g'(Q)QK, QK >= \|\text{diag}[\mathrm{Q^T \Lambda Q, K}]\|_{\mathrm{F}}^2 \geq 0$. In other words, the projected Hessian of F at Q is positive semidefinite.*

Proof. The result follows directly from (14) and the definition of Frobenius inner product (10). □

The strict inequality is not true in the above theorem. In fact, if we denote $\Omega := \text{diag}[X, K] = \text{diag}\{\omega_1, \ldots, \omega_n\}$, then

$$\omega_i = \sum_{s=1}^{i-1} x_{si} k_{si} - \sum_{t=i+1}^{n} x_{it} k_{it}. \tag{18}$$

Let X be fixed. Since $\sum_{i=1}^n \omega_i = 0$, the system $\omega_i = 0$ for $i = 1, \ldots, n$ contains only $n - 1$ independent equations in the $n(n-1)/2$ unknowns k_{ij}. We should be able to find a nontrivial skew symmetric matrix K that makes $\Omega = 0$. However, we now show that only those matrices X at which $\beta(Q) = 0$ are the possible *asymptotically stable* equilibrium point for (16).

THEOREM 3.4. *If $\beta(Q) \neq 0$ at a stationary point Q, then there exists a skew symmetric matrix K such that $\langle g'(Q)QK, QK\rangle < 0$. Thus, Q cannot be a local minimum of* (9).

Proof. Transforming similarly by a permutation matrix if necessary, we may assume that $\beta(Q)$ is of the form

$$\beta(Q) = \text{diag}\{\beta_1 I_{n_1}, \ldots, \beta_k I_{n_k}\}, \tag{19}$$

where I_{n_i} is the $n_i \times n_i$ identity matrix and $\beta_1 > \cdots > \beta_k$. Since $[Q^T \Lambda Q, \beta(Q)] = 0$, it follows that $X = Q^T \Lambda Q$ must be block diagonal of the form

$$X = \text{diag}\{X_{11}, \ldots, X_{kk}\}, \tag{20}$$

where each X_{ii} is a real symmetric matrix of size $n_i \times n_i$. Define $E := Q\beta(Q)Q^T$. Since $[\Lambda, E] = 0$ and all entries of Λ are distinct, E is a diagonal matrix whose entries

$E = \operatorname{diag}(e_1, \ldots, e_n)$ are simply a permutation of those of $\beta(Q)$. Note that Q^T is the orthogonal matrix of eigenvectors of X. So Q must also have the same structure as X. For any $n \times n$ skew symmetric matrix $K = [K_{ij}]$ partitioned in the same way as in (20) that satisfies $K_{ii} = 0$ for all $i = 1, \ldots, k$, it is easy to check that $\operatorname{diag}[Q^T \Lambda Q, K] = 0$. The projected Hessian now becomes

$$\begin{aligned} \langle g'(Q)QK, QK\rangle &= -\langle [\beta(Q), K], [Q^T \Lambda Q, K]\rangle \\ &= -\langle E\tilde{K} - \tilde{K}E, \Lambda\tilde{K} - \tilde{K}\Lambda\rangle \\ &= -2\sum_{i<j} (\lambda_i - \lambda_j)(e_i - e_j)\tilde{k}_{ij}^2, \end{aligned} \tag{21}$$

where $\tilde{K} = [\tilde{k}_{ij}] := QKQ^T$ is still a skew symmetric matrix with the same structure as K. From (21), we pick up values of $\tilde{k}_{ij}$ so that $\langle g'(Q)QK, QK\rangle < 0$. □

Theorem 3.4 implies that at a stationary point Q where $\beta(Q) \neq 0$, there exists a certain direction along which the functional value F is increasing. The corresponding equilibrium point $X = Q^T \Lambda Q$, therefore, has at least one unstable (repelling) direction. To converge to such an unstable equilibrium point, the descent flow $X(t)$ must stay on very special directrices that form a measure zero, nowhere dense subset in R^n. From the numerical analysis standpoint, this kind of equilibrium point will never be realized because computation along the directrix can easily be derailed by round-off errors and hence pushed away from the unstable equilibrium point.

4. Numerical examples. Since $\alpha(X)$ is a diagonal matrix, all diagonal matrices on $\mathcal{M}(\Lambda)$ are equilibrium points for (16). Thus one should avoid using Λ as the initial value X_0. One way to generate a reasonable initial value is by defining $X_0 := Q^T \Lambda Q$ with Q a random orthogonal matrix. There are many techniques for generating such a Q [1], [19].

All the following computations are done on a DECstation 5000/200 with double precision. We display all the numbers with only five digits so as to fit the data comfortably in the running text.

Example 1. To simulate reasonable test data, we start with a randomly generated symmetric matrix

$$M_0 = \begin{bmatrix} 4.3792\times10^{-1} & 4.3055\times10^{-1} & 1.2086\times10^{+0} & 1.0968\times10^{+0} & 1.4616\times10^{+0} \\ 4.3055\times10^{-1} & 1.0388\times10^{+0} & 1.5021\times10^{+0} & 7.2134\times10^{-1} & 1.4543\times10^{-1} \\ 1.2086\times10^{+0} & 1.5021\times10^{+0} & 1.5396\times10^{-2} & 9.7239\times10^{-1} & 7.2076\times10^{-1} \\ 1.0968\times10^{+0} & 7.2134\times10^{-1} & 9.7239\times10^{-1} & 1.8609\times10^{+0} & 1.2622\times10^{+0} \\ 1.4616\times10^{+0} & 1.4543\times10^{-1} & 7.2076\times10^{-1} & 1.2622\times10^{+0} & 1.4024\times10^{+0} \end{bmatrix}.$$

The diagonal entries

$$a = [4.3792\times10^{-1}, 1.0388\times10^{+0}, 1.5396\times10^{-2}, 1.8609\times10^{+0}, 1.4024\times10^{+0}]$$

and eigenvalues

$$\lambda = [-1.4169\times10^{+0}, -5.6698\times10^{-1}, 4.3890\times10^{-1}, 1.4162\times10^{+0}, 4.8842\times10^{+0}]$$

of M_0 are used as the test data. Now we randomly generate an orthogonal matrix

$$Q_1 = \begin{bmatrix} -6.4009\times10^{-1} & -5.3594\times10^{-1} & -1.8454\times10^{-1} & -3.3375\times10^{-2} & -5.1757\times10^{-1} \\ 2.1804\times10^{-1} & -1.2359\times10^{-1} & -5.0336\times10^{-1} & -8.2193\times10^{-1} & 9.0802\times10^{-2} \\ -7.2099\times10^{-1} & 5.6072\times10^{-1} & 1.4302\times10^{-2} & -2.4876\times10^{-1} & 3.2199\times10^{-1} \\ 2.8417\times10^{-3} & -1.9828\times10^{-1} & 8.4401\times10^{-1} & -4.9375\times10^{-1} & -6.7297\times10^{-2} \\ -1.5134\times10^{-1} & -5.8632\times10^{-1} & 3.0406\times10^{-3} & 1.3284\times10^{-1} & 7.8464\times10^{-1} \end{bmatrix},$$

and define $X_0 = Q_1^T \Lambda Q_1$. We use the subroutine ODE in [18] as the integrator where local control parameters ABSERR and RELERR in ODE are set to be 10^{-12}. We examine the output values at a time interval of 1, and assume the path has reached an equilibrium point when two consecutive output points are within a distance of 10^{-10}. At $t \approx 11$, the gradient flow converges to the matrix

$$M_1 = \begin{bmatrix} 4.3792\times10^{-1} & 2.6691\times10^{-1} & -1.9178\times10^{-1} & -6.1356\times10^{-1} & -1.5920\times10^{+0} \\ 2.6691\times10^{-1} & 1.0388\times10^{+0} & -7.2845\times10^{-1} & -8.6726\times10^{-1} & -1.9618\times10^{+0} \\ -1.9178\times10^{-1} & -7.2845\times10^{-1} & 1.5396\times10^{-2} & -6.3601\times10^{-1} & 1.6256\times10^{-1} \\ -6.1356\times10^{-1} & -8.6726\times10^{-1} & -6.3601\times10^{-1} & 1.8609\times10^{+0} & 1.5032\times10^{+0} \\ -1.5920\times10^{+0} & -1.9618\times10^{+0} & 1.6256\times10^{-1} & 1.5032\times10^{+0} & 1.4024\times10^{+0} \end{bmatrix}.$$

The flow in theory should stay in the surface $\mathcal{M}(\Lambda)$. The eigenvalues of M_1 are checked to agree with λ to 10 digits.

If we use another random orthogonal matrix

$$Q_2 = \begin{bmatrix} -4.7879\times10^{-1} & 8.7948\times10^{-2} & -4.1424\times10^{-3} & 3.0041\times10^{-1} & 8.2022\times10^{-1} \\ -4.1099\times10^{-1} & -5.7368\times10^{-1} & -6.8750\times10^{-1} & -7.0455\times10^{-2} & -1.5607\times10^{-1} \\ 1.7225\times10^{-1} & 6.1511\times10^{-1} & -6.1521\times10^{-1} & -4.2281\times10^{-1} & 1.8634\times10^{-1} \\ -7.1440\times10^{-1} & 2.5656\times10^{-1} & 3.2325\times10^{-1} & -5.0226\times10^{-1} & -2.5895\times10^{-1} \\ 2.4860\times10^{-1} & -4.6795\times10^{-1} & 2.1060\times10^{-1} & -6.8830\times10^{-1} & 4.4845\times10^{-1} \end{bmatrix},$$

then at $t \approx 13$, the gradient flow converges to the matrix

$$M_2 = \begin{bmatrix} 4.3792\times10^{-1} & -1.4087\times10^{+0} & 4.8811\times10^{-1} & -2.0882\times10^{+0} & 1.2285\times10^{+0} \\ -1.4087\times10^{+0} & 1.0388\times10^{+0} & 2.3067\times10^{-1} & 1.1160\times10^{+0} & -8.8543\times10^{-1} \\ 4.8811\times10^{-1} & 2.3067\times10^{-1} & 1.5396\times10^{-2} & -7.2958\times10^{-2} & 7.2054\times10^{-1} \\ -2.0882\times10^{+0} & 1.1160\times10^{+0} & -7.2958\times10^{-2} & 1.8609\times10^{+0} & -3.7601\times10^{-1} \\ 1.2285\times10^{+0} & -8.8543\times10^{-1} & 7.2054\times10^{-1} & -3.7601\times10^{-1} & 1.4024\times10^{+0} \end{bmatrix},$$

whose eigenvalues again agree reasonably well with λ.

Example 2. Under the same stopping criterion, we repeat the experiment in Example 1 with 2,000 test data. The diagonal entries a and eigenvalues λ are generated from symmetric matrices with normal distribution entries. The orthogonal matrices Q are generated from the QR decomposition of other stochastically independent (nonsymmetric) random matrices [19]. We collect the length of integration required for reaching convergence in each case. This length should be inherent only to the individual problem data (and the stopping criterion), but should be independent of the machine used in computation. The histogram of the lengths is presented in Fig. 2 where for better display the frequency distribution is plotted in its natural logarithm. When there is no distribution for a particular length (so the logarithm is negative infinity), the plot is left blank. As can be seen, about 77% of the cases converge with the length of integration less than 7 and about 93% converge with length less than 17. The maximal length of integration that occurred in this test is 296. It is perhaps also interesting to note that all the 2,000 cases converge to a desirable solution. Confirming our previous argument over Theorem 3.4, none of the cases gets trapped at a point where $F(Q) \neq 0$, although this kind of equilibrium points do exist.

Example 3. In this example, we experiment with the case when multiple eigenvalues $\lambda = [1, 1, 1, 1, 4]$ are present. We take $a = [1.0749, 1.3309, 1.1197, 2.3035, 2.1709]$. Using the random orthogonal matrix

$$Q = \begin{bmatrix} -4.8713\times10^{-2} & -1.3354\times10^{-1} & 9.4639\times10^{-1} & -1.1419\times10^{-1} & 2.6666\times10^{-1} \\ 9.8790\times10^{-1} & -4.3307\times10^{-2} & 7.2187\times10^{-2} & -5.1681\times10^{-2} & -1.1955\times10^{-1} \\ -6.9873\times10^{-2} & -4.2957\times10^{-1} & 1.8176\times10^{-1} & 6.5185\times10^{-1} & -5.9384\times10^{-1} \\ 3.0930\times10^{-2} & -8.5347\times10^{-1} & -2.5445\times10^{-1} & -8.1527\times10^{-2} & 4.4637\times10^{-1} \\ 1.2584\times10^{-1} & 2.5953\times10^{-1} & -3.6892\times10^{-2} & 7.4347\times10^{-1} & 6.0225\times10^{-1} \end{bmatrix},$$

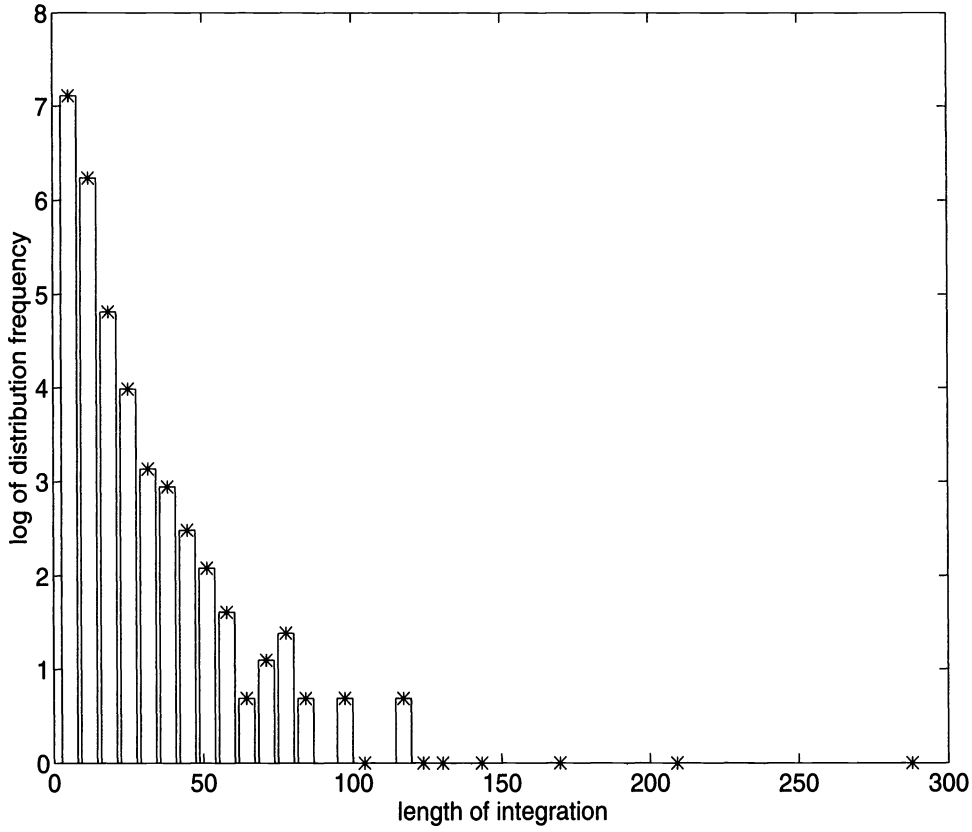


FIG. 2. *Histogram on the lengths of integration required for convergence.*

we find that a limit point exists at

$$M = \begin{bmatrix} 1.0749\text{x}10^{+0} & -1.5748\text{x}10^{-1} & -9.4707\text{x}10^{-2} & 3.1254\text{x}10^{-1} & 2.9622\text{x}10^{-1} \\ -1.5748\text{x}10^{-1} & 1.3309\text{x}10^{+0} & 1.9903\text{x}10^{-1} & -6.5679\text{x}10^{-1} & -6.2250\text{x}10^{-1} \\ -9.4707\text{x}10^{-2} & 1.9903\text{x}10^{-1} & 1.1197\text{x}10^{+0} & -3.9499\text{x}10^{-1} & -3.7437\text{x}10^{-1} \\ 3.1254\text{x}10^{-1} & -6.5679\text{x}10^{-1} & -3.9499\text{x}10^{-1} & 2.3035\text{x}10^{+0} & 1.2354\text{x}10^{+0} \\ 2.9622\text{x}10^{-1} & -6.2250\text{x}10^{-1} & -3.7437\text{x}10^{-1} & 1.2354\text{x}10^{+0} & 2.1709\text{x}10^{+0} \end{bmatrix}$$

when $t \approx 41$.

Example 4. In this example, we consider the case when multiple diagonal entries $a = [1, 1, 1, 1, 1]$ are present. Using $\lambda = [1.9747, 2.3050, 3.8938, -0.8128, -2.3608]$ and the random orthogonal matrix

$$Q = \begin{bmatrix} -3.3399\text{x}10^{-1} & 2.6628\text{x}10^{-1} & -2.3522\text{x}10^{-1} & -6.6904\text{x}10^{-1} & 5.6089\text{x}10^{-1} \\ -3.5191\text{x}10^{-1} & -8.8924\text{x}10^{-1} & 2.0821\text{x}10^{-1} & -1.9279\text{x}10^{-1} & 6.9964\text{x}10^{-2} \\ 2.6488\text{x}10^{-1} & -3.3460\text{x}10^{-1} & -8.6998\text{x}10^{-1} & 1.8120\text{x}10^{-1} & 1.6787\text{x}10^{-1} \\ 6.2901\text{x}10^{-1} & -1.2402\text{x}10^{-1} & 2.8591\text{x}10^{-2} & -6.7641\text{x}10^{-1} & -3.6141\text{x}10^{-1} \\ 5.4662\text{x}10^{-1} & -1.0493\text{x}10^{-1} & 3.7899\text{x}10^{-1} & 1.5765\text{x}10^{-1} & 7.2230\text{x}10^{-1} \end{bmatrix},$$

we find a limit point exists at

$$M = \begin{bmatrix} 1.0000\times10^{+0} & -1.4905\times10^{+0} & 1.1257\times10^{-1} & -1.4301\times10^{-1} & -1.6216\times10^{+0} \\ -1.4905\times10^{+0} & 1.0000\times10^{+0} & -8.1015\times10^{-2} & -4.5784\times10^{-1} & -7.5669\times10^{-1} \\ 1.1257\times10^{-1} & -8.1015\times10^{-2} & 1.0000\times10^{+0} & 1.4749\times10^{+0} & -2.1841\times10^{+0} \\ -1.4301\times10^{-1} & -4.5784\times10^{-1} & 1.4749\times10^{+0} & 1.0000\times10^{+0} & 4.3081\times10^{-1} \\ -1.6216\times10^{+0} & -7.5669\times10^{-1} & -2.1841\times10^{+0} & 4.3081\times10^{-1} & 1.0000\times10^{+0} \end{bmatrix},$$

when $t \approx 8$.

5. Conclusion. The Schur–Horn theorem guarantees that the inverse eigenvalue problem of constructing a Hermitian matrix with prescribed diagonal entries and eigenvalues always has a solution. The numerical methods described in [12] will not work generally for finding such a solution. We propose two methods. The lift-and-project method makes a connection with the Wielandt–Hoffman theorem. The gradient flow method can be integrated by any available ordinary differential equation solver. We show the gradient flow method always converges. Numerical experiment seems to suggest that the method works reasonably well.

Acknowledgments. The author wishes to thank Professor Leonid Faybusovich for pointing out a discussion in a paper by Professor Roger Brockett [20].

REFERENCES

[1] T. W. Anderson, I. Olkin, and L. G. Underhill, *Generation of random orthogonal matrices*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 625–629.

[2] B. C. Arnold, *Majorization and the Lorenz Order: A Brief Introduction*, Lecture Notes in Statist. 43, Springer-Verlag, Berlin, 1987.

[3] A. M. Bloch, *Steepest descent, linear programming, and Hamiltonian flows*, Contemp. Math., 114 (1990), pp. 77–88.

[4] J. P. Boyle and R. L. Dykstra, *A method for finding projections onto the intersection of convex sets in Hilbert space*, in Advances in Order Restricted Statistical Inference, Lecture Notes in Statist., Vol. 37, R. Dykstra, T. Robertston, and F. Wright, eds., Springer-Verlag, Berlin, 1986, pp. 28–47.

[5] F. Brauer and J. A. Nohel, *Qualitative Theory of Ordinary Differential Equations*, W. A. Benjamin, New York, 1969.

[6] R. W. Brockett, *Least squares matching problems*, Linear Algebra Appl., 122/123/124(1989), pp. 761–777.

[7] ———, *Dynamical systems that sort lists and solve linear programming problems*, in Proc. 27th IEEE Conference on Decision and Control, IEEE, (1988), pp. 799-803; Linear Algebra Appl., 146 (1991), pp. 79–91.

[8] W. Cheney and A. Goldstein, *Proximity maps for convex sets*, Proc. Amer. Math. Soc., 10(1959), pp. 448–450.

[9] M. T. Chu and K. R. Driessel, *The projected gradient method for least squares matrix approximations with spectral constraints*, SIAM J. Numer. Anal., 27 (1990), pp. 1050–1060.

[10] M. T. Chu, *Numerical methods for inverse singular value problems*, SIAM J. Numer. Anal., 29 (1992), pp. 885–903.

[11] F. Deutsch, *Von Neumann's alternating method: the rate of convergence*, Approximation Theory IV, C. Chui, L. Schumaker, and J. Ward, eds., Academic Press, New York, 1983, pp. 427–434.

[12] S. Friedland, J. Nocedal, and M. L. Overton, *The formulation and analysis of numerical methods for inverse eigenvalue problems*, SIAM J. Numer. Anal. 24 (1987), pp. 634–667.

[13] S. P. Han, *A successive projection method*, Math. Programming, 40 (1988), pp. 1–14.

[14] A. Horn, *Doubly stochastic matrices and the diagonal of a rotation matrix*, Amer. J. Math., 76(1956), pp. 620–630.

[15] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1991.

[16] A. W. Marshall and I Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.

[17] L. MIRSKY, *Matrices with prescribed characteristic roots and diagonal elements*, J. London Math. Soc., 33 (1958), pp. 14–21.

[18] L. F. SHAMPINE AND M. K. GORDON, *Computer Solution of Ordinary Differential Equations: The Initial Value Problem*, W. H. Freeman, San Francisco, 1975.

[19] G. W. STEWART, *The efficient generation of random orthogonal matrices with an application to condition estimators*, SIAM J. Numer. Anal., 17 (1980), pp. 403–404.

[20] R. W. BROCKETT, *Differential geometry and the design of gradient algorithms*, Proc. Symp. Pure Math., 54 (1993), pp. 69–92.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 218–231, January 1995

015

APPLICATION OF THE SMITH NORMAL FORM TO THE STRUCTURE OF LATTICE RULES*

J. N. LYNESS† AND P. KEAST‡

Abstract. Two independent approaches to the theory of the lattice rule have been exploited at length in the literature. One is based on the generator matrix A of the lattice Λ whose elements provide the abscissas of Q. The other, based on the t-cycle form $Q(\Lambda)f$ of Sloan and Lyness [*Math. Comput.*, 52 (1989), pp. 81–94], leads to a canonical form for Q. In this paper, a close connection between these approaches is demonstrated. This connection reflects the close relation between the Kronecker decomposition theorem for Abelian groups and the Smith normal form of an integer matrix. It is shown that the invariants of the canonical form of $Q(\Lambda)f$ coincide with the elements of the Smith normal form of $B = (A^T)^{-1}$, the reciprocal lattice generator matrix. This fact may be used to provide a straightforward solution to the previously intransigent problem of identifying and removing a repetition in the general t-cycle form.

Key words. Smith normal form, lattice rule, multidimensional quadrature, good lattice points

AMS subject classification. 65D32

1. Background and introduction.

A lattice rule is a multidimensional quadrature rule for integrating over an s-dimensional hypercube. In this section we provide a brief introduction to the theory followed by an outline of the contents of the rest of the paper. Without loss of generality we shall take the hypercube of integration to be $[0,1)^s$.

A *lattice* Λ is an infinite array of points. The standard definition demands that these points satisfy: (a) $\mathbf{p}, \mathbf{q} \in \Lambda$ implies $\mathbf{p}-\mathbf{q} \in \Lambda$ and (b) there exists no limit point; that is, there exists a positive $\epsilon(\Lambda)$ such that $|\mathbf{p}-\mathbf{q}| \geq \epsilon(\Lambda)$ unless $\mathbf{p} = \mathbf{q}$.

Of special note is the *s-dimensional unit lattice* Λ_0 that comprises all points whose components are all integers. An s-dimensional lattice, which contains the s-dimensional unit lattice as a sublattice, is known as an *integration lattice.* A *lattice rule* $Q(\Lambda)$ is defined only in terms of an integration lattice Λ. It has an abscissa set comprising all the points of this integration lattice Λ that lie in $[0,1)^s$ and applies an equal weight to each abscissa.

Matrix algebra is a useful tool for determining some of the properties of lattices in general and integration lattices in particular. A key concept is the *generator matrix* A of Λ. This is an $s \times s$ matrix, whose rows $\mathbf{a}_r$ are elements of Λ having the property that the lattice comprises all points $\mathbf{p}$ of the form

$$\mathbf{p} = \sum_{r=1}^{s} \lambda_r \mathbf{a}_r = \lambda A, \tag{1.1}$$

where λ_i are integer and $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_s)$.

The $s \times s$ unit matrix I is a generator matrix of the *unit lattice* Λ_0.

* Received by the editors August 12, 1991; accepted for publication (in revised form) by John Gilbert, October 13, 1993.

† Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439 (lyness@mcs.anl.gov). The work of this author was supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy contract W-31-109-Eng-38.

‡ Department of Mathematics, Statistics, and Computing Science, Dalhousie University, Halifax, Nova Scotia B3H 3J5, Canada. The work of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant OGP0002699.

Corresponding to an s-dimensional lattice Λ is its *reciprocal lattice* $\Lambda^{\perp}$; this plays a key role in the discretisation error theory of $Q(\Lambda)$ [Lyness 1989]. In the present context, it is convenient to define $\Lambda^{\perp}$ as the lattice having a generator matrix $B = (A^T)^{-1}$. It is relatively straightforward to show that the condition for Λ to be an integration lattice is that B, the generator matrix of $\Lambda^{\perp}$, be an integer matrix, i.e., every element of B be an integer. From this it follows immediately that every nonzero element of an integration lattice generator matrix A is a rational whose denominator is a factor of N, given by

$$N = |\det A|^{-1} = |\det B|. \tag{1.2}$$

N is of course an integer and it can be shown that it coincides with the number of lattice points in the hypercube $[0,1)^s$. These N points form the abscissa set of the *lattice rule* $Q(\Lambda)$ based on the integration lattice Λ. This rule applies an equal weight $1/N$ to each abscissa.

A generator matrix of a lattice Λ is not unique. A cursory examination of (1.1) above shows that the same set of points $\mathbf{p}$ is obtained when $\mathbf{a}_r$ is replaced by $-\mathbf{a}_r$, when $\mathbf{a}_r$ and $\mathbf{a}_t$ are interchanged, and when $\mathbf{a}_r$ is replaced by $\mathbf{a}_r + \mathbf{a}_t$ with $t \neq r$. These modifications to expression (1.1) may be effected by elementary integer row operations on the generator matrix A. Any such individual operation has the same effect on A as premultiplication by an integer matrix of very simple structure. This matrix is one of a type termed unimodular.

A *unimodular matrix* V is a square integer matrix of determinant ± 1. The product of unimodular matrices is itself unimodular. The inverse of a unimodular matrix is also unimodular, and elementary integer row (column) operations on a general matrix may be accomplished by pre (post) multiplication by a unimodular matrix. In particular, when A is a generator matrix of Λ, so is VA, where V is any unimodular matrix and all other generator matrices of Λ are of this form. (Note that the lattice generated by AV is generally different from Λ.)

The approach to the theory of lattice rules based on lattice generator matrices is described in more detail in [Lyness 1989], where the connection between the reciprocal lattice and the accuracy of the rule is described. A key result is that an integration lattice can be specified *uniquely* by the Hermite normal form of B. This has led to results about the number of distinct lattice rules [Lyness and Sørevik 1989] and the structure of embedded lattice rules [Lyness, Sørevik, and Keast 1991], and has provided much of the underlying theory needed to construct a complete search for good (cost-effective) lattice rules [Lyness and Sørevik 1991].

However, the original approach to the theory of lattice rules (e.g., see [Sloan 1985] and [Sloan and Lyness 1989]) is quite different in character. It appears to make no use whatever of matrix algebra, relying on a notation that is more appropriate to a quadrature rule. This notation is a development of a standard form for the *number theoretic rule*

$$Qf = \frac{1}{N}\sum_{j=1}^{N} f\left(\left\{\frac{j\mathbf{z}}{N}\right\}\right) \qquad \mathbf{z} \in \Lambda_0, \tag{1.3}$$

where $\{\mathbf{x}\}$ has its conventional meaning as the vector whose components are the fractional parts of those of $\mathbf{x}$. Specifically,

$$\{\mathbf{x}\} \in [0,1)^s; \quad \{\mathbf{x}\} - \mathbf{x} \in \Lambda_0. \tag{1.4}$$

It is customary to define $\bar{f}(\mathbf{x})$ as a unit periodic extension of $f(\mathbf{x})$ that coincides with $f(\mathbf{x})$ in the hypercube $[0,1)^s$. Thus $\bar{f}(\mathbf{x}) = f(\{\mathbf{x}\})$ and, using this notation, (1.3) takes the form

$$Qf = \frac{1}{N}\sum_{j=1}^{N} \bar{f}\left(\frac{j\mathbf{z}}{N}\right) \qquad \mathbf{z} \in \Lambda_0. \tag{1.5}$$

Number theoretic rules have been the subject of continuous and thorough investigation since their introduction by Korobov in 1959 [Korobov 1959]. A recent survey of this work appears in [Niederreiter 1988].

The number theoretic rule is itself a lattice rule. The key to understanding form (1.3) or (1.5) is to note the use of $\bar{f}$ in place of f. It appears that this has the effect of taking a set of N points arranged at equal intervals along a line in R^s, and translating each point individually so as to end up with a set of points that are distributed in $[0,1)^s$. It can be readily shown that these points are part of an s-dimensional integration lattice Λ and comprise all the points of Λ that lie in $[0,1)^s$. One motivation for this paper is to illuminate the connection between this lattice Λ and any of its generator matrices A.

In the rest of this paper, t is a positive integer, D is a diagonal $t \times t$ matrix whose elements d_i are positive integers, and Z is a $t \times s$ integer matrix whose rows are the vectors $\mathbf{z}_i$. What we term a *t-cycle D-Z form* of an s-dimensional lattice rule is an expression in the form of the right-hand side of

$$Qf = \frac{1}{d_1 d_2 \ldots d_t}\sum_{j_1=1}^{d_1}\sum_{j_2=1}^{d_2}\cdots\sum_{j_t=1}^{d_t} f\left(\left\{\frac{j_1\mathbf{z}_1}{d_1} + \frac{j_2\mathbf{z}_2}{d_2} + \cdots + \frac{j_t\mathbf{z}_t}{d_t}\right\}\right). \tag{1.6}$$

It is shown in Theorem 2.1 of [Sloan and Lyness 1989] that, so long as $t \geq 1$ and D is not singular, this form represents a lattice rule. This may be done by showing that all points lie on a lattice, that all points are in $[0,1)^s$, and that each point is assigned equal weight. The first two items are trivial. The third is straightforward, but leads us to one of the problems associated with this approach. It may well happen that the same point occurs more than once in the summation. However, if this happens, then every point is repeated the same number of times. The form is termed *k-repetitive*, or simply repetitive, when each point is repeated $k > 1$ times, in which case $\nu(Q)$, the number of abscissas required by the rule, is given by $d_1 d_2 \ldots d_t/k$ which is of course an integer. Unfortunately, the proof given in [Sloan and Lyness 1989] is not constructive. No immediate way of determining k from the elements of D and Z was then available.

It is important to bear in mind that the same rule may be represented using many different D-Z forms. Several examples of this are given in [Sloan and Lyness 1989]. Relatively trivial variants may be obtained by replacing any particular $\mathbf{z}_i$ in (1.6) by $k\mathbf{z}_i$ where k is any integer prime to d_i. When several components are linearly dependent, clearly they can be recombined into fewer components following the rules of linear algebra. However, operations reducing the number of components and changing their form are possible when all components are linearly independent. A very simple example of this is provided by the product trapezoidal rule. Thus, so long as d_1 and d_2 are mutually prime, we have

$$Qf = \frac{1}{d_1 d_2}\sum_{j_1=1}^{d_1}\sum_{j_2=1}^{d_2} f\left(\left\{\frac{j_1(1,0)}{d_1} + \frac{j_2(0,1)}{d_2}\right\}\right) = \frac{1}{d_1 d_2}\sum_{j=1}^{d_1 d_2} f\left(\left\{\frac{j(d_2,d_1)}{d_1 d_2}\right\}\right). \tag{1.7}$$

However, this equation is *not* valid unless $\gcd(d_1, d_2) = 1$.

It appears then that the same rule may be expressed in a t-cycle D-Z form in many different ways employing different values of t. The rule Q is defined to be of rank $r = r(Q)$ if it can be expressed in an r-cycle D-Z form, but not in an $(r-1)$-cycle D-Z form. An r-cycle form of Q is termed a *minimal* form. The principal result in [Sloan and Lyness 1989] is as follows. It is possible to express a rule of rank r in a nonrepetitive minimal form in such a way that $d_{i+1}|d_i$ $i = 1, 2, \ldots, r-1$. When this is the case, the $\mathbf{z}_i$ are linearly independent, and the elements d_i are known as *invariants*. These are unique; that is, every lattice rule Q has a unique rank and unique set of invariants.

This nomenclature is taken from group theory. The abscissa set of a lattice rule forms an Abelian group under addition modulo 1. The t-cycle D-Z form (1.6) corresponds to an expression of this group as a direct sum of cyclic groups in accordance with the famous decomposition theorem of Kronecker in 1877. See also [Hartley and Hawkes 1970, pp. 153–162]. Many of the results of Sloan and Lyness are obtained as applications of the group theory based on this theorem.

Any minimal D-Z form of Q in which the nonzero elements of D are the invariants is known as a *canonical form*. In this case the corresponding vectors $\mathbf{z}_i$ in (1.6) are linearly independent, but are not specified uniquely.

A trivial modification of this definition of a canonical form is employed in §4. It is clear from (1.6) that when any $d_j = 1$, the corresponding sum (over one element) may be omitted, whatever the corresponding vector $\mathbf{z}_j$ may be. In the sequel, on occasion, there will arise naturally what we term an s-cycle canonical form, where the rank of the rule is $r \leq s$. Such a canonical form has $d_j = 1$, $j \in [r+1, s]$; we may obtain a standard r-cycle canonical form by removing the final $s - r$ rows of D and Z and the final $s - r$ columns of D.

In this section, we have described very briefly the two principal approaches to lattice rule theory. These are through the lattice generator matrix A; see (1.1) and through the t-cycle D-Z form (1.6), respectively. This description provides a proper background and a coherent list of definitions.

In §2 we outline the theory as it exists for relating one approach to the other. This is not a long section, as this problem has not been treated seriously before. Sections 3 and 4 contain new results, based on the Smith normal form of a matrix having rational elements. In §3, we show how to obtain a canonical form directly from B, a generator matrix of the reciprocal lattice. In §4 we describe a shorter calculation to obtain a canonical form from a possibly repetitive form that bypasses the explicit calculation of either B or A. Section 5 contains numerical examples.

2. Some relations between the two approaches. In the preceding section, we described two distinct ways of specifying a lattice rule. One requires a single generator matrix A; the other requires a pair of matrices, D and Z. Sections 3 and 4 are concerned with developing an elegant connection between these different specifications of the same rule and between different D-Z specifications of the same rule. This section is devoted to the somewhat pedestrian methods currently available.

We note first that the lattice Λ formed by Qf in (1.6) includes all t points $\mathbf{z}_i/d_i$ $i = 1, 2, \ldots, t$ together with all points generated by them. These t points by themselves may not happen to generate an integration lattice. However, the expression (1.6) uses $\bar{f}$ in place of f. This implies that the fractional part of any of these t points also lies on the lattice Λ. The effect of this is that the lattice must contain the points of Λ_0,

and so in total includes all points of the form

$$\mathbf{p} = \sum_{i=1}^{t} j_i \mathbf{z}_i / d_i + \sum_{i=1}^{s} k_i \mathbf{e}_i, \tag{2.1}$$

and any point expressible in this form is a member of Λ. (Here, as is conventional, $\mathbf{e}_i$ is the ith unit s-vector.) In other words, the lattice Λ is generated by the rows of the $(t+s) \times s$ matrix

$$A^* = \begin{pmatrix} D^{-1}Z \\ I \end{pmatrix}. \tag{2.2}$$

However, as mentioned above, the same lattice is generated by any matrix obtainable from A^* by using elementary integer row operations. These have the same effect as premultiplying A^* by a unimodular $(t+s) \times (t+s)$ matrix V^*. Thus Λ is generated by the rows of any $s \times s$ matrix A satisfying

$$\begin{pmatrix} A \\ 0 \end{pmatrix} = V^* \begin{pmatrix} D^{-1}Z \\ I \end{pmatrix}. \tag{2.3}$$

A natural approach is to put A^* in upper triangular form, but any construction that results in t zero rows is sufficient.

Occasionally, one can pick out s rows from (2.2) by inspection. The following lemma may justify such a result.

LEMMA 2.1. *Let $Q(\Lambda)$ be given by an s-cycle D-Z representation, and set $\tilde{A} = D^{-1}Z$; then if $\Lambda(\tilde{A})$ is an integration lattice, it is the integration lattice of Q.*

Proof. $\Lambda(\tilde{A})$ is generated by the rows of $\tilde{A}$. Thus it includes all points of the form

$$\mathbf{p} = \sum_{i=1}^{s} j_i \mathbf{z}_i / d_i \qquad \mathbf{j} \in \Lambda_0. \tag{2.4}$$

Since $\tilde{A}$ is an integration lattice, it includes all points $\mathbf{e}_i$ $i = 1, 2, \ldots, s$. Thus, specification (2.4) coincides with specification (2.1) of the lattice Λ. □

For example, if Z is known to be unimodular, the following theorem allows us to write down a generator matrix directly.

THEOREM 2.2. *Let $Q(\Lambda)$ be given in an s-cycle D-Z representation with Z unimodular. Then this representation is nonrepetitive, and $A = D^{-1}Z$ is a generator matrix of Λ.*

Proof. Clearly,

$$B = (A^T)^{-1} = D(Z^T)^{-1},$$

being the product of two integer matrices, is an integer matrix. Thus, A generates an integration lattice and, in view of the previous lemma, is the generator matrix of Λ. Moreover, since $|\det Z| = 1$, we find

$$d_1 d_2 \ldots d_s = \det D = |\det A|^{-1} = N,$$

where N is the number of distinct abscissas used by $Q(\Lambda)$. Thus, the D-Z form is not repetitive. □

The reverse process, that of obtaining an s-cycle D-Z form of $Q(\Lambda)$ from a given generator matrix A of the integration lattice Λ is also straightforward. Let $\mathbf{a}_r$ be a

row of A and d_r be the smallest integer (or any integer) for which $\mathbf{z}_r = d_r\mathbf{a}_r \in \Lambda_0$. Then an s-cycle D-Z specification is given by the $s \times s$ matrix Z whose rows are $\mathbf{z}_r$ and $D = \text{diag}\{d_1, d_2, \ldots, d_s\}$. (Unfortunately, this simple approach gives, in general, a highly repetitive D-Z form.)

We believe that Theorem 2.2 is new and in simple examples may be helpful in recognizing a nonrepetitive form. But what is particularly noticeable in the results of this section is the absence of any *general* procedure for avoiding or recognizing a repetitive form or for producing a canonical form. A new way of carrying out these tasks, which leads directly to a canonical form, is given in the next section.

3. Reduction of B to D-Z form. We noted earlier that the same lattice may have many different generator matrices. These are related by elementary integer *row* operations. In particular, when B is any generator of a reciprocal lattice $\Lambda^\perp$, this same lattice is also generated by $B' = VB$ when V is any unimodular matrix. Successive row operations may be used to put B into *upper triangular lattice form* (utlf), in which all elements are nonnegative, and the largest element in any column lies on the diagonal. This is essentially the Hermite normal form. It has been exploited in previous papers to count the number of lattice rules, to obtain information about sublattices and superlattices, and to form the basis of a search program for good lattice rules (see, e.g., [Lyness, Sørevik, and Keast 1991].

As mentioned earlier, integer column operations applied to B (or *post*multiplication by unimodular matrices) result in a matrix that represents a different lattice. Nevertheless, if one allows column operations as well as row operations, one may diagonalize B. There are generally several ways of doing this though, of course, any such diagonal form has the same determinant (or product of nonzero diagonal elements). This procedure is significantly more involved than the procedure for the Hermite normal form but is reasonably straightforward. Since elementary operations may be used to interchange rows and columns, it is apparent that we may rearrange the order of these diagonal elements. However, there are, in addition, generally different possibilities for the set of diagonal elements. For example, the matrix given by

$$B = \begin{pmatrix} 7 & 14 & 21 \\ 35 & 73 & 117 \\ 7 & 20 & 66 \end{pmatrix} \tag{3.1}$$

can be reduced to diagonal form in many ways by using unimodular matrices. Two ways are as follows:

$$\begin{pmatrix} 5 & -1 & 0 \\ 1 & 0 & 0 \\ 9 & -2 & 1 \end{pmatrix} \begin{pmatrix} 7 & 14 & 21 \\ 35 & 73 & 117 \\ 7 & 20 & 66 \end{pmatrix} \begin{pmatrix} 2 & 1 & 5 \\ -1 & 0 & -4 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 21 \end{pmatrix} \tag{3.2}$$

and

$$\begin{pmatrix} 11 & -2 & 0 \\ -38 & 7 & 0 \\ 9 & -2 & 1 \end{pmatrix} \begin{pmatrix} 7 & 14 & 21 \\ 35 & 73 & 117 \\ 7 & 20 & 66 \end{pmatrix} \begin{pmatrix} -1 & -8 & 5 \\ 1 & 7 & -4 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 21 & 0 \\ 0 & 0 & 21 \end{pmatrix}, \tag{3.3}$$

where the pre and postmultiplying matrices are unimodular.

Apart from sign changes and from reordering the diagonal elements, these are, in fact, the only possibilities for diagonalizing this particular matrix B by unimodular transformations. This may be shown from theory developed in the nineteenth century.

The *Smith normal form* of B, denoted by $\mathrm{snf}(B)$, is a diagonalization of B using integer elementary row and column operations in which the nonzero diagonal entries satisfy $d_{j,j}/d_{i,i} = \text{integer}$ for all $j \geq i$. If the restriction that the diagonal entries be in nondecreasing order is removed, then the diagonal form is not unique. However, any ordering can be achieved by pre and postmultiplication by permutation matrices, which are unimodular.

THEOREM 3.1 ([Smith, 1861]). *Given a $t \times s$ matrix $\tilde{A}$ whose elements are rational numbers, there exist unimodular matrices V and U of sizes $t \times t$ and $s \times s$, respectively, such that*

$$\delta = snf(\tilde{A}) = V\tilde{A}U \tag{3.4}$$

is a $t \times s$ diagonal matrix having $\bar{t}$ nonzero elements which are rationals satisfying

$$\delta_{i+1,i+1}/\delta_{i,i} = \text{integer} \qquad i = 1, 2, \ldots, \bar{t} - 1. \tag{3.5}$$

The matrix δ is unique and is known as the Smith normal form of $\tilde{A}$. (But the matrices V and U are not unique.)

A convenient, accessible modern reference to this theory, which contains a brief proof of this theorem, is found in [Schrijver 1986, pp. 50–51] A deeper treatment, set in the appropriate number theory context, appears in Newman 1972. Algorithms to obtain the Hermite normal form and the Smith normal form have been published; see, for example, [Bradley 1971] and [Kannan and Bachem 1979]. In addition, [Maple V 1991] contains a procedure for finding D for integer matrices (but not U or V), while the group theory language Cayley [Cannon 1984] has the facility to compute D, U, and V.

The key theorem of this paper, which is a simple application of the theorem defining the Smith normal form, follows.

THEOREM 3.2. *Let $Q(\Lambda)$ be an s-dimensional lattice rule and let B be a generator matrix of the reciprocal lattice $\Lambda^{\perp}$. Then an s-cycle canonical form of $Q(\Lambda)$ is given by Z and D, where*

$$D = \mathrm{snf}(B) = VBU \quad \text{and} \quad Z = U^T, \tag{3.6}$$

U and V being unimodular.

Proof. Note that since B is an integer matrix, the elements of D are integers. Let us consider the lattice rule $Q(\Lambda')$ whose D-Z form comprises these particular matrices D and Z. Since Z is unimodular, we may invoke Theorem 2.2 to establish that $D^{-1}Z$ is a generator matrix of Λ'. This being so, since V^T is unimodular $A = V^T D^{-1} Z$ is also a generator matrix of Λ' and so, by elementary manipulation B is a generator matrix of $\Lambda'^{\perp}$. Since the lattice generated by B is unique and its reciprocal is unique, Λ' coincides with Λ in the theorem. To establish the theorem, we note that, in view of Theorem 3.1, the elements of D have the divisibility property required for a canonical form. □

COROLLARY 3.3. *Every lattice rule $Q(\Lambda)$ has an s-cycle canonical form with Z a unimodular matrix.*

COROLLARY 3.4. *The invariants (and rank) of $Q(\Lambda)$ coincide with the nonunit elements (and their number) of the Smith normal form of any generator matrix B of the reciprocal lattice of Λ.*

Let us now return to the numerical example. The lattice rule $Q(\Lambda)$, whose reciprocal lattice is generated by B in (3.1), is of rank 2, has invariants $n_1 = n_2 = 21$, and may be expressed in canonical D-Z form with $\mathbf{z}_1 = (-8, 7, 0)$ and $\mathbf{z}_2 = (5, -4, 1)$.

Note that by means of permutation matrices (which are unimodular), we can rearrange the order of the diagonal elements in the Smith normal form. And, if we abandon the divisibility property, we can usually find other sets of diagonal elements. In example (3.1), only the two possibilities arise, because these two diagonal matrices are the only ones with determinant 441 that have the correct Smith normal form. The diagonal matrix diag{3,3,49}, for example, cannot be obtained from B since it has Smith normal form diag{1,3,147}. This fact is worth mentioning, because when $D = VBU$ and D is diagonal but not necessarily in Smith normal form, the rule $Q(\Lambda)$ is also defined by D and $Z = U^T$. This is also an s-cycle nonrepetitive form but is not necessarily canonical.

It is almost self evident that the Smith normal form of the reciprocal of any nonsingular square matrix M is the reciprocal of the Smith normal form of M. It follows that the Smith normal form of the generator matrix $A = (B^T)^{-1}$ is the reciprocal of D in (3.6). One can write down immediately the correspondents of Theorem 3.2 and Corollary 3.4. These are given in Theorem 3.5 and Corollary 3.6.

THEOREM 3.5. *Let A be a generator matrix of Λ, and let $\delta = \mathrm{snf}(A) = VAU$, V and U being unimodular, then an s-cycle canonical form of $Q(\Lambda)$ is given by $D = \delta^{-1}$ and $Z = U^{-1}$.*

Naturally, the unimodular matrices V and U occurring here are the transposes of the inverses of those in (3.6).

COROLLARY 3.6. *The invariants (and rank) of $Q(\Lambda)$ coincide with the inverses of the nonunit elements (and their number) of the Smith normal form of any generator matrix A of Λ.*

Theorems 3.2 and 3.5 were discovered independently by Langtry in [Langtry 1995].

In this section we have provided a general method for obtaining a D-Z canonical form from a generator matrix A. In §2 we showed how to derive a generator matrix from a general t-cycle D-Z form. Taken together, we have an algorithm for finding the rank, invariants, and canonical form of any lattice rule expressed in t-cycle D-Z form. It is convenient to list the main steps of this algorithm here.

ALGORITHM I

(a) Construct the $(t+s) \times s$ matrix $\binom{\tilde{A}}{I}$, where $\tilde{A} = D^{-1}Z$.

(b) Put this matrix into upper triangular form; i.e., construct an $s \times s$ matrix A such that

$$\begin{pmatrix} A \\ 0 \end{pmatrix} = V^* \begin{pmatrix} D^{-1}Z \\ I \end{pmatrix}. \tag{3.7}$$

(Note that A is a generator matrix of Λ.)

(c) Construct the Smith normal form of A; thus $\bar{\delta} = VAU$.

Then $D_c = \bar{\delta}^{-1}$ and $Z_c = U^{-1}$.

If one simply requires the invariants (perhaps to determine the rank), one does not need to calculate U. However, in the Smith normal form reduction, the calculation of U (or of U^{-1} or both) can be effected in situ.

This algorithm is essentially the same as one proposed in [Langtry 1995]. We give a numerical example in §5 comparing it with another algorithm.

4. A shorter algorithm for the canonical form. We have used Algorithm I described above extensively as a service routine. Since most of our work has been theoretical in nature and speed is no object, we have found it quite satisfactory.

However, further analysis, developed in this section, leads to a faster and we believe a more elegant algorithm that is given at the end of this section.

A casual inspection of Algorithm I suggests that there may be room for improvement. To fix ideas, suppose that s is very much larger than t. The algorithm starts with a diagonal $t \times t$ matrix and a full $t \times s$ matrix and ends up with a diagonal $r \times r$ matrix and a full $r \times s$ matrix, where the rank $r \leq t$. In between, only integer row and column operations take place. One may be left wondering whether there is really any need to bring in the larger $(t+s) \times s$ matrix, operate on this to get an $s \times s$ matrix; then operate on this to end up with the smaller matrices? To be more specific, suppose $t = 1$. If the rule is not repetitive, it is already in canonical form. In the unlikely event that one applied Algorithm I, it would approach this almost trivial problem by forming an $(s+1) \times s$ matrix and putting it in upper triangular form. The present authors are not suggesting that a user would actually use this algorithm in such a trivial case. The challenge is to provide an algorithm which, when presented with a trivial situation, carries out its task correspondingly quickly. Algorithm II, at the end of this section, is more streamlined. It uses almost exclusively $t \times s$ matrices.

We start the theory by considering the Smith normal form

$$\delta = \mathrm{snf}(\tilde{A}) = V\tilde{A}U \tag{4.1}$$

of the matrix $\tilde{A} = D^{-1}Z$. We recall from Theorem 3.1 that δ is a $t \times s$ diagonal matrix whose only nonzero elements are rationals satisfying

$$\delta_{i+1,i+1}/\delta_{i,i} = \text{integer} \qquad i = 1, 2, \ldots, \bar{t} - 1, \tag{4.2}$$

where $\bar{t} \leq \min(s,t)$ is the number of nonzero elements of δ.

LEMMA 4.1. *Let the nonzero diagonal elements of* δ *in* (4.1), *expressed in* their lowest terms, *be*

$$\delta_{i,i} = m_i/n_i \qquad i = 1, 2, \ldots, \bar{t}. \tag{4.3}$$

Then

$$n_{i+1} | n_i \qquad i = 1, 2, \ldots, \bar{t} - 1. \tag{4.4}$$

Proof. The proof is elementary. From (4.2) we have

$$\frac{m_{i+1}}{n_{i+1}} \frac{n_i}{m_i} = \text{integer}.$$

Since n_{i+1} has no factor in common with m_{i+1}, it follows that n_{i+1} divides n_i. □

To proceed, we introduce s equations, each of which is an identity, and rewrite (4.1) in the form

$$\begin{pmatrix} V & 0 \\ 0 & U^{-1} \end{pmatrix} \begin{pmatrix} \tilde{A} \\ I \end{pmatrix} U = \begin{pmatrix} \delta \\ I \end{pmatrix}. \tag{4.5}$$

Here, as previously, I and U are $s \times s$ matrices and δ is a $t \times s$ diagonal matrix. The $\bar{t}$ nonzero diagonal elements of δ satisfy (4.3) and (4.4) above, with $(m_i, n_i) = 1$. We note that the $(t+s) \times (t+s)$ matrix on the left is unimodular. The thrust of the next lemma will be to provide a reduction in which the rational elements $\delta_{i,i}$ are replaced by integer inverse elements $1/n_i$.

LEMMA 4.2. *For all m,n such that* $(m,n) = 1$, *there exists a* 2×2 *unimodular matrix* V *such that*

$$V \begin{pmatrix} m/n \\ 1 \end{pmatrix} = \begin{pmatrix} 1/n \\ 0 \end{pmatrix}. \tag{4.6}$$

Proof. Since $(m,n) = 1$, there exist integers α, β such that $\alpha m + \beta n = 1$. It is trivial to verify that

$$V = \begin{pmatrix} \alpha & \beta \\ -n & m \end{pmatrix} \tag{4.7}$$

satisfies (4.6) and has unit determinant. □

COROLLARY 4.3. *Let* Δ *be the* $(t+s) \times s$ *matrix on the right-hand side of* (4.5), *its elements satisfying* (4.3) *and* (4.4). *Then there exists a* $(t+s) \times (t+s)$ *unimodular matrix* $V^{(w)}$ *such that* $V^{(w)}\Delta$ *differs from* Δ *only in the (w,w) element, which is replaced by* $1/n_w$ *and in the* $(w+t, w)$ *element, which is replaced by zero.*

Proof. $V^{(w)}$ differs from the unit matrix I only in that the four elements required to carry out row operations on rows w and $t+w$ are replaced by the four in Lemma 4.2, with m_w and n_w replacing m and n. □

THEOREM 4.4. *Given a* $t \times s$ *rational-valued matrix* $\tilde{A}$, *there exists an* $s \times s$ *unimodular matrix* U *and a* $(t+s) \times (t+s)$ *unimodular matrix* $\tilde{V}$ *having the property that*

$$\tilde{V} \begin{pmatrix} \tilde{A} \\ I \end{pmatrix} = \begin{pmatrix} \tilde{\delta} \\ J \end{pmatrix} U^{-1}, \tag{4.8}$$

where $\tilde{\delta}$ *is a diagonal* $t \times s$ *matrix whose nonzero elements satisfy*

$$\tilde{\delta}_{ii} = 1/n_i \qquad i = 1, 2, \ldots, \bar{t} \leq t$$

with integer n_i, *where*

$$n_{i+1} | n_i \qquad i = 1, 2, \ldots, \bar{t} - 1, \tag{4.9}$$

and each row of J *either is* 0 *or is* $\mathbf{e}_u$ *with* $u > \bar{t}$.

Proof. As mentioned before, (4.1) is equivalent to (4.5). We may premultiply successively by $V^{(1)}, V^{(2)}, \ldots, V^{(\bar{t})}$, these being defined in Corollary 4.3. The effect on the left-hand side of (4.3) is to replace $\begin{pmatrix} V & 0 \\ 0 & U^{-1} \end{pmatrix}$ by

$$\tilde{V} = V^{(\bar{t})} V^{(\bar{t}-1)} \ldots V^{(2)} V^{(1)} \begin{pmatrix} V & 0 \\ 0 & U^{-1} \end{pmatrix}, \tag{4.10}$$

which is obviously a $(t+s) \times (t+s)$ unimodular matrix. The effect on the right-hand side is to successively replace the only nonzero element in the wth row by $1/n_w$ and the $(w+t)$th row by zero leaving a matrix of the form given by the left member of the right-hand side of (4.8). This establishes the theorem. □

Our major result follows simply from Theorem 4.4.

THEOREM 4.5. *Let* $Q(\Lambda)$ *be given in a t-cycle D-Z form, and let* $\tilde{A} = D^{-1}Z$. *Let the Smith normal form of* $\tilde{A}$ *be* $\delta = V\tilde{A}U$ *and the nonzero elements of* δ *be* $\delta_{i,i} = m_i/n_i$, $i = 1, 2, \ldots, \bar{t}$ *in their lowest terms. Then an s-cycle canonical form of* $Q(\Lambda)$ *is given by*

$$D_c = \operatorname{diag}\{n_1, n_2, \ldots, n_{\bar{t}}, 1, \ldots, 1\} \quad \textit{and} \quad Z_c = U^{-1}. \tag{4.11}$$

Proof. As discussed in §2, the lattice Λ is generated by the rows of $\begin{pmatrix} \tilde{A} \\ I \end{pmatrix}$.

Since this is invariant under premultiplication by a unimodular matrix, this lattice is generated by the rows of the $(t+s) \times s$ matrix on the left-hand side of (4.8), which coincides with the $(t+s) \times s$ matrix on the right-hand side of (4.8). The $\bar{t}$ zero rows of J clearly play no part in this lattice generation and may be removed. The other $s - \bar{t}$ rows may be reordered in a natural way. We may identify D_c^{-1} and Z_c with the matrices remaining on the right-hand side of (4.8). □

In the expression for D_c in (4.11), there are $s - \bar{t}$ unit elements displayed. Besides these, some of the integers denoted by n_i may also be unity. Rank r is, of course, the number of nonunit diagonal elements in D_c and may be less than $\bar{t}$, which itself by definition cannot exceed $\min(s,t)$.

With this theorem at hand, we summarize the main steps of an algorithm in which an s-dimensional lattice rule $Q(\Lambda)$ given in a t-cycle D-Z form is put into canonical form.

ALGORITHM II

(a) Construct the $t \times s$ matrix $\tilde{A} = D^{-1}Z$.

(b) Construct the Smith normal form $\delta = V\tilde{A}U$ of $\tilde{A}$. This is a diagonal $t \times s$ matrix.

(c) Put the elements of δ in their lowest terms, that is, $\delta_{i,i} = \dfrac{m_i}{n_i}$ with $(m_i, n_i) = 1$. Then a canonical form of $Q(\Lambda)$ is given by

$$D_c = \operatorname{diag}\{n_1, n_2, \ldots, n_{\bar{t}}, 1, \ldots, 1\} \quad \text{and} \quad Z_c = U^{-1}.$$

If one simply requires the invariants (perhaps to determine the rank), one does not need to calculate U. However, in the Smith normal form reduction, the calculation of U (or of U^{-1} or both) can be effected in situ.

(d) If one requires a generator matrix of Λ or of $\Lambda^{\perp}$, one calculates $A = D_c^{-1}U^{-1}$, or $B = D_cU^T$.

5. Numerical examples. The first example in this section illustrates the solution of the same simple problem using in turn Algorithms I and II.

Example 1. Find a canonical form of

$$Qf = \frac{1}{81} \sum_{j_1=1}^{9} \sum_{j_2=1}^{9} \bar{f}\left(\frac{j_1(0,8,4)}{9} + \frac{j_2(6,5,7)}{9} \right). \tag{5.1}$$

Here we have

$$D = \begin{pmatrix} 9 & 0 \\ 0 & 9 \end{pmatrix}, \quad Z = \begin{pmatrix} 0 & 8 & 4 \\ 6 & 5 & 7 \end{pmatrix},$$

and step (a) of both algorithms requires

$$\tilde{A} = D^{-1}Z = \begin{pmatrix} 0 & 8/9 & 4/9 \\ 6/9 & 5/9 & 7/9 \end{pmatrix}. \tag{5.2}$$

Example 1 *using Algorithm* I. In step (b) of Algorithm I we reduce $\begin{pmatrix} \tilde{A} \\ I \end{pmatrix}$ to upper triangular form to find

$$\begin{pmatrix} A \\ 0 \end{pmatrix} = \begin{pmatrix} 3/9 & 0 & 0 \\ 0 & 1/9 & 5/9 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = V^* \begin{pmatrix} 0 & 8/9 & 4/9 \\ 6/9 & 5/9 & 7/9 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{5.3}$$

The 5×5 unimodular matrix V^* is not retained. The reduction is carried out using integer elementary row operations.

Step (c) requires us to find $\bar{\delta} = \mathrm{snf}(A) = VAU$. Thus

$$\bar{\delta} = \begin{pmatrix} 1/9 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1 \end{pmatrix} = V \begin{pmatrix} 3/9 & 0 & 0 \\ 0 & 1/9 & 5/9 \\ 0 & 0 & 1 \end{pmatrix} U. \tag{5.4}$$

This step is carried out using unimodular row and column operations. The 3×3 matrices U and V are not calculated, but U^{-1} is assembled as the calculation proceeds in a standard way and is given below. Thus we find

$$D_c = \bar{\delta}^{-1} = \begin{pmatrix} 9 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad Z_c = U^{-1} = \begin{pmatrix} 0 & 1 & 5 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{5.5}$$

Since this is of rank 2, the third component may be discarded, giving

$$Qf = \frac{1}{27} \sum_{j_1=1}^{9} \sum_{j_2=1}^{3} \bar{f}\left(\frac{j_1(0,1,5)}{9} + \frac{j_2(1,0,0)}{3} \right).$$

There are, in fact, many alternate choices for $\mathbf{z}_1$ and $\mathbf{z}_2$.

Example 1 *using Algorithm* II. This is shorter. Starting as before from (5.2), we proceed immediately to the Smith normal form. Thus

$$\tilde{\delta} = \begin{pmatrix} 1/9 & 0 & 0 \\ 0 & 4/3 & 0 \end{pmatrix} = V\tilde{A}U = V \begin{pmatrix} 0 & 8/9 & 4/9 \\ 6/9 & 5/9 & 7/9 \end{pmatrix} U. \tag{5.6}$$

As before, our program retained U^{-1} in situ. In accordance with step (c) of the algorithm, we obtain D_c from the denominators of $\tilde{\delta}$ by disregarding the 4 in $\tilde{\delta}_{2,2} = 4/3$ and filling in the diagonal with units. This gives

$$D_c = \begin{pmatrix} 9 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad Z_c = U^{-1} = \begin{pmatrix} -6 & 11 & 1 \\ -2 & 3 & 0 \\ 1 & -1 & 0 \end{pmatrix}. \tag{5.7}$$

This corresponds to (5.5) obtained using Algorithm I. Different vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ are obtained, but the resulting rule is the same and is in canonical form.

A minor variant of this problem, mentioned briefly in §2, is that of finding the integration lattice of lowest order which contains some specified points. In this problem one simply defines $\tilde{A}$ as an array of these points. There is no need to define D and Z. The following example illustrates this.

Example 2. Find a canonical D-Z form of the three-dimensional integration lattice of lowest order that includes the five points

$$\mathbf{z}_j = \left(\frac{1}{3j-1}, \frac{1}{3j}, \frac{1}{3j+1}\right) \qquad j = 1,2,3,4,5. \tag{5.8}$$

We shall determine in passing whether the lattice generated by these five points is an integration lattice; that is, are the unit vectors $\mathbf{e}_i$ already included? The matrix $\tilde{A}$ in part (a) of our algorithm above is

$$\tilde{A} = \begin{pmatrix} 1/2 & 1/3 & 1/4 \\ 1/5 & 1/6 & 1/7 \\ 1/8 & 1/9 & 1/10 \\ 1/11 & 1/12 & 1/13 \\ 1/14 & 1/15 & 1/16 \end{pmatrix} = \frac{1}{720720} \begin{pmatrix} 360360 & 240240 & 180180 \\ 144144 & 120120 & 102960 \\ 90090 & 80080 & 72072 \\ 65520 & 60060 & 55440 \\ 51480 & 48048 & 45045 \end{pmatrix}. \tag{5.9}$$

We now construct the Smith normal form. This is a 5×3 diagonal matrix with principal diagonal given by

$$\delta = V\tilde{A}U = \operatorname{diag}\{1/720720, 1/280, 3/20\}, \tag{5.10}$$

and the inverse of the matrix U used in the reduction

$$Z_c = U^{-1} = \begin{pmatrix} 385164 & 148148 & 99 \\ -33120301 & -12739230 & -8513 \\ 11831180 & 4550687 & 3041 \end{pmatrix}. \tag{5.11}$$

The *invariants* are given by the denominators in δ. That is,

$$D_c = \operatorname{diag}\{720720, 280, 20\}. \tag{5.12}$$

D_c and Z_c give a canonical form in D, Z notation.

Note that the lattice generated by the five rows of $\tilde{A}$ in (5.9) is not an integration lattice. We know this because the diagonal elements of δ in (5.10), in their lowest terms, are not all inverse integers. Ignoring the numerator 3 in the element $\delta_{33} = 3/20$ has the effect of increasing the density of lattice points by this factor.

6. Concluding remarks. From a technical point of view, the results in this paper merely show how to carry out various standard tasks relating to the manipulation of lattice rules. The tool is a standard technique to obtain the Smith normal form of an integer matrix. Using this normal form, we can readily find a Sloan–Lyness canonical form of $Q(\Lambda)$ from a generator matrix of Λ. And we can determine whether a given form of $Q(\Lambda)$ is repetitive by reducing it to a canonical form.

However, we believe that this paper has wider implications. The Smith normal form of an integer matrix is in fact the link between two apparently almost independent approaches to the theory of lattice rules. This is because the Smith normal form is a standard tool in the proof of the Kronecker decomposition theorem. The referee has pointed out that there is a sense in which this paper is, in effect, traversing a part of the proof of the decomposition theorem. In our opinion the principal virtue of the theorems in this paper is that they unite these two parts of the same theory to their mutual benefit.

REFERENCES

G. H. BRADLEY (1971), *Algorithms for Hermite and Smith normal matrices and linear diophantine equations*, Math. Comput., 25, pp. 897–907.

J. J. CANNON (1984), *An introduction to the group theory language, Cayley*, Computational Group Theory, Proceedings of the London Mathematical Society Symposium on Computational Group Theory, M. D. Atkinson, ed., Academic Press, New York.

B. HARTLEY AND T. O. HAWKES (1970), *Rings, Modules and Linear Algebra*, Chapman and Hall, Ltd., London.

R. KANNAN AND A. BACHIEM (1979), *Polynomial algorithms for computing the Smith and Hermite normal forms of an integer matrix*, SIAM J. Comput., 8, pp. 499–507.

N. M. KOROBOV (1959), *The approximate computation of multiple integrals*, Dokl. Akad. Nauk. SSSR, 124, pp. 1207–1210. (In Russian.)

T. N. LANGTRY (1995), *The determination of canonical forms for lattice quadrature rules*, J. Comput. Appl. Math., to appear.

J. N. LYNESS (1989), *An introduction to lattice rules and their generating matrices*, IMA J. Numer. Anal., 9, pp. 405–419.

J. N. LYNESS AND T. SØREVIK (1989), *The number of lattice rules*, BIT, 29, pp. 527–534.

J. N. LYNESS AND T. SØREVIK (1991), *A search program for finding optimal integration lattices*, Computing, 47, pp. 103–120.

J. N. LYNESS, T. SØREVIK, AND P. KEAST (1991), *Notes on integration and integer sublattices*, Math. Comput., 56, pp. 243–255.

MAPLE V (1991), *Maple V Language Reference Manual*, Springer-Verlag, New York.

M. NEWMAN (1972), *Integral Matrices*, Academic Press, New York.

H. NIEDERREITER (1988), *Quasi-Monte Carlo methods for multidimensional numerical integration*, in International Series of Numerical Mathematics, Vol. 85, Numerical Integration III, G. Hämmerlin and H. Brass, eds., Birkhauser-Verlag, Basel, pp. 157–171.

A. SCHRIJVER (1986), *Theory of Linear and Integer Programming*, Wiley & Sons, New York.

I. H. SLOAN (1985), *Lattice methods for multiple integration*, J. Comput. Appl. Math., 12, pp. 131–143.

I. H. SLOAN AND J. N. LYNESS (1989), *The representation of lattice quadrature rules as multiple sums*, Math. Comput., 52, pp. 81–94.

H. J. S. SMITH (1861), *On systems of linear indeterminate equations and congruences*, Philos. Trans. Roy. Soc. London (A), 151, pp. 293–326.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 232–234, January 1995

016

STRICT APPROXIMATION OF MATRICES*

K. ZIĘTAK†

Abstract. This paper describes a mechanism that includes the well-known strict approximation of a real vector which can be applied in the case of spectral approximation to define a unique strict spectral approximant of a matrix. For this purpose a new ordering is introduced.

Key words. strict approximation, lexicographic ordering, minimal elements, approximation of matrices

AMS subject classifications. 41A29, 41A65, 15A99

1. Introduction.

It is well known that if $||\cdot||$ is a norm on a finite-dimensional linear vector space and $\mathcal{K}$ is a nonempty closed convex subset of this space, then the minimum of $||a||$ over a in $\mathcal{K}$ is attained. Moreover, it follows from the convexity of the norm that the set $\mathcal{S}_1$ of minimizers is convex.

If the norm is strictly convex, then the set $\mathcal{S}_1$ must consist of a single point. This is true because if a and b are minimizers in $\mathcal{K}$, then the midpoint $\frac{1}{2}(a+b)$ is again in $\mathcal{K}$, and the norm of this point is lower than the common value of $||a||$ and $||b||$, so that a and b cannot be two distinct minimizers.

When the norm is not strictly convex, there may be more than one minimizer. It can be difficult to compute an element that is not uniquely defined, and for this reason it is useful to define a tie-breaking mechanism to distinguish exactly one minimizer.

For the case of the approximation of a real vector in the ℓ_∞ norm, such a tie-breaking mechanism has been introduced by Rice [4], who called the resulting approximant a *strict approximant.* This is a special type of Chebyshev approximation; therefore we call it a *strict Chebyshev approximation* (compare [1]).

In this paper we introduce a generalization of this mechanism that can be applied to the important problem of the best approximation of a matrix in an operator norm.

Let $\mathcal{M}$ be a nonempty closed convex set in the space $C^{m\times n}$ of $m \times n$ matrices, and let C be a given member of this space. Let $||\cdot||$ be the spectral norm of a matrix. The norm $||A||$ is equal to the square root of the largest eigenvalue of the matrix $A^H A$.

The problem of spectral approximation of a matrix C consists of minimizing $||B - C||$ over B in $\mathcal{M}$. This is, of course, equivalent to minimizing $||A||$ over A in the nonempty closed convex set

$$\mathcal{K} := \{B - C : B \in \mathcal{M}\}. \tag{1}$$

Because all the matrices $\operatorname{diag}(1, \alpha, \ldots, \alpha)$ with $-1 \le \alpha \le 1$ have the spectral norm equal to 1, this norm is not strictly convex. Therefore the set of all spectral approximations of C may contain more than one point.

We describe a tie-breaking mechanism that includes strict Chebyshev approximation and can be applied in the case of spectral approximation to define a unique strict spectral approximant.

* Received by the editors June 4, 1992; accepted for publication (in revised form) by H. F. Weinberger, September 22, 1993.

† Institute of Computer Science, University of Wrocław, ul. Przesmyckiego 20, 51-151 Wrocław, Poland (zietak@ii.uni.wroc.pl).

2. Lexicographic minimization. We first consider the general problem of minimizing some norm $||a||$ over a nonempty closed convex set $\mathcal{K}$ in a finite-dimensional linear vector space. As we mentioned at the beginning of the Introduction, the set $\mathcal{S}_1$ of minimizers for this problem is convex and nonempty. However, it may contain more than one point if the norm is not strictly convex. We define the following algorithm for shrinking the set of minimizers.

Denote the original norm to be minimized by $||\cdot||_1$. Introduce a finite sequence of other norms $||\cdot||_2, \ldots, ||\cdot||_\ell$ with the property that the norm $||\cdot||_1 + \cdots + ||\cdot||_\ell$ is strictly convex. Let $\mathcal{S}_1$ be the set of minimizers of $||a||_1$ on $\mathcal{K}$. Since $\mathcal{S}_1$ may contain more than one point, we attempt to break the tie between the elements of $\mathcal{S}_1$ by minimizing $||a||_2$ on the convex set $\mathcal{S}_1$. Let $\mathcal{S}_2$ denote the (nonempty) set of minimizers for this problem.

We continue this process by induction to obtain the sequence $\mathcal{K} \supseteq \mathcal{S}_1 \supseteq \mathcal{S}_2 \supseteq \cdots \supseteq \mathcal{S}_\ell$ of nonempty convex sets, where $\mathcal{S}_j$ is the set of minimizers of $||a||_j$ over a in $\mathcal{S}_{j-1}$.

There is another simple way to characterize the last set $\mathcal{S}_\ell$. We define the total ordering $a \succeq b$ to mean that for some $0 \leq k \leq \ell$, we have $||a||_j = ||b||_j$ for $j \leq k$ and $||a||_{k+1} > ||b||_{k+1}$ if $k < n$. For obvious reasons we call this *the lexicographic ordering with respect to the sequence of norms* $||\cdot||_1, \ldots, ||\cdot||_\ell$. Then $\mathcal{S}_\ell$ is the set of minimal elements of $\mathcal{K}$ with respect to this ordering.

We observe the following simple but useful fact, which states that our sequence of norms succeeds in breaking the tie in the minimization of $||\cdot||_1$.

THEOREM. *If the norm* $||\cdot||_1 + \cdots + ||\cdot||_\ell$ *is strictly convex, then the set* $\mathcal{S}_\ell$ *consists of a single point.*

Proof. Since the values of the first $\ell - 1$ norms are fixed on $\mathcal{S}_{\ell-1}$, the set $\mathcal{S}_\ell$ can be characterized as the set of minimizers of the norm $||\cdot||_1 + \cdots + ||\cdot||_\ell$ on the convex set $\mathcal{S}_{\ell-1}$. Because this norm is strictly convex, $\mathcal{S}_\ell$ consists of a single point. This completes the proof. □

Let $\mathcal{K}$ be determined as in (1) and let $t = \min\{m, n\}$. To define the strict spectral approximation, we let $\ell = t$ and construct the sequence of norms as follows. Let the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_t \geq 0$ of an m by n matrix A be ordered. They are defined by saying that the eigenvalues of the Hermitian positive semidefinite matrix $A^H A$ are $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_t^2$. Then σ_1 is the spectral norm of A. We denote $\sigma(A) := [\sigma_1, \ldots, \sigma_t]^T \in R^t$. The function

$$||A||_k := \{\sigma_1^2 + \cdots + \sigma_k^2\}^{1/2} \tag{2}$$

is a norm for each $k, k \leq t$. This is a particular case of the Ky Fan $p - k$ norm for $p = 2$ (see [2, p. 195]). For $k = t$ (2) is the Frobenius norm.

We now define the nested sequence of minimizing sets as above, with $\mathcal{S}_0 = \mathcal{K}$, and the set $\mathcal{S}_k$ of the minimizers of $||A||_{k-1}$ over $\mathcal{S}_{k-1}$. Because $||A||_{k-1}$ has a fixed value on $\mathcal{S}_{k-1}$ when $k > 1$, the set $\mathcal{S}_k$ can alternatively be defined as the set of minimizers of the singular value σ_k on $\mathcal{S}_{k-1}$. Let B from $\mathcal{M}$ be such that $B - C$ belongs to $\mathcal{S}_t$. Then B is called a *strict spectral approximant* of a given matrix C.

From (2) we see that $||A||_t$ is the Frobenius norm of A, and it is therefore strictly convex because this is just the Euclidean norm of the vector whose entries are the elements of A. Consequently, the sum of the norms $||\cdot||_k$ for k from 1 to t is also strictly convex, and from the above theorem we obtain the following corollary, which is our principal result.

COROLLARY 1. *In a given nonempty closed convex set* $\mathcal{M}$ *of* $m \times n$ *complex matrices there is exactly one strict spectral approximant* B *to any given matrix* C.

Let us consider the lexicographic ordering $\succeq$ with respect to the sequence of the norms (2). It is easily seen that $A_1 \succeq A_2$ if and only if $\sigma(A_1) \geq \sigma(A_2)$, where $\geq$ denotes the ordinary lexicographic ordering in R^t. Therefore we obtain the following corollary.

COROLLARY 2. *A matrix $B \in \mathcal{M}$ is the strict spectral approximation of C if and only if the vector $\sigma(B - C)$ is minimal with respect to the ordinary lexicographic ordering in the set $\{\alpha : \alpha = \sigma(B - C), B \in \mathcal{M}\}$.*

The characterization given in the Corollary 2 was used in [7] as the definition of the strict spectral approximation for the case in which $\mathcal{M}$ is a linear subspace of real matrices.

Remark. If the vector space is R^n with the ℓ_∞ norm, applying the above algorithm with the sequence of norms

$$||a||_k := \max_{1 \leq j_1 \leq \cdots \leq j_k \leq n} \{|a_{j_1}|^2 + \cdots + |a_{j_k}|^2\}^{1/2}$$

leads to strict Chebyshev approximation and the Theorem gives another proof of the well-known fact that the strict Chebyshev approximant exists and is unique (see, for example, [5, p. 243]; compare [3]).

At the end we mention that for the special case of square matrices with $\mathcal{M}$ the convex set of positive Hermitian matrices, it is easily verified that the strict spectral approximation coincides with the c_p-minimal positive approximant of Rogers and Ward [6].

Acknowledgments. The first draft of this paper was based on the results obtained in the report [7]. The author would like to express her deep gratitude to Professor H. F. Weinberger for many helpful criticisms and valuable suggestions that substantially improved the presentation and enabled her to obtain much more general results. Also, she is very grateful to Professor G. H. Golub for making possible her participation at the XII Householder Symposium where the first version of this paper was presented.

REFERENCES

[1] C. S. DURIS AND M. G. TEMPLE, *A finite step algorithm for determining the "strict" Chebyshev solution to $Ax = b$*, SIAM J. Numer. Anal., 10 (1973), pp. 690–699.

[2] R. A. HORN AND CH. R. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.

[3] R. HUOTARI AND WU LI, *The continuity of metric projection in $\ell_\infty(n)$, the Polya algorithm, the strict best approximation, and tubularity of convex sets*, J. Math. Anal. Appl., 182 (1994), pp. 836–856.

[4] J. R. RICE, *Tchebycheff approximation in a compact metric space*, Bull. Amer. Math. Soc., 68 (1962), pp. 405–410.

[5] J. R. RICE, *The Approximation of Functions*, Vol. II, Addison-Wesley, Reading, MA, 1969.

[6] D. D. ROGERS AND J. D. WARD, *C_p-minimal positive approximants*, Acta Sci. Math. (Szeged), 43 (1981), pp. 109–115.

[7] K. ZIĘTAK, *Strict spectral approximation of matrices*, Report N–218, Institute of Computer Science, University of Wrocław, Poland, 1992.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 235–253, January 1995

017

A CARTESIAN PARALLEL NESTED DISSECTION ALGORITHM*

MICHAEL T. HEATH† AND PADMA RAGHAVAN‡

Abstract. This paper is concerned with the distributed parallel computation of an ordering for a symmetric positive definite sparse matrix. The purpose of the ordering is to limit fill and enhance concurrency in the subsequent Cholesky factorization of the matrix. A geometric approach to nested dissection is used based on a given Cartesian embedding of the graph of the matrix in Euclidean space. The resulting algorithm can be implemented efficiently on massively parallel, distributed memory computers. One unusual feature of the distributed algorithm is that its effectiveness does not depend on data locality, which is critical in this context, since an appropriate partitioning of the problem is not known until after the ordering has been determined. The ordering algorithm is the first component in a suite of scalable parallel algorithms currently under development for solving large sparse linear systems on massively parallel computers.

Key words. parallel algorithms, sparse linear systems, ordering, Cartesian coordinates, nested dissection, Cholesky factorization

AMS subject classifications. 65F, 65W

1. Introduction.

The ordering of the equations and unknowns in a sparse system of linear equations can have a dramatic effect on the computational work and storage required for solving the system by direct methods. The reason is that most sparse systems suffer fill during the factorization process; that is, matrix entries that are initially zero become nonzero during the computation and the amount of such fill depends strongly on the ordering of the rows and columns of the matrix. Thus, ordering sparse matrices for efficient factorization is an important step in solving many large-scale computational problems in science and engineering, such as finite element structural analysis. In general, finding an ordering that minimizes fill is a very difficult combinatorial problem (NP-complete) [22]. Practical sparse factorization algorithms are therefore based on heuristically chosen orderings that are reasonably effective in limiting fill, but much less costly to compute than the optimum. Some of the most commonly used ordering heuristics are minimum degree, nested dissection, and various schemes for reducing the bandwidth or profile of the matrix [5].

In addition to determining fill, the ordering also affects the potential parallelism that can be exploited in factoring the matrix. These two considerations—reducing fill and enhancing parallelism—are largely compatible, but by no means coincident objectives. Sparsity and parallelism are positively correlated to some extent, since sparsity implies a lack of interconnections among matrix elements that often translates into computational subtasks that can be executed independently on different processors. This relationship is extremely complicated, however, and parallel efficiency depends on many other considerations as well, such as load balance and communication traffic. Thus, for example, minimum degree is in many cases the most effective heuristic known for limiting fill, but may produce orderings for which the natural load balance is uneven in parallel factorization. As another example, band-oriented methods,

* Received by the editors October 8, 1992; accepted for publication (in revised form) by J. R. Gilbert, October 13, 1993. This research was supported by the Defense Advanced Research Projects Agency through Army Research Office contract DAAL03-91-C-0047.

† Department of Computer Science and National Center for Supercomputing Applications, University of Illinois, 405 N. Mathews Ave., Urbana, Illinois 61801 (heath@ncsa.uiuc.edu).

‡ National Center for Supercomputing Applications, University of Illinois, 405 N. Mathews Ave., Urbana, Illinois 61801. Current address: Department of Computer Science, University of Tennessee, Knoxville, Tennessee 37996 (padma@cs.utk.edu).

however effective they may or may not be in limiting fill, tend to inhibit rather than promote concurrency in the factorization.

In this paper we are concerned with the problem of computing fill-reducing orderings for symmetric positive definite sparse matrices that will enable efficient Cholesky factorization on large-scale, distributed-memory parallel computers. Perhaps the most important consideration is that the ordering itself be computed in parallel on the same multiprocessor machine. Most previous work on parallel sparse matrix factorization has focused on the more costly (and more easily parallelized) numeric phases and has simply assumed that an appropriate and effective ordering could be precomputed on a serial machine (see [7] for a survey of this work). Such an approach is not scalable, however, as any such serial phase will eventually become a bottleneck as the problem size and number of processors grow. We therefore seek a distributed parallel ordering algorithm that can be integrated on the same machine with the subsequent parallel numeric computation and maintain reasonable efficiency over a wide range of parallel architectures and number of processors. Additional issues that concern us are the fill (and hence work and storage) that results from a given ordering, and also the resulting concurrency, load balance, and communication traffic in computing the Cholesky factor on such a parallel computer.

Designing an efficient, scalable, distributed ordering algorithm for sparse matrices presents a formidable challenge. The best serial ordering algorithms have evolved over an extended period of time and are very efficient. Much of this efficiency results from sophisticated data structures and algorithmic refinements that are difficult to extend to a distributed parallel setting. Moreover, many of these algorithms involve inherently serial precedence constraints and have relatively little computation over which to amortize the communication necessary in a parallel implementation. Perhaps most daunting of all, we seem to have a bootstrapping problem in that the efficiency of most distributed parallel algorithms depends on having a high degree of data locality, but we do not know how to partition our problem and distribute it across the processors until after we have an ordering. We therefore propose an ordering algorithm that lends itself to a distributed parallel implementation whose effectiveness does not depend on initial data locality.

2. Background. Throughout this paper we assume familiarity with numerous basic concepts in sparse matrix computations. Such background material can be found, for example, in the textbook by George and Liu [5]. In particular, we use the standard graph model for sparse Gaussian elimination, which we briefly explain here. The graph of an $n \times n$ symmetric matrix A is an undirected graph having n vertices, with an edge between two vertices i and j if the corresponding entry a_{ij} is nonzero in the matrix. We use the notation $G = (V, E)$ to denote the vertex and edge sets, respectively, of a graph G. The structural effect of Gaussian elimination on the matrix is easily described in terms of the corresponding graph. The fill introduced into the matrix as a result of eliminating a variable adds fill edges to the corresponding graph so that the neighbors of the eliminated vertex become a clique. A small example graph and corresponding matrix A are shown in Fig. 1. Also shown is the fill in the Cholesky factor L of the example matrix, where $A = LL^T$.

2.1. Nested dissection. Nested dissection is a divide-and-conquer strategy for ordering sparse matrices, originally due to Alan George [3]. Let V_s be a set of vertices (called a separator) whose removal, along with all edges incident on vertices in V_s, disconnects the graph into two remaining subgraphs, $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$. If the matrix is reordered so that the vertices within each subgraph are numbered

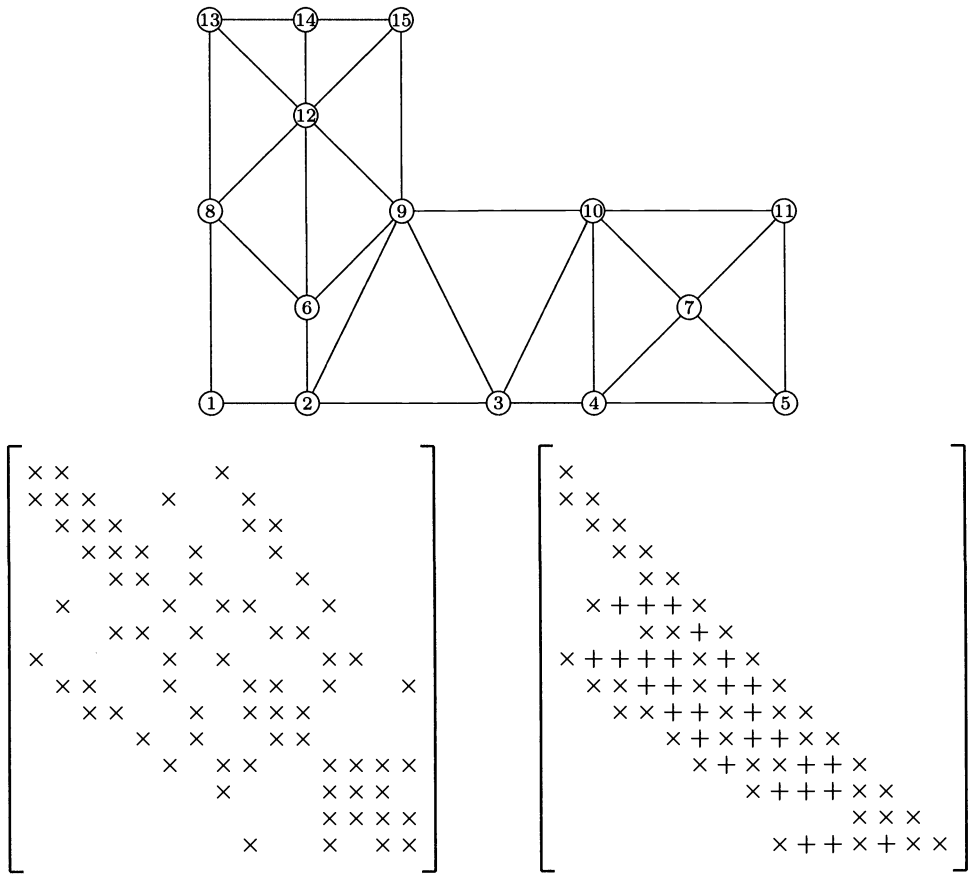


FIG. 1. *Example of finite element graph (top) and the nonzero patterns of the corresponding sparse matrix (left) and its Cholesky factor (right) with fill indicated by +.*

contiguously and the vertices in the separator are numbered last, then the matrix has the bordered block diagonal form

$$A = \begin{bmatrix} A_1 & 0 & S_1 \\ 0 & A_2 & S_2 \\ S_1^T & S_2^T & A_s \end{bmatrix}.$$

The significance of the above partitioning of the matrix is twofold: first, the zero blocks are preserved in the factorization, thereby limiting fill; second, factorization of the matrices A_1 and A_2 can proceed independently, thereby enabling parallel execution on separate processors. This idea can be applied recursively, breaking each subgraph into smaller and smaller pieces with successive separators, giving a nested sequence of dissections of the graph that inhibit fill and promote concurrency at each level.

Figure 2 shows our original example reordered by nested dissection. In the subsequent Cholesky factorization, the reordered matrix suffers considerably less fill than with the original ordering and also permits greater parallelism. For example, columns 1, 2, 3, 7, and 8 of the Cholesky factor depend on no prior columns, and hence can be computed simultaneously, whereas in the original ordering every column of the Cholesky factor depends on the immediately preceding column.

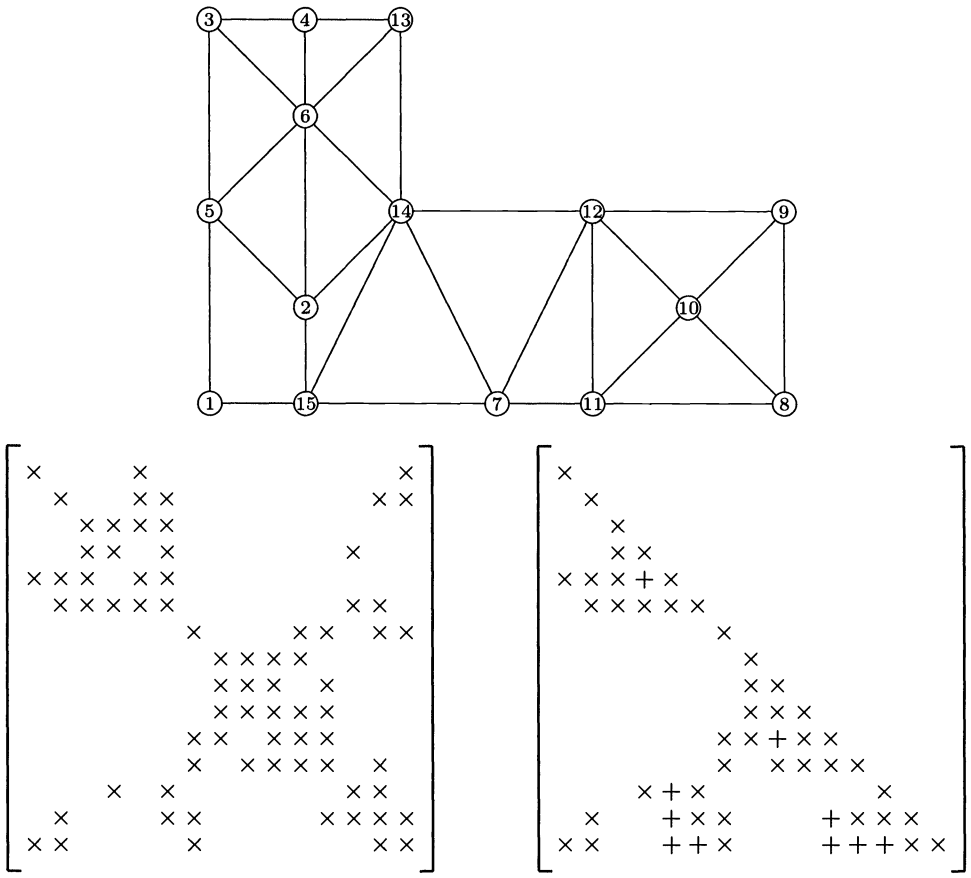


FIG. 2. *Finite element graph reordered by nested dissection (top) and the nonzero patterns of the corresponding sparse matrix (left) and its Cholesky factor (right) with fill indicated by +.*

The effectiveness of nested dissection in limiting fill depends on the size of the separators that split the graph, with smaller separators obviously being better. For planar problems (e.g., two-dimensional finite difference or finite element grids), suitably small separators can usually be found [10]. For problems in dimensions higher than two, or for highly irregular problems with less localized connectivity, nested dissection tends to be less effective, but so do most other ordering heuristics, which explains why iterative methods are often preferred over direct methods in such circumstances. In this paper we focus on problems for which an embedding of the graph in the two-dimensional Euclidean plane is given, but whose graph is not necessarily planar. Such a problem might result, for example, from two-dimensional finite element structural analysis. Indeed, our test problems are obtained from standard commercial structural analysis packages that routinely supply Cartesian coordinates for the vertices. Our approach generalizes to three dimensions in a reasonably straightforward manner [17].

In addition to the size of a separator, the relative sizes of the resulting subgraphs is also important. Maximum benefit from the divide-and-conquer approach is obtained when the remaining subgraphs are of about the same size; an effective nested dissection algorithm should not permit an arbitrarily skewed ratio between the sizes of the pieces.

In a parallel setting, this criterion takes on additional significance in that it largely determines the load balance of the computational subtasks assigned to individual processors. Thus, the choice of separators should take into account both size and balance.

Nested dissection algorithms differ primarily in the heuristics used for choosing separators. A typical approach to automatic nested dissection for irregular graphs [4] involves first finding a "peripheral" vertex, generating a level structure based on the connectivity of the graph, and then choosing a "middle" level of vertices as the separator. Such an approach is difficult to implement efficiently on a distributed parallel computer for a number of reasons, including the necessary serialization of some of the steps, and the communication required to assess the connectivity of the graph, especially *before* the graph has been partitioned so that data locality can be maintained (i.e., contiguous pieces are assigned to individual processors and "nearby" pieces assigned to "nearby" processors). More recent heuristics for computing graph separators include spectral methods [9], [16] and methods based on geometric projections and mappings [13]–[15], [20]. These may have greater potential for parallel implementation, but this has yet to be demonstrated in practice. An explicitly parallel implementation of the Kernighan–Lin algorithm for computing graph separators can be found in [6].

In this paper we present a nested dissection algorithm based on a new approach to computing separators, one that is designed to be effective in a distributed parallel environment. Our approach differs from standard graph-theoretic methods for computing separators in that it uses an embedding of the vertices in Cartesian space to facilitate an efficient parallel implementation that does not depend on initial data locality. For this reason, coordinate information has previously been used in other contexts to partition problems (particles, grids, etc.) across multiple processors. For example, such a coordinate-based "recursive bisection" approach has been used for load balancing of parallel computations [2, p. 430], [21] and for domain decomposition in solving partial differential equations on regular [1] and irregular [19] domains. In matrix factorization, however, potential fill must also be taken into account, in addition to the numerical balance of the partitioning. A key feature of our approach is that it permits one to exploit the tradeoff between fill and parallelism by finding better separators when unbalanced partitions are acceptable.

2.2. Cartesian representation. One motivation for our use of a Cartesian representation of the graph is to make the data "self identifying." This will be important when we consider implementing the algorithm on distributed memory parallel computers. In particular, the data can be scattered randomly across the local memories of the processors, yet we can still tell where (geographically) any given piece of data lies within the overall problem, without needing any communication to establish context. In effect, this approach makes the distributed memory "content addressable," thereby reducing much of the problem of computing separators to relatively simple counting and searching operations, which can be done very effectively in a distributed manner.

For each vertex $v \in V$, we assume that we are given a pair of Cartesian coordinates, which we denote by $x(v)$ and $y(v)$, representing the horizontal and vertical coordinate directions, respectively, in the Euclidean plane. One might wish to apply a rotation to the coordinate system to place the graph into some more advantageous orientation; we assume that this has already been done, if desired. As several authors have observed, e.g., [21], one possible way to determine a good orientation would be to compute the axis of minimum inertia of the vertices as a collection of points in

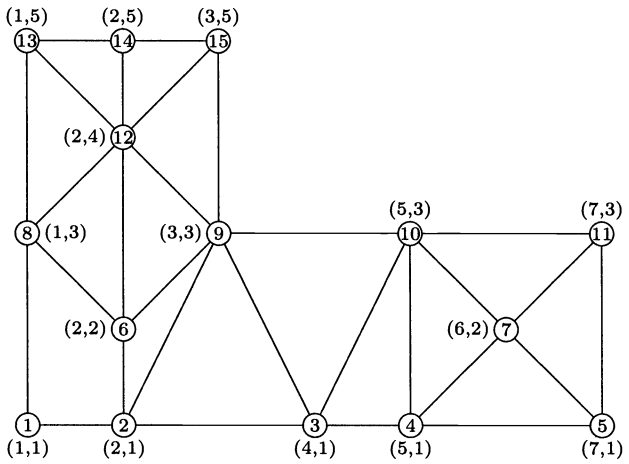


FIG. 3. *Finite element graph with Cartesian coordinates of nodes shown.*

the plane. For ease of handling of list structures, we use an integer representation of the original coordinate values. Treating each coordinate dimension separately, the original coordinate values are sorted and then each coordinate value is represented by its position in the sorted sequence. Figure 3 shows our example graph with Cartesian coordinates for the nodes.

3. Cartesian separators. We now describe our strategy for computing a vertex separator in a Cartesian labeled graph $G = (V, E)$. Let s be a coordinate value chosen in one of the two coordinate dimensions, say x. We refer to s as a "separating value" because it is used to dissect the graph along the given coordinate dimension. Let U_1, U_2, and U_s be the sets of all vertices whose x coordinate is less than s, greater than s, and equal to s, respectively. This partitioning of the nodes in the graph does not necessarily give us a vertex separator, because there may still be paths connecting vertices in U_1 and U_2. However, any such path must contain an edge that "straddles" the separating value s. Let E_s be the set of all such straddle edges, i.e.,

$$E_s = \{(u_1, u_2) \in E : u_1 \in U_1, u_2 \in U_2\}.$$

For each edge $(u_1, u_2) \in E_s$, arbitrarily select one of its two associated vertices for inclusion in the set C_s, which we refer to as the "correction set" for V_s. We now define the following sets:

$$V_1 = U_1 \setminus C_s, \quad V_2 = U_2 \setminus C_s, \quad V_s = U_s \cup C_s.$$

The set V_s is a vertex separator for the graph, since each vertex in V_1 is connected only to vertices in V_s or other vertices in V_1, and similarly for V_2. We refer to such a separator as a "Cartesian separator"; henceforth, when we use the term separator we will mean a Cartesian separator.

We illustrate these concepts for the example of Fig. 3. Using $s = 3$ as a separating

value in the x dimension, we get the initial sets

$$U_1 = \{1, 2, 6, 8, 12, 13, 14\}, \quad U_2 = \{3, 4, 5, 7, 10, 11\}, \quad U_s = \{9, 15\}.$$

The set of straddle edges is the singleton set $E_s = \{(2, 3)\}$. Choosing one of the endpoints of this edge, say node 2, we get the correction set $C_s = \{2\}$. Thus, the final subgraphs and separator are given by

$$V_1 = \{1, 6, 8, 12, 13, 14\}, \quad V_2 = \{3, 4, 5, 7, 10, 11\}, \quad V_s = \{2, 9, 15\}.$$

It is not difficult to devise graphs for which even the best Cartesian separator is much larger than necessary. For example, a one-dimensional grid wound into a spiral in the plane will be cut many times by any bisecting line, but can be separated evenly by removing a single vertex. Similarly, a planar graph consisting of n concentric squares whose corresponding corners are connected can be separated evenly by removing only four vertices, yet any bisecting line will cut $2n$ edges, giving a separator of size n. However, we have found Cartesian separators to be very effective for separating graphs that arise in practice (see the computational results in §7 below and also in [17]). In the next sections we proceed to discuss the two main subproblems in computing a Cartesian separator:

(i) Determining an appropriate choice for the separating value s;
(ii) Determining the correction set C_s.

3.1. Choosing a separating value. As we observed earlier, the two main criteria for choosing a separator are that the separator be small and that the resulting subgraphs be well balanced (i.e., about equal in size). These criteria are generally in conflict, so there is a tradeoff between them. In choosing a separating value s for computing a Cartesian separator in a given dimension, the balance between the sizes of the resulting subgraphs is determined by the relative numbers of vertices having coordinates less than s or greater than s in that dimension. Thus, we can attain any desired degree of balance, including optimal balance, simply by counting vertices (assuming that vertices are chosen appropriately for the correction set C_s to maintain the initial balance). Determining the size of a Cartesian separator, on the other hand, is more difficult, since the initial set U_s of vertices with coordinates equal to s is merely an initial approximation that must be augmented by the correction set C_s, whose size is not so easily determined. In seeking a small separator we will, for efficiency, merely estimate the eventual separator size rather than compute it exactly. For a given coordinate value s, we define the quantity

$$\eta(s) = |U_s| + |E_s|,$$

where the sets U_s and E_s are as defined previously. Clearly, $\eta(s)$ is an upper bound on the separator size; it may be an overestimate because a single vertex may "cover" more than one straddle edge in E_s, so that $|C_s|$ may be smaller than $|E_s|$. Nevertheless, $\eta(s)$ is sufficiently accurate for our purposes, and we will use it as an estimate for the separator size in seeking an approximate minimum.

The desired balance between the two subgraphs resulting from a single dissection is given by a user-specified quantity, α, $0 < \alpha < 1$, which is interpreted as a limit on the relative proportion between the sizes of the two subgraphs. Specifically, we require that the separating value s be chosen so that

$$\alpha|V| \leq |U_i| \leq (1 - \alpha)|V|, \qquad i = 1, 2.$$

A value of $\alpha = \frac{1}{3}$, for example, means that one subgraph can be at most twice the size of the other. There may be many potential separating values that satisfy this balance condition, with some values resulting in smaller separators than others. We choose the value s that minimizes the estimate $\eta(s)$ for the separator size. We handle the special case $\alpha = \frac{1}{2}$ separately, since it requires perfect balance (as closely as possible) regardless of the resulting separator size, and hence the estimate $\eta(s)$ need not be computed.

We illustrate these concepts for the example of Fig. 3, working with the x dimension. If $\alpha = \frac{1}{3}$, then a separating value of either $s = 3$ or $s = 4$ satisfies the balance criterion. Calculating the estimated separator size for each of these values, we get $\eta(3) = 3$ and $\eta(4) = 2$, so that we would choose $s = 4$ as the best separating value in this case. If $\alpha = \frac{1}{5}$ instead, then any separating value in the interval [2], [5] would satisfy the balance criterion, but the estimated separator sizes would still give $s = 4$ as the best choice.

We now sketch an algorithm for computing a separating value that minimizes the approximate separator size subject to the specified balance constraint. Our algorithm is formulated in terms of traversing sorted lists and computing suitable counts. This relatively simple serial algorithm serves to introduce appropriate terminology, notation, and data structures, providing a framework for our subsequent development of a distributed parallel algorithm. For definiteness, assume that we are working with the x coordinate dimension; similar definitions and procedures are also applicable to the y dimension. In general, we process both dimensions in the same fashion and use whichever yields the smaller separator. When this procedure for computing separators is used repeatedly in nested dissection, a different coordinate dimension may be selected at each stage.

For a given graph $G = (V, E)$, the vertices in V are maintained in a *vertex list*, in increasing order of their x coordinate values. The vertex list is traversed to compute a *vertex count list* of counts of vertices in G at each coordinate value, in increasing order in the x dimension. The vertex count list is traversed in increasing order and the cumulative count of vertices incremented until the first value is found, say a, that satisfies the balance condition. Traversal of the list then continues until a value is found at which the balance condition is no longer satisfied; we denote by b the last value at which the balance condition was still satisfied. Alternatively, depending on which would give the smallest expected running time, b could instead be found by traversing the vertex count list in decreasing order from the top. In either case, we will have identified the block $[a, b]$ of potential separating values, all of which satisfy the balance condition.

We must now compute the estimate $\eta(i)$ for each value $i \in [a, b]$. Let (u, v) be an edge in E, with $x(u) \leq x(v)$. Such an edge can be thought of as beginning at $x(u)$ and ending at $x(v)$. Let $\beta(i)$ and $\varepsilon(i)$ denote the number of edges that begin and end, respectively, at i. Edges in E are maintained in an *edge list* in increasing order of the x coordinates of their associated vertices. An edge (u, v) is entered into the ordered edge list at positions given by $x(u)$ and $x(v)$, where $x(u) < x(v)$, and marked respectively as a begin and an end entry. The edge list is traversed to compute an *edge count list* containing the β and ε values. If $\kappa(i)$ denotes the number of edges that cross i, then $\kappa(i) = \kappa(i-1) + \beta(i-1) - \varepsilon(i)$. This fact is used to compute $\kappa(i)$ for each value in the block $[a, b]$ by traversing the edge count list. We note also that the size of the initial approximation to the separator, $|U_i|$, can be computed for each coordinate value i by scanning the vertex count list. Finally, we note that for each coordinate

value i, our estimate for the final separator size is given by $\eta(i) = |U_i| + \kappa(i)$. Having computed the value of $\eta(i)$ for each $i \in [a, b]$, we select the coordinate value s with the minimum value of $\eta(s)$ as the separating value for that dimension. A separating value is similarly computed for the other coordinate dimension, and the one yielding the smaller estimated separator size is selected as the separating value for computing a Cartesian separator.

3.2. Constructing a separator. Having chosen a separating value s in one of the coordinate dimensions, we now proceed to construct a Cartesian separator. Again, for definiteness, assume that we have chosen the x coordinate dimension. According to our earlier definition, the desired separator V_s is the union of the initial approximate separator U_s and the correction set C_s. The set U_s is easily computed using the vertex list. The construction of the correction set C_s requires that we compute the set E_s of edges that straddle the separating value s. A simple way to search for these straddle edges would be to traverse the edge list in increasing order up to value s. For each beginning edge (u, v), with $x(u) \leq x(v)$, we add the edge to E_s if $x(v) > s$. Upon reaching value s in traversing the edge list, we have completed the computation of E_s. We initialize the set V_s to be U_s, then for each edge in E_s such that neither of its endpoints is already in V_s, we augment V_s by one of those endpoints. The choice of which endpoint to include in V_s can be made arbitrarily, or the choice can be governed by requiring that the balance condition be maintained.

In the worst case, the computational cost of this simple algorithm for finding straddle edges is proportional to the number of edges in the subgraph. This cost can be reduced by using the concept of a *group tree* [18], which enables more efficient searching for intervals that contain a given point s. In the group tree approach, we associate each edge in the subgraph with the interval whose endpoints are the coordinate values in the given dimension of the corresponding pair of vertices. The two resulting group trees (one for each coordinate dimension) are formed initially for the entire graph G, and thereafter can be modified easily for use in the searches at successive levels of nested dissection. Not counting this initialization cost, the cost of finding the straddle edges for a given subgraph G_i using a group tree search is proportional to $\log_2 |V_i|$ plus the number of edges found. This is a substantial improvement over the cost of the simpler algorithm described earlier, which is linear in $|E_i|$. With initialization costs included, the two methods have the same order of complexity, but the group tree approach may still provide a substantial savings in the constant factor, especially in the balanced case ($\alpha = \frac{1}{2}$).

4. Cartesian nested dissection. Having described an algorithm for computing a Cartesian separator for a given graph, we can use the algorithm repeatedly to derive an algorithm for Cartesian nested dissection to order a sparse matrix. The most natural way to implement such an algorithm is to invoke the separator algorithm recursively on successively smaller subgraphs of the initial graph. For reasons that will become clear later, we take a breadth-first approach, dealing with all of the subgraphs at a given level of dissection before moving on to the next level.

We introduce some notation here that we will find useful later on in formulating the parallel algorithm. For any given level l of the nested dissection process, we let $\mathcal{G}_l$ denote the set of subgraphs of the initial graph at level l. We begin at level 0 with $\mathcal{G}_0 = \{G\}$, where $G = (V, E)$ is the graph of the given sparse matrix to be ordered. The vertices and edges of G are scanned to construct the working vertex and edge lists, and these lists are used in turn to generate the corresponding count lists. A separating coordinate value s and Cartesian separator V_s are then computed for G as described

previously, which yields two subgraphs G_1 and G_2. The vertices in the separator V_s are numbered $|V| - |V_s| + 1$ through $|V|$, completing level 0 of the dissection process. At level 1, we apply the Cartesian separator algorithm to each of the two subgraphs in $\mathcal{G}_1 = \{G_1, G_2\}$. Working lists are constructed for each subgraph, and separating coordinate values s_1 and s_2 and corresponding Cartesian separators V_{s_1} and V_{s_2} are computed. The vertices in the two separators are numbered and the four remaining subgraphs are then similarly processed at level 2, and so on. This process continues until all vertices in the original graph have been numbered. At most $\log_2(|V|)$ levels of nested dissection are required to number all of the vertices, since the lth level results in 2^l subgraphs.

4.1. Serial complexity. We now estimate the serial time complexity of the foregoing Cartesian nested dissection algorithm. Consider a Cartesian labeled graph $G = (V, E)$ with N vertices and M edges. We assume that any subgraph of G has at least as many edges as vertices. To compute the cost of ordering G we compute bounds for the cost of initialization and the cost of each level of dissection.

In the initialization step, vertices are sorted in increasing order of both x and y coordinate values. The complexity of this step is $O(N \log_2 N)$ using heap sort. These sorted lists are used to construct the working lists of vertices, in time proportional to N. The sorted lists of vertex coordinates are also used to construct the edge lists in time proportional to M. A group tree is constructed for each dimension by mapping edges to intervals. Each group tree can have at most N groups. Entering an interval into a group tree takes time proportional to $\log_2 N$. The cost of forming group trees is therefore proportional to $M \log_2 N$. The overall cost of the initialization step is therefore $O(M \log_2 N)$.

The cost of separating a subgraph $G_i = (V_i, E_i)$ is given by the sum of the costs of computing a separating value and then constructing and numbering the corresponding separator. Computing a separating value that satisfies the balance condition requires the formation and traversal of the vertex count lists. The cost of forming these lists is proportional to $|V_i|$. The cost of traversing them depends on the number of actual coordinate values in the graph, which is obviously at most $|V_i|$. Computing the estimate η for the separator size requires the formation and traversal of the edge count lists, resulting in cost proportional to $|E_i|$. Computing the set of edges that straddle the separating value involves searching one of the group trees and deleting edges selected. This can be accomplished in time proportional to $\log_2 N$ and the number of edges found. Computing and numbering the actual separator can be performed in time proportional to the size of the separator, which is much smaller than $|V_i|$. The cost of separating G_i is therefore of the form $c_s|E_i|$, where c_s is a small constant. The cost of separating all subgraphs at level l of nested dissection is therefore given by

$$c_s \sum_{G_i \in \mathcal{G}_l} |E_i| \leq c_s M.$$

It remains to estimate the costs of forming working lists and group trees for each resulting subgraph. Each list for G_i can easily be decomposed into two lists, one for each resulting subgraph, in time proportional to the length of the list. This is possible since it can be decided which subgraph an entity belongs to by a simple comparison of the appropriate coordinate value with the separating value. Such a decomposition will yield lists that are still in increasing order of the respective coordinate value since the original sorted order is not affected by deletions. Accordingly, this cost is $O(|E_i|)$. A group tree for G_i can be decomposed into group trees for each of the

resulting subgraphs. Each interval in a group tree for G_i is examined. It can be easily determined if the interval lies in G_{i_1} or G_{i_2} by comparing it with the separating value. The interval can then be added to the appropriate group tree. Including the overhead of allocating and initializing groups, the cost is proportional to $|E_i|$. Over all subgraphs in $\mathcal{G}_l$, the total cost of updating lists and group trees is therefore $c_u M$, where c_u is a small constant.

It follows from the above paragraphs that the cost of one level of nested dissection is $c_o M$, where c_o is a suitable constant. Thus, a single initialization step followed by at most $\log_2 N$ levels of nested dissection results in a serial time complexity of $O(M \log_2 N)$.

5. Computing separators in parallel. We now adapt the Cartesian separator algorithm for use on a distributed memory parallel computer. Our goal is to distribute the computation evenly across the processors while keeping the volume and frequency of interprocessor communication low. For the resulting parallel algorithm to be scalable, both higher and lower order costs should be shared among all processors, and all data structures should be distributed across all memories. The distributed parallel algorithm will have the same general form as the serial algorithm, but the work of forming lists and counting and searching will be shared by all of the processors. In effect, each processor will own a portion of the data and will be responsible for any counting or searching involving that portion. Coordinating such joint activities among the processors and reporting the results will obviously require some interprocessor communication, but we try to limit this for good efficiency.

Let the number of processors be P. We assume that the set of vertices V of the original graph is distributed among the processors so that each processor has approximately $|V|/P$ vertices. For ease of implementation, we distribute the set of edges E among the processors so that each edge is assigned to a processor holding one of the two vertices at its endpoints. This may not result in an even distribution of edges for all graphs, but for most graphs arising in practice, such as finite element graphs, the number of edges on each processor will be at most a constant times $|E|/P$. In mapping the problem data to processor memories, we make no assumption that locality is preserved, nor do we assume any correlation between the topology of the graph and the topology of the processor interconnection network. Indeed, the parallel algorithm we propose tends to perform best with a random data distribution, since such a distribution tends to balance the computational load in forming and searching the various lists required.

The data distribution described above results in each processor's having vertices and edges at almost all coordinate values, but not having all of the vertices and edges associated with any one coordinate value. As a consequence, in determining a separating value, vertex and edge count lists must be accumulated over all processors and traversed in increasing order of coordinate values to identify a separating value satisfying a balance condition and/or minimizing η, and finally this computed separating value must be disseminated to all processors. Obviously, these steps require several phases of interprocessor communication, as well as a significant amount of computation. For effective parallelization, we will distribute lower order costs, such as computing separating values over subgraphs at a given level of nested dissection, as well higher order costs, such as constructing the vertex and edge count lists, across all of the processors, and will also try to minimize communication costs.

In dealing with distributed data structures, we will adopt the notation that the portion of a given entity that resides on processor π_k will be indicated by appending

(π_k) to the usual notation for the global object in question. Thus, for example, $V_i(\pi_k)$ denotes the portion of vertices in subgraph G_i that reside on processor π_k, and so on.

5.1. Computing separating values in parallel. We now describe the process of computing separating values in parallel. We formulate the computation in terms of various global list operations, each of which requires a communication pattern akin to parallel prefix. For many distributed memory parallel computers, the startup cost for communication is relatively high, and therefore it pays to minimize the number of messages required to send a given volume of data. For this reason, we will concatenate together all of the data to be exchanged among processors over all of the subgraphs at a given level of nested dissection, so that a single set of communications will suffice for computing a global list operation. Grouping communications in this manner represents a substantial saving in the number of messages over computing the separating value for each subgraph individually, which would incur a separate round of communication for each. This is one reason we chose a breadth-first rather than a depth-first approach.

As we have seen, the determination of appropriate separating values requires node counts for each of a series of coordinate values. In a parallel setting, the necessary count information is distributed over all of the processors. Thus, for each coordinate value, the counts must be accumulated across the processors, the resulting separating values computed, and this information must then be made available to all of the processors. These three steps are required for each subgraph in $\mathcal{G}_l = \{G_1, \ldots, G_r\}$ at a given level l of nested dissection, and each step requires global communication. To reduce the number of messages, and hence the total communication startup overhead, we will combine all of the relevant data for all of the subgraphs at a given level for each communication step. Of course, for good parallel efficiency, we must also share the computational work among all of the processors as well.

We first consider the process of accumulating count information across all processors. We will allocate this global *accumulation* among the processors by making each processor responsible for a block of coordinate values. Let L denote the set of coordinate values along a given dimension over all subgraphs in $\mathcal{G}_l$, and let L be partitioned into P contiguous blocks of values, $L(0), \ldots, L(P-1)$, such that each block covers about the same number of vertices (which is always possible for reasonably well-behaved graphs). Processor π_k will be responsible for accumulating the counts for each value in block $L(k)$ for all G_i. Each processor initially has counts over all the coordinate values L, but only for its own portion of each subgraph, whereas we want each processor π_k to contain the counts over each entire subgraph, but only for its own assigned block of coordinate values $L(k)$.

The best implementation of such a global information exchange operation depends on the interconnection network among the processors. One example is dimensional exchange in a hypercube network, in which processors exchange data pairwise in successive dimensions of the hypercube. In the current context, the exchange of information between each pair of processors involves splitting and merging their respective lists. The lists are structured so that they can be merged in time proportional to the sum of their sizes. After d steps, where d is the dimension of the hypercube, each processor has the desired information, namely, processor π_k contains the counts over all subgraphs for the kth block of coordinate values.

The pairwise accumulation process described above effectively spreads the work of accumulating counts for the coordinate values across all of the processors, but we must still traverse the resulting count lists and compute the cumulative vertex counts

to determine a separating value for each subgraph. The set of coordinate values spanned by an individual subgraph G_i may intersect more than one block of values $L(k)$, and hence the corresponding count lists may be spread over multiple processors. Thus, the necessary list traversals and cumulative vertex counts will require further interprocessor communication. For a given subgraph G_i, let $L_i = \{l_i, \ldots, r_i\}$ be the ordered set of coordinate values spanned by V_i, and let $L_i(k) = L_i \cap L(k) = \{l_i(\pi_k), \ldots, r_i(\pi_k)\}$. To determine if a separating value lies within $L_i(k)$, processor π_k requires a cumulative count of vertices in G_i over all previous coordinate values $l_i, \ldots, l_i(\pi_k) - 1$. Furthermore, processor π_k requires such cumulative counts over all subgraphs in $\mathcal{G}_l$.

Computation of the required cumulative counts is an example of a parallel prefix computation, which can be implemented in a number of ways, with the best choice dependent on the interconnection network among the processors. Our implementation, which we refer to as *cascading*, is again based on dimensional exchange and requires $\log_2 P$ communication steps and $P \log_2 P$ messages. An alternate implementation of parallel prefix can reduce the number of messages required, but it does not reduce the number of steps and requires nonneighbor communication in a hypercube.

Once cumulative counts have been cascaded, each processor can now determine, for each subgraph, the set of values within block $L(k)$ that satisfies the balance condition. These sets of values must then be aggregated over all processors to arrive at the full set of values satisfying the balance condition for each subgraph. This global *aggregation* of sets can again be computed by a dimensional exchange process having d steps, at step i of which each processor exchanges information with its neighbor in the ith dimension and the information received is combined with previous information by set union.

For each subgraph in $\mathcal{G}_l$, the above three-stage process determines a block of coordinate values satisfying the balance condition. A similar three-stage process is used for each subgraph to compute a value that minimizes η. Each processor can then determine the final separating value for each subgraph by making a local comparison of the computed separating values in each coordinate dimension.

5.2. Constructing separators in parallel. Having determined a separating value, we must now construct a separator for each subgraph. Portions of each separator can be computed locally, but communication would be required to compute the complete sets. However, we can avoid some of the overhead that would be required by taking a different approach in which the processors cooperate to number their portions of each separator without ever forming the set union explicitly. Since the numbering of vertices within a single separator is arbitrary, we adopt the convention that the vertices in $V_{s_i}(\pi_k)$ are numbered after those in $V_{s_i}(\pi_{k-1})$ for $0 < k < P$. To determine the range of numbers to use for each processor's portion, a variant of the previous cascade algorithm is used.

The fact that the union over all processors is not explicitly constructed may result in a separator that is somewhat larger than strictly necessary. Consider two edges having a common vertex and residing on different processors. In the serial case, the common vertex could be selected to cover both edges, but in the distributed case a different vertex may be selected from each edge, thereby increasing the size of the separator.

6. Parallel Cartesian nested dissection. The algorithm given in the previous section computes a set of separators for all of the subgraphs at a given level of nested dissection. Thus, the algorithm could be applied repeatedly, beginning with

the original graph G, to produce a complete nested dissection ordering in at most $\log_2(|V|)$ steps. In a distributed parallel setting, however, it may be advantageous not to follow this process all the way to the end, since each step requires a significant amount of communication. Instead, the dissection process can be stopped as soon as a level has been reached at which there are at least as many subgraphs as processors. The data can then be reorganized to place whole subgraphs on each processor, so that a serial ordering algorithm can be applied to the remaining subgraphs on each processor from that point on, with no further communication required. We now describe such a two-phase approach in greater detail.

The first phase consists of carrying out the first D levels of Cartesian nested dissection as described earlier, where D is the first level at which the number of subgraphs is at least tP, with $t \geq 1$ a parameter specified by the user. The choice $t = 1$ yields less overall communication, since it shifts more of the work to the second, communication-free phase. However, a choice of $t > 1$, by producing more subgraphs than the number of processors, may allow more flexibility in achieving a good load balance across processors during the second phase. Thus, there is a problem-dependent trade-off in choosing a value for t. Whatever the choice for t, after D steps the Cartesian nested dissection process is stopped, and we must then redistribute the problem data so that each subgraph is assigned in its entirety to only one processor. This redistribution step requires a significant amount of global communication, which must be taken into account in assessing the total cost of the algorithm.

The necessary redistribution of problem data can be accomplished by a variant of the pairwise accumulation algorithm described earlier. In our earlier use of pairwise accumulation, we used the blocks of coordinate values, $L(0), \ldots, L(P-1)$, as a means of organizing the accumulation so that at each step of dimensional exchange the computation would be shared among processors and the resulting data would be assigned to processors in a systematic way. For purposes of redistributing problem data between the global and local phases of the hybrid ordering algorithm, numerical accumulation is not required, but we can still use the same organization as pairwise accumulation to direct the flow of data to the appropriate destinations. Specifically, the labels of the subgraphs to be redistributed, $G(0), \ldots, G(P-1)$, play the same role that the coordinate blocks played previously.

6.1. Parallel complexity. We now provide estimates of the communication and computational complexity of the parallel Cartesian nested dissection algorithm for a graph $G = (V, E)$ with N vertices and M edges using P processors. We assume that each processor holds at most cN/P vertices and cM/P edges, where c is a small constant. In the remainder of this section, the letter c is used to denote a suitable constant.

We estimate the communication complexity in terms of N_{msgs}, the number of messages communicated by each processor. Communication is limited to the distributed phase comprising D levels of nested dissection, where $D \approx \log_2(tP)$ and t is a small constant. At each level of distributed nested dissection, a few accumulation, cascading, and global aggregation operations are performed. Each of these operations involves $\log_2 P$ messages per processor. Over D levels, this amounts to $O((\log_2 P)^2)$ messages per processor. Since redistribution is simply a variant of pairwise accumulation, it also requires $\log_2 P$ messages. Accordingly,

$$N_{\text{msgs}} \leq c\,(\log_2 P)^2.$$

To estimate the computational complexity, we observe that the cost of a single

level of nested dissection is proportional to the maximum number of edges on a processor, excluding the overhead associated with pairwise accumulation, cascading, and global aggregation operations. The one-time cost of redistribution must also be taken into account. But for these exceptions, the cost of nested dissection would amount to $c(M/P)\log_2 N$. The overhead associated with cascading and global aggregation operations is proportional to the amount of information communicated. For these operations, the lists communicated contain a few values for each graph at that level of nested dissection. The communication volume is of the form $c\log_2 P|\mathcal{G}_l|$ for each level l. Since $|\mathcal{G}_l|$ doubles for each successive level of nested dissection, the communication volume is given by

$$c(\log_2 P)\{1+2+4+8+\cdots+tP\} \leq 2ctP(\log_2 P).$$

It can be seen from this result that the associated cost is $O(P\log_2 P)$. Using accumulation (without explicit merging at each stage) results in $O(1)$ overhead for each pairwise communication step. At the end of $\log_2 P$ such steps, each processor π_k must merge count information over values in $L(\pi_k)$. Recall that the sets $L(\pi_k)$ were chosen so that each contains approximately N/P vertices. Therefore, the cost of merging is proportional to N/P. Likewise, there is only a constant overhead associated with the redistribution operation, since a processor simply forwards a portion of a received message. Following redistribution, new data structures must be set up on each processor for use in further processing, but this work is perfectly parallel and spread more or less evenly across the processors. Thus, the parallel arithmetic complexity is $O((M/P)\log_2 N)$.

7. Test results. In this section we present some empirical test results for the parallel Cartesian nested dissection algorithm. In Table 1 we show the number of vertices and edges for two types of test problems. The first type, labeled Gxxx, are regular square grids of the given size; for example, G100 is a 100×100 square grid. The second type, labeled Lx, are L-shaped finite element problems generated by ANSYS, a standard commercial software package for finite element analysis. A small example is shown in Fig. 4. These L-shaped graphs are highly irregular; for example, in L3 the ratio of the longest-to-shortest edge is 3420, which is comparable to the number of nodes.

TABLE 1
Description of test problems.

Problem	N	M
G100	10,000	19,000
G200	40,000	79,000
G300	90,000	179,400
G400	160,000	319,200
L3	12,864	37,983
L6	25,728	76,086
L12	42,880	127,170

We give test results for the Cartesian nested dissection (CND) algorithm using two different options. By CND-bal we mean the CND algorithm using only the "exact" balance criterion $\alpha = \frac{1}{2}$, and by CND-opt we mean the CND algorithm using the approximately optimal separator size within the balance range permitted by a value of $\alpha = \frac{1}{3}$. The latter choice for α is heuristic; it is simply intended to give the algorithm some freedom to reduce the separator size, yet not allow the splitting of the graph

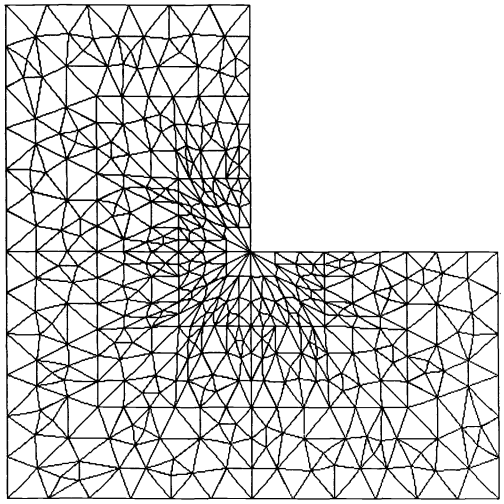


FIG. 4. *Example of small L-shaped graph.*

to become too skewed. We note that this value has been used in theoretical work on graph separators [10]. CND-bal does not require estimation or optimization of the separator size, and hence is less costly to compute than CND-opt. CND-bal should produce well-balanced subgraphs but may suffer a great deal of fill. CND-opt, on the other hand, incurs much less fill but may not maintain good balance. We have also implemented a hybrid algorithm that uses CND-opt for the highest levels of nested dissection to keep those critical separators small, then switches over to the cheaper CND-bal for the remaining levels of dissection. We do not provide results for this hybrid approach, however, as they simply fall between those for pure CND-opt and CND-bal, mimicking one or the other more closely depending on the crossover point chosen for switching criteria. For comparison with CND-bal and CND-opt, we also give results for two well-known serial ordering algorithms, automatic nested dissection (AND) [5], and multiple minimum degree (MMD) [11].

Tables 2 and 3 compare the orderings with respect to sparsity preservation by considering the resulting number of nonzeros in the Cholesky factor L and the total number of floating point operations required to compute L. There is no need for a sparsity comparison for the regular grids, since CND-bal produces theoretically ideal orderings for such problems. For the L-shaped problems, we see that CND-bal compares well with AND, and that CND-opt compares reasonably well with MMD, which is usually considered the best heuristic known for irregular problems.

TABLE 2
Thousands of nonzeros in Cholesky factor L.

Problem	CND-bal	CND-opt	AND	MMD
L3	462	401	458	381
L6	957	858	949	779
L12	2444	1819	2112	1476

Tables 4 and 5 compare the orderings with respect to two theoretical measures of parallelism, namely, the height of the elimination tree (see, e.g., [12]) and the work, measured in millions of floating point operations, along the critical path in

TABLE 3
Millions of floating-point operations to compute L.

Problem	CND-bal	CND-opt	AND	MMD
L3	22	14	24	13
L6	49	35	55	27
L12	278	120	219	66

the elimination tree (essentially tree height weighted by the number of floating point operations at each node). These measures have commonly been used to give a rough idea of the potential running time of parallel sparse Cholesky factorization using a given ordering. Both measures are rather pessimistic, however, in that they do not take into account the available data parallelism, nor the differing abilities of dense kernels to exploit it. Nevertheless, we see that CND-opt produces shorter elimination trees than AND or MMD, and the critical cost for CND-opt is also very competitive with the other orderings. We expect the elimination trees produced by CND-bal to be very well balanced, but the larger separators incurred can cause the total height of the tree and the critical cost to be significantly higher than those for the other three orderings.

TABLE 4
Elimination tree height.

Problem	CND-bal	CND-opt	AND	MMD
L3	632	441	581	580
L6	672	668	675	915
L12	1626	995	1444	1397

TABLE 5
Work along critical path.

Problem	CND-bal	CND-opt	AND	MMD
L3	11	2.7	11	3.0
L6	13	6.8	21	4.6
L12	134	31.0	77	13.0

Tables 6 and 7 show the ordering times for the CND algorithm using various numbers of processors P on an iPSC/860 hypercube multicomputer. The blank entries in the tables indicate cases that were not run because the problem would not fit in memory for that number of processors. We cannot give comparative results for AND and MMD, since they are not parallel algorithms. In Table 6 we show results only for CND-bal, since it already produces ideal orderings for square grids, and hence there is no need to use the optimal criterion. As expected for any fixed problem size, we see a diminishing gain as more processors are used. Yet, in light of our previous experience with sparse matrix algorithms on such parallel machines, we find it encouraging that we continue to see any speedup at all as we reach as many as 128 processors. In particular, these results suggest that communication costs are not growing unreasonably as the number of processors increases.

It should be noted that all of these test problems are relatively small, as even the largest problems still fit on only four processors. The size of our test problems was limited by the logistic difficulties of generating large problems, transferring them across national networks, and getting them into and out of the parallel machines through the relatively primitive and cumbersome parallel input/output (I/O) facilities currently

TABLE 6
Time in seconds for ordering regular grids.

P	G100	G200	G300	G400
1	2.4	12.3	36.7	
2	2.1	8.3	24.9	
4	1.1	5.1	12.0	22.8
8	0.6	2.6	6.9	11.2
16	0.4	1.6	3.6	5.9
32	0.3	1.0	2.0	3.5
64	0.3	0.7	1.3	2.1
128	0.3	0.5	0.9	1.4

TABLE 7
Time in seconds for ordering L-shaped graphs.

	CND-bal			CND-opt		
P	L3	L6	L12	L3	L6	L12
1	9.1			20.0		
2	5.9	14.6		10.1	19.8	
4	4.0	8.9	15.2	6.9	13.2	25.7
8	2.1	4.4	8.5	4.4	8.7	19.1
16	1.3	2.5	4.7	3.0	5.5	11.1
32	0.9	1.6	3.0	2.3	3.7	8.9
64	0.7	1.1	2.0	1.8	2.8	6.2
128	0.6	0.9	1.6	1.5	2.4	5.0

available. Eventually the algorithm we have developed will be integrated into an overall distributed parallel software environment, such as a structural analysis package, so that the problem can be generated and solved in place on the parallel machine, with problem size limited only by the total memory available on the entire ensemble of processors. Our preliminary results with much smaller problems encourage us to expect the CND algorithm to be very effective in such an environment.

8. Future work. We are encouraged by our results to date, but a considerable amount of work remains to be done along these lines. More extensive experimentation is needed, both in solving much larger and more diverse problems and in comparing the results with other competing algorithms. The ordering algorithm could be extended in several ways. For example, it may compute a separator that is unnecessarily large, and it would be desirable to reduce the separator to one of minimal size. We would also like to experiment with random sampling techniques to reduce the computational cost of the algorithm. Another area for further research is the use of rotations, conformal mappings, or other transformations of the input graph that might enhance the effectiveness of the Cartesian nested dissection algorithm.

The ordering algorithm has recently been generalized to handle problems in three dimensions and also nonsymmetric matrices [17]. The subsequent numeric factorization and triangular solution are developed in [8]. We are currently engaged in integrating the entire suite of algorithms into a usable software library format, porting it to additional parallel machines, and exploring its use in conjunction with software packages for specific applications areas, such as finite element structural analysis.

Acknowledgements. We wish to thank John Gilbert, Esmond Ng, and three anonymous referees for helpful comments that improved the presentation of this paper.

REFERENCES

[1] M. J. BERGER AND S. H. BOKHARI, *A partitioning strategy for nonuniform problems on multiprocessors*, IEEE Trans. Computers, C-36 (1987), pp. 570–580.

[2] G. C. FOX, M. A. JOHNSON, G. A. LYZENGA, S. W. OTTO, J. K. SALMON, AND D. W. WALKER, *Solving Problems on Concurrent Processors*, Vol. 1, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[3] J. A. GEORGE, *Nested dissection of a regular finite element mesh*, SIAM J. Numer. Anal., 10 (1973), pp. 345–363.

[4] J. A. GEORGE AND J. W.-H. LIU, *An automatic nested dissection algorithm for irregular finite element problems*, SIAM J. Numer. Anal., 15 (1978), pp. 1053–1069.

[5] ———, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

[6] J. R. GILBERT AND E. ZMIJEWSKI, *A parallel graph partitioning algorithm for a message-passing multiprocessor*, Internat. J. Parallel Programming, 16 (1987), pp. 427–449.

[7] M. T. HEATH, E. NG, AND B. W. PEYTON, *Parallel algorithms for sparse linear systems*, SIAM Rev., 33 (1991), pp. 420–460.

[8] M. T. HEATH AND P. RAGHAVAN, *Distributed solution of sparse linear systems*, Tech. Report UIUCDCS-R-93-1793, Department of Computer Science, University of Illinois, Urbana, February 1993.

[9] B. HENDRICKSON AND R. LELAND, *An improved spectral graph partitioning algorithm for mapping parallel computations*, Tech. Report SAND92-1460, Sandia National Laboratories, Albuquerque, NM, September 1992.

[10] R. LIPTON AND R. TARJAN, *A separator theorem for planar graphs*, SIAM J. Appl. Math., 36 (1979), pp. 177–199.

[11] J. W.-H. LIU, *Modification of the minimum degree algorithm by multiple elimination*, ACM Trans. Math. Software, 11 (1985), pp. 141–153.

[12] ———, *The role of elimination trees in sparse factorization*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 134–172.

[13] G. MILLER, S. TENG, W. THURSTON, AND S. VAVASIS, *Automatic mesh partitioning*, in Workshop on Sparse Matrix Computations: Graph Theory Issues and Algorithms, Institute for Mathematics and Its Applications, Springer-Verlag, New York, Berlin, 1992.

[14] G. L. MILLER, S. TENG, AND S. A. VAVASIS, *A unified geometric approach to graph separators*, in Proceedings of the 32nd Annual Symposium on Foundations of Computer Science, IEEE, 1991, pp. 538–547.

[15] G. L. MILLER AND W. THURSTON, *Separators in two and three dimensions*, in Proc. 22nd Ann. ACM Symp. Theory of Comput., New York, 1990, ACM, pp. 300–309.

[16] A. POTHEN, H. D. SIMON, AND K.-P. LIOU, *Partitioning sparse matrices with eigenvectors of graphs*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 430–452.

[17] P. RAGHAVAN, *Line and plane separators*, Tech. Report UIUCDCS-R-93-1794, Department of Computer Science, University of Illinois, Urbana, February 1993.

[18] J. D. ULLMAN, *Computational Aspects of* VLSI, Computer Science Press, Rockville, MD, 1984.

[19] C. T. VAUGHAN, *Structural analysis on massively parallel computers*, Comput. Systems Engrg., 2 (1991), pp. 261–267.

[20] S. A. VAVASIS, *Automatic domain partitioning in three dimensions*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 950–970.

[21] R. D. WILLIAMS, *Performance of dynamic load balancing algorithms for unstructured mesh calculations*, Concurrency: Practice and Experience, 3 (1991), pp. 457–481.

[22] M. YANNAKAKIS, *Computing the minimum fill-in is* NP*-complete*, SIAM J. Alg. Disc. Meth., 2 (1981), pp. 77–79.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 254–277, January 1995

018

ON THE ITERATIVE SOLUTION OF HERMITE COLLOCATION EQUATIONS*

YU-LING LAI†, APOSTOLOS HADJIDIMOS‡, ELIAS N. HOUSTIS‡, AND JOHN R. RICE‡

Abstract. Collocation methods based on bicubic Hermite piecewise polynomials have been proven to be effective techniques for solving general second order linear elliptic partial differential equations (PDEs) with mixed boundary conditions [*ACM Trans. Math. Software*, 11 (1985), pp. 379–412]. The corresponding system of discrete collocation equations is generally nonsymmetric and nondiagonally dominant. Methods for their iterative solution are not known and are currently solved using Gauss elimination with scaling and partial pivoting. Point iterative methods like those in ITPACK [Tech. Report CNA-216, Center for Numerical Analysis, Univ. of Texas at Austin, April 1988] do not converge even for the collocation equations obtained from the discretization of model PDE problems. The development of efficient iterative solvers for these collocation equations is necessary for the case of three-dimensional PDE problems and their parallel solution, since direct solvers tend to be space bound and their parallelization is difficult. In this paper block iterative methods are developed and analyzed for the collocation equations corresponding to elliptic PDEs defined on a rectangle and subject to uncoupled mixed boundary conditions. For these types of PDE problems certain boundary degrees of freedom of the collocation approximation can be predetermined symbolically [Houstis, Mitchell, and Rice]. The remaining equations are called "interior" collocation equations. The system of all discrete equations is referred to as "general" collocation equations. Papatheodorou [*Math. Comp.*, 41 (1983), pp. 511–525] was first to determine the exact parameters of accelerated overrelaxation (AOR)-type iterative methods for the case of "interior" collocation equations associated with a model problem. This paper generalizes the results of Papatheodorou for the "interior" collocation equations and presents new results for a particular class of "general" collocation equations. Specifically, in the case of a model elliptic PDE problem with uncoupled mixed boundary conditions, analytic expressions are derived for the eigenvalues of the block Jacobi iteration matrix based on a new partitioning of the interior collocation matrix, and the optimal overrelaxation factors are determined for the block successive overrelaxation (SOR) iterative method. A number of numerical results are presented to verify the theoretical analysis of the block SOR method and to compare its convergence behavior with those of the block Jacobi, Gauss–Seidel and the optimal AOR of Papatheodorou. Furthermore, the authors compare the time and memory complexity of the block SOR, LINPACK Band GE, and generalized minimal residual (GMRES) mathematical software for solving the Hermite collocation equations obtained from the discretization of several PDE problems. The numerical results indicate that the block SOR is an efficient method for solving these equations.

Key words. elliptic partial differential equations, collocation methods, SOR iterative method

AMS subject classifications. 65N35, 65N05, 65F10

1. Introduction.

We consider the discrete equations obtained from applying the collocation method based on bicubic Hermite piecewise polynomials to discretize a general second order linear elliptic partial differential equation of the form

$$Lu \equiv au_{xx} + cu_{yy} + du_x + eu_y + fu = g, \qquad (x, y) \in R, \tag{1}$$

subject to the boundary conditions

$$Bu \equiv \alpha u + \beta \frac{\partial u}{\partial n} = \delta, \qquad (x, y) \in \partial R, \tag{2}$$

* Received by the editors December 31, 1992; accepted for publication (in revised form) by A. Greenbaum, September 27, 1993. This work was supported by Air Force Office of Scientific Research grant 91-F49620 and National Science Foundation grant CCR 86-19817.

† Department of Mathematics, Purdue University, West Lafayette, Indiana 47907.

‡ Department of Computer Sciences, Purdue University, West Lafayette, Indiana 47907 (hadjidim@cs.purdue.edu, enh@cs.purdue.edu, jrr@cs.purdue.edu).

where R is a general domain and the coefficients and the right-hand sides in (1)–(2) may depend on x and y. It is an *open problem* to find iterative methods to solve these equations.

The main objective of this paper is to theoretically and experimentally analyze iterative methods for the solution of Hermite collocation equations associated with the PDE (1) on a rectangular domain with Dirichlet or Neumann conditions on parts of the boundary. A "natural" ordering of the collocation equations and unknowns [11] leads to a banded coefficient matrix that is generally nonsymmetric and nondiagonally dominant and whose diagonal elements are almost all zero. Thus a straightforward application of the classic point iterative methods to solve these equations is not possible. These systems are currently solved by Gauss elimination with scaling and partial pivoting [3]. Some "customized" direct and iterative solvers have been developed for solving the Hermite collocation equations for special elliptic PDE operators and boundary conditions on the unit square [2], [1].

The iterative solution of the Hermite collocation equations was first addressed in [12] and [17] for the case of *interior* Hermite collocation applied on the Poisson problem with Dirichlet boundary conditions defined on the unit square. The iterative methods were based on a special reordering of the equations and the unknowns, which resulted to a block tridiagonal coefficient matrix. In this paper we extend the iterative approaches proposed in [12] and [17] for a class of "general" Hermite collocation equations. These extensions are based on a new partitioning of the corresponding "interior" collocation matrix that allows us to derive analytically the eigenvalues of the corresponding block Jacobi iteration matrix and determine the optimal overrelaxation factor of the SOR iterative method [21], [22]. In addition, we improve several of Papatheodorou's theoretical results for the "interior" collocation equations. We present experimental results that show that the SOR method converges well, as expected, for the model problem and also for a more general PDE. Comparisons are made with other iterative methods (preconditioned conjugate gradient) and direct solvers; the SOR method is seen to be the most efficient.

The organization of this paper is as follows. In §2, we introduce a notation for defining the various block partitionings of collocation coefficient matrices used in the spectral analysis of the Jacobi iteration matrix. Moreover, we give a brief description of the Hermite collocation method. In §3, we describe the new ordering of the interior collocation equations and the unknowns, present a block partitioning of the coefficient matrix, and carry out the spectral analysis of the corresponding Jacobi iteration matrix. These results are applicable for Dirichlet model problems on the unit square. In §4, we establish similar results to those in §3 for general Hermite collocation equations. Moreover, the spectral analysis of the Jacobi iteration matrix associated with a new partitioning of interior collocation matrix is used to analyze the convergence of the block SOR for model problems with some types of uncoupled mixed boundary conditions on a rectangle. In §5, we use the results of §4 to study the convergence analysis of the block SOR method. Moreover, we make some comparisons concerning the two block Jacobi iteration matrices and develop the corresponding optimal block SOR iterative method. Finally, in §6 we study the numerical behavior of several block iteration methods including optimal and adaptive SOR, Jacobi, and Gauss–Seidel. We verify some of the theoretical results obtained in this paper. In addition, we compare the performance of the optimal block SOR solution, three preconditioning conjugate gradient methods based on GMRES software and the LINPACK Band GE solver. Data are given for the iterations, time, and

memory required to solve for a model PDE problem with several types of boundary conditions and a more general PDE problem. The numerical results indicate that the block SOR method developed is an efficient alternative for solving the Hermite collocation equations obtained from the discretization of general elliptic PDEs on rectangular regions subject to uncoupled mixed boundary conditions.

2. Preliminaries. In this section we introduce some special notation for partitioning matrices and give a brief formulation of the Hermite collocation discretization approach for the PDE problems considered in this paper. First, we introduce the block form

$$[A|B] = \left[\begin{array}{cc|cc} a_{11} & a_{12} & b_{11} & b_{12} \\ a_{21} & a_{22} & b_{21} & b_{22} \end{array}\right],$$

which we subsequently use to construct the $(2n) \times (2n)$ matrix

$$[A|B]_{\otimes(2n)} = \begin{bmatrix} a & B & & & \\ & A & B & & \\ & & \ddots & & \\ & & & A & B \\ & & & & A \; b \end{bmatrix}, \quad a = \begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix}, \quad b = \begin{bmatrix} b_{12} \\ b_{22} \end{bmatrix}.$$

Note that if all a_{ij} and b_{ij} are $(2m) \times (2m)$ matrices, then A and B are matrices of 2×2 block form and of order $4m$. So the new matrix is of order $4mn$.

Next we present a brief description of the Hermite collocation method for PDE problems defined on the unit square. For simplicity, we consider the case of a uniform grid with spacing $h = 1/n$, where n is the number of subintervals in the $x-$ and $y-$ directions and with a set of nodal points bordering the unit square at a distance h. For this mesh the coordinates of the nodal points are (x_i, y_j), where $x_i = (i-1)h$, $y_j = (j-1)h$, $i, j = 0, 1, \ldots, (n+2)$. It is known that the basis functions of Hermite cubic piecewise approximate space are generated by the following two cubic polynomials on (0,1): $\phi(t) := (1-t)^2(1+2t)$, $\psi(t) := t(1-t)^2$. Specifically, the one-dimensional Hermite cubic piecewise basis functions are

$$B_{2i-1} = \begin{cases} \phi(-\sigma_i) & \text{if } t_{i-1} \le t \le t_i, \\ \phi(\sigma_i) & \text{if } t_i \le t \le t_{i+1}, \\ 0 & \text{otherwise,} \end{cases} \qquad B_{2i} = \begin{cases} -h\psi(-\sigma_i) & \text{if } t_{i-1} \le t \le t_i, \\ h\psi(\sigma_i) & \text{if } t_i \le t \le t_{i+1}, \\ 0 & \text{otherwise,} \end{cases}$$

where $i = 1, 2, \ldots, (n+1)$, $\sigma_i(t) = (t - t_i)/h$, and $t = x$ or y. In this paper we consider the "fast" determination of an approximate solution to (1)–(2) defined as

$$u_n(x, y) = \sum_{i,j=1}^{2(n+1)} \alpha_{ij} B_i(x) B_j(y). \tag{3}$$

The unknown coefficients (degrees of freedom (dof)) α_{ij} are determined so that $Lu_n = f$ is satisfied exactly at interior collocation points (the four Gauss points of each mesh subrectangle) and u_n satisfies the boundary conditions at the two Gauss points of each boundary subinterval plus at the four corner points. This system of algebraic equations with respect to α_{ij} is called general Hermite collocation equations.

It is worth noticing that the structure of the collocation coefficient matrix depends on the numbering of the collocation points and the numbering of the four unknowns

associated with each node (x_i, y_j). Based on the definition of the basis functions one can easily show that the values of the unknowns coincide with the values of the approximate solution and its partial derivatives at each nodal point. Specifically, using the standard derivative operator notation, we have $\alpha_{2i-1,2j-1} = u_n(x_i, y_j)$, $\alpha_{2i,2j-1} = D_x u_n(x_i, y_j)$, $\alpha_{2i-1,2j} = D_y u_n(x_i, y_j)$, $\alpha_{2i,2j} = D^2_{xy} u_n(x_i, y_j)$.

For Dirichlet or Neumann conditions on part of a rectangular boundary, some of the unknowns can be determined symbolically from the boundary collocation equations. In this case, we are left with a linear system of size $(4n^2) \times (4n^2)$; its coefficient matrix is referred to as the "interior" collocation matrix and the whole procedure is called the "interior" collocation method [11]. In the case of mixed boundary conditions such a priori elimination of unknowns is not possible symbolically.

3. The case of interior collocation equations. The analysis of the iterative solution of the "interior" Hermite collocation equations associated with a Dirichlet model problem on the unit square was first considered in [12] and [17]. It is based on the spectral analysis of the corresponding Jacobi matrix under a new block partitioning of the "interior" collocation coefficient matrix. The results obtained in [17] assume that $n = 2^l$, where n is the number of subintervals in each coordinate direction. In this section we generalize the spectral analysis of the Jacobi matrix for *any* $n \geq 2$. Furthermore, we extend the ordering and partitioning introduced in [17] for uncoupled boundary conditions on part of the boundary. The main result is presented in Theorem 3.2.

3.1. The reordering and partitioning of interior collocation equations. The ordering of collocation equations is crucial for the convergence of iterative methods. Here we extend the reordering proposed in [17] for the interior Hermite collocation equations corresponding to the PDE problem (1)–(2) with the "uncoupled boundary conditions"

$$u = \delta \text{ on } \partial R_1 \subset \partial R, \qquad \frac{\partial u}{\partial n} = \delta \quad \text{on } \partial R_2 = \partial R - \partial R_1.$$

In Fig. 1(c) we depict the structure of the collocation equations for a specific mesh and the so-called "natural" ordering [11] described in Fig. 1(a). This system is generally banded and is neither symmetric nor diagonally dominant. Papatheodorou [17] introduced an ordering of the interior collocation equations and unknowns so that the coefficient matrix of the resulted system has nonzero diagonals. This ordering is depicted in Fig. 1(b) for the mesh $n_x = n_y = 3$ assuming Dirichlet conditions on $x = 0$ and $y = 1$ and Neumann conditions on $x = 1$ and $y = 0$. The structure of the corresponding linear system is shown in Fig. 1(d).

From the above example we can easily conclude that for the general case ($n_x \neq n_y$) and any uncoupled boundary conditions, the interior collocation coefficient matrix A associated with the new ordering has the following block structure:

(4)
$$A = \begin{bmatrix} x & x & x & 0 & & & & \\ x & x & x & 0 & & & & \\ 0 & x & x & x & x & 0 & & \\ 0 & x & x & x & x & 0 & & \\ & & \ddots & & \ddots & & \ddots & \\ & & 0 & x & x & x & x & 0 \\ & & 0 & x & x & x & x & 0 \\ & & & & 0 & x & x & x \\ & & & & 0 & x & x & x \end{bmatrix},$$

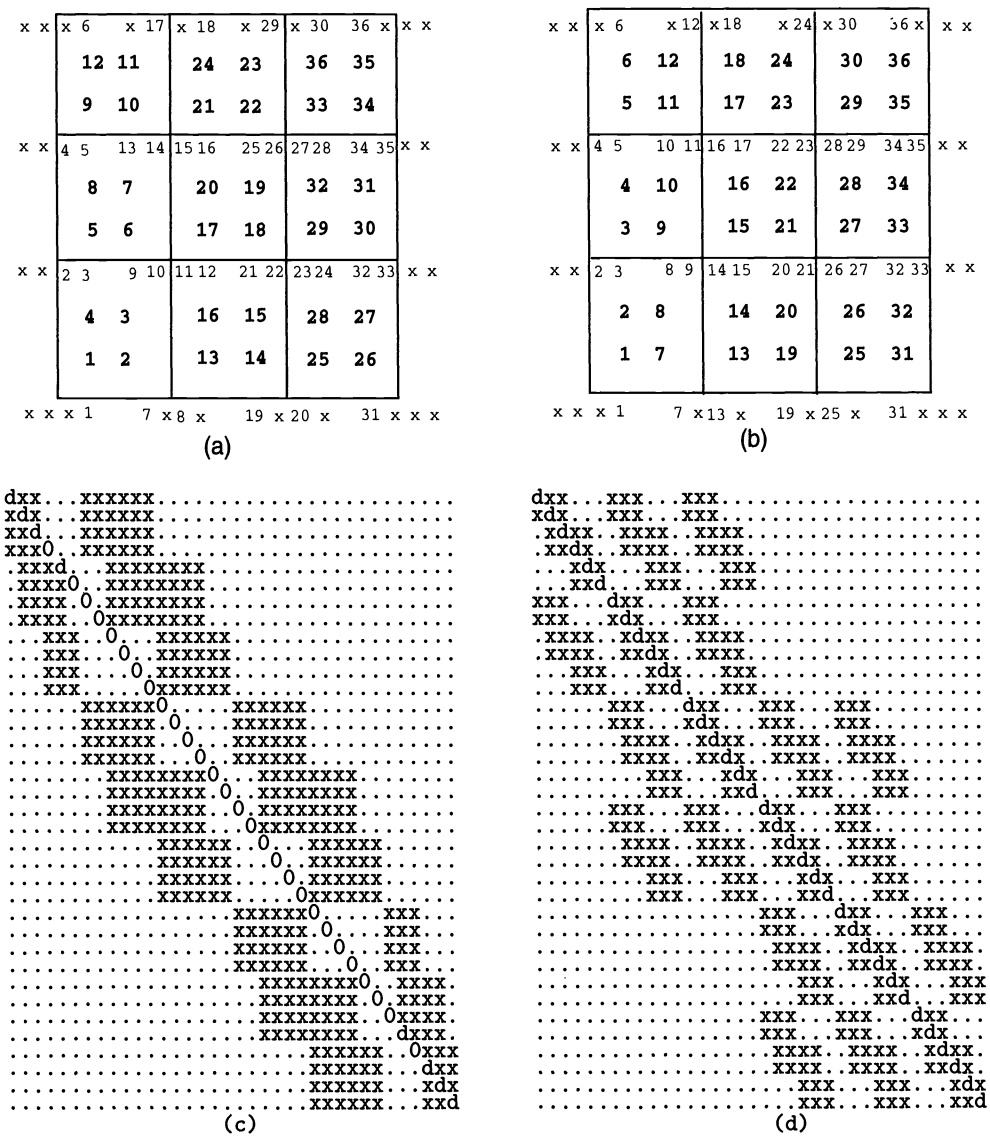


FIG. 1. *An example of the structures of interior collocation equations for the mesh* $n_x = n_y = 3$ *using "natural" ordering* (a) *and Papatheodorou's ordering* (b) *of equations and unknowns. The* 36 *collocation points are shown in bold; the numbering indicates the ordering of the equations. The unknowns are associated with nodal points and numbered in light type. Those unknowns eliminated symbolically are denoted by* x. *In* (c) *and* (d) *we display the structures of the collocation matrices associated with the orderings* (a) *and* (b), *respectively, where* d *denotes a nonzero diagonal element.*

where x denotes a $(2n_y) \times (2n_y)$ matrix. Apparently, under this partitioning, A is a block 2-cyclic consistently ordered matrix (cf. [21, p. 102]) of order $4n_xn_y$. This property motivates us to explore the use of block iterative methods to solve the Hermite collocation system. Specifically, we are interested in studying some special cases of these systems in which the block Jacobi iteration matrix associated with the partitioning in (4) has certain properties as described in [17]. In the next section we briefly present the generalization of the results in [17] for $n_x, n_y \geq 2$ following a much

simpler analytical approach.

3.2. Spectral analysis of the Jacobi matrix. We consider the interior Hermite collocation coefficient matrix for the case of a Poisson equation on a rectangle with Dirichlet boundary conditions and a uniform grid. To simplify the notation in the theoretical results that follow, we use n and m instead of n_x and n_y, respectively. In this case the collocation coefficient matrix is of the form

$$A = \left[\begin{array}{cc|cc} A_1 & A_2 & A_3 & -A_4 \\ A_3 & A_4 & A_1 & -A_2 \end{array}\right]_{\otimes(2n)} \tag{5}$$

with each A_i being of order $2m$. Note that the partitioning in (4) allows us to write A as

$$A = \begin{bmatrix} D_1 & -U_1 & & & \\ -L_1 & D_1 & -U_1 & & \\ & \ddots & \ddots & \ddots & \\ & & -L_1 & D_1 & -U_1 \\ & & & -L_1 & \bar{D}_1 \end{bmatrix}, \tag{6}$$

where

$$D_1 = \begin{bmatrix} A_2 & A_3 \\ A_4 & A_1 \end{bmatrix}, \quad \bar{D}_1 = \begin{bmatrix} A_2 & -A_4 \\ A_4 & -A_2 \end{bmatrix}, \quad -L_1 = \begin{bmatrix} 0 & A_1 \\ 0 & A_3 \end{bmatrix}, \quad -U_1 = \begin{bmatrix} -A_4 & 0 \\ -A_2 & 0 \end{bmatrix}. \tag{7}$$

In the subsequent analysis we assume that D_1, $\bar{D}_1$ are nonsingular. Furthermore, we introduce the matrices

$$R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} = \begin{bmatrix} A_2 & A_3 \\ A_4 & A_1 \end{bmatrix}^{-1} \begin{bmatrix} -A_4 & A_1 \\ -A_2 & A_3 \end{bmatrix}, \tag{8}$$

$$\begin{bmatrix} R_{31} \\ R_{32} \end{bmatrix} = \begin{bmatrix} A_2 & -A_4 \\ A_4 & -A_2 \end{bmatrix}^{-1} \begin{bmatrix} A_1 \\ A_3 \end{bmatrix}, \tag{9}$$

and note that

$$\begin{bmatrix} -A_4 & A_1 \\ -A_2 & A_3 \end{bmatrix} = \begin{bmatrix} 0 & -I \\ -I & 0 \end{bmatrix} \begin{bmatrix} A_2 & A_3 \\ A_4 & A_1 \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}. \tag{10}$$

From the relations (8) and (10), we obtain

$$R^{-1} = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} R \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}.$$

Consequently, we have

$$R^{-1} = \begin{bmatrix} R_{11} & -R_{12} \\ -R_{21} & R_{22} \end{bmatrix}. \tag{11}$$

If R_{21} is invertible then from (11), R_{11} $(= R_{21}R_{22}R_{21}^{-1})$ is similar to R_{22} and the following lemma holds.

LEMMA 3.1. *If R_{21} is nonsingular, then $R_{31} = -R_{21}^{-1}$.*

Proof. First we observe that (8) implies

$$\begin{bmatrix} A_2 & A_3 \\ A_4 & A_1 \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} = \begin{bmatrix} -A_4 & A_1 \\ -A_2 & A_3 \end{bmatrix}. \tag{12}$$

From (9) we have that $A_2R_{31} - A_4R_{32} = A_1$ and $A_4R_{31} - A_2R_{32} = A_3$. If we use the previous expressions for A_1 and A_3 in (12), we obtain

$$\begin{bmatrix} A_2 & -A_4 \\ A_4 & -A_2 \end{bmatrix} \begin{bmatrix} I & -R_{32} \\ 0 & -R_{31} \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} = \begin{bmatrix} A_2 & -A_4 \\ A_4 & -A_2 \end{bmatrix} \begin{bmatrix} 0 & R_{31} \\ I & R_{32} \end{bmatrix}.$$

Since $\bar{D}_1$ is invertible, the above equation can be simplified as follows:

$$\begin{bmatrix} I & -R_{32} \\ 0 & -R_{31} \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} = \begin{bmatrix} 0 & R_{31} \\ I & R_{32} \end{bmatrix}.$$

Comparing both sides, we readily obtain that $-R_{31}R_{21} = I$. Hence our assertion is established. □

The block partitioning (6) of A corresponds to the splitting $A = D - L - U$ where $D = \mathrm{diag}(D_1, \ldots, D_1, \bar{D}_1)$ and L and U are strictly lower and upper triangular matrices. Let $J = D^{-1}(L + U)$ be the block Jacobi matrix associated with this partitioning. A calculation using (7)–(9), and Lemma 3.1 shows that the spectrum $\sigma(J)$ of J satisfies $\sigma(J) = \sigma(J_1) \cup \{0\}$ because the first and last block columns of J are zero columns.

$$J_1 = \begin{bmatrix} 0 & R_{21} & 0 & & & & & \\ R_{12} & 0 & 0 & R_{11} & & & & \\ R_{22} & 0 & 0 & R_{21} & 0 & & & \\ & 0 & R_{12} & 0 & 0 & R_{11} & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & 0 & R_{12} & 0 & 0 & R_{11} \\ & & & & R_{22} & 0 & 0 & R_{21} \\ & & & & & 0 & -R_{21}^{-1} & 0 \end{bmatrix}. \tag{13}$$

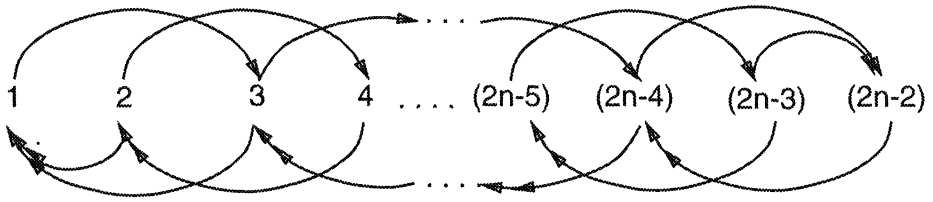
Note that J_1 has only $(n-1)$ diagonal blocks. Using (11) we obtain that

$$J_1^{-1} = \begin{bmatrix} 0 & -R_{21} & R_{22} & & & & & \\ R_{21}^{-1} & 0 & 0 & 0 & & & & \\ 0 & 0 & 0 & -R_{21} & R_{22} & & & \\ & R_{11} & -R_{21} & 0 & 0 & 0 & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & R_{11} & -R_{21} & 0 & 0 & 0 \\ & & & & 0 & 0 & 0 & -R_{21} \\ & & & & & R_{11} & -R_{21} & 0 \end{bmatrix}. \tag{14}$$

Then from (13) and (14) we have that

$$J_1 + J_1^{-1} = \begin{bmatrix} 0 & 0 & R_{22} & & & & & \\ R^* & 0 & 0 & R_{11} & & & & \\ R_{22} & 0 & 0 & 0 & R_{22} & & & \\ & R_{11} & 0 & 0 & 0 & R_{11} & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & R_{11} & 0 & 0 & 0 & R_{11} \\ & & & & R_{22} & 0 & 0 & 0 \\ & & & & & R_{11} & -R^* & 0 \end{bmatrix}, \tag{15}$$

where $R^* = R_{12} + R_{21}^{-1}$. Consider now the directed graph associated with $J_1 + J_1^{-1}$.



From this graph it is seen that a similarity permutation transformation transforms $J_1 + J_1^{-1}$ to

$$\bar{J} = \left[\begin{array}{ccccc|ccccc} 0 & R_{22} & & & & & & & & \\ R_{22} & 0 & R_{22} & & & & & & & \\ & \ddots & \ddots & \ddots & & & & 0 & & \\ & & R_{22} & 0 & R_{22} & & & & & \\ & & & R_{22} & 0 & & & & & \\ \hline R^* & & & & & 0 & R_{11} & & & \\ & 0 & & & & R_{11} & 0 & R_{11} & & \\ & & \ddots & & & & \ddots & \ddots & \ddots & \\ & & & 0 & & & & R_{11} & 0 & R_{11} \\ & & & & -R^* & & & & R_{11} & 0 \end{array}\right]. \tag{16}$$

Let $K = \text{tridiag}(1, 0, 1)$ be of order $(n-1)$. Note that from (11), R_{22} is similar to R_{11} so we have that $\sigma(\bar{J}) = \sigma(G)$, where $G = K \otimes R_{22}$. The symbol $\otimes$ denotes Kronecker product (cf. [10] and also [15] where tensor products were used for the first time in connection with discretized PDE problems). Some of its properties used here are $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ and $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$. For the first property to hold, it is assumed that AC and BD are well defined, while for the second one, it is assumed that A and B are square and nonsingular. The existence of nonsingular matrices X and Y such that $KX = XD_K$ and $R_{22}Y = YJ_R$ ($D_K = \text{diag}(2\cos\frac{\pi}{n}, \ldots, 2\cos\frac{(n-1)\pi}{n})$) and J_R being the Jordan canonical form of R_{22} implies that $G(X \otimes Y) = (X \otimes Y)(D_K \otimes J_R)$. Note that $D_K \otimes J_R$ is upper triangular and the nonsingularity of X and Y implies that $X \otimes Y$ is nonsingular. Hence we conclude that $\sigma(G) = \bigcup_{k=1}^{n-1} \{2\rho\cos\frac{k\pi}{n} | \rho \in \sigma(R_{22})\}$. The above observations are summarized in the following theorem.

THEOREM 3.2. *Let J be the block Jacobi iteration matrix corresponding to* (5) *based on the partitioning in* (4) *and assume that relations* (7), (8), *and* (9) *hold. Then the spectrum of J is given by the expression*

$$\sigma(J) = \{0\} \bigcup_{k=1}^{n-1} \left\{ \mu | \mu + \frac{1}{\mu} = 2\rho \cos \frac{k\pi}{n}, \quad \rho \in \sigma(R_{22}) \right\}. \tag{17}$$

The following observations are a direct consequence of this theorem.

Remark 1. Zero is an eigenvalue of J of multiplicity $4m$.

Remark 2. The corresponding result in [17] can be obtained as a special case of Theorem 3.2. To see this, denote by $\bar{R}$ the corresponding matrix R in [17] and assume that 2^l is the order of J. Then the corresponding result in [17] can be stated as follows: For every $\mu \in \sigma(J)$ such that $\mu \neq 0$ then $\mu + \frac{1}{\mu} = \frac{2}{\rho} \cos\theta$, where $\rho \in \sigma(\bar{R}_{11})$ and $\theta = (2m-1)\pi/2^k$, $m = 1, 2, \ldots, 2^k$, $k = 1, 2, \ldots, l$. If we set $n = 2^l$ in Theorem 3.2, then we can show that $\bar{R}_{11} R_{22} = -I$ and

$$\{\frac{k\pi}{n} | k = 1, 2, \ldots, (n-1)\} = \left\{ \frac{(2m-1)\pi}{2^k} | m = 1, 2, \ldots, 2^k, k = 1, 2, \ldots, l \right\},$$

which proves our assertion.

4. The case of general collocation equations. In this section we consider the reordering, partitioning, and the spectral analysis of the block Jacobi matrix corresponding to the general collocation equations obtained from the discretization of a model PDE problem on the unit square. In the case of uncoupled boundary conditions, the solution of the general collocation equations turns out to be equivalent to the solution of the corresponding interior collocation ones. Thus, it is sufficient to carry out the spectral analysis of the block Jacobi matrix associated with the interior collocation equations. This analysis is quite different from the one presented in the previous section. It is based on a new partitioning of the general and interior collocation equations. This partitioning leads to an always convergent block Jacobi iteration matrix. It is worth remembering that this was not the case for the block Jacobi matrix of §3. Our main result is presented in Theorem 4.3, which is exploited further in §5 to derive the optimal SOR method for the Poisson equation problem under (i) Dirichlet and (ii) Neumann boundary conditions.

4.1. The reordering and partitioning of the general collocation equations. As in the case of interior collocation, we observe that the use of the "natural" ordering of the general collocation equations and unknowns yields the collocation matrices whose structure is not suitable at least for point iterative solvers. To overcome this difficulty, we propose a new ordering that is depicted in Fig. 2 for a specific mesh. In the general case of mixed boundary conditions, the collocation coefficient matrix corresponding to this ordering has no zeros on the diagonal.

With this new ordering scheme, the general collocation matrix has the block tridiagonal structure as in (18)(a), where x denotes a $2(n_y+1) \times 2(n_y+1)$ matrix. The whole matrix is block 2-cyclic consistently ordered of order $4(n_y+1)(n_x+1)$. Since interior collocation matrices may be produced from general collocation matrices by eliminating symbolically the uncoupled boundary conditions, the partitioning in

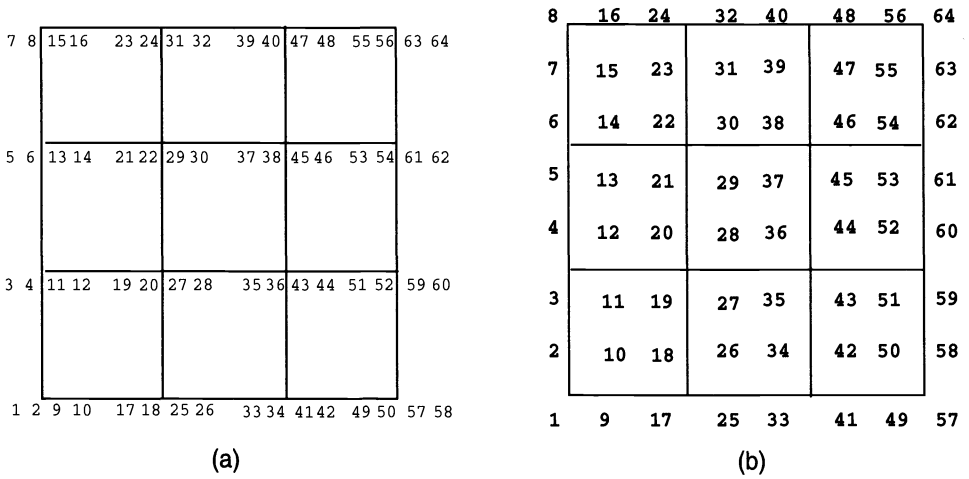
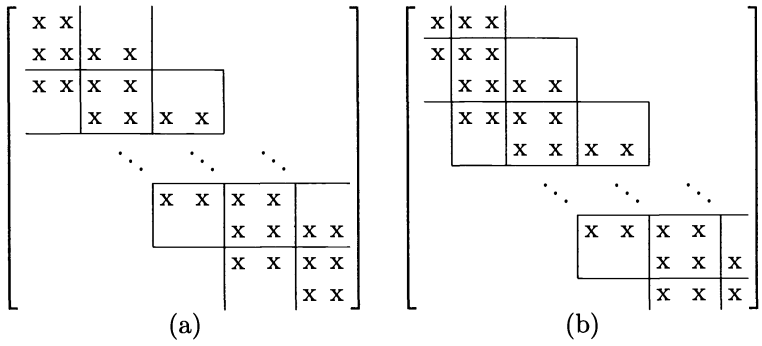


FIG. 2. *The new numbering of the general collocation unknowns* (a) *and equations* (b) *for a mesh with* $n_x = n_y = 3$. *Note that there are* 28 *equations associated with the boundary collocation points and the corresponding set of unknowns that are not eliminated symbolically.*

(18)(a) leads to the decomposition (18)(b) of the interior collocation matrix in (4).

(18)



For most of the cases we have studied, we have found experimentally that block adaptive SOR iterative methods (cf. [9]) and some standard block iterative methods do not converge under the partitioning defined in (4). However, they do converge if the new partitioning (18)(b) is used. We investigate this issue in §5. Next we obtain some useful properties of the block Jacobi iteration matrix associated with the partitioning defined in (18)(b).

4.2. Spectral analysis of the Jacobi matrix. First we apply the block partitioning in (18)(b) to the interior collocation matrix (5) and consider the corresponding splitting $A = D - L - U$. If we assume that A_1 and A_2 of (5) are nonsingular, then D is invertible and the Jacobi matrix associated with the above splitting is $J = D^{-1}(L+U)$. Moreover, it is clear that the spectra of J and $J' = (L+U)D^{-1}$ are the same. Since J' is easier to study, we turn our attention to $\sigma(J')$. The block partitioning and the

definition of J' imply that

$$J' = \left[\begin{array}{cc|cc|cc|cc} & & 0 & P & Q & & & \\ & & P-Q & 0 & 0 & & & \\ \hline & & & & 0 & 0 & P & Q \\ & & & & Q & P & 0 & 0 \\ \hline & & & & & & \ddots & \\ \hline & & & & 0 & 0 & P & Q \\ & & & & Q & P & 0 & 0 \\ \hline & & & & & & 0 & 0 \quad P-Q \\ & & & & & & Q & P \quad 0 \end{array}\right], \tag{19}$$

where $P = -\frac{1}{2}(A_3A_1^{-1} + A_4A_2^{-1})$, $Q = -\frac{1}{2}(A_3A_1^{-1} - A_4A_2^{-1})$. Since P and Q are $2m \times 2m$ matrices, it is not an easy task to find $\sigma(J')$ directly. Instead, we determine $\sigma(J')$ when P and Q are real scalars and use this result to find $\sigma(J')$ in the general case.

LEMMA 4.1. *If P and Q are real scalars, then the eigenvalues μ of J' in* (19) *are either $\mu = \pm(P-Q)$ or satisfy the equation $\mu^2 - 2Q\mu\cos\theta + Q^2 - P^2 = 0$, where $\theta = \frac{k\pi}{n}$, $k = 1, 2, \ldots, (n-1)$.*

Proof. This proof is based on the analysis in [4, pp. 218–230] that has been successfully used in [20] and [13]. For this reason we retain the notation established in [4]. For the sake of convenience, we assume that $PQ(P \pm Q) \neq 0$ without loss of generality. The problem of determining the eigenvalues and eigenvectors of J' is equivalent to solving the boundary value problem of the matrix difference equation

$$\begin{aligned} &B_0 Z_{k-1} + (B_1 - \mu I)Z_k + B_2 Z_{k+1} = 0, \qquad k = 1, 2, \ldots, n, \\ &z_{2,0} = -z_{1,1}, \qquad z_{1,n+1} = -z_{2,n}, \\ &B_0 = \begin{bmatrix} 0 & 0 \\ 0 & Q \end{bmatrix}, \quad B_1 = \begin{bmatrix} 0 & P \\ P & 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix}, \quad Z_k = \begin{bmatrix} z_{1,k} \\ z_{2,k} \end{bmatrix}, \end{aligned} \tag{20}$$

where μ is an eigenvalue of J'. This can be solved by the nonmonic matrix polynomial theory. The nonmonic matrix polynomial that corresponds to (20) is given by

$$L(\lambda) := B_2\lambda^2 + (B_1 - \mu I)\lambda + B_0 = \begin{bmatrix} Q\lambda^2 - \mu\lambda & P\lambda \\ P\lambda & Q - \mu\lambda \end{bmatrix}. \tag{21}$$

From Theorem 8.3 in [4], we know that the general solution of (20) is given by

$$Z_k = X_F J_F^k g, \qquad k = 0, 1, 2, \ldots, \tag{22}$$

where (X_F, J_F) (cf. [4, Chaps. 1, 7]) is a Jordan pair of the matrix polynomial $\mathrm{L}(\lambda)$, $g \in \mathbf{C}^n$ and n is the degree of $\det(\mathrm{L}(\lambda))$. From (21) it is readily obtained that

$$\det(L(\lambda)) = -\lambda(Q\mu\lambda^2 - (\mu^2 + Q^2 - P^2)\lambda + Q\mu). \tag{23}$$

We distinguish two cases according to whether or not μ is zero.
If $\mu = 0$, then it can be proved that $0 \notin \sigma(J')$ (see [14, Lem. 4.1]).
If $\mu \neq 0$, then the eigenvalues of $L(\lambda)$ are given by the expressions

$$\lambda_0 = 0, \lambda_1 = \frac{\mu^2 + Q^2 - P^2 + \sqrt{(\mu^2 + Q^2 - P^2)^2 - 4Q^2\mu^2}}{2Q\mu},$$

$$\lambda_2 = \frac{\mu^2 + Q^2 - P^2 - \sqrt{(\mu^2 + Q^2 - P^2)^2 - 4Q^2\mu^2}}{2Q\mu}.$$

It is clear from (23) that $\lambda_1\lambda_2 = 1$ and $(\lambda_1 + \lambda_2)Q\mu = \mu^2 + Q^2 - P^2$.

If $\lambda_1 \neq \lambda_2$, the eigenvectors of $L(\lambda)$ associated with λ_i , i=0,1,2, are

$$x_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad x_1 = \begin{bmatrix} w_1 \\ 1 \end{bmatrix}, \quad x_2 = \begin{bmatrix} w_2 \\ 1 \end{bmatrix}, \quad w_1 = \frac{\mu\lambda_1 - Q}{P\lambda_1}, \quad w_2 = \frac{\mu\lambda_2 - Q}{P\lambda_2}.$$

Since all the eigenvalues of $L(\lambda)$ have only one eigenvector each, the finite Jordan pair is

$$X_F = \begin{bmatrix} 1 & w_1 & w_2 \\ 0 & 1 & 1 \end{bmatrix}, \quad J_F = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_2 \end{bmatrix}, \quad g = \begin{bmatrix} g_0 \\ g_1 \\ g_2 \end{bmatrix}.$$

It is easy to check that the vectors Z_k defined by (22) satisfy the matrix difference equation (20). For the vector $g = [g_1, g_2]^T$ that satisfies the boundary conditions in (20), we have

$$\begin{bmatrix} 1 + w_1\lambda_1 & 1 + w_2\lambda_2 \\ \lambda_1^n + w_1\lambda_1^{n+1} & \lambda_2^n + w_2\lambda_2^{n+1} \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{24}$$

If $g = [0,0]^T$, then $Z_k = 0$, $k = 0, 1, 2, \ldots$. Since there exists a nonzero solution to (24), we must have

$$(1 + w_1\lambda_1)(1 + w_2\lambda_2)(\lambda_2^n - \lambda_1^n) = 0. \tag{25}$$

If we assume $1 + w_i\lambda_i = 0$, then we get $\lambda_i = \frac{Q-P}{\mu}$. Moreover, solving $Q\mu\lambda_i^2 - (\mu^2 + Q^2 - P^2)\lambda_i + Q\mu = 0$ with respect to μ, we obtain $\mu = \pm(Q - P)$ for $P \neq 0$. This implies $\lambda_1 = \lambda_2 = \pm 1$ which contradicts the assumption $\lambda_1 \neq \lambda_2$. Hence from (25), we conclude $\lambda_1^n - \lambda_2^n = 0$ and determine that $\lambda_1 = e^{i\theta}$, $\lambda_2 = e^{-i\theta}$, $\theta = \frac{k\pi}{n}$, $k = 1, 2, \ldots, n-1$ since $\lambda_1\lambda_2 = 1$. It is worth noticing that for each pair of λ's there are two μ's obtained from equation $\mu^2 - 2Q\mu\cos\theta + Q^2 - P^2 = 0$.

In the case of $\lambda_1 = \lambda_2$, similar analysis leads to the determination of $\mu = \pm(Q-P)$ which completes the proof (see [14, Lem. 4.1]). □

It is difficult to determine $\det(L(\lambda))$ explicitly when P and Q are real matrices. This is due to the fact that (21) is not a 2×2 matrix. Thus, applying the above analysis to obtain $\sigma(J')$ is not an easy task. Instead, we determine μ from each known λ from the scalar case. Specifically, we can show that for $\lambda = e^{\frac{k\pi}{n}i}$, the equation $\det(L(\lambda)) = 0$ can be simplified into

$$\det\left(\begin{bmatrix} Qe^{\frac{k\pi}{n}i} - \mu I & P \\ P & Qe^{-\frac{k\pi}{n}i} - \mu I \end{bmatrix}\right) = 0, \tag{26}$$

which is equivalent to determining the eigenvalues of

$$S_k = \begin{bmatrix} Qe^{\frac{k\pi}{n}i} & P \\ P & Qe^{-\frac{k\pi}{n}i} \end{bmatrix}.$$

To eliminate the complex numbers involved, we perform the similarity transformation $R_k S_k R_k^{-1}$, where

$$R_k = \begin{bmatrix} I & -e^{\frac{k\pi}{n}i} I \\ iI & ie^{\frac{k\pi}{n}i} I \end{bmatrix}.$$

Then the problem at hand is transformed into that of determining the spectrum $\sigma(T_k)$ of

$$T_k = \begin{bmatrix} (Q-P)\cos\frac{k\pi}{n} & (Q-P)\sin\frac{k\pi}{n} \\ -(Q+P)\sin\frac{k\pi}{n} & (Q+P)\cos\frac{k\pi}{n} \end{bmatrix}.$$

Lemma 4.1 gives the basic idea as to how to tackle the matrix problem case. Lemma 4.2 states the corresponding result. Its proof is presented in [14, Lem. 4.2].

LEMMA 4.2. *Let J' be the matrix in* (19) *with P and Q being real matrices. Then its spectrum is given by* $\sigma(J') = \cup_{k=1}^{n-1}\sigma(T_k) \cup \sigma(P-Q) \cup \sigma(Q-P)$.

Note that (19) implies that $Q-P = A_4A_2^{-1}$ and $Q+P = -A_3A_1^{-1}$. We now present our main result for the general collocation equations.

THEOREM 4.3. *Let J be the block Jacobi iteration matrix corresponding to* (5) *with the partitioning in* (18)(b). *Then its spectrum is given by*

$$\sigma(J) = \bigcup_{k=1}^{n-1} \sigma(T_k) \cup \sigma(A_4A_2^{-1}) \cup \sigma(-A_4A_2^{-1}),\ T_k = \begin{bmatrix} A_4A_2^{-1}\cos\frac{k\pi}{n} & A_4A_2^{-1}\sin\frac{k\pi}{n} \\ A_3A_1^{-1}\sin\frac{k\pi}{n} & -A_3A_1^{-1}\cos\frac{k\pi}{n} \end{bmatrix}.$$

Remark. Note that (5) was obtained from a particular class of general collocation equations after eliminating some unknowns symbolically.

5. Iterative methods for the solution of a model problem. In this section we consider the collocation equations obtained by the discretization of the model PDE problem with Dirichlet or Neumann boundary conditions defined on the unit square. Using the analysis of the previous sections, we derive the eigenvalue spectra of the block Jacobi iteration matrices J_1 and J_2 corresponding to the block partitionings in (4) and (18)(b). Then the analysis of the optimal SOR method for the Dirichlet problem is made and optimal results are obtained for the method based on the (18)(b). For the block SOR method based on the (4) partitioning, optimal results are already known [8]. We conclude this section with the analysis of the optimal SOR method for Neumann boundary conditions.

5.1. The Dirichlet case. We consider the iterative solution of the interior collocation equations associated with the Dirichlet boundary value problem

$$u_{xx} + u_{yy} = f \quad \text{in } R = (0,1)\times(0,1), \qquad u = g \text{ on } \partial R \tag{27}$$

and a uniform mesh ($h_x = 1/n_x = h_y$). After applying Papatheodorou's ordering scheme (see Fig. 1) and factoring out $(1/9h^2)$, the collocation matrix is the same as the matrix A in (5). For this particular problem, the entries of A_i, $i = 1,2,3,4$, are independent of h and have the same structure as A. Specifically, we have

$$A_i = \left[\begin{array}{cc|cc} a_1 & a_2 & a_3 & -a_4 \\ a_3 & a_4 & a_1 & -a_2 \end{array}\right]_{\otimes(2n)}. \tag{28}$$

The values of a_j, $j = 1,2,3,4$, corresponding to the A_i's are listed below (see [17]).

	a_1	a_2	a_3	$-a_4$
A_1	$-24-18\sqrt{3}$	$-12-8\sqrt{3}$	24	$-3-\sqrt{3}$
A_2	$-12-8\sqrt{3}$	$-3-2\sqrt{3}$	$3-\sqrt{3}$	0
A_3	24	$3-\sqrt{3}$	$-24+18\sqrt{3}$	$12-8\sqrt{3}$
A_4	$3+\sqrt{3}$	0	$-12+8\sqrt{3}$	$3-2\sqrt{3}$

To find analytic expressions for the elements of $\sigma(J)$, some preliminary analysis is needed.

LEMMA 5.1. *Let the matrices A and B be defined as*

$$A = \left[\begin{array}{cc|cc} a_1 & a_2 & a_3 & -a_4 \\ a_3 & a_4 & a_1 & -a_2 \end{array}\right]_{\otimes(2n)}, \qquad B = \left[\begin{array}{cc|cc} b_1 & b_2 & b_3 & -b_4 \\ b_3 & b_4 & b_1 & -b_2 \end{array}\right]_{\otimes(2n)},$$

and suppose that B is nonsingular and $a_2 b_4 \neq a_4 b_2$. Then the generalized eigenproblem (cf. [5, pp. 251–266]) $A^T x = \lambda B^T x$ *has eigenvalues λ given by the following expressions*:

(i) $\lambda = \frac{a_2+a_4}{b_2+b_4}$ *associated with the eigenvector* $x = [1, 1, -1, -1, \ldots]^T$.

(ii) $\lambda = \frac{a_2-a_4}{b_2-b_4}$ *associated with the eigenvector* $x = [1, -1, 1, -1, \ldots]^T$.

(iii) λ *satisfies the equation*

$$\frac{f_1(\lambda)f_2(\lambda) - f_3(\lambda)f_4(\lambda)}{f_1(\lambda)f_4(\lambda) - f_2(\lambda)f_3(\lambda)} = \cos\theta, \quad \theta = \frac{k\pi}{n}, \quad k = 1, 2, \ldots, (n-1),$$

with associated eigenvector $x = [\rho_1 + \rho_2 g, w_1\rho_1 + w_2\rho_2 g, \ldots, \rho_1^n + \rho_2^n g, w_1\rho_1^n + w_2\rho_2^n g]^T$, *where* $f_i(\lambda) = a_i - \lambda b_i$, $i = 1, 2, 3, 4$, $\rho_1 = e^{i\theta}$, $\rho_2 = e^{-i\theta}$,

$$w_1 = \frac{\rho_1 f_2(\lambda) - f_4(\lambda)}{f_2(\lambda) - \rho_1 f_4(\lambda)}, \quad w_2 = \frac{1}{w_1} \quad \text{and} \quad g = -\frac{w_1 f_2(\lambda) + f_4(\lambda)}{w_2 f_2(\lambda) + f_4(\lambda)}.$$

Proof. The solution of the eigenproblem $A^T x = \lambda B^T x$ is equivalent to solving the matrix difference equation $B_0 Z_{k-1} + B_1 Z_k + B_2 Z_{k+1} = 0$, $k = 1, 2, \ldots, n$, where

$$B_0 = \begin{bmatrix} -f_4(\lambda) & -f_2(\lambda) \\ 0 & 0 \end{bmatrix}, \quad B_1 = \begin{bmatrix} f_2(\lambda) & f_4(\lambda) \\ f_3(\lambda) & f_1(\lambda) \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 & 0 \\ f_1(\lambda) & f_3(\lambda) \end{bmatrix},$$

with the boundary conditions

$$B_0 Z_0 = 0, \qquad \begin{bmatrix} 0 & 0 \\ f_1(\lambda) & f_3(\lambda) \end{bmatrix} Z_{n+1} = \begin{bmatrix} 0 & 0 \\ -f_2(\lambda) - f_4(\lambda) & -f_1(\lambda) - f_2(\lambda) \end{bmatrix} Z_n.$$

Then, the assertions of this lemma follow from an analysis analogous to that in the proof of Lemma 4.1. The details are presented in [14, Lem. 5.1]. □

LEMMA 5.2. *Let A_i, $i = 1, 2, 3, 4$, be the matrices in* (28). *Then there exists a nonsingular matrix X such that $A_4^T X = A_2^T X D$ and $A_3^T X = A_1^T X \bar{D}$, where*

$$\begin{aligned} D &= \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{2n}) \\ &= \operatorname{diag}\left(\frac{3-2\sqrt{3}}{3+2\sqrt{3}}, \frac{3-2\sqrt{3}}{-3-2\sqrt{3}}, \alpha_1^+, \alpha_1^-, \ldots, \alpha_{n-1}^+, \alpha_{n-1}^-\right), \end{aligned} \tag{29}$$

$$\begin{aligned} \bar{D} &= \operatorname{diag}(\bar{\lambda}_1, \bar{\lambda}_2, \ldots, \bar{\lambda}_{2n}) \\ &= \operatorname{diag}\left(\frac{9-7\sqrt{3}}{9+7\sqrt{3}}, \frac{15-9\sqrt{3}}{-15-9\sqrt{3}}, \beta_1^+, \beta_1^-, \ldots, \beta_{n-1}^+, \beta_{n-1}^-\right), \end{aligned} \tag{30}$$

$$\alpha_k^{\pm} = \frac{3\sqrt{3} \pm \sqrt{43 + 40\cos\theta_k - 2\cos^2\theta_k}}{(-28 - 16\sqrt{3}) + (\sqrt{3}+1)\cos\theta_k},$$

$$\beta_k^{\pm} = \frac{(37 + 8\cos\theta_k) \pm 3\sqrt{3}\sqrt{43 + 40\cos\theta_k - 2\cos^2\theta_k}}{(-64 - 36\sqrt{3}) + (19 + 9\sqrt{3})\cos\theta_k} \quad \textit{and} \quad \theta_k = \frac{k\pi}{n}.$$

Proof. The proof is given in [14, Lem. 5.2]. □

5.1.1. Spectra of the block Jacobi iteration matrices. Let J_1 and J_2 be the block Jacobi iteration matrices associated with the partitionings in (4) and (18)(b) of the interior collocation matrix. We now derive analytic expressions for $\sigma(J_1)$ and $\sigma(J_2)$.

First from Lemma 5.2, we have that A_1 and A_2 are nonsingular and $A_4A_2^{-1}A_3A_1^{-1}$ is invertible. Therefore, the blocks of R in (8) can be found explicitly. Specifically, we get $R_{22} = (A_1 - A_4A_2^{-1}A_3)^{-1}(-A_4A_2^{-1}A_1 + A_3)$. Then, an obvious calculation shows that $\sigma(R_{22}) = \sigma((-A_4A_2^{-1}A_1 + A_3)(A_1 - A_4A_2^{-1}A_3)^{-1}) = \sigma((-A_4A_2^{-1} + A_3A_1^{-1})(I - A_4A_2^{-1}A_3A_1^{-1})^{-1})$. After applying Lemma 5.2, we conclude that

$$\sigma(R_{22}) = \left\{ \frac{\bar{\lambda}_i - \lambda_i}{1 - \bar{\lambda}_i\lambda_i}, \quad i = 1, 2, \ldots, 2n \right\},$$

since $A_2^{-T}A_4^T$ and $A_1^{-T}A_3^T$ commute. Lemma 5.2 implies that $X^TA_4A_2^{-1}(X^T)^{-1} = D$ and $X^TA_3A_1^{-1}(X^T)^{-1} = \bar{D}$. By a similarity transformation using $\mathrm{diag}(X^T, X^T)$ and an obvious permutation of rows and columns, it is seen that T_k of Theorem 4.3 is similar to $\mathrm{diag}(D_1, D_2, \ldots, D_n)$, where

$$D_i = \begin{bmatrix} \lambda_i \cos\frac{k\pi}{n} & \lambda_i \sin\frac{k\pi}{n} \\ \bar{\lambda}_i \sin\frac{k\pi}{n} & -\bar{\lambda}_i \cos\frac{k\pi}{n} \end{bmatrix}.$$

So, we have

$$\sigma(T_k) = \{\mu | \mu^2 - (\bar{\lambda}_i - \lambda_i)\mu\cos\frac{k\pi}{n} - \bar{\lambda}_i\lambda_i = 0, i = 1, 2, \ldots, 2n\}.$$

Combining the above results with those of Theorems 3.2 and 4.3, we conclude that

$$\sigma(J_1) = \{0\} \cup \left\{ \bigcup_{k=1}^{n-1} \{\mu | \mu + \frac{1}{\mu} = 2\frac{\bar{\lambda}_i - \lambda_i}{1 - \bar{\lambda}_i\lambda_i}\cos\frac{k\pi}{n}, \quad i = 1, 2, \ldots, 2n\} \right\}, \tag{31}$$

$$\begin{aligned} \sigma(J_2) = {} & \{\pm\lambda_1, \ldots, \pm\lambda_{2n}\} \cup \\ & \left\{ \textstyle\bigcup_{k=1}^{n-1} \{\mu | \mu^2 - (\bar{\lambda}_i - \lambda_i)\mu\cos\frac{k\pi}{n} - \bar{\lambda}_i\lambda_i = 0, \quad i = 1, 2, \ldots, 2n\} \right\}, \end{aligned} \tag{32}$$

where $\lambda_i, \bar{\lambda}_i$ are the ones of Lemma 5.2. Note that $0 \in \sigma(J_1)$ with multiplicity $4n$.

It can be proven that for $i = 1, \ldots, (2n)$, λ_i and $\bar{\lambda}_i$ of (29) and (30) are real numbers with magnitudes less than 1. This implies that $\mu + \frac{1}{\mu}$ is real and has absolute value less than 2 and all eigenvalues of J_1, except 0, are complex and lie on the circumference of the unit circle. Therefore, the spectral radius $\rho(J_1)$ of J_1 is equal to 1. On the other hand, based on (32), we can show that the spectral radius $\rho(J_2)$ of J_2 is equal to

$$\rho(J_2) = a := \frac{1}{2}\left((\lambda_3 - \bar{\lambda}_3)\cos\frac{\pi}{n} + \sqrt{(\lambda_3 - \bar{\lambda}_3)^2\cos^2\left(\frac{\pi}{n}\right) + 4\lambda_3\bar{\lambda}_3} \right) \tag{33}$$

and bounded above by $|\bar{\lambda}_3|$, where λ_3 and $\bar{\lambda}_3$ are those of Lemma 5.2. Thus, we conclude that for any discretization grid size n, the following relation $\rho(J_2) < |\bar{\lambda}_3| < \rho(J_1) = 1$ holds. Consequently, for the model problem in §5.1, the Jacobi iterative method associated with the partitioning (18)(b) converges. This is not the case for the partitioning (4).

5.1.2. Optimal SOR. The optimal SOR method with the Jacobi matrix J_1 has been derived in [8]. In this paper we consider the determination of SOR overrelaxation parameter when the Jacobi matrix is J_2. Recall that J_2 is consistently ordered weakly cyclic of index 2. Therefore, the algorithm of Young and Eidson [23] (see also [22, pp. 194–200]) can be applied to determine the optimal SOR method. To apply this algorithm the hull (smallest convex polygon) of $\sigma(J_2)$ is required. From (32) we obtain

$$\mu = \left((\lambda_j - \bar{\lambda}_j)\cos\frac{k\pi}{n} \pm \sqrt{\left((\lambda_j - \bar{\lambda}_j)\cos\frac{k\pi}{n}\right)^2 + 4\lambda_j\bar{\lambda}_j} \right)/2. \tag{34}$$

For real μ we have already found that $\max|\mu| = a$ in (33). However, μ may be a complex number only when $\lambda_j\bar{\lambda}_j < 0$. Furthermore, all complex eigenvalues of J_2 associated with them must lie on the circumference of the circle centered at (0,0) with radius $\sqrt{-\lambda_j\bar{\lambda}_j}$. Let b be the maximum value of $\sqrt{-\lambda_j\bar{\lambda}_j}$ among those j such that $-\bar{\lambda}_j\lambda_j > 0$. Then

$$b = \max_k \sqrt{-\alpha_k^- \beta_k^-},\ \frac{k\pi}{n} \in \left(\cos^{-1}\left(\frac{122 - 54\sqrt{3}}{59}\right),\ \cos^{-1}(10 - 6\sqrt{3}) \right) \tag{35}$$

can be obtained from (29)–(30). It follows that all complex eigenvalues of J_2 lie inside or on the circumference of the circle with center at (0,0) and radius b. On the other hand, from (33) we have $a = \rho(J_2) \in \sigma(J_2)$. If n is even then $k = \frac{n}{2}$ in (34) implies that $bi \in \sigma(J_2)$, where i is the imaginary unit. Thus, the ellipse with semiaxes a and b is the optimal capturing ellipse of $\sigma(J_2)$ and the optimal value of ω_{opt} is given by

$$\omega = \frac{2}{1 + (1 + b^2 - a^2)^{1/2}},\quad \rho(\mathcal{L}_\omega) = \left(\frac{a+b}{1 + (1 + b^2 - a^2)^{1/2}}\right)^2 \tag{36}$$

where $\mathcal{L}_\omega$ is the associated block SOR iteration matrix with overrelaxation parameter ω. However, if n is odd then $bi \notin \sigma(J_2)$. In this case the value of ω in (36) is still a very good approximation to ω_{opt}, because b is only slightly greater than the imaginary semiaxis of the corresponding optimum-capturing ellipse and tends to the optimal one ($b = 0.0237973$) when $n \to \infty$. Two examples of $\sigma(J_2)$ for each of the two cases of n even and odd are depicted in Fig. 3.

5.2. The Neumann case. Here we consider the iterative solution of the interior collocation equations associated with the following Neumann boundary value problem and a uniform mesh

$$u_{xx} + u_{yy} = f \text{ in } R = (0,1) \times (0,1), \qquad \partial u/\partial n = g \quad \text{on } \partial R. \tag{37}$$

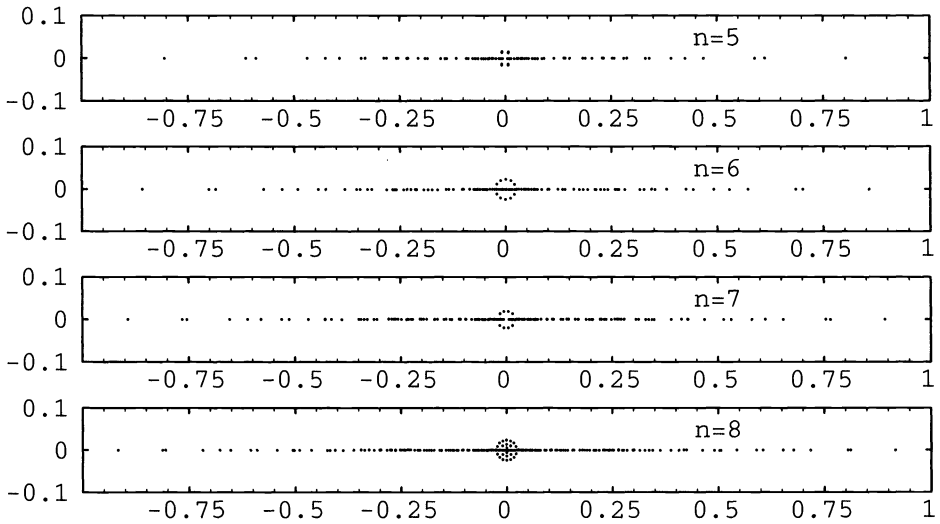


FIG. 3. *The spectrum $\sigma(J_2)$ of the Jacobi matrix J_2 associated with the partitioning in* (18)(b) *of the interior collocation matrix.*

For the analysis below we introduce a similar notation to that in §2, namely,

$$[A|B]_{\otimes(\tilde{2n})} = \begin{bmatrix} a & B & & & \\ & A & B & & \\ & & \ddots & & \\ & & & A & B \\ & & & & A \quad b \end{bmatrix}, \quad a = \begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix}, \quad b = \begin{bmatrix} b_{11} \\ b_{21} \end{bmatrix}$$

that differs only in the definition of the vectors a and b.

Using Papatheodorou's ordering of §3.1 and factoring out $(1/9h^2)$, the interior collocation matrix has the form

$$A = \left[\begin{array}{cc|cc} A_1 & A_2 & A_3 & -A_4 \\ A_3 & A_4 & A_1 & -A_2 \end{array}\right]_{\otimes(\tilde{2n})}.$$

For this particular problem, the entries of A_i, $i = 1, 2, 3, 4$, are independent of h and have the same structure as before, namely,

$$A_i = \left[\begin{array}{cc|cc} a_1 & a_2 & a_3 & -a_4 \\ a_3 & a_4 & a_1 & -a_2 \end{array}\right]_{\otimes(\tilde{2n})}.$$

The values of a_j corresponding to A_i are those given in §5.1. Following the analysis developed in §4.2, we obtain that the corresponding block Jacobi iteration matrix J'

is given by

$$J' = \left[\begin{array}{cc|cc|cc|cc|cc} & & 0 & P & Q & & & & & \\ & & P+Q & 0 & 0 & & & & & \\ \hline & & & & 0 & 0 & P & Q & & \\ & & & & Q & P & 0 & 0 & & \\ \hline & & & & & & \ddots & & & \\ \hline & & & & & & 0 & 0 & P & Q \\ & & & & & & Q & P & 0 & 0 \\ \hline & & & & & & & & 0 & 0 \\ & & & & & & & & Q & P \end{array}\right],$$

where P and Q are defined in the same way as in (19). Using the similarity transformation $SJ'S^{-1}$ with $S = \mathrm{diag}(1, 1, -1, -1, 1, 1, \ldots)$, J' is transformed to J''. The matrix J'' is of exactly the same structure as J' in (19) with the only difference being that $-Q$ is replaced by Q. On the other hand, we have

$$P_1 \left[\begin{array}{cc|cc} a_1 & a_2 & a_3 & -a_4 \\ a_3 & a_4 & a_1 & -a_2 \end{array}\right]_{\otimes(\tilde{2n})}, \qquad P_2 = \left[\begin{array}{cc|cc} a_2 & a_1 & a_4 & -a_3 \\ a_4 & a_3 & a_2 & -a_1 \end{array}\right]_{\otimes(2n)},$$

with $P_1 = \mathrm{diag}(I, -I, I, -I, \ldots)$ and $P_2 = \mathrm{diag}(1, -I_2, I_2, -I_2, I_2, \ldots, (-1)^n)$, where I is the 2×2 identity matrix and $I_2 = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$. Now applying Lemmas 4.2, 5.1, and 5.2, we get

$$\begin{aligned} \sigma(J') \;=\; & \{\pm\bar{\lambda}_1, \ldots, \pm\bar{\lambda}_{2n}\} \cup \\ & \left\{\bigcup_{k=1}^{n-1} \{\mu \mid \mu^2 - (\bar{\lambda}_i - \lambda_i)\mu \cos\tfrac{k\pi}{n} - \bar{\lambda}_i\lambda_i = 0,\ i = 1, 2, \ldots, 2n\}\right\}, \end{aligned} \tag{38}$$

where

$$(\lambda_1, \lambda_2, \ldots, \lambda_{2n}) = \left(\frac{15 - 7\sqrt{3}}{-15 - 7\sqrt{3}}, \frac{9 - 9\sqrt{3}}{9 + 9\sqrt{3}}, \alpha_1^+, \alpha_1^-, \ldots, \alpha_{n-1}^+, \alpha_{n-1}^-\right),$$

$$(\bar{\lambda}_1, \bar{\lambda}_2, \ldots, \bar{\lambda}_{2n}) = \left(\frac{48 - 18\sqrt{3}}{-48 - 18\sqrt{3}}, \frac{18\sqrt{3}}{-18\sqrt{3}}, \beta_1^+, \beta_1^-, \ldots, \beta_{n-1}^+, \beta_{n-1}^-\right),$$

$$\alpha_k^{\pm} = \frac{3\sqrt{3} \pm \sqrt{43 + 40\cos\theta_k - 2\cos^2\theta_k}}{(-28 - 16\sqrt{3}) + (\sqrt{3} + 1)\cos\theta_k},$$

$$\beta_k^{\pm} = \frac{(37 + 8\cos\theta_k) \pm 3\sqrt{3}\sqrt{43 + 40\cos\theta_k - 2\cos^2\theta_k}}{(-64 - 36\sqrt{3}) + (19 + 9\sqrt{3})\cos\theta_k}, \qquad \theta_k = \frac{k\pi}{n}.$$

From (38), it is clear that all the eigenvalues of J except ± 1, which are simple, have magnitudes less than 1. Therefore $\rho(J') = 1$ and $\mathrm{index}(I - J') = 1$ (i.e., $\mathrm{rank}(I - J')^2 = \mathrm{rank}(I - J')$). Moreover, J' is a block 2-cyclic matrix. The analysis in [7] and §5.3 allows us to obtain the optimal SOR overrelaxation parameter for n even and its "close" approximation for n odd by means of the formulas (36). Note that b is exactly the same as in the Dirichlet case while $a = \bar{\lambda}_3$.

6. Numerical results. In this section we present numerical results to verify the formulas and the convergence behavior of various iterative methods considered in this paper. We also compare the time and space performance of optimal SOR, LINPACK Band GE, and GMRES software [19] for solving the interior and general collocation equations. All numerical computations were carried out on a Sun Sparc IPX 4/470 with 32 Mbytes of main memory in double precision. The execution times estimated are given in seconds and the space is measured in words.

First, we attempt to verify numerically the formulas (32) and (38). For this we choose $n_x = n_y = 3$ and find the eigenvalues of the block Jacobi iteration matrices J_2 and J' by using the subroutine EVLRG from the IMSL/MATH library. The eigenvalues are presented in Tables 1 and 2, respectively. They agree with those obtained from the formulas (32) and (38) at least up to the the number of the decimal digits displayed in these tables.

TABLE 1
The 36 eigenvalues of the Jacobi matrix J_2 for $n_x = n_y = 3$.

±0.5726	±0.3272	±0.3169	±0.2411	±0.2136	±0.1741
±0.1238	±0.0858	±0.0718	±0.0718	±0.0526	±0.0499
±0.0374	±0.0263	±0.0260	±0.0123	±0.0079	±0.0014

TABLE 2
The 36 eigenvalues of the Jacobi matrix J' for $n_x = n_y = 3$.

±1.000	±0.753	±0.732	±0.573	±0.401	±0.366
±0.327	±0.317	±0.214	±0.212	±0.179	±0.126
±0.058	±0.037	±0.026	±0.012	±0.001	±0.001

Second, we verify some of the convergence results obtained in this paper. For this we apply the INTERIOR and GENERAL HERMITE COLLOCATION subroutines from the ELLPACK system [18] to discretize several PDE problems on the unit square. For the solution of these equations, we have developed three new solution modules in ELLPACK based on block AOR, SOR, and adaptive SOR methods (cf. [9]) and new indexing modules based on the orderings introduced in this paper. Depending on the initial value of ω_0 selected for the adaptive SOR, we introduce the following notation: SOR_1 if $\omega_0 = 1.0$, SOR_2 if ω_0 is equal to the optimal ω for a model problem, and SOR_3 if ω_0 is the final adaptive ω found by solving the same problem on a coarser mesh unless $n_x = n_y = 2$ in which case we take $\omega_0 = 1.0$. Since SOR_1 gives no better performance than SOR_2 and SOR_3, we do not include it in the tables that follow. Throughout, we denote the semi-optimal SOR with ω the optimal value for a model problem by SOR_0. We have implemented the adaptive procedure used by the ITPACK routines [16]. For completeness, we note that the AOR method for the solution of $Ax = b$ is defined by

$$(D - rL)x_{n+1} = [(1-\omega)D + (\omega - r)L + \omega U]x_n + \omega b,$$

assuming the splitting $A = D - L - U$. Its convergence properties depend on the choice of the pair of parameters (ω, r) [6]. For comparison purposes we use AOR with $(\omega, r) = (0.5, 1.0)$, which is the optimal pair of parameters found and used by Papatheodorou in [17].

The iterative solvers implemented depend on the block partitioning of the collocation coefficient matrix. In this study we consider three different matrix partitionings

TABLE 3

The convergence behavior of four block iterative methods for solving the interior and general collocation equations obtained by discretizing the equation $u_{xx} + u_{yy} = f$ *with Dirichlet boundary conditions* $(u = g)$. *The functions* f *and* g *are selected so that* $u(x, y) = \phi(x)\phi(y)$, *where* $\phi(x) = 0$, *if* $x \leq 0.35$, *or if* $x \geq 0.65$, *otherwise* $\phi(x)$ *is a quintic polynomial determined so that it has two continuous derivatives. In* (a) *The* AOR *is based on* P_I *and the rest on the* P_{II} *block partitioning of the interior collocation matrix. In* (b) *all methods implemented use the partitioning* (18)(a). *The initial vector* $x_0 = [.5, .5, \ldots, .5]^T$ *was used in all iterative methods. The data displayed include number of iterations required to achieve specified tolerance, maximum discretization error, and the exact or estimated value of the* SOR *parameter* ω.

(a)

Interior Hermite collocation									
mesh	AOR (0.5,1.0)		Jacobi		Gauss–Seidel		Optimal SOR		
size	iter	error	iter	error	iter	error	ω_{opt}	iter	error
2 x 2	17	1.21	9	1.21	6	1.21	1.0314	6	1.21
4 x 4	17	1.28e−2	29	1.28e−1	15	1.28e−1	1.1786	9	1.28e−1
8 x 8	41	7.56e−2	94	7.55e−2	48	7.56e−2	1.4271	19	7.56e−2
16 x 16	200	2.59e−2	305	2.63e−2	154	2.62e−2	1.6536	40	2.59e−2

(b)

General Hermite collocation							
mesh	Jacobi		Gauss–Seidel		Optimal SOR		
size	iter	error	iter	error	ω_{opt}	iter	error
2 x 2	12	1.19	7	1.19	1.0314	6	1.19
4 x 4	32	1.28e−1	18	1.28e−1	1.1786	11	1.28e−1
8 x 8	104	7.56e−2	56	7.56e−2	1.4271	21	7.57e−2
16 x 16	344	2.63e−2	182	2.61e−2	1.6536	46	2.59e−2

(c)

	Interior collocation						General collocation				
mesh	Optimal SOR			AdaptiveSOR$_3$			Optimal SOR		AdaptiveSOR$_3$		
size	ω_{opt}	iter	error	ω	iter	error	iter	error	ω	iter	error
2 x 2	1.0314	6	1.21	1.0131	6	1.21	6	1.19	1.0176	7	1.19
4 x 4	1.1786	9	1.28e−1	1.0131	15	1.28e−1	11	1.12e−1	1.0176	15	1.28e−1
8 x 8	1.4271	19	7.57e−2	1.2685	32	7.56e−2	21	7.57e−2	1.2829	31	7.57e−2
16 x 16	1.6536	40	2.59e−2	1.5821	59	2.59e−2	46	2.59e−2	1.6528	68	2.59e−2

depicted below for a specific mesh size $n_x = n_y = 3$.

$$P_I = \left[\begin{array}{cc|c|c} x & x & x & \\ x & x & x & \\ \hline & x & x\ x & x \\ & x & x\ x & x \\ \hline & & x & x\ x \\ & & x & x\ x \end{array}\right], \quad P_{II} = \left[\begin{array}{c|cc|cc|c} x & x & x & & & \\ \hline x & x & x & & & \\ & x & x & x & x & \\ & x & x & x & x & \\ \hline & & & x & x & x \\ & & & & x & x \end{array}\right], \quad P_{III} = \left[\begin{array}{c|c|c|c|c|c} x & x & x & & & \\ \hline x & x & x & & & \\ \hline & x & x & x & x & \\ \hline & x & x & x & x & \\ \hline & & & x & x & x \\ \hline & & & x & x & x \end{array}\right].$$

They are denoted by P_I, P_{II}, and P_{III} where each x denotes a 6×6 matrix and has the same structure as the global one. P_I and P_{II} correspond to partitionings in (4) and (18)(b), respectively.

The efficiency of the block iterative methods depends on the time required to solve the linear subsystems $D_i x = b$, where D_i is the ith block diagonal element of A. In general we expect the bandwidth of the matrices D_i to be small. However, for the block partitionings above, the upper and lower bandwidth of some D_i's is $(2n+2)$. For these D_i's, instead of solving the corresponding linear subsystem $D_i x = b$ directly, we solve the transformed one $PD_iP^{-1}y = Pb$, where $y = Px$ and $P =$

TABLE 4

The time and memory complexity of five solvers for solving the discrete equations obtained by applying the INTERIOR HERMITE COLLOCATION procedure to the equation $u_{xx} + u_{yy} = f$ *with Dirichlet boundary conditions. The function* f *is selected so that* $u(x,y) = 10\phi(x)\phi(y)$, *where* $\phi(x) = e^{-100(x-0.1)^2}(x^2 - x)$. *The optimal* SOR *is based on* P_{II} *block partitioning and Band* GE *was applied with partial pivoting and "natural ordering" of the equations. The* GMRES *software was applied with three different preconditioners consisting of the diagonal matrices of the matrices* P_I, P_{II}, *and* P_{III}. *The Band* GE *could not run for mesh size* 128×128 *on the machine used due to memory limitations. The initial value* x_0 *used for all iterative methods was determined using the multigrid-type approach. The times include the cost of estimating* x_0.

(a)

		Optimal SOR				Band GE		
mesh	equations	time	iter	workspace	error	time	workspace	error
2x2	16	0.02	5	264	2.905e−1	0.02	464	2.905e−1
4x4	64	0.14	10	1136	1.456e−1	0.07	2624	1.456e−1
8x8	256	1.02	19	4704	1.563e−2	0.53	16640	1.563e−2
16x16	1024	6.22	27	19136	6.083e−4	5.03	115712	6.082e−4
32x32	4096	50.35	57	77184	5.795e−5	60.77	856064	5.795e−5
64x64	16384	360.28	99	310016	2.035e−6	797.75	6569984	2.035e−6
128x128	65536	3031.63	213	1242624	1.263e−7	NA	NA	NA

(b)

GMRES (restarted every 50 steps)								
			PREC1		PREC2		PREC3	
mesh	equations	error[1]	time	iter	time	iter	time	iter
2x2	16	2.905e−1	0.02	3	0.03	6	0.03	7
4x4	64	1.456e−1	0.20	15	0.16	10	0.24	18
8x8	256	1.563e−2	1.87	28	1.15	18	1.80	28
16x16	1024	6.082e−4	19.52	64	9.56	33	14.89	48
32x32	4096	5.766e−5	108.25	79	48.96	36	83.65	66
64x64	16384	2.056e−6	1255.03	244	371.66	66	559.91	107
128x128	65536	1.1400e−7	9134.46	400[2]	2571.77	106	5685.47	282

[1] Approximately the same error is found by using any of the three preconditioners as long as the same stopping criterion is satisfied.

[2] At this step the stopping criterion was not satisfied. The corresponding error was 1.18e−7

$[e_1, e_{n+1}, e_2, e_{n+2}, \ldots, e_n, e_{2n}]$ with e_i being the standard unit vectors. It is easy to see that the bandwidth of PD_iP^{-1} is only 5. Thus the transformed diagonal subsystem can be solved much faster using Band GE without pivoting.

In the following tables we display the maximum discretization error $||u - u_h||_\infty$ based on a 65×65 grid, where u is the exact solution of the PDE problem and u_h is the computed Hermite cubic piecewise polynomial solution given by (3). To compare the efficiency among various iterative solvers considered, we assume the same initial solution x_0 and the same stopping criterion, namely,

$$\frac{||x_{n+1} - x_n||_\infty}{||x_{n+1}||_\infty} < \text{eps} = 5 \times 10^{-6}.$$

Table 3(a) indicates the convergence of four block iterative methods applied to the system of interior collocation equations corresponding to different mesh sizes. The AOR implemented is based on the partitioning P_I, while the rest of the block methods are based on the partitioning P_{II}. The optimal parameters of AOR used are

TABLE 5

The performance and convergence data of optimal SOR *and adaptive* SOR_3 *based on* P_{II} *block partitioning and the Band* GE *with partial pivoting and "natural ordering" for solving the discrete equations obtained by applying INTERIOR HERMITE COLLOCATION procedure to solve the Poisson equation* $u_{xx}+u_{yy}=f$ *with Neumann boundary conditions (Tables* (a) *and* (b)*) and uncoupled mixed boundary conditions (Table* (c)*). The function* f *is selected as in Table* 3*. The data displayed include number of iterations required to achieve specified tolerance, maximum discretization error, the exact and estimated value of the* SOR *parameter* ω *used, and execution times. For mesh size* 128×128*, the Band* GE *could not run on the machine used due to memory limitations.*

(a)

	Optimal SOR				Adaptive SOR_3				Band GE	
mesh	ω_{opt}	time	iter	error	ω	time	iter	error	time	error
2x2	1.2926	0.03	10	2.48	1.091	0.02	9	2.48	0.02	2.48
4x4	1.3042	0.18	13	3.22e−1	1.091	0.25	23	3.22e−1	0.07	3.22e−1
8x8	1.5498	1.17	22	1.40e−1	1.436	1.64	31	1.40e−1	0.52	1.40e−1
16x16	1.7392	9.56	46	4.76e−2	1.704	12.21	58	4.69e−2	5.01	4.76e−2
32x32	1.8550	79.09	94	1.40e−2	1.800	82.21	94	1.15e−2	58.03	1.40e−2
64x64	1.9153	664.84	197	2.18e−3	1.600	467.76	125	6.21e−3	797.97	2.20e−3
128x128	1.9413	7746.36	599	8.05e−4	1.800	1584.63	77	5.24e−3	NA	NA

(b)

mesh	Optimal SOR		Adaptive SOR_2				GMRES(50)		
size	iter	error	ω	iter	error	time	iter	error	time
2x2	10	2.48	1.2926	10	2.48	0.02	7	2.48	0.02
4x4	13	3.22e−1	1.3042	13	3.22e−1	0.17	12	3.22e−1	0.17
8x8	22	1.40e−1	1.5498	22	1.40e−1	1.17	19	1.40e−1	1.21
16x16	46	4.76e−2	1.600	55	4.63e−2	11.39	35	4.76e−2	10.23
32x32	94	1.40e−2	1.800	93	1.14e−2	80.49	91	1.40e−2	113.95
64x64	197	2.18e−3	1.600	125	6.15e−3	481.64	194	2.19e−3	1188.78
128x128	599	8.05e−4	1.800	77	5.19e−3	1592.89	684	8.02e−4	14505.91

(c)

with boundary condition $u=g_1$ at $x=0$ or $y=1$ and $u_n=g_2$ at $x=1$ or $y=0$										
	SOR_0				Adaptive SOR_3				Band GE	
mesh	ω	time	iter	error	ω	time	iter	error	time	error
2x2	1.162	0.02	9	1.22	1.150	0.03	11	1.22	0.00	1.22
4x4	1.2414	0.27	26	1.31e−1	1.150	0.35	32	1.31e−1	0.07	1.31e−1
8x8	1.4885	1.64	31	7.38e−2	1.494	2.21	43	7.38e−2	0.53	7.40e−2
16x16	1.6964	12.48	60	2.60e−2	1.750	14.83	70	2.59e−2	5.02	2.57e−2
32x32	1.8304	68.5	75	7.78e−3	1.900	100.65	116	7.44e−3	59.15	7.28e−2
64x64	1.903	520.75	150	1.27e−3	1.600	437.67	108	2.52e−3	794.17	1.14e−3
128x128	1.9364	5773.44	434	4.35e−4	1.800	1601.95	81	2.07e−3	NA	NA

$(\omega, r) = (0.5, 1.0)$ according to the analysis in [17]. The optimal SOR parameter ω_{opt} was obtained based on the analysis presented in §5. The data in these tables suggest that the block SOR has the fastest convergence. Table 3(b) depicts the convergence behavior of three of the four iterative methods considered in Table 3(a) for the general collocation equations. AOR (0.5,1.0) is not efficient for this type of equation, so we do not present any data here. It is worth noticing that the spectral analysis of the Jacobi iteration matrix for interior and general collocation equations has shown that ω_{opt} is the same for both cases. The data in these tables suggest that the block SOR has the fastest convergence. Table 3(c) depicts the convergence data (number of iterations and discretization error) of optimal SOR and adaptive SOR_3 for both general and interior collocation equations. These data suggest that the adaptive SOR behaves almost as the optimal SOR for the model problem considered for relative coarse meshes.

TABLE 6

The performance and convergence data of SOR_0 (ω *takes the optimal values for the Dirichlet model problem in Table* 3), *Adaptive* SOR_3, *Band* GE, *and* GMRES (*restarting every* 50 *steps*) *for solving the interior collocation equations obtained from the discretization of the equation* $[2+(y-1)e^{-y^4}]u_{xx}+[1+\frac{1}{(1+4x^2)}]u_{yy}+5[x(x-1)+(y-0.3)(y-0.7)]u = f$, *with boundary conditions* ($u = g$). *The functions* f *and* g *are selected so that* $u(x,y) = \frac{x+y^2}{1+2x} + (1+x)(y-1)e^{-y^4} + 5(x+y)\cos(xy)$. *All applied solvers were based on* P_{II} *block structure with multigrid-type initialization. The data displayed include maximum discretization error and execution times.*

mesh	Band GE		Adaptive SOR_3		SOR_0		GMRES(50)	
size	time	error	time	error	time	error	time	error
2x2	0.05	7.67e−3	0.03	7.67e−3	0.0	7.67e−3	0.02	7.67e−3
4x4	0.25	1.57e−3	0.17	1.57e−3	0.12	1.57e−3	0.15	1.57e−3
8x8	1.80	1.24e−4	0.84	1.25e−4	0.67	1.24e−4	0.97	1.24e−4
16x16	15.95	8.61e−6	3.05	1.24e−5	4.59	8.62e−6	8.07	8.61e−6
32x32	66.21	6.06e−7	12.15	9.30e−6	31.58	6.06e−7	70.23	6.06e−7
64x64	849.99	4.35e−9	56.58	8.92e−6	216.13	8.58e−9	466.88	1.26e−8

Table 4(a) depicts the time and memory complexity of optimal SOR and the LINPACK Band GE with partial pivoting, and Table 4(b) depicts GMRES [19] under three different preconditioners to solve the interior collocation equations associated with a model problem under different mesh sizes. In SOR and GMRES, the initial guess of the solution corresponding to an $n \times n$ mesh is estimated from the previous collocation approximation based on an $(n/2) \times (n/2)$ mesh. Throughout we refer to this scheme of initialization as the multigrid-type initialization. The execution times of iterative methods include the total time to estimate the initial guess. The direct solver is applied to the system obtained using the natural ordering while the block SOR utilizes the above-mentioned transformations to diagonal subsystems. These subsystems were solved using Band GE *without* pivoting. It should be added that generally Band GE with partial pivoting is necessary to solve the general collocation systems. In these experiments, which are simply the block diagonal matrices associated with the block matrices P_I, P_{II}, and P_{III} of the collocation matrix, we consider right preconditioning for GMRES. We refer to the matrices as PREC1, PREC2, and PREC3. The GMRES procedure is restarted every 50 steps and the stopping criterion is set to be $\frac{||b-Ax_n||_2}{||b-Ax_0||_2} < \text{eps} = 5 \times 10^{-5}$. The data suggest that the iterative methods have much smaller memory requirements. This of course was expected. However, we were surprised that the time efficiency of the optimal SOR was better than the rest of the solvers considered and occurred at a level of relatively coarse meshes. In the case of GMRES, the preconditioner based on the block diagonal matrix corresponding to P_{II} block structure had the best performance.

Table 5 compares the performance and convergence behavior of optimal SOR, adaptive SOR_3, Band GE, and GMRES(50) for a model problem with Neumann (Tables 5(a) and 5(b)) and uncoupled boundary conditions (Table 5(c)). Again we observe that for fine meshes optimal SOR outperforms the rest of methods, especially with respect to the accuracy, with GMRES(50) being the slowest. All iterative solvers used the previous most accurate solution to estimate their initial solution.

Table 6 indicates the performance of SOR (ω takes the optimal values for the Dirichlet model problem in Table 3), adaptive SOR_3, Band GE, and GMRES (restarted every 50 steps) for solving the interior collocation equations obtained from the discretization of a general elliptic PDE with Dirichlet boundary conditions on the unit square. All applied solvers were based on P_{II} block structure. The multigrid-type

approach was used to start the iterations. It is clear that the semi-optimal SOR is the fastest for fine meshes without affecting the discretization error. Adaptive SOR_3 appears to affect the discretization error.

REFERENCES

[1] B. Bialecki, G. Fairweather, and K. R. Bennett, *Fast direct solvers for piecewise Hermite bicubic orthogonal spline collocation equations*, SIAM J. Numer. Anal., 29 (1992), pp. 156–173.

[2] W. R. Dyksen, *Tensor product generalized* ADI *method for separable elliptic problems*, SIAM J. Numer. Anal., 24 (1987), pp. 59–76.

[3] W. R. Dyksen and J. R. Rice, *The importance of scaling for the Hermite bicubic collocation equations*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 707–719.

[4] I. Gohberg, P. Lancaster, and L. Rodman, *Matrix Polynomials*, Academic Press, New York, 1982.

[5] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, 1983.

[6] A. Hadjidimos, *Accelerated overrelaxation method*, Math. Comp., 32 (1978), pp. 149–157.

[7] ———, *On the optimization of the classical iterative schemes for the solution of complex singular linear systems*, SIAM J. Alg. Disc. Meth., 6 (1985), pp. 555–566.

[8] A. Hadjidimos, T. S. Papatheodorou, and Y. G. Saridakis, *Optimal block iterative schemes for certain large sparse and non-symmetric linear systems*, Linear Algebra Appl., 110 (1988), pp. 285–318.

[9] A. Hageman and D. M. Young, *Applied Iterative Methods*, Academic Press, New York, 1981.

[10] P. R. Halmos, *Finite Dimensional Vector Spaces*, Princeton University, Princeton, NJ, 1958.

[11] E. N. Houstis, W. Mitchell, and J. R. Rice, *Collocation software for second order elliptic partial differential equations*, ACM Trans. Math. Software, 11 (1985), pp. 379–412.

[12] E. N. Houstis, T. S. Papatheodorou, and R. Barourt, *On the iterative solution of collocation equations*, 10th IMACS World Congress Proceedings, 1982, Montreal, pp. 98–100.

[13] S.-B. Kim, A. Hadjidimos, E. N. Houstis, and J. R. Rice, *Multi-Parameterized Schwartz Splittings*, Tech. Report, CSD-TR-92-073, Computer Science Department, Purdue University, West Lafayette, IN, 1992.

[14] Y.-L. Lai, A. Hadjidimos, E. N. Houstis, and J. R. Rice, *On the Iterative Solution of Hermite Collocation Equations* , Tech. Report, CSD-TR-92-094, Computer Science Department, Purdue University, West Lafayette, IN, 1992.

[15] R. E. Lynch, J. R. Rice, and D. H. Thomas, *Direct solution of partial difference equations by tensor product methods*, Numer. Math., 6 (1964), pp. 185–199.

[16] T. C. Oppe, W. D. Joubert, and D. R. Kincaid, *A Package for Solving Large Sparse Linear Systems by Various Iterative Methods*, Tech. Report CNA-216, Center for Numerical Analysis, University of Texas at Austin, TX, April 1988.

[17] T. S. Papatheodorou, *Block* AOR *iteration for non-symmetric matrices*, Math. Comp., 41 (1983), pp. 511–525.

[18] J. R. Rice and R. F. Boisvert, *Solving Elliptic Problems Using ELLPACK*, Springer-Verlag, New York, 1985.

[19] Y. Saad and M. H. Schultz, GMRES: *A generalized minimal residual algorithm for solving non-symmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[20] W. P. Tang, *Schwarz Splitting and Template Operator*, Ph.D. Thesis, Computer Science Department, Stanford University, Stanford, CA, 1987.

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

[22] D. M. Young, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.

[23] D. M. Young and H. E. Eidson, *On the Determination of the Optimal Relaxation Factor for the* SOR *Method When the Eigenvalues of the Jacobi Method are Complex*, Report CNA-1, Center of Numerical Analysis, University of Texas at Austin, TX, 1990.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 278–288, January 1995

019

PERTURBATIONS, SINGULAR VALUES, AND RANKS OF PARTIAL TRIANGULAR MATRICES*

LEIBA RODMAN† AND HUGO J. WOERDEMAN†

Abstract. Questions regarding minimal representations of perturbations of discrete systems lead to the study of perturbations of lower triangular partial matrices and their minimal rank completions. Distance to the set of lower triangular partial matrices having minimal ranks smaller than a given integer is given in terms of (suitably generalized) singular numbers. Minimal ranks of lower triangular partial matrices in an arbitrary small neighborhood of a given lower triangular partial matrix are identified. The results are applied to minimal representations of discrete systems.

Key words. minimal representation, discrete time system, partial matrices, singular values, minimal rank

AMS subject classifications. 15A18, 15A99

1. Introduction.

Consider the finite horizon discrete system θ given by

$$x_{k+1} = A_k x_k + B_k u_k, \quad x_0 = 0,\ k = 0, 1, \ldots, n-1;$$
$$y_{k+1} = C_k x_{k+1}.$$

Here $\{u_k\}_{k=0}^{n-1}$ is a sequence of vector inputs, $\{x_k\}_{k=0}^{n}$ is the sequence of vector states in the state space $\mathbf{R}^p$, and $\{y_k\}_{k=1}^{n}$ is the sequence of vector outputs; A_k, B_k, C_k are matrices of appropriate sizes. The dimension p of the state space is called the *order* of the system θ. In [W] the question was addressed: What is the smallest order among all systems with the same input-output behaviour as θ? In this paper we address the question whether the order may be lowered even further in case one allows a slight perturbation of the input-output behaviour. This leads to the study of minimal ranks of perturbations of lower triangular partial matrices.

Recall that a *lower triangular partial matrix* is defined to be a block matrix of the form

$$\mathcal{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ \vdots & \vdots & & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix}, \tag{1.1}$$

where for $i \geq j$, A_{ij} is a $p_i \times m_j$ matrix with entries in F (F is either the field of real numbers or the field of complex numbers), and the entries of the blocks A_{ij} $(i < j)$ are independent free variables that take values in F. The blocks A_{ij} $(i \geq j)$ are thought of as given, or specified, blocks, while the blocks A_{ij} $(i < j)$ are considered unspecified and will be often designated by question marks. Note that the set of lower triangular partial matrices of a given size forms a vector space. A block matrix $B = [B_{ij}]_{i,j=1}^{n}$, where B_{ij} is a $p_i \times m_j$ matrix with entries in F, is called a *completion* of the lower triangular partial matrix (1.1) if $B_{ij} = A_{ij}$ for $i \geq j$ (in other words, B is obtained from A by specifying the unspecified entries in A). The lowest possible rank of completion of A is called the *minimal rank* of A and will be denoted $mr(A)$.

* Received by the editors October 26, 1992; accepted for publication (in revised form) by R. A. Horn, August 29, 1993.

† Department of Mathematics, College of William and Mary, Williamsburg, Virginia 23187-8795 (lxrodm@cs.wm.edu, hugo@cs.wm.edu). The research of the first author was partially supported by National Science Foundation grant DMS 91-23841. The research of the second author was partially supported by National Aeronautics and Space Administration contract NAS1-18347.

In this paper we start with the study of the behaviour of $mr(\mathcal{A})$ under perturbations of $\mathcal{A}$. One of the main results is Theorem 2.1, which gives a formula for the distance from a given lower triangular partial matrix to the set of lower triangular partial matrices having prescribed minimal rank. This result is obtained using the spectral norm (maximal singular value). In §3 we show that for a large class of unitarily invariant norms (different from multiples of the spectral norm) the formula given by Theorem 2.1 does not work. This formula is related to the celebrated Arveson distance formula (see [A]). In §4 we describe all minimal ranks of lower triangular partial matrices that are arbitrarily close to a given one. Finally, we return in the last section to the application of the results to the problem of minimal representation of perturbed discrete systems.

Throughout the paper, we denote by $s_k(X)$, $k = 1, 2, \ldots$ the singular values of the matrix X arranged in nonincreasing order.

2. Distance to the closest partial triangular matrix with prescribed minimal rank. Let

$$\mathcal{A} = \begin{bmatrix} A_{11} & ? & \cdots & ? \\ A_{21} & A_{22} & \cdots & ? \\ \vdots & \vdots & & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix} \tag{2.1}$$

be a lower triangular partial matrix as in (1.1). We define the norm $\|\mathcal{A}\|_p$ as follows (the subscript p stands for partial):

$$\|\mathcal{A}\|_p := \max \left\{ \left\| \begin{bmatrix} A_{i1} & \cdots & A_{ii} \\ \vdots & & \vdots \\ A_{n1} & \cdots & A_{ni} \end{bmatrix} \right\|, \ i = 1, \ldots, n \right\}. \tag{2.2}$$

Here $\|B\|$ denotes the largest singular value of the matrix B. It is easy to check that $\|\cdot\|_p$ is a norm for lower triangular partial matrices. We call a lower triangular partial matrix $\mathcal{B}$ an *ϵ-perturbation* of $\mathcal{A}$ if $\|\mathcal{B} - \mathcal{A}\|_p \leq \epsilon$. Let

$$\beta_k(\mathcal{A}) = \inf\{\epsilon > 0 | \text{ there exists an } \epsilon - \text{perturbation } \mathcal{B} \text{ of } \mathcal{A} \text{ with } mr(\mathcal{B}) < k\}.$$

We shall prove the following result.

THEOREM 2.1. *Let $\mathcal{A}$ be a lower triangular partial matrix. Then for all k,*

$$\beta_k(\mathcal{A}) = \max_{i=1,\ldots,n} s_k(A^{(i)}), \tag{2.3}$$

where

$$A^{(i)} = \begin{bmatrix} A_{i1} & \cdots & A_{ii} \\ \vdots & & \vdots \\ A_{n1} & \cdots & A_{ni} \end{bmatrix}, \qquad i = 1, \ldots, n.$$

Theorem 2.1 can be regarded as an extension to the lower triangular partial matrices framework of the well-known fact that $s_k(X)$ coincides with the distance from X to the set of matrices having rank less than k.

The proof is based on the following result obtained in [GSRW].

PROPOSITION 2.2. *For $k = 1, 2, \ldots$, we have*

$$\inf s_k(A_c) = \max_{i=1,\ldots,n} s_k(A^{(i)}),$$

where the infimum is taken over all completions A_c of $\mathcal{A}$.

Proof of Theorem 2.1. Let $\epsilon > 0$. By Proposition 2.2 there exists a completion A of $\mathcal{A}$ such that

$$s_k(A) < \frac{\epsilon}{2} + \max_{i=1,\ldots,n} s_k(A^{(i)}).$$

Choose $B = (B_{ij})_{i,j=1}^n$ with rank k so that

$$\|A - B\| < s_k(A) + \frac{\epsilon}{2}.$$

Let $\mathcal{B}$ denote the lower triangular partial matrix

$$\mathcal{B} = \begin{bmatrix} B_{11} & ? \cdots & ? \\ \vdots & \ddots & \vdots \\ & & ? \\ B_{n1} & \cdots & B_{nn} \end{bmatrix}.$$

Then $mr(\mathcal{B}) < k$ and

$$\|\mathcal{A} - \mathcal{B}\|_p \leq \|A - B\| < \epsilon + \max_{i=1,\ldots,n} s_k(A^{(i)}).$$

This proves the inequality $\leq$ in (2.3).

Conversely, let $\mathcal{B}$ be a lower triangular partial matrix with $mr(\mathcal{B}) < k$. Then, using Weyl's inequality (note that $s_k^2(C) =$ the kth eigenvalue of C^*C), we obtain

$$\begin{aligned} s_k(A^{(i)}) &= s_k(B^{(i)} + (A^{(i)} - B^{(i)})) \\ &\leq s_k(B^{(i)}) + s_1(A^{(i)} - B^{(i)}) \\ &= 0 + s_1(A^{(i)} - B^{(i)}) \leq \|\mathcal{A} - \mathcal{B}\|_p. \end{aligned}$$

Taking the maximum over $i = 1, \ldots, n$, we obtain

$$\max_{i=1,\ldots,n} s_k(A^{(i)}) \leq \|\mathcal{A} - \mathcal{B}\|_p.$$

Since $\mathcal{B}$ was arbitrary with $mr(\mathcal{B}) < k$, we consequently obtain the desired inequality $\geq$ in (2.3). □

In view of Theorem 2.1 and Proposition 2.2, the numbers $\beta_k(\mathcal{A})$ can be alternatively described by

$$\beta_k(\mathcal{A}) = \inf s_k(A_c), \tag{2.4}$$

where the infimum is taken over all completions A_c of $\mathcal{A}$. The equality (2.4) can be interpreted as follows. For a given norm $|||\cdot|||$ on the vector space V of lower triangular partial matrices $\mathcal{A} = [A_{ij}]_{i \geq j}$, where the size of A_{ij} is $p_i \times m_j$, define the s-numbers $s_k(\mathcal{A}; |||\cdot|||)$ as the distance (in the norm $|||\cdot|||$) from $\mathcal{A} \in V$ to the set of lower triangular partial matrices having minimal rank less than k. In this notation,

$$\beta_k(\mathcal{A}) = s_k(\mathcal{A}; \|\cdot\|_p),$$

where the norm $\|\cdot\|_p$ is defined by (2.2). Another natural norm on V is

$$\|\mathcal{A}\|_c = \min\{\|A_c\| : A_c \text{ is a completion of } \mathcal{A}\}$$

(again, $\|B\|$ stands for the maximal singular value of B). It turns out that

$$s_k(\mathcal{A}; \|\cdot\|_p) = s_k(\mathcal{A}; \|\cdot\|_c) \qquad (k = 1, 2, \ldots).$$

Indeed, in view of (2.4) we need only to verify that

$$s_k(\mathcal{A}; \|\cdot\|_c) = \inf\{s_k(A_c) : A_c \text{ is a completion of } \mathcal{A}\}.$$

But this equality follows easily upon unraveling the definition of $s_k(\mathcal{A}; \|\cdot\|_c)$.

3. Other norms. In this section we consider general unitarily invariant norms, not just the spectral norm, as in the previous section. The main result here is that Theorem 2.1 is not valid for a large class of unitarily invariant norms.

Recall that a norm $|||\cdot|||$ on the set of $m \times n$ matrices is called *unitarily invariant* if $|||X||| = |||UXV|||$ for any $m \times n$ matrix X and any unitary matrices U and V. It is well known ([SS, §II.3], [GK, Chap. III]) that any unitarily invariant norm is given by a symmetric gauge function $\Phi(\alpha_1, \ldots, \alpha_k)$, $k = \min(m, n)$:

$$|||X||| = \Phi(s_1(X), \ldots, s_k(X)).$$

Since we must work with unitarily invariant norms of matrices of various sizes, it will be convenient to assume that

$$|||X||| = \Phi(s_1(X), s_2(X), \ldots, s_n(X), \ldots), \tag{3.1}$$

where $\Phi(\alpha_1, \alpha_2, \ldots, \alpha_n, \ldots)$ is a symmetric gauge function defined on the set of non-increasing sequences $\{\alpha_i\}_{i=1}^{\infty}$ of nonnegative numbers such that only finitely many of the α_i's are different from 0. By convention, $s_k(X) = 0$ if $k >$ rank X.

From now on, we assume that the norm $|||\cdot|||$ is given by (3.1).

Given a matrix norm $|||\cdot|||$, we define for a matrix A:

$$s_{|||\cdot|||,k}(A) = \min\{|||A - B||| : \text{rank } B < k\},$$

$k = 1, 2, \ldots$. For a lower triangular partial matrix $\mathcal{A}$ as in (2.1), we let

$$|||\mathcal{A}||| = \max_{i=1,\ldots,n} |||A^{(i)}|||, \tag{3.2}$$

where

$$A^{(i)} = \begin{bmatrix} A_{i1} & \cdots & A_{ii} \\ \vdots & & \vdots \\ A_{n1} & \cdots & A_{ni} \end{bmatrix}.$$

Define also the s-numbers as in the previous section with respect to the norm (3.2):

$$s_k(\mathcal{A}; |||\cdot|||) = \inf\{\epsilon > 0 \mid \text{there is a lower triangular partial matrix with } mr(\mathcal{B}) < k \text{ and } |||\mathcal{A} - \mathcal{B}||| \leq \epsilon\}$$

for $k = 1, 2, \ldots$.

PROPOSITION 3.1. *For any lower triangular partial matrix $\mathcal{A}$, we have*

$$s_1(\mathcal{A}; |||\cdot|||) = \max_{i=1,\dots,n} s_{|||\cdot|||,1}(A^{(i)}) \tag{3.3}$$

and

$$s_k(\mathcal{A}; |||\cdot|||) \geq \max_{i=1,\dots,n} s_{|||\cdot|||,k}(A^{(i)}) \quad \text{for } k > 1. \tag{3.4}$$

Proof. The equality (3.3) is trivial since zero is the only lower triangular partial matrix with minimal rank less than one. To prove (3.4), analyzing the proof of the inequality $\geq$ in (2.3), it is seen that we only need

$$s_{|||\cdot|||,k}(A+B) \leq s_{|||\cdot|||,k}(A) + s_{|||\cdot|||,1}(B) \tag{3.5}$$

for all matrices A and B having the same size. In turn, (3.5) easily follows from the triangle inequality. (If $\hat{A}$ has rank $< k$ and $|||A-\hat{A}||| < s_{|||A-\hat{A}|||,k}(A)+\epsilon$, then we have that $\hat{A}+0$ is a matrix of rank $< k$ at most with distance $\epsilon + s_{|||\cdot|||,k}(A) + s_{|||\cdot|||,1}(B)$ from $A+B$.) □

It turns out that for a large class of norms of the form (3.1) that are not multiples of the spectral norm, the equality in (3.4) cannot be guaranteed (in contrast with Theorem 2.1). To state this result precisely, we introduce the following definition. A norm $|||\cdot|||$ given by (3.1) will be called *regular* (or, more exactly, *q-regular*) if there exists an integer $q \geq 2$ with the property that for every pair of sequences $\alpha_1 \geq \cdots \geq \alpha_q \geq 0$, $\beta_1 \geq \cdots \geq \beta_q \geq 0$ such that $\alpha_j \geq \beta_j$ $(j = 1, \dots, q)$ and $\alpha_j > \beta_j$ for at least one j $(1 \leq j \leq q)$, the inequality $\Phi(\alpha_1, \dots, \alpha_q, 0, 0, \dots) > \Phi(\beta_1, \dots, \beta_q, 0, 0, \dots)$ is valid. Many commonly used norms are regular, e.g.,

$$\Phi(\alpha_1, \dots, \alpha_m, 0, 0, \dots) = \left(\sum_{j=1}^{k} \alpha_j^p \right)^{1/p},$$

$p \geq 1$, $k \geq 2$ fixed, but the spectral norm is not. An example of a nonregular nonspectral norm that coincides with the spectral norm on the set of matrices of rank not exceeding m is given by

$$\Phi(\alpha_1, \dots, \alpha_m, 0, \dots) = \max \left(\alpha_1, (m-1)^{-1} \sum_{j=1}^{m} \alpha_j \right).$$

(This example was communicated to us by C.-K. Li [L].)

THEOREM 3.2. *Assume that $|||\cdot|||$ is a q-regular norm. Then there exists a lower triangular partial matrix $\mathcal{A} = [A_{ij}]_{i,j=1}^{n}$ such that*

$$s_q(\mathcal{A}; |||\cdot|||) > \max_{i=1,\dots,n} s_{|||\cdot|||,q}(A^{(i)})). \tag{3.6}$$

The proof is based on the following lemma.

LEMMA 3.3. *Let* $|||\cdot|||$ *be as in Theorem* 3.2. *Then for every pair of matrices* Y, Z *of sizes* $q \times q$ *and* $m \times q$, *respectively, such that* $Z \neq 0$, *the inequality*

$$|||\begin{bmatrix} Y \\ Z \end{bmatrix}||| > |||Y||| \tag{3.7}$$

holds.

Proof. Let $\tau_1^2 \geq \tau_2^2 \geq \cdots$ and $\sigma_1^2 \geq \sigma_2^2 \geq \cdots$ be the eigenvalues of $\begin{bmatrix} Y \\ Z \end{bmatrix}^* \begin{bmatrix} Y \\ Z \end{bmatrix}$ and Y^*Y, respectively (we assume $\tau_j = \sigma_j = 0$ for $j > q$). Clearly, $\tau_j \geq \sigma_j$ $(j = 1, \ldots, q)$ and $\sum_{j=1}^{q} \tau_j^2 > \sum_{j=1}^{q} \sigma_j^2$ (by comparing the traces of $\begin{bmatrix} Y \\ Z \end{bmatrix}^* \begin{bmatrix} Y \\ Z \end{bmatrix}$ and Y^*Y). Thus, $\tau_j > \sigma_j$ for some j, and since $|||\cdot|||$ is q-regular, the inequality (3.7) follows. □

Proof of Theorem 3.2. Without loss of generality we assume $\Phi(1, 0, 0, \ldots) = 1$. Let

$$\mathcal{A} = \begin{bmatrix} a & 1 & ? \\ I_{q-1} & 0 & a^* \\ 0 & 0 & 1 \end{bmatrix}, \tag{3.8}$$

where a is a $1 \times (q-1)$ row of the form $a = [0 \cdots 0x]$, $|x| > 1$. Thus,

$$A^{(1)} = \begin{bmatrix} a & 1 \\ I_{q-1} & 0 \\ 0 & 0 \end{bmatrix}, \qquad A^{(2)} = \begin{bmatrix} I_{q-1} & 0 & a^* \\ 0 & 0 & 1 \end{bmatrix}. \tag{3.9}$$

Let Ω be the set of all matrices

$$S = \begin{bmatrix} s_1 & s_2 \\ s_3 & s_4 \\ s_5 & s_6 \end{bmatrix}, \tag{3.10}$$

such that rank $S < q$ and

$$|||A^{(1)} - S||| = s_{|||\cdot|||,q}(A^{(1)}).$$

(The partition in (3.10) is consistent with the partition of $\mathcal{A}$ in (3.8).) By Lemma 3.3, $s_5 = 0$, $s_6 = 0$ for every $S \in \Omega$. Also, $s_4 \neq 0$. Indeed, for every matrix

$$T = \begin{bmatrix} t_1 & t_2 \\ t_3 & 0 \\ 0 & 0 \end{bmatrix}$$

(where t_3 is $(q-1) \times (q-1)$ and t_2 is 1×1) of rank $< q$, we have that at least one of t_3 or t_2 is singular. Say, t_3 is singular. Then $s_1(I_{q-1} - t_3) \geq 1$ and therefore $|||A^{(1)} - T||| \geq 1$. On the other hand, letting b be the $(q-1) \times 1$ column $[0 \cdots 0x^{-1}]^T$, we have

$$\operatorname{rank} \begin{bmatrix} a & 1 \\ I_{q-1} & b \end{bmatrix} = q - 1$$

and

$$|||\begin{bmatrix} a & 1 \\ I_{q-1} & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} a & 1 \\ I_{q-1} & b \\ 0 & 0 \end{bmatrix}||| < 1.$$

This verifies that $s_4 \neq 0$ for every $S \in \Omega$. Since Ω is closed and compact, we conclude that for every $\epsilon > 0$,

$$s_{|||\cdot|||,q}(A^{(1)}) < \min\left\{ |||A^{(1)} - T||| : T = \begin{bmatrix} t_1 & t_2 \\ t_3 & t_4 \\ t_5 & t_6 \end{bmatrix}, \text{ rank } T < q \text{ and at least one of the inequalities } \|t_4\| \leq \epsilon \text{ and } \|[t_5\ t_6]\| \geq \epsilon \text{ holds.} \right\}. \tag{3.11}$$

Analogous considerations for $A^{(2)}$ (note that $A^{(2)}$ is the adjoint of $A^{(1)}$ after certain permutations of columns and rows) lead to the following inequality for every $\epsilon > 0$:

$$s_{|||\cdot|||,q}(A^{(2)}) < \min\left\{ |||A^{(2)} - T||| : T = \begin{bmatrix} t_1 & t_2 & t_3 \\ t_4 & t_5 & t_6 \end{bmatrix}, \text{ rank } T < q \text{ and at least one of the inequalities } \|t_4\| \leq \epsilon \text{ and } \left\| \begin{bmatrix} t_5 \\ t_6 \end{bmatrix} \right\| \geq \epsilon \text{ holds.} \right\}. \tag{3.12}$$

(The partition of T in (3.12) is consistent with the partition of $A^{(2)}$ in (3.9).)

Let now

$$\mathcal{B}_m = \begin{bmatrix} b_1^{(m)} & b_2^{(m)} & ? \\ b_3^{(m)} & b_4^{(m)} & b_7^{(m)} \\ b_5^{(m)} & b_6^{(m)} & b_8^{(m)} \end{bmatrix}; \qquad m = 1, 2, \ldots,$$

be a sequence of lower triangular partial matrix having the same sizes of corresponding blocks as A and such that $mr(\mathcal{B}_m) < q$ and such that

$$|||\mathcal{A} - \mathcal{B}_m||| \to s_q(\mathcal{A}; |||\cdot|||)$$

as $m \to \infty$.

Passing to a subsequence, if necessary, we can assume that the limits $b_j = \lim_{m\to\infty} b_j^{(m)} (j = 1, \ldots, 8)$ exist. Clearly rank $B_m^{(1)} < q$ and rank $B_m^{(2)} < q$. If $b_5 \neq 0$ or $b_6 \neq 0$, then by (3.11),

$$s_q(\mathcal{A}; |||\cdot|||) > \max\{s_{|||\cdot|||,q} A^{(1)}, s_{|||\cdot|||,q}(A^{(2)})\}.$$

If $b_5 = 0$, $b_6 = 0$, then by (3.12) the same inequality follows. □

4. Minimal ranks of perturbed partial triangular matrices. In this section the norm under consideration is the spectral norm. Let $\mathcal{A}$ be a lower triangular partial matrix. Define the set

$$M(\mathcal{A}) = \{k : \text{ for every } \epsilon > 0 \text{ there exists an } \epsilon\text{-perturbation } \mathcal{A}_\epsilon \text{ of } \mathcal{A} \text{ such that } mr(\mathcal{A}_\epsilon) = k\}.$$

THEOREM 4.1. *We have*

$$M(\mathcal{A}) = \{\alpha, \alpha + 1, \ldots, \beta - 1, \beta\},$$

where

$$\alpha = \max \left\{ \operatorname{rank} \begin{bmatrix} A_{i1} & \cdots & A_{ii} \\ \vdots & & \vdots \\ A_{n1} & \cdots & A_{ni} \end{bmatrix} : 1 \leq i \leq n \right\}$$

and β is some integer.

The proof is based on the following lemma.

LEMMA 4.2. *Let $\mathcal{A}$ and $\mathcal{A}'$ be lower triangular partial matrices with the same pattern of specified entries and such that $\mathcal{A}$ and $\mathcal{A}'$ differ in only one specified entry. Then*

$$|mr(\mathcal{A}) - mr(\mathcal{A}')| \leq 1.$$

Proof. Let B' be a minimal rank completion of $\mathcal{A}'$, so that rank $B' = mr(\mathcal{A}')$. Let B be a completion of $\mathcal{A}$ that coincides with B' in the unspecified entries. Then B and B' differ in only one entry, and hence $|\text{rank}B - \text{rank}B'| \leq 1$. So

$$mr(\mathcal{A}) \leq \text{rank}B \leq \text{rank}B' + 1 = mr(\mathcal{A}') + 1.$$

Analogously, $mr(\mathcal{A}') \leq mr(\mathcal{A}) + 1$. □

Proof of Theorem 4.1. That $k \notin M(\mathcal{A})$ for every $k < \alpha$ follows from the definition of α and from the fact that a rank of a matrix can only become larger if the matrix is perturbed in a sufficiently small neighborhood. Furthermore, we have

$$s_{\alpha+1} \begin{bmatrix} A_{i1} & \cdots & A_{ii} \\ \vdots & & \vdots \\ A_{n1} & \cdots & A_{ni} \end{bmatrix} = 0 \quad \text{for } i = 1, \ldots, n,$$

and therefore by Theorem 2.1, $\beta_{\alpha+1}(\mathcal{A}) = 0$. In other words, for every $\epsilon > 0$ there exists an ϵ-perturbation $\mathcal{A}_\epsilon$ of $\mathcal{A}$ with $mr(\mathcal{A}_\epsilon) \leq \alpha$. Since for small $\epsilon > 0$ we obviously have $mr(\mathcal{A}_\epsilon) \geq \alpha$, it follows that $\alpha \in M(\mathcal{A})$.

Now let $\mathcal{B}_i$ be an ϵ-perturbation of $\mathcal{A}$ with $mr(\mathcal{B}_i) = k_i$, $i = 1, 2$. Say $k_1 \leq k_2$. There exists a continuous path $\mathcal{B}(t)$, $0 \leq t \leq 1$, of lower triangular partial matrices such that $\mathcal{B}(t)$ is an ϵ-perturbation of $\mathcal{A}$ for all t, and for some partition $0 < t_1 < t_2 < \cdots < t_p < 1$, we have that $\mathcal{B}(t)$ differs from $\mathcal{B}(s)$ in exactly one specified entry if $t_i \leq s < t \leq t_{i+1}$, for $i = 0, \ldots, p$ (by definition, $t_0 = 0$, $t_{p+1} = 1$). By Lemma 4.2, we have

$$|mr\mathcal{B}(t_i) - mr\mathcal{B}(t_{i+1})| \leq 1, \; i = 0, \ldots, p.$$

Therefore, for every integer k such that $k_1 \leq k \leq k_2$, there is at least one index i such that $mr\mathcal{B}(t_i) = k$. This completes the proof of Theorem 4.1. □

It is possible that for every $\epsilon > 0$ there exists a lower triangular partial matrix $\mathcal{B}$ in an ϵ-neighborhood of $\mathcal{A}$ such that $mr(\mathcal{B}) < mr(\mathcal{A})$ (in contrast with the well-known property of the rank of a matrix) as illustrated by the following simple example:

$$mr \begin{bmatrix} 1 & ? \\ 0 & 1 \end{bmatrix} = 2, \qquad mr \begin{bmatrix} 1 & ? \\ \epsilon & 1 \end{bmatrix} = 1 \quad (\epsilon \neq 0).$$

It is interesting to identify the integer β in Theorem 4.1. We have a lower bound.

PROPOSITION 4.3. *Let $\mathcal{A} = [A_{ij}]_{i,j=1}^{n}$ be a lower triangular partial matrix, where A_{ij} is $p_i \times m_j$. Then*

$$\beta \geq \sum_{i=1}^{n} \gamma_i - \sum_{i=1}^{n-1} \operatorname{rank} \begin{bmatrix} A_{i+1,1} & \cdots & A_{i+1,i} \\ \vdots & & \vdots \\ A_{n1} & \cdots & A_{ni} \end{bmatrix}, \tag{4.1}$$

where

$$\gamma_i = \min \left\{ m_i + \operatorname{rank} \begin{bmatrix} A_{i,1} & \cdots & A_{i,i-1} \\ \vdots & & \vdots \\ A_{n,1} & \cdots & A_{n,i-1} \end{bmatrix}, \; p_i + \operatorname{rank} \begin{bmatrix} A_{i+1,1} & \cdots & A_{i+1,i} \\ \vdots & & \vdots \\ A_{n,1} & \cdots & A_{n,i-1} \end{bmatrix} \right\}.$$

Proof. By [CJRW], the maximal rank of completions of the lower triangular partial matrix

$$\mathcal{A}_i = \begin{bmatrix} A_{i1} & \cdots & A_{i,i-1} & ? \\ A_{i+1,1} & \cdots & A_{i+1,i-1} & A_{i+1,i} \\ \vdots & & \vdots & \\ A_{n1} & \cdots & A_{n,i-1} & A_{n,i} \end{bmatrix}$$

is equal to γ_i. It is easy to see that the set of $p_i \times m_i$ matrices X for which

$$\operatorname{rank} \begin{bmatrix} A_{i1} & \cdots & A_{i,i-1} & X \\ A_{i+1,1} & \cdots & A_{i+1,i-1} & A_{i+1,i} \\ \vdots & & \vdots & \\ A_{n1} & \cdots & A_{n,i-1} & A_{n,i} \end{bmatrix} = \gamma_i \tag{4.2}$$

is dense. Therefore, for every $\epsilon > 0$ there exists $\mathcal{B} = [B_{ij}]$ in the ϵ-neighborhood of $\mathcal{A}$ such that $B_{ij} = A_{ij}$ if $i > j$ and (4.2) holds with $X = B_{ii}$. Now the formula for the minimal rank of lower triangular partial matrices [W] shows that $mr(\mathcal{B})$ coincides with the right-hand side of (4.1). □

We conjecture that equality holds in (4.1).

5. Application to minimal representation of discrete systems. Consider the finite horizon discrete system

$$\begin{aligned} x_{k+1} &= A_k x_k + B_k u_k, \qquad x_0 = 0, \; k = 0, 1, \ldots, n-1; \\ y_{k+1} &= C_k x_{k+1}, \end{aligned}$$

with $\{u_k\}_{k=0}^{n-1}$ the sequence of vector inputs in $\mathbf{R}^n$, $\{x_k\}_{k=0}^{n}$ the sequence of vector states in the state space $\mathbf{R}^p$, and $\{y_k\}_{k=1}^{n}$ the sequence of vector outputs $\mathbf{R}^m$. As stated in the introduction, the dimension p of the state space is called the *order* of the system. The input-output map is given by

(5.1)

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} C_0B_0 & 0 & 0 & 0 \cdots & 0 \\ C_1A_1B_0 & C_1B_1 & 0 & 0 \cdots & 0 \\ C_2A_2A_1B_0 & C_2A_2B_1 & C_2B_2 & 0 \cdots & 0 \\ \vdots & \vdots & & & \vdots \\ C_{n-1}A_{n-1}\cdots A_1B_0 & C_{n-1}A_{n-1}\cdots A_2B_1 & & \cdots & C_{n-1}B_{n-1} \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_{n-1} \end{bmatrix}.$$

With any matrix X whose lower triangular part is equal to

$$\mathcal{A} = \begin{bmatrix} C_0B_0 & & & \\ C_1A_1B_0 & C_1B_1 & & \\ \vdots & \vdots & \ddots & \\ C_{n-1}A_{n-1}\cdots A_1B_0 & C_{n-1}A_{n-1}\cdots A_2B_1 & \cdots & C_{n-1}B_{n-1} \end{bmatrix}, \tag{5.2}$$

we can associate a system with the same input-output behavior as the original system, as follows. Write $q = \operatorname{rank} X$, and make a rank decomposition

$$X = \begin{bmatrix} F_0 \\ \vdots \\ F_{n-1} \end{bmatrix} \begin{bmatrix} G_0 & \dots & G_{n-1} \end{bmatrix}$$

of X, where F_i and G_i are of sizes $m \times q$ and $q \times n$, respectively. Then the system

$$\begin{aligned} x_{k+1} &= x_k + G_k u_k,\ x_0 = 0,\ k = 0, \dots, n-1, \\ y_{k+1} &= F_k x_{k+1}, \end{aligned}$$

has the same input-output behaviour as the original system (for the continuous analog of this argument see Proposition 4.1 in [GK2]). This leads to a problem of choosing a completion X of (5.2) of minimal possible rank. The references [KW], [GK1], and [GK2] give a more complete background on system-theoretic applications of the minimal rank completion problem. If the original system is allowed to be slightly perturbed, then it is exactly the problem solved in previous sections. Thus, the result of Theorem 4.1 (concerning α) can be applied to such systems.

THEOREM 5.1. *Given the finite horizon discrete system*

$$\theta_1 = \begin{cases} x_{k+1} = A_k x_k + B_k u_k,\ x_0 = 0,\ k = 0, \dots, n-1, \\ y_{k+1} = C_k x_{k+1}. \end{cases}$$

Let

$$\alpha = \max_{0 \le i \le n-1} \left\{ \operatorname{rank} \begin{bmatrix} C_iA_i\cdots A_1B_0 & \cdots & C_iB_i \\ \vdots & & \vdots \\ C_{n-1}A_{n-1}\cdots A_1B_0 & \cdots & C_{n-1}A_{n-1}\cdots A_{n-1-i}B_i \end{bmatrix} \right\}.$$

Then for every $\epsilon > 0$, there exists a finite horizon discrete system

$$\theta_2 = \begin{cases} x'_{k+1} = A'_k x'_k + B'_k u_k, \quad x'_0 = 0, \quad k = 0, \dots, n-1, \\ y'_{k+1} = C'_k x'_{k+1}, \end{cases}$$

of order α such that when θ_1 and θ_2 are fed the same inputs $\{u_1, \dots, u_{n-1}\}$, their outputs $\{y_1, \dots, y_n\}$ and $\{y'_1, \dots, y'_n\}$, respectively, satisfy

$$\|y_i - y'_i\| \le \epsilon \left\| \begin{bmatrix} u_0 \\ \vdots \\ u_{i-1} \end{bmatrix} \right\|, \qquad i = 1, \dots, n-1.$$

Moreover, α is the smallest number with this property.

Proof. Let $\epsilon > 0$. According to Theorem 4.1 there exists a lower triangular partial matrix $\mathcal{D} = [D_{ij}]_{1 \le i \le j \le n}$ such that $mr(\mathcal{D}) = \alpha$ and

$$\|\mathcal{D} - \mathcal{A}\|_p < \epsilon,$$

where $\mathcal{A}$ is defined in (5.2). Using the reasoning given before the theorem, we know that there exists a finite horizon discrete system θ_2 of order α whose input/output map is given by

$$\begin{bmatrix} D_{11} & 0 & \cdots & 0 \\ D_{21} & D_{22} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ D_{n1} & D_{n2} & \cdots & D_{nn} \end{bmatrix}.$$

Comparing now the outputs $\{y_1, \ldots, y_n\}$ and $\{y_1', \ldots, y_n'\}$ of the systems θ_1 and θ_2, respectively, both with inputs $\{u_0, \ldots, u_{n-1}\}$, we obtain

$$y_i - y_i' = [C_{i-1}A_{i-1} \cdots A_1 B_0 \ \cdots \ C_{i-1}B_{i-1}] \begin{bmatrix} u_0 \\ \vdots \\ u_{i-1} \end{bmatrix} - [D_{i1} \cdots D_{ii}] \begin{bmatrix} u_0 \\ \vdots \\ u_{i-1} \end{bmatrix}.$$

But then, since $\|\mathcal{A} - \mathcal{D}\|_p < \epsilon$, we obtain

$$\|y_i - y_i'\| \le \epsilon \left\| \begin{bmatrix} u_0 \\ \vdots \\ u_{i-1} \end{bmatrix} \right\|, \qquad i = 1, \ldots, n.$$

The final statement in the theorem follows immediately from the fact that α is the smallest number in $M(\mathcal{A})$ (see Theorem 4.1). □

REFERENCES

[A] W. Arveson, *Interpolation in nest algebras*, J. Funct. Anal., 3 (1975), pp. 208–233.

[CJRW] N. Cohen, C. R. Johnson, L. Rodman, and H. J. Woerdeman, *Ranks of completion of partial matrices*, H. Dym, S. Goldberg, M. A. Kaashoek, P. Lancaster, eds., Operator Theory: Advances and Applications, 40 (1989), pp. 165–185.

[GK1] I. Gohberg and M. A. Kaashoek, *Time varying linear systems with boundary conditions and integral operators* I, *The transfer operator and its properties*, Integral Equations Operator Theory, 7 (1984), pp. 325–391.

[GK2] ———, *Minimal representations of semiseparable kernels and systems with separable boundary conditions*, J. Math. Anal. Appl., 124 (1987), pp. 436–458.

[GK] I. Gohberg and M. G. Krein, *Introduction to the theory of linear nonselfadjoint operators*, Trans. Math. Monographs, Vol. 18, Amer. Math. Soc., Providence, RI, 1969.

[GRSW] I. Gohberg, L. Rodman, T. Shalom, and H. J. Woerdeman, *Singular values of completions of partial block triangular matrices*, Linear Multilinear Algebra, 33 (1992), pp. 233–250.

[KW] M. A. Kaashoek and H. J. Woerdeman, *Unique minimal rank extensions of triangular operators*, J. Math. Anal. Appl., 131 (1988), pp. 501–516.

[L] C.-K. Li, private communication, 1992.

[SS] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.

[W] H. J. Woerdeman, *The lower order of lower triangular operators and minimal rank extensions*, Integral Equations Operator Theory, 10 (1987), pp. 859–879.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 289–306, January 1995

020

EXTERNAL DESCRIPTIONS AND STAIRCASE FORMS IN IMPLICIT SYSTEMS*

VASSILIS SYRMOS†, PETR ZAGALAK‡, AND VLADIMIR KUČERA‡

Abstract. This paper studies the relationship between staircase forms and normal external descriptions in implicit systems. The authors show how to compute the proper and nonproper controllability indices of an implicit system using the reachability Hessenberg form and the corresponding normal external description. The normal external description of the system is computed using embedding techniques. Finally, the differences and similarities between the normal external descriptions computed are shown using the reachability and controllability Hessenberg forms of an implicit system.

Key words. normal external description, staircase form, unimodular matrix

AMS subject classifications. 15A22, 65F30, 93B10, 93B20

1. Introduction.

The use of input-output descriptions in system theory has a long history. Kučera in [5] extensively studied structural properties of linear systems using the polynomial approach as a tool. Furthermore, he showed [5], [6] how to formulate design problems in terms of Diophantine equations, the solutions to which characterize feedback controllers. Although, this body of work focused on the detailed structure and solution of Diophantine equations, the computational aspects of the problem were not addressed.

The results proposed in this paper tackle the following problem. Given an implicit realization $\{E, A, B\}$ how can we compute in a computationally efficient manner, the proper and nonproper controllability indices [7] of such a realization from the stairs of the generalized Hessenberg form. Although computing the reachability and controllability indices of such systems using the Hessenberg form has been studied by van Dooren [17], [18], there is no information on how to compute the proper and nonproper controllability indices. To connect the length of the stairs to these indices we first compute a minimal polynomial basis for the controllability pencil using the Hessenberg form. If this minimal basis also satisfies some further properties, it will be called a normal external description of the system (cf. §2). The computation of the normal external description is motivated by the fact that one can read out the proper and nonproper controllability indices. As a result, once a relation between the Hessenberg form and the computed normal external description is established, we are able to directly use the Hessenberg form to compute and distinguish between the proper and nonproper controllability indices. To further motivate these results we draw some attention to the pole placement problem in implicit systems. The most detailed and unified approach to date is given in [20]. The results in [20] exploit this fine distinction between the controllability indices to state Rosenbrock's theorem for implicit systems. The computation of such state-feedback controllers is based strictly on the polynomial approach and the solution of a Diophantine equation involving normal external descriptions. Therefore, our results can be viewed as a first step in

* Received by the editors November 19, 1992; accepted for publication (in revised form) by P. van Dooren October 6, 1993. This research was supported by National Science Foundation contract NCR-9210408 and by University of Hawaii Research Council Funds.

† Department of Electrical Engineering, University of Hawaii at Manoa, 2540 Dole Street, Holmes Hall, Honolulu, Hawaii 96822 (hellas@euclid.eng.hawaii.edu).

‡ Institute of Information Theory and Automation, Czech Academy of Sciences, P. O. Box 18, 182 08 Prague 8, Czech Republic (kucera@utia.cas.cz).

solving the same problem from a computational point of view. Another application of the normal external description is the computation of coprime matrix fraction descriptions in implicit systems, which can be generalized using the results of this paper and also by extending the algorithm in [13]. These applications are not discussed in this paper since such results are still far from complete.

There are two important tools in this problem; one is the staircase (reachability, controllability) form of the implicit system $\{E, A, B\}$ and the other is the use of embedding techniques. In particular, we embed the controllability pencil in a square unimodular matrix, which we invert. The proposed embedding technique ensures the minimality of the controllability and reachability chains of the system, and therefore the feedback invariants are preserved under the embedding. In addition, the structure of the computed normal external description shows how to compute the proper and nonproper controllability indices of the system using the length of the stairs of the reachability Hessenberg form.

Finally, we show the differences and similarities of the controllability and reachability Hessenberg forms in computing a normal external description. All these ideas are clarified in a simple example.

2. Preliminaries and basic concepts. We shall consider a linear, time invariant system described by

$$E\dot{x} = Ax + Bu, \tag{2.1}$$

where E, A are $n \times n$ matrices, where E is generally singular and B is an $n \times m$ matrix over $\mathbb{R}$, the field of real numbers. Without any loss of generality, assume that rank $B = m$. Furthermore, we assume that the system is regular, i.e., $\det(sE - A) \not\equiv 0$.

We first recall the definitions of reachability, controllability given in [7]–[9], [11], [16], [19].

DEFINITION 2.1 (controllability; see [19]). *The system* (2.1) *is said to be controllable if the pencil* $[sE - A \;\; -B]$ *has neither finite nor infinite* (*except those of order* 1) *elementary divisors.*

Here, we use the concept of controllability in the sense of Verghese [19].

DEFINITION 2.2 (reachability; see [16]). *The system* (2.1) *is said to be reachable if it is controllable and* rank $[E \;\; B] = n$.

We use these definitions of controllability and reachability to connect the proper and nonproper controllability indices with the block Hessenberg form. For more details concerning these definitions, see [8] and references therein. It easily follows that reachability implies controllability, but not vice versa.

The description of system (2.1) we have used up to now is very often called the description in an internal form. On the other hand, the relationship

$$[sE - A \;\; -B] \begin{bmatrix} N(s) \\ D(s) \end{bmatrix} = 0$$

shows that the vector

$$\begin{bmatrix} N(s) \\ D(s) \end{bmatrix},$$

where $N(s)$, $D(s)$ are polynomial matrices, reflects the input-output or external behavior of (2.1). Hence, under the assumption of the controllability of (2.1), a basis of $\ker[sE - A, -B]$ serves as a generator of all possible input-output vectors $\left[\begin{smallmatrix} N(s) \\ D(s) \end{smallmatrix}\right]$ or,

in other words, such a basis describes the system as well as the description (2.1). We refer to such a basis as an external description of (2.1).

DEFINITION 2.3 (normal external description). *Given E, A, B in (2.1), then the polynomial matrices $N(s)$ and $D(s)$ of respective sizes $n \times m$ and $m \times m$ form a normal external description if the following is true.*

(i) $\left[\begin{smallmatrix} N(s) \\ D(s) \end{smallmatrix}\right]$ *is a minimal polynomial basis of* $\operatorname{Ker}[sE - A, -B]$;

(ii) $N(s)$ *is a minimal polynomial basis of* $\operatorname{Ker} P(sE - A)$ *where P is a maximal annihilator of B;*

(iii) $\left[\begin{smallmatrix} N(s) \\ D(s) \end{smallmatrix}\right]$ *is nonincreasingly column-degree ordered, i.e., $c_1 \geq c_2 \geq \cdots \geq c_m$, where c_i stands for the degree of the column i of $\left[\begin{smallmatrix} N(s) \\ D(s) \end{smallmatrix}\right]$.*

Remark 2.1. A normal external description of (2.1) is not unique. If $N(s), D(s)$ and $N'(s), D'(s)$ form normal external descriptions of (2.1), then they are related by

$$\begin{bmatrix} N(s) \\ D(s) \end{bmatrix} = \begin{bmatrix} N'(s) \\ D'(s) \end{bmatrix} U(s),$$

where $U(s)$ is a unimodular matrix.

A normal external description can now be used for a definition of controllability and reachability indices. We define

$$k_i = \deg_{c_i} N(s)$$

and

$$c_i = \deg_{c_i} \begin{bmatrix} N(s) \\ D(s) \end{bmatrix}$$

for $i = 1, 2, \ldots, m$ where $\deg_{c_i}(.)$ denotes the degree of the ith column.

DEFINITION 2.4. *The integers $r_i = 1 + k_i, i = 1, 2, \ldots, m$ are said to be the reachability indices of* (2.1), *while the integers $c_i, i = 1, 2, \ldots, m$ are said to be the controllability indices of* (2.1). *When $c_i > k_i$, then c_i is said to be a proper controllability index, otherwise it is said to be a nonproper one.*

There are many other ways to define the reachability and controllability indices and we refer the reader to [7], [9], and [11] for details. The one used here allows the connection between the proper, nonproper indices and the stairs of the Hessenberg form. Furthermore, the generalized version of Rosenbrock's theorem is based on these indices and therefore for state-feedback applications it is important to distinguish between them and compute them.

THEOREM 2.1. *The system* (2.1) *is reachable if and only if*

$$\sum_{i=1}^{m} r_i = n \tag{2.2}$$

and controllable if and only if

$$\sum_{i=1}^{m} c_i = \operatorname{rank} E. \tag{2.3}$$

Under the assumption of reachability (see Definition 2.2) there exist unitary transformations Q, Z such that [1], [12], [17], [18]

$$Q[sE - A \quad -B] \begin{bmatrix} Z & 0 \\ 0 & I_m \end{bmatrix} = [sE_r - A_r \quad -B_r],$$

where (E_r, A_r, B_r) are in lower generalized block Hessenberg form; that is

$$A_r = \begin{bmatrix} A_{1,1} & A_{1,2} & \dots & 0 \\ A_{2,1} & A_{2,2} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ A_{k-1,1} & \dots & A_{k-1,k-1} & A_{k-1,k} \\ A_{k1} & \dots & A_{k,k-1} & A_{k,k} \end{bmatrix}, \quad B_r = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ A_{k,k+1} \end{bmatrix}, \tag{2.4}$$

$$E_r = \begin{bmatrix} E_{1,1} & 0 & \dots & 0 \\ E_{2,1} & E_{2,2} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ E_{k-1,1} & \dots & E_{k-1,k-1} & 0 \\ E_{k1} & \dots & E_{k,k-1} & E_{k,k} \end{bmatrix}, \tag{2.5}$$

where [17] we have the following.

$A_{j,j+1}$ are of dimensions $t_j \times t_{j+1}$, $j \in \{1, 2, \dots, k\}$ and have full row rank t_j. Moreover, $A_{k,k+1}$ has full row rank t_k, but from our assumption that B is of full column rank ($t_k = t_{k+1} = m$) it follows that $A_{k,k+1}$ is a square nonsingular matrix of dimensions $m \times m$. This guarantees the absence of finite unreachable modes.

$E_{j,j}$, are of dimensions $t_j \times t_j$ $j \in \{1, 2, \dots, k-1\}$ and have full rank t_j. This guarantees the absence of infinite unreachable modes. Furthermore, $\dim \ker E_{kk} = \dim \ker E = \rho$.

The matrices $E_{j,j}$ are chosen *lower triangular*, that is $E_{j,j} = R_{j,r}$, where $R_{j,r}$ are lower triangular and the matrices $A_{j,j+1}$ are chosen *lower triangular in the bottom corner*, [12], i.e.,

$$A_{j,j+1} = [S_{1,j+1} \;\; 0],$$

where $S_{1,j+1}$ is a lower triangular matrix of dimensions $t_j \times t_j$. We put the matrices in $A_{j,j+1}$ in lower triangular form at the bottom corner for reasons that will be clarified in §3.

The reachability indices of (2.1) are computed as follows:

$$t_{i+1} - t_i \text{ reachability indices } r_j \text{ of order } k - i,$$

where $t_0 = 0$.

We mention here that the reachability Hessenberg form is obtained by first compressing the rows of B and second by performing the staircase algorithm in the pencil $M(sE - A) = s\bar{E} - \bar{A}$, where M is the maximal annihilator of B. This remark will be further explained when we study the controllability Hessenberg form.

In the presence of infinite elementary divisors, that is when there exist unreachable modes at infinity, there exist unitary transformations Q, Z such that [17], [18]

$$Q[sE - A \;\; -B] \begin{bmatrix} Z & 0 \\ 0 & I_m \end{bmatrix} = \begin{bmatrix} sE_\infty - A_\infty & 0 & 0 \\ * & sE_r - A_r & -B_r \end{bmatrix} = [sE_h - A_h \;\; -B_h],$$

where the pencil $[sE_r - A_r \;\; -B_r]$ has dimensions $o_r \times o_r + m$ and contains the reachable part of the system. In addition (E_r, A_r, B_r) are in the form described by (2.4) and

(2.5). The pencil $(sE_\infty - A_\infty)$ has dimensions $o_\infty \times o_\infty$, contains the structure of the unreachable part at infinity, and is of the form

$$\begin{bmatrix} -S_{1,\infty} & 0 & \cdots & 0 \\ sJ_{1,\infty} - N_{1,\infty} & -S_{2,\infty} & \cdots & 0 \\ * & sJ_{2,\infty} - N_{2,\infty} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ * & \cdots & -S_{l,\infty} & 0 \\ * & \cdots & sJ_{l,\infty} - N_{l,\infty} & -S_{l+1,\infty} \end{bmatrix}, \tag{2.6}$$

where [17] the following are true.

$J_{j,\infty}$ are of dimensions $s_{j+1} \times s_j$, $j \in \{1, 2, \ldots, l\}$ and have full column rank s_j.

$S_{j,\infty}$ are of dimensions $s_j \times s_j$ $j \in \{1, 2, \ldots, l\}$ and have full rank s_j.

The matrices $S_{j,\infty}$ are chosen *lower triangular* and the matrices $J_{j,\infty}$ are chosen *lower triangular in the top corner*, i.e.,

$$J_{j,\infty} = \begin{bmatrix} R_{j,\infty} \\ 0 \end{bmatrix},$$

where $R_{j,\infty}$ is a lower triangular matrix of dimensions $s_j \times s_j$.

The infinite elementary divisors are computed as follows:

$$s_{i+1} - s_i \text{ infinite elementary divisors } d_j \text{ of order } l - i + 1,$$

where $s_0 = 0$.

We denote $\dim \ker E_r$ as ρ_r and $\dim \ker E_\infty$ as ρ_∞.

To this end we propose the controllability Hessenberg form of the system (E, A, B). In this case the initial step is different from the one in the reachability form. Here, we treat the controllability pencil as an augmented nonsquare pencil, that is, $(s[E\ 0] - [A\ B])$. Applying the staircase algorithm proposed in [17], on the augmented pencil we get

$$Q\,[s[E\ 0] - [A\ B]]\,Z = [sE_c - A_c],$$

$$A_c = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & 0 & 0 \\ A_{2,1} & A_{2,2} & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ A_{k-1,1} & \cdots & A_{k-1,k-1} & A_{k-1,k} & 0 \\ A_{k1} & \cdots & A_{k,k-1} & A_{k,k} & A_{k,k+1} \end{bmatrix}, \tag{2.7}$$

$$E_c = \begin{bmatrix} E_{1,1} & 0 & \cdots & 0 & 0 \\ E_{2,1} & E_{2,2} & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ E_{k-1,1} & \cdots & E_{k-1,k-1} & 0 & 0 \\ E_{k1} & \cdots & E_{k,k-1} & E_{k,k} & 0 \end{bmatrix}, \tag{2.8}$$

where [17] the following hold.

$A_{j,j+1}$ are of dimensions $s_j \times t_{j+1}$, $j \in \{1, 2, \ldots, k\}$ and have full row rank s_j.

$E_{j,j}$, are of dimensions $s_j \times t_j$ $j \in \{1, 2, \ldots, k-1\}$ and have full column rank t_j.

The matrices

$$E_{jj} = \begin{bmatrix} R_{j,c} \\ 0 \end{bmatrix},$$

where $R_{j,c}$ are chosen *lower triangular* and the matrices $A_{j,j+1}$ are chosen *lower triangular in the bottom corner*, i.e.,

$$A_{j,j+1} = [S_{1,j+1} \;\; 0],$$

where $S_{1,j+1}$ is a lower triangular matrix of dimensions $s_j \times s_j$.

Notice that in the controllability form the initial compression is performed on the columns of $[E\ 0]$ as opposed to the reachability form where the initial compression is performed on the rows of B.

The following theorem is due to van Dooren [17] and we briefly review it here.

THEOREM 2.2. *Let (E, A, B) be a controllable system, then there are*

$t_{i+1} - s_i$ controllability indices c_j of order $k - i$ and

$s_{i+1} - t_{i+1}$ infinite elementary divisors d_j of order $k - i$,

where $s_0 = 0$ and t_i, s_i $i = 0, 1, \ldots, k-1$ are the dimensions of the stairs in the pencil $[sE_c - A_c]$.

In the rest of this section we show that, under the assumption of reachability, we can use the stairs of the reachability Hessenberg form to compute not only the controllability indices but also to distinguish whether they are *proper* or *nonproper*. Before showing how to do that, we present two ancillary results.

LEMMA 2.1. *Let (E, A, B) be reachable. Then the system (E, A, B) has only trivial chains at infinity or equivalently has only infinite elementary divisors of order one. Moreover, it has*

$$q = s_k - t_k$$

nonproper controllability indices that are equal to the number of trivial chains at infinity.

Proof. The proof follows easily from the condition rank $[E\ B] = n$ and Theorem 2.2. □

Since (E, A, B) has only trivial chains at infinity it follows from Theorem 2.2 that

$$s_i = t_i, \qquad i = 1, \ldots, k-1.$$

Furthermore, we assume that (E, A, B) has no column minimal indices of order zero (rank $B = m$). This assumption implies that

$$s_k = t_{k+1}.$$

Therefore, $q = t_{k+1} - t_k$.

LEMMA 2.2. *Let (E, A, B) be a reachable system, then*

$$t_k = m.$$

Proof. Since

$$\sum_{i=1}^{k+1} t_i = n + m \quad \text{and} \quad \sum_{\substack{i=1 \\ i \neq k}}^{k+1} t_i = n,$$

then

$$\sum_{i=1}^{k+1} t_i - \sum_{\substack{i=1 \\ i \neq k}}^{k+1} t_i = t_k = m. \qquad \square$$

The main result of this section can now be stated using Lemmas 2.1 and 2.2 as follows.

THEOREM 2.3. *Let (E, A, B) be a reachable system then there are $t_{i+1} - t_i$ controllability indices c_j of order $k - i$, $i = 0, 1, \ldots, k-1$ from which*

$$q = t_{k+1} - t_k$$

are nonproper.

This result just gives the order of the controllability indices and also shows how many of them are proper and how many are nonproper in a quantitative way. However, it does not reveal the finer structure that we are interested in. In particular, we want to know the order of the proper and also the nonproper ones. This extra information is contained in every normal external description, as we have already mentioned, but still is not obvious from the Hessenberg form. Therefore, in the next section we show how we can reveal this structure from the Hessenberg form. To do that we use the Hessenberg form to compute the normal external description; therefore we are able to relate the column degrees of the external description to the stairs of the Hessenberg form. Once this relation is known then we show how to use the stairs of the Hessenberg form to get a qualitative information about the controllability indices.

3. Computing a normal external description.

3.1. Reachability Hessenberg forms, normal external descriptions. The procedure described in this section is very similar to the one proposed in [15] for state-variable systems. However, there are fine differences at the final step of the algorithm that are quite important since we are able to extract all the information needed. To compute a normal external description $N(s)$, $D(s)$, we first embed the pencil $[sE_h - A_h \quad -B_h]$ in a unimodular pencil as follows. Assume a matrix $C_h \in \mathbb{R}^{m \times n}$

$$C_h = [* \quad C_r], \tag{3.1}$$

where C_r is of the form

$$C_r = \begin{bmatrix} C_1 & & & 0 \\ * & C_2 & & \\ * & * & \ddots & \\ * & * & \cdots & C_k \end{bmatrix}, \tag{3.2}$$

where C_1 is a square nonsingular matrix of dimensions $t_1 \times t_1$ and $C_i \in \mathbb{R}^{(t_i - t_{i-1}) \times t_i}$ for $i = 2, \ldots, k$. Moreover, C_i for $i = 2, \ldots, k$ satisfy the following:

$$\begin{bmatrix} A_{i-1,i} \\ C_i \end{bmatrix}, \qquad i = 2, \ldots, k \tag{3.3}$$

are nonsingular matrices. In particular, we can select C_i's as follows:

$$C_i = [* \quad S_{2,i}], \qquad i = 1, 2, \ldots, k, \tag{3.4}$$

where $S_{1,i}$'s are *lower triangular* matrices of dimensions $(t_i - t_{i-1}) \times (t_i - t_{i-1})$; thus

$$\begin{bmatrix} A_{i-1,i} \\ C_i \end{bmatrix} = \begin{bmatrix} S_{1,i} & 0 \\ * & S_{2,i} \end{bmatrix} = S_{i,r}, \qquad i = 2, \ldots, k,$$

and $C_1 = S_{1,r}$. At this point we mention that $S_{i,r}$ is lower triangular. This is due to the fact that we put $A_{i-1,i}$ in lower triangular form at the bottom corner in the Hessenberg form. The form of $S_{i,r}$, i.e., lower triangular, is exploited later in the algorithm, computing a normal external description. It is pointed out that the selection of C_i is similar to the one proposed in [2] and is far from unique.

Due to the selection of the matrix C_r, the pencil

$$U_r(s) = \begin{bmatrix} sE_r - A_r & -B_r \\ -C_r & 0 \end{bmatrix} \tag{3.5}$$

is unimodular. That is $\det U_r(s) = \text{const} \neq 0$. Therefore, the inverse of $U_r(s)$ exists and is a polynomial matrix. By denoting the inverse of $U(s)$ as $V(s)$ and partitioned as

$$V_r(s) = \begin{bmatrix} F_r(s) & N_r(s) \\ G_r(s) & D_r(s) \end{bmatrix}, \tag{3.6}$$

it is clear that the pair $N_r(s)$, $D_r(s)$ is a basis for the kernel of $[sE_r - A_r \quad -B_r]$. Hence the problem of determining a normal external description has been reduced to the inversion of the unimodular pencil

$$U(s) = \begin{bmatrix} sE_h - A_h & B_h \\ C_h & 0 \end{bmatrix}.$$

Remark 3.1 (see [15]). The selection of the sizes of C_i is not arbitrary. The lengths of the Jordan chains of the infinite elementary divisors of $U(s)$ must be kept minimal, namely, equal to the number of stairs in the Hessenberg form. This guarantees the desired relationship between the column minimal indices of the pencil $[sE - A \quad -B]$ and the infinite elementary divisors of the unimodular pencil. This fact is crucial to obtain a normal external description. A different relationship will not produce a normal external description. Therefore, though the selection of C_i is not unique, the size of the blocks must comply with the rule used in (3.2).

PROPOSITION 3.1. *Let* (2.1) *be regular, i.e.,* $\det(sE - A) \not\equiv 0$ *and the pair* $N_r(s)$, $D_r(s)$ *described as in* (3.6). *Then, we have the following:*

$N_r(s)$, $D_r(s)$ *are right coprime*;

$D_r(s)$ *and* $sE_r - A_r$ *have the same nonunity invariant polynomials*;

$N_r(s)$, $D_r(s)$ *are column-degree ordered.*

Proof. The first two statements can be shown in a similar way as in [15]. The fact that $N_r(s)$, $D_r(s)$ are column-degree ordered will be shown constructively from the proposed algorithm. □

The next step in this section is twofold. First, we determine in detail the structure of the normal external description using the Hessenberg form, and, second, we show that the resulting normal external description is indeed column-degree ordered. We point out that the main concern here is the structure of the resulting normal description rather than the inversion of the unimodular matrix $U(s)$, which can be computed as in [2]. This is because our goal is to determine the relation between the

stairs of the Hessenberg form and the column degrees of the computed normal external description. Finally, this relation shows how to compute the proper/nonproper controllability indices using only the stairs of the Hessenberg form.

It is clear that such a method will provide computationally efficient techniques for the computation of the pair $N(s)$, $D(s)$. For this purpose we use the block Hessenberg form of the pair (E, A, B). By defining P to be a permutation matrix we can bring the pencil

$$P^T \begin{bmatrix} sE_r - A_r & B_r \\ C_r & 0 \end{bmatrix}$$

to the following form

(3.7)

$$\begin{bmatrix} -S_{1,r} & 0 & \dots & 0 & 0 \\ sJ_{1,r} - N_{1,r} & -S_{2,r} & \dots & 0 & 0 \\ * & sJ_{2,r} - N_{2,r} & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ * & \dots & sJ_{k-1,r} - N_{k-1,r} & -S_{k,r} & 0 \\ * & \dots & * & sJ_{k,r} - N_{k,r} & -S_{k+1,r} \end{bmatrix} = sJ_r - S_r,$$

where

$$S_{1,r} = C_1,\ S_{k+1,r} = A_{k,k+1},\ J_{k,r} = E_{kk},\ J_{i,r} = \begin{bmatrix} R_{i,r} \\ 0 \end{bmatrix} \quad \text{for } i = 1, \dots, k-1$$

and

$$S_{i,r} = \begin{bmatrix} A_{i-1,i} \\ C_i \end{bmatrix} \quad \text{for } i = 2, \dots, k.$$

The infinite elementary divisors of the unimodular pencil $U(s)$ are computed as follows [17]

$$t_{i+1} - t_i \text{ infinite elementary divisors } \delta_j \text{ of order } k - i + 1.$$

Therefore

$$\begin{bmatrix} I & 0 \\ 0 & P^T \end{bmatrix} U(s) = \begin{bmatrix} sJ_\infty - S_\infty & 0 \\ * & sJ_r - S_r \end{bmatrix} = sJ - S. \tag{3.8}$$

The lower triangular structure of $(sJ - N)$ implies a lower triangular structure of its inverse description as follows

$$\begin{bmatrix} sJ_\infty - S_\infty & 0 \\ * & sJ_r - S_r \end{bmatrix} \begin{bmatrix} N_\infty(s) & 0 \\ * & V_r(s) \end{bmatrix} = I, \tag{3.9}$$

where

$$V_r(s) = \begin{bmatrix} F_r(s) & N_r(s) \\ G_r(s) & D_r(s) \end{bmatrix}. \tag{3.10}$$

Due to this structure a normal external description for the system (E_h, A_h, B_h) is of the form

$$\begin{bmatrix} N(s) \\ D(s) \end{bmatrix} = \begin{bmatrix} N_\infty(s) & 0 \\ * & N_r(s) \\ * & D_r(s) \end{bmatrix}. \tag{3.11}$$

This lower triangular structure implies that we can compute the $N_\infty(s)$ and $(N_r(s), D_r(s))$ independently and, in the sequel, compute the off-diagonal terms. Equation (3.9) shows that

$$N_\infty(s) = (sJ_\infty - S_\infty)^{-1}.$$

Since $l+1$ is the infinite elementary divisor in $(sJ_\infty - S_\infty)$, then

$$S_\infty^{-1}(s\bar{N}_\infty - I)^{-1} = \tag{3.12}$$

$$\begin{bmatrix} \hat{N}^\infty_{0,1} & 0 & \cdots & 0 & 0 \\ \hat{N}^\infty_{1,1}s + O(s^0) & \hat{N}^\infty_{0,2} & \cdots & 0 & 0 \\ \hat{N}^\infty_{2,1}s^2 + O(s) & \hat{N}^\infty_{1,2}s + O(s^0) & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{N}^\infty_{l-1,1}s^{l-1} + O(s^{l-2}) & \hat{N}^\infty_{l-2,2}s^{l-2} + O(s^{l-3}) & \cdots & \hat{N}^\infty_{0,l} & 0 \\ \hat{N}^\infty_{l,1}s^l + O(s^{l-1}) & \hat{N}^\infty_{l-1,2}s^{l-1} + O(s^{l-2}) & \cdots & \hat{N}^\infty_{1,l}s + O(s^0) & \hat{N}^\infty_{0,l+1} \end{bmatrix},$$

where $O(s^i)$ denotes polynomial terms of power less or equal to i and

$$\hat{N}^\infty_{0,i} = S^{-1}_{i,\infty}.$$

The procedure for computing the expression in (3.12) is parallel to that proposed in [15]. In addition, the form of all the submatrices in (3.14) can be derived by using similar techniques to the ones used in [15].

Though we can argue that the computation of $N_r(s)$ and $D_r(s)$ can be performed in parallel to that in [15], we analytically compute the normal external description since in this case fine differences in the algorithm are important to the determination of the proper and nonproper controllability indices.

The largest reachability index in the pencil $[sE_r - A_r \;\; -B_r]$ is k [17]. This implies that the largest infinite elementary divisor in $(sJ_r - S_r)$ is equal to $k+1$. Then

$$\begin{aligned} U_r(s)^{-1} = V_r(s) &= -S_r^{-1}(I + s\bar{N}_r + \cdots + s^k\bar{N}_r^k)P^T \\ &= V_{0,r} + V_{1,r}s + \cdots + V_{k,r}s^k, \end{aligned} \tag{3.13}$$

where

$$\bar{N}_r = J_r S_r^{-1} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ \bar{N}_{1,r} & 0 & \cdots & 0 & 0 \\ * & \bar{N}_{2,r} & \ddots & 0 & 0 \\ * & * & \ddots & \vdots & \vdots \\ * & * & \cdots & \bar{N}_{k,r} & 0 \end{bmatrix}$$

and

$$\bar{N}_{i,r} = \begin{bmatrix} T_{i,r} \\ 0_{t_{i+1}-t_i} \end{bmatrix}, \qquad T_{i,r} = R_{i,r}S_{i,r}^{-1}.$$

By denoting $N_r(s)$ and $D_r(S)$ as

$$\begin{aligned} N_r(s) &= N_{0,r} + sN_{1,r} + s^2N_{2,r} + \cdots + s^kN_{k,r}, \\ D_r(s) &= D_{0,r} + sD_{1,r} + s^2D_{2,r} + \cdots + s^kD_{k,r}, \end{aligned}$$

we can compute

$$N_{i,r} = -[I_{o_r} \;\; 0]V_{i,r}\begin{bmatrix} 0 \\ I_m \end{bmatrix}, \tag{3.14}$$

$$D_{i,r} = -[0 \;\; I_m]V_{i,r}\begin{bmatrix} 0 \\ I_m \end{bmatrix}, \qquad i = 1, 2, \ldots, k. \tag{3.15}$$

In the sequel, we show that the pair $N_r(s)$, $D_r(s)$ is column-degree ordered. For this purpose we use the special form of the matrices $\bar{N}$ and S to compute the leading column coefficients of $N(s)$ and $D(s)$. It can be shown that the powers of $\bar{N}$ have the form

$$\bar{N}_r^i = \begin{bmatrix} 0 & 0 \\ \hat{N}_{i,r} & 0 \end{bmatrix}, \qquad i = 1, 2, \ldots, k, \tag{3.16}$$

where $\hat{N}_{i,r}$ are lower block triangular matrices and have dimensions $o_i \times o_i$ where $o_i = o_r + m - \sum_{j=1}^{i} t_j$. The block diagonal elements of $\hat{N}_{i,r}$ are computed as follows:

$$\hat{N}_{i,j}^r = \prod_{q=i+j-1}^{j} \bar{N}_{q,r}, \qquad j + i = 2, 3, \ldots, k+1. \tag{3.17}$$

Using the special form of the powers of $\bar{N}$ we see that

$$S_r^{-1}(s\bar{N}_r - I)^{-1} = \tag{3.18}$$

$$\begin{bmatrix} \hat{N}_{0,1}^r & 0 & \cdots & 0 & 0 \\ \hat{N}_{1,1}^r s + O(s^0) & \hat{N}_{0,2}^r & \cdots & 0 & 0 \\ \hat{N}_{2,1}^r s^2 + O(s) & \hat{N}_{1,2}^r s + O(s^0) & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{N}_{k-1,1}^r s^{k-1} + O(s^{k-2}) & \hat{N}_{k-2,2}^r s^{k-2} + O(s^{k-3}) & \cdots & \hat{N}_{0,k}^r & 0 \\ \hat{N}_{k,1}^r s^k + O(s^{k-1}) & \hat{N}_{k-1,2}^r s^{k-1} + O(s^{k-2}) & \cdots & \hat{N}_{1,k}^r s + O(s^0) & \hat{N}_{0,k+1}^r \end{bmatrix}.$$

The pair $N_r(s)$, $D_r(s)$ can be easily computed by carefully applying the appropriate column permutations in (3.20). In particular, the pair $N_r(s)$, $D_r(s)$ has the form

$$\begin{bmatrix} N_r(s) \\ D_r(s) \end{bmatrix} = \tag{3.19}$$

$$\begin{bmatrix} H_{0,1}^r & 0 & \cdots & 0 & 0 \\ H_{1,1}^r s + O(s^0) & H_{0,2}^r & \cdots & 0 & 0 \\ H_{2,1}^r s^2 + O(s) & H_{1,2}^r s + O(s^0) & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ H_{k-1,1}^r s^{k-1} + O(s^{k-2}) & H_{k-2,2}^r s^{k-2} + O(s^{k-3}) & \cdots & H_{0,k}^r & 0 \\ H_{k,1}^r s^k + O(s^{k-1}) & H_{k-1,2}^r s^{k-1} + O(s^{k-2}) & \cdots & H_{1,k}^r s + O(s^0) & H_{0,k+1}^r \end{bmatrix},$$

where

$$H_{0,i}^r = S_{i,r}^{-1}\begin{bmatrix} 0_{t_{i-1}} \\ I_{t_i - t_{i-1}} \end{bmatrix}, \quad i = 1, 2, \ldots, k+1, \quad t_0 = 0 \tag{3.20}$$

and

$$H_{i,j}^r = \hat{N}_{i,j}^r H_{0,i}, \quad i,j \in Z^+, \qquad i+j = 2,\dots,k+1. \tag{3.21}$$

Equation (3.21) shows that $D_r(s)$ is column-degree ordered. The next step is to find the highest column coefficient of $D_r(s)$, $D_{r,hc}$. From (3.21) we know that

$$D_{r,hc} = [\, H_{k,1}^r \quad H_{k-1,2}^r \quad \cdots \quad H_{1,k}^r \quad H_{0,k+1}^r \,],$$

and therefore if we determine the structure of $H_{k,i}^r$, for $i = 1,2,\dots,k+1$ we have actually computed $D_{r,hc}$. We note here that $H_{0,k+1}^r$ does not occur in our case, since we assumed that B has full column rank. Using (3.23) we can compute each $H_{k+1-j,j}^r$ for $j = 1,2,\dots,k$ as follows:

$$H_{k+1-j,j}^r = S_{k+1,r}^{-1} \left\{ \prod_{q=k}^{j} \begin{bmatrix} T_{q,r} \\ 0_{t_{q+1}-t_q} \end{bmatrix} \right\} \begin{bmatrix} 0 \\ I_{t_j - t_{j-1}} \end{bmatrix}, \qquad j = 1,2,\dots,k,$$

which results in

$$D_{r,hc} = S_{k+1,r}^{-1} T_{k,r} \begin{bmatrix} D_{t,1} & 0 & \cdots & 0 & 0 \\ * & D_{t,2} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ * & * & \cdots & D_{t,k-1} & 0 \\ * & * & \cdots & * & I_{t_k - t_{k-1}} \end{bmatrix}, \tag{3.22}$$

where $D_{t,i}$ are full rank lower triangular matrices of dimensions $t_i - t_{i-1} \times t_i - t_{i-1}$, $i = 1,\dots,k-1$. This is due to the property that $T_{i,r}$, $i = 1,2,\dots,k-1$ enjoy; namely, they are lower triangular matrices. The matrix $T_{k,r}$ is singular and

$$\dim \ker T_{k,r} = \dim \ker R_{k,r} S_{k,r}^{-1} = \dim \ker E_{k,k} = \rho_r$$

since $S_{k,r}$ is nonsingular. Moreover, $T_{k,r}$ is of the form

$$T_{k,r} = \begin{bmatrix} \bar{T}_{k,r} & 0 \\ 0 & 0 \end{bmatrix},$$

where $\bar{T}_{k,r}$ is a full rank triangular matrix of dimensions $(t_k - \rho_r) \times (t_k - \rho_r)$.

The structure of $D_{r,hc}$ reveals the correspondence between the length of the stairs of the Hessenberg form and the column degrees of the normal external description and is given as follows.

THEOREM 3.1. *Let (E,A,B) be reachable. Then the controllability indices of (E,A,B) are given as follows:*

If $t_k - \rho_r < t_i - t_{i-1}$, then there are

$$t_k - \rho_r \quad \text{proper controllability indices of order} \quad k+1-i$$

and

$$(t_i - t_{i-1}) - (t_k - \rho_r) \quad \text{nonproper controllability indices of order} \quad k-i;$$

if $t_k - \rho_r \geq t_i - t_{i-1}$, then there are

$$t_i - t_{i-1} \quad \text{proper controllability indices of order} \quad k+1-i,$$

where t_i, $i = 1, 2, \ldots, k$ are the dimensions of the stairs in the pencil $[sE_r - A_r \quad -B_r]$ and $\rho_r = \dim \ker E_r$.

Proof. Since $N_r(s)$ is column degree order, we examine the highest coefficient matrix of

$$[H^r_{k-1,1}s^{k-1} + O(s^{k-2}) \quad H^r_{k-2,2}s^{k-2} + O(s^{k-3}) \quad \cdots \quad H^r_{0,k}],$$

which is given by

$$[H^r_{k-1,1} \quad H^r_{k-2,2} \quad \cdots \quad H^r_{0,k}]$$

and

$$H^r_{k-j,j} = S^{-1}_{k,r}\left\{\prod_{q=k-1}^{j}\begin{bmatrix} T_{q,r} \\ 0_{t_{q+1}-t_q}\end{bmatrix}\right\}\begin{bmatrix} 0 \\ I_{t_j - t_{j-1}}\end{bmatrix}, \qquad j = 1, 2, \ldots, k-1;$$

$$H^r_{0,k} = S^{-1}_{k,r}\begin{bmatrix} 0_{t_{k-1}} \\ I_{t_k - t_{k-1}}\end{bmatrix}.$$

Therefore, the highest coefficient matrix is given by

$$N_{r,hc} = S^{-1}_{k,r}\begin{bmatrix} N_{t,1} & 0 & \cdots & 0 & 0 \\ * & N_{t,2} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ * & * & \cdots & N_{t,k-1} & 0 \\ * & * & \cdots & * & I_{t_k - t_{k-1}}\end{bmatrix},$$

where $N_{t,i}$ are full rank lower triangular matrices of dimensions $t_i - t_{i-1} \times t_i - t_{i-1}$, $i = 0, 1, \ldots, k-1$. This is due to the property that $T_{i,r}$, $i = 1, 2, \ldots, k-1$ enjoy; namely, they are lower triangular matrices of full rank.

The proof now follows easily from the definition of the controllability indices and the form of $N_{r,hc}$ and $D_{r,hc}$ (see (3.22)). □

Note that in the case when E is nonsingular $\rho_r = 0$ and $t_k \geq \mathfrak{t}_i - t_{i-1}$, for all $i \in \{1, 2, \ldots, k\}$ since

$$t_1 \leq t_2 \leq \cdots \leq t_{k-1} \leq t_k$$

and therefore the controllability indices coincide with the reachability indices.

3.2. Controllability Hessenberg forms, normal external descriptions. In this section we use again, as before, the idea of embedding the pair (E_c, A_c) in a unimodular matrix $U_c(s)$ by selecting an appropriate matrix C_c. The construction of C_c is along the same lines of those for the reachability Hessenberg form. An obvious difference in this case is that $C_c \in \mathbb{R}^{m\times(n+m)}$ as opposed to the selection of C_r where $C_r \in \mathbb{R}^{m\times n}$. In addition the structure of the C_c is now dictated by the stairs of the controllability form. Therefore, the unimodular pencil under investigation is

$$U_c(s) = s\begin{bmatrix} E_c \\ 0\end{bmatrix} - \begin{bmatrix} A_c \\ C_c\end{bmatrix}. \tag{3.23}$$

Since the procedure for computing a normal external description is very similar to the one studied for the case of the reachability form, we only streamline the differences of the two procedures.

We denote

$$V_c(s) = U_c(s)^{-1} = \begin{bmatrix} F_c(s) & N_c(s) \\ G_c(s) & D_c(s) \end{bmatrix}. \tag{3.24}$$

An important difference between $D_r(s)$ (cf. §3.1) and $D_c(s)$ is the dimensions. From the structure of the reachability Hessenberg form we can see that $D_r(s) \in \mathbb{R}^{m\times m}[s]$, while the structure of the controllability Hessenberg form implies that $D_c(s) \in \mathbb{R}^{t_{k+1}\times m}[s]$. It is quite simple to determine the structure of the highest coefficient matrix $D_{c,hc}$ of $D_c(s)$ using similar arguments as in §3.1. Here, we simply present the form without giving any details;

$$D_{c,hc} = S_{k+1,c}^{-1} \begin{bmatrix} D_{t,1} & 0 & \cdots & 0 & 0 \\ * & D_{t,2} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ * & * & \cdots & D_{t,k-1} & 0 \\ * & * & \cdots & * & D_{t,k} \\ * & * & \cdots & * & * \end{bmatrix}, \tag{3.25}$$

where $D_{t,i}$ are full rank lower triangular matrices of dimensions $t_i - t_{i-1} \times t_i - t_{i-1}$, $i = 1, \ldots, k$. This is due to the fact that the initial compression in the controllability Hessenberg is performed on $[E \;\; 0]$, as opposed to the reachability Hessenberg form where the initial compression is performed on B. This difference results in a non-square highest coefficient matrix for $D_c(s)$. Using the controllability Hessenberg form results in loss of information, in the transformed system, about the input matrix. In particular, even in the case when the system is reachable, although we can determine the controllability indices of the system and also see how many nonproper/proper controllability indices exist, we cannot distinguish them. Note that in the case of a controllable, but not reachable system, even this is not true. Here is an example that clarifies these ideas. Let

$$[sE - A \;\; -B] = \begin{bmatrix} -1 & s & 0 & 0 & 0 \\ 0 & -1 & s & 0 & 1 \\ 0 & 0 & -1 & 1 & 0 \end{bmatrix}.$$

This reachable system has only one nonproper controllability index and one infinite elementary divisor of order one. Now let

$$[sE - A \;\; -B] = \begin{bmatrix} -1 & s & 0 & 0 & 0 & 0 \\ 0 & -1 & s & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \end{bmatrix}.$$

This controllable but not reachable system has the same controllability indices as the above-mentioned system; however it has two infinite elementary divisors of order one.

That is, the normal external description of the transformed system does not provide this extra information, as in the case of the reachability Hessenberg form. Therefore, we must bring the normal external description back to the original system coordinates and read out the column degrees.

Another remark that we would like to make is that the dimensions of the $D_{t,i} \in \mathbb{R}^{(t_i - t_{i-1})\times(t_i - t_{i-1})}$ give the number of the controllability indices. The order of the controllability indices easily follows from the corresponding powers of s; that is, the

$D_{t,i} \in \mathbb{R}^{(t_i - t_{i-1}) \times (t_i - t_{i-1})}$ corresponds to $t_i - t_{i-1}$ controllability indices of order $(k + 1 - i)$.

Finally, as we have already mentioned, the index of nilpotency of $\bar{N}$ is minimal. Although this is true, the algorithm can produce misleading results when $\bar{N}$ is ill conditioned. However, the algorithm provides a direct and reliable solution to the problem in the case of a well-conditioned $\bar{N}$.

4. An example. Consider the following reachable implicit linear system.

$$E = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

4.1. Reachability Hessenberg forms, normal external descriptions. The reachability Hessenberg form of $\{E, A, B\}$ is

$$E_r = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad A_r = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad B_r = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix};$$

the steps of which are $t_1 = t_2 = 3$. Therefore, the system $\{E, A, B\}$ has three reachability indices of order two; namely, $r_1 = r_2 = r_3 = 2$. We can now use the procedure described in Theorem 3.1 to compute the controllability indices of $\{E, A, B\}$ as follows:

$$t_2 - \rho_r = 1 < 3 = t_1 - t_0,$$

hence there are

$$t_2 - \rho_r = 1$$

proper controllability index of order $k + 1 - i = 2$; namely,

$$c_1 = 2$$

and

$$t_1 - t_0 - (t_2 - \rho_r) = 2$$

nonproper controllability indices; namely,

$$c_2 = c_3 = 1.$$

The computed normal external description for $\{E_r, A_r, B_r\}$, using the algorithm described in §3, is

$$N_r(s) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & s \end{bmatrix}, \qquad D_r(s) = \begin{bmatrix} s^2 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Note that

$$D_{r,hc} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

as expected.

A normal external description for $\{E, A, B\}$ is

$$N(s) = \begin{bmatrix} 0 & 0 & -s \\ 0 & 0 & -1 \\ s & 0 & 0 \\ 1 & 0 & 0 \\ 0 & s & 0 \\ 0 & 1 & 0 \end{bmatrix}, \qquad D(s) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ s^2 & 0 & 1 \end{bmatrix}.$$

4.2. Controllability Hessenberg forms, normal external descriptions. The controllability Hessenberg form of $\{E, A, B\}$ is

$$E_c = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \qquad A_c = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The steps of the controllability Hessenberg form are $s_1 = 1$, $s_2 = 5$, $t_1 = 1$, $t_2 = 3$, $t_3 = 5$. Therefore, $\{E, A, B\}$ has $t_1 - t_0 = 1$ controllability index of order $k + 1 - i = 2$; namely, $c_1 = 2$ and $t_2 - t_1 = 2$ controllability indices of order $k + 1 - i = 1$; namely, $c_2 = c_3 = 1$. Furthermore, we know that there exist $t_3 - t_2 = 2$ *nonproper* controllability indices.

The computed normal external description for the pair $\{E_c, A_c\}$ is

$$N_c(s) = \begin{bmatrix} 1 & 0 & 0 \\ s & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad D_c(s) = \begin{bmatrix} s^2 & 1 & 0 \\ 0 & s & 0 \\ 0 & 0 & s \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

and the highest order coefficient matrix is

$$D_{c,hc} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

A normal external description for $\{E, A, B\}$ is

$$N(s) = \begin{bmatrix} 0 & -s & 0 \\ 0 & -1 & 0 \\ s & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & s \\ 0 & 0 & 1 \end{bmatrix}, \qquad D(s) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ s^2 & 1 & 0 \end{bmatrix};$$

which shows that $c_1 = 2$ is a proper controllability index and $c_2 = c_3 = 1$ are nonproper controllability indices.

All the algorithms in this paper have been implemented in MATLAB [10].

5. Conclusions. In this paper we presented a computationally efficient method for computing normal external descriptions using staircase forms and embedding techniques in implicit systems. The structure of the computed normal external description revealed the relationship between the reachability Hessenberg form and the proper/nonproper controllability indices of the system. That is, it was shown how to compute and distinguish the proper/nonproper controllability indices from the length of the stairs of the reachability Hessenberg form. Finally, we studied the controllability Hessenberg form and the corresponding computed normal external description and compared it with the one computed using the reachability Hessenberg form of the system.

Future research along these lines can be performed by developing an alternative algorithm for the computation of normal external descriptions in implicit systems by generalizing Patel's algorithm in [13]. This observation is motivated by the fact that in the regular state-space case ($E = I$), the algorithm in [15] and Patel's algorithm use the same principles. As a matter of fact, both algorithms start from the Hessenberg form.

REFERENCES

[1] T. BEELEN, P. VAN DOOREN, AND M. VERHAGEN, *A class of fast staircase algorithms for state-space systems*, Proc. American Control Conf., Seattle, WA, 1986.

[2] T. BEELEN AND P. VAN DOOREN, *A pencil approach for embedding a polynomial matrix into a unimodular matrix*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 77–89.

[3] G. D. FORNEY, JR., *Minimal bases of rational vector spaces, with applications to multivariable linear systems*, SIAM J. Control, 13 (1975), pp. 493–520.

[4] N. KARCANIAS AND H. ELIOPOULOU, *A classification of minimal polynomial bases for singular systems*, MTNS '89, Amsterdam, Holland, 1989, pp. 255–262.

[5] V. KUČERA, *Analysis and design of discrete linear control systems*, Prentice-Hall, Englewood Cliffs, NJ, 1991.

[6] V. KUČERA AND P. ZAGALAK, *Constant solutions of polynomial equations*, Internat. J. Controls, 53 (1991), pp. 495–502.

[7] J. J. LOISEAU, K. ÖZÇALDIRAN, M. MALABRE, AND N. KARCANIAS, *A feedback classification of singular systems*, Kybernetica, 27 (1991), pp. 289–305.

[8] F. L. LEWIS, *A survey of linear singular systems*, Circ. Syst. Signal Process., 5 (1986), pp. 3–36.

[9] M. MALABRE, V. KUČERA, AND P. ZAGALAK, *Reachability and controllability indices for linear descriptor systems*, Systems Control Letters, 15 (1990), pp. 119–123.

[10] C. MOLER, *MATLAB User's Guide*, The Mathworks Inc., Sherborn, MA, 1980.

[11] K. ÖZÇALDIRAN, *Control of Descriptor Systems*, Ph.D. thesis, Georgia Institute of Technology, Atlanta, 1985.

[12] C. C. PAIGE, *Properties of numerical algorithms related to computing controllability*, IEEE Trans. Automat. Control, 26 (1981), pp. 130–138.

[13] R. V. PATEL, *Computation of matrix fraction descriptions of linear time-invariant systems*, IEEE Trans. Automat. Contr., AC-26 (1991), pp. 148–161.

[14] H. H. ROSENBROCK, *State-space and Multivariable Theory*, Thomas Nelson, London, 1970.

[15] V. L. SYRMOS AND P. ZAGALAK, *Computing normal external descriptions and feedback design*, Linear Algebra Appl., 189 (1993), pp. 613–639.

[16] E. L. YIP AND E. F. SINCOVEC, *Solvability, controllability, and observability of continuous descriptor systems*, IEEE Trans. Automat. Contr., AC-26 (1981), pp. 702–705.

[17] P. VAN DOOREN, *The computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 27 (1979), pp. 103–141.

[18] ———, *The generalized eigenstructure problem in linear system theory*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 111–129.

[19] G. C. VERGHESE, P. VAN DOOREN, AND T. KAILATH, *Properties of the system matrix of a generalized state-space system*, Internat. J. Control, 30 (1979), pp. 235–243.

[20] P. ZAGALAK AND J. J. LOISEAU, *Invariant factors assignment in linear systems*, SINS'92 Texas, Ft. Worth, TX, 1992.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 307–311, January 1995

021

A REMARK ON MINC'S MAXIMAL EIGENVECTOR BOUND FOR POSITIVE MATRICES*

GEOFF A. LATHAM†

Abstract. A maximal eigenvector bound resembling that of Minc [*SIAM J. Math. Anal.*, 7 (1970), pp. 424–427] is derived for structured nonnegative matrices which, when applied to certain positive matrices, improves on the estimate derived from Minc's bound.

Key words. nonnegative matrices, positive matrices, Perron–Frobenius theory, maximal eigenvector bounds

AMS subject classifications. 15A48, 15A42, 15A45, 15A18

1. Introduction. It is a classical result of Frobenius that any nonnegative matrix $A \in \mathbb{R}^{n\times n}$ with positive spectral radius $\rho(A)$ possesses a nonnegative eigenvector $\boldsymbol{x}$, called a *maximal eigenvector*, corresponding to the so-called *maximal eigenvalue* $\rho(A)$, i.e., $A\boldsymbol{x} = \rho(A)\boldsymbol{x}$. The importance of the theory of nonnegative matrices, and this result in particular, derives from the fact that, in many applications, the requirement for physical solutions leads naturally to a nonnegativity constraint, both for the matrix A and solutions of systems of equations involving A. In some applications (cf. [1]), it is important to obtain estimates of the ratios of components of a maximal eigenvector in terms of easily computable functions of the matrix elements. Independently, the problem of estimating the ratio $\gamma = \max_{i,j} x_i/x_j$ for a positive maximal eigenvector has been examined theoretically in some detail [3]–[5].

In this note, a new estimate is derived for an upper bound of the ratio of maximal eigenvector components. This bound is derived using a technique due to Minc [3]. As well as being applicable to all positive matrices, the bound holds for certain structured nonnegative matrices. In this sense, the bound has slightly wider applicability than its counterpart in [3]: Examples of the use of the estimate to bound γ are also given.

2. A maximal eigenvector bound. The main result of this note is the following theorem.

THEOREM 1. *Let $A \in \mathbb{R}^{n\times n}$, $n \geq 2$, be a nonnegative matrix and $\boldsymbol{x}$ any maximal eigenvector of A satisfying $A\boldsymbol{x} = \rho(A)\boldsymbol{x}$ with $\rho(A) > 0$. For fixed distinct indices $\underline{i}$ and $\bar{i}$, let $\underline{I} \ni \underline{i}$ and $\overline{I} \ni \bar{i}$ be any index sets such that $x_j \leq x_{\underline{i}}$ for all $j \in \underline{I}$ and $x_j \geq x_{\bar{i}}$ for all $j \in \overline{I}$. Let $\overline{I}_0 = \{j \mid a_{\bar{i}j} = 0\}$. Assume that*

(i) *$a_{ij}x_j > 0$ for all $j \notin \underline{I} \cup \overline{I}_0$,*

(ii) *$a_{\underline{i}j} > 0$ for some $j \in \overline{I}$,*

$x_{\underline{i}} > 0$ and $x_{\bar{i}} > 0$. Then the ratio $\chi = x_{\bar{i}}/x_{\underline{i}}$ satisfies the estimate

$$\chi \leq \tfrac{1}{2}\big(M + \sqrt{M^2 + 4\alpha}\big), \tag{1}$$

where

$$M = \max_{j\notin \underline{I}\cup\overline{I}_0} \frac{a_{\bar{i}j}}{a_{\underline{i}j}} \quad \text{and} \quad \alpha = \frac{\sum_{j\in\underline{I}} a_{\bar{i}j}}{\sum_{j\in\overline{I}} a_{\underline{i}j}}. \tag{2}$$

* Received by the editors January 11, 1993; accepted for publicaton (in revised form) by R. Horn, October 20, 1993.

† Centre for Mathematics and its Applications, Australian National University, Canberra ACT 0200, Australia (gal851@cisr.anu.edu.au).

Before proving Theorem 1, we first state the useful elementary estimate that underlies the results in [3]. It proves to be the key to our proof of Theorem 1.

LEMMA 1. *If $q_1, q_2, \dots, q_n$ are positive numbers, then*

$$\min_i \frac{p_i}{q_i} \le \frac{p_1 + \cdots + p_n}{q_1 + \cdots + q_n} \le \max_i \frac{p_i}{q_i}, \tag{3}$$

for any real numbers $p_1, p_2, \dots, p_n$. Equality holds on either side of (3) *if and only if all the ratios p_i/q_i are equal.*

For a proof of this lemma, see [2, p. 79] or [1].

Proof of Theorem 1. Because $x_{\underline{i}} > 0$, the eigenvalue equation can be used to form the quotient χ giving

$$\begin{aligned}
\chi = \frac{x_{\overline{i}}}{x_{\underline{i}}} &= \frac{\sum_j a_{\overline{i}j} x_j}{\sum_j a_{\underline{i}j} x_j}, \\
&= \frac{\sum_{j \in \underline{I}} a_{\overline{i}j} x_j}{\sum_j a_{\underline{i}j} x_j} + \frac{\sum_{j \notin \underline{I} \cup \overline{I}_0} a_{\overline{i}j} x_j}{\sum_j a_{\underline{i}j} x_j}, \\
&\le \frac{\sum_{j \in \underline{I}} a_{\overline{i}j} x_{\underline{i}}}{\sum_{j \in \overline{I}} a_{\underline{i}j} x_{\overline{i}}} + \frac{\sum_{j \notin \underline{I} \cup \overline{I}_0} a_{\overline{i}j} x_j}{\sum_{j \notin \underline{I} \cup \overline{I}_0} a_{\underline{i}j} x_j}, \\
&\le \frac{\sum_{j \in \underline{I}} a_{\overline{i}j}}{\sum_{j \in \overline{I}} a_{\underline{i}j}} \frac{1}{\chi} + \max_{j \notin \underline{I} \cup \overline{I}_0} \frac{a_{\overline{i}j}}{a_{\underline{i}j}}, \quad \text{(using Lemma 1)}
\end{aligned} \tag{4}$$

from which we arrive at the quadratic inequality

$$\chi^2 - M\chi - \alpha \le 0, \tag{5}$$

where M and α are as given in (2). If γ_+ denotes the positive root of the left-hand side of (5), then

$$\chi \le \gamma_+ \quad \text{where} \quad \gamma_+ = \tfrac{1}{2}\big(M + \sqrt{M^2 + 4\alpha}\big), \tag{6}$$

which proves the theorem. □

Remark 1. Only certain nonnegative matrices with special structure satisfy the conditions of the theorem. In particular, (i) means that zero entries in row $\underline{i}$ can only lie in columns from $\underline{I} \cup \overline{I}_0$, while (ii) requires at least one nonzero entry in row $\underline{i}$ in some column from $\overline{I}$. If A is nonnegative and irreducible, it is a result of Frobenius that the maximal eigenvector $\boldsymbol{x}$ is positive, and so condition (i) reduces to $a_{\underline{i}j} > 0$ for all $j \notin \underline{I} \cup \overline{I}_0$, and the requirements $x_{\underline{i}} > 0$ and $x_{\overline{i}} > 0$ can be dropped. Furthermore, if A is positive, then $\overline{I}_0$ is empty, and conditions (i) and (ii) are automatically satisfied indicating that the theorem applies for *any* positive matrix. For nonnegative matrices that are not positive, there is only a slight loss of generality by omitting $\overline{I}_0$ from the theorem altogether, since this affects only the structural assumption (i) and not the size of M.

Remark 2. The most obvious application of Theorem 1 is the estimation of an upper bound for $x_{\overline{i}}/x_{\underline{i}}$ for some given fixed indices $\underline{i} \ne \overline{i}$. In this case, with no prior knowledge of the ordering of the components in the maximal eigenvector, one can simply take $\overline{I} = \{\overline{i}\}$ and $\underline{I} = \{\underline{i}\}$, in which case condition (ii) of the theorem reduces to $a_{\underline{i}\overline{i}} > 0$.

Remark 3. It is evident that, for given fixed $\underline{i}$ and $\overline{i}$, the best bound in (1) (or (6)) is obtained by optimizing the choice of index sets. The smallest upper bound in (1) will be obtained for that choice of index sets that simultaneously minimizes M

and α. However, these requirements can be contradictory. Requiring M to be small suggests taking $\underline{I}$ as large as possible; yet requiring α to be small suggests taking $\underline{I}$ as small as possible and $\overline{I}$ as large as possible (see (2) and Example 2 below).

Remark 4. Usually, a complete knowledge of the index sets $\underline{I}$ and $\overline{I}$, as defined in the theorem, is not available. However, for certain matrices, such index sets are easily identified (see §2.2 below). The more information that is known about the ordering of maximal eigenvector components, then the better is the potential of (1) to give tighter upper bounds, with the best situation occurring when the complete ordering of the maximal eigenvector is known.

2.1. Positive matrices. If A is positive, then the application of Lemma 1 to (4) yields

$$\chi \leq \max_j \frac{a_{\bar{i}j}}{a_{\underline{i}j}}, \tag{7}$$

which is a result of Minc [3, p. 425], [4, p. 42]. Because $\gamma_+ > M$, it is clear that the bound given by γ_+ in (6) will exceed that in (7) whenever $\underline{I}$ does not contain all indices j for which the maximum on the right-hand side of (7) is attained. However, if the reverse is true, then it is possible for (6) to give a better estimate of χ than does (7). For example, if the maximum in (7) occurs only in the column $j = \underline{i}$, then (6), applied with $\underline{I} = \{\underline{i}\}$, can give a better estimate.

If $\overline{I} \subseteq \{i \mid x_i = \max_j x_j\}$ and $\underline{I} \subseteq \{i \mid x_i = \min_j x_j\}$, then both (6) and (7) give bounds for $\gamma = \max_{i,j} x_i/x_j$. Of course, to apply these estimates directly requires a knowledge of these index sets, something which for arbitrary positive A is not available. Hence, overestimating in (6) gives

$$\gamma \leq \max_{i \neq j} \gamma_+, \tag{8}$$

where the maximum is taken over all $\underline{I} = \{i\}$, $\overline{I} = \{j\}$ for $i \neq j$, while overestimating in (7) gives

$$\gamma \leq \max_{i,j,k} \frac{a_{ij}}{a_{kj}}, \tag{9}$$

which are two easily computable bounds for γ. The estimate (9) is again due to Minc [3, p. 425], [4, p. 42]. The right-hand side of (8) exceeds the right-hand side of (9). This is because the largest M as given by (2) and taken over all $\underline{I} = \{i\}$, $\overline{I} = \{j\}$ (and $\overline{I}_0$ empty), equals the right-hand side of (9). Since $\gamma_+ > M$, the largest γ_+ must exceed the largest M, which is the right-hand side of (9).

Remark 5. Of course, if a single index i, for which x_i is, respectively, either a maximum or minimum, is known, then the maxima in (8) and (9) may be taken over all other indices with this index fixed and equal to $\underline{i}$ if x_i is a minimum, or $\bar{i}$ if x_i is a maximum.

2.2. Ordered matrices. One class of matrices for which it is easy to identify, at least partially, the ordering of maximal eigenvector components and hence suitable index sets $\underline{I}$ and $\overline{I}$, are those whose column entries are ordered across rows. An easy result that aids in the identification of maximal eigenvector ordering for this class is the following lemma.

LEMMA 2. *If A is nonnegative and $A\boldsymbol{x} = \rho(A)\boldsymbol{x}$ with $\rho(A) > 0$ and $\boldsymbol{x} > 0$, then*

for any index i^* satisfying $a_{i^*j} > 0$ for all j,

$$\min_j \frac{a_{ij} - a_{i'j}}{a_{i^*j}} \le \frac{x_i - x_{i'}}{x_{i^*}} \le \max_j \frac{a_{ij} - a_{i'j}}{a_{i^*j}}, \tag{10}$$

for all i and i'. Moreover equality can occur on either side of (10) *if and only if $a_{ij} - a_{i'j} = \lambda a_{i^*j}$ for all j and some $\lambda \in \mathbb{R}$.*

Proof. Because $x_{i^*} > 0$, the eigenvalue equation can be used to form the quotient

$$\frac{x_i - x_{i'}}{x_{i^*}} = \frac{\sum_j (a_{ij} - a_{i'j}) x_j}{\sum_j a_{i^*j} x_j}.$$

Since $\boldsymbol{x} > 0$ and $a_{i^*j} > 0$ for all j, an application of Lemma 1 then gives (10). Lemma 1 also implies that equality can hold on either side of (10) if and only if $(a_{ij} - a_{i'j})/a_{i^*j}$ is constant for all j. □

3. Examples. Here we illustrate the use of (1), and in the case of positive matrices, compare the estimates from (1) and (7).

Example 1. Consider estimating x_3/x_2 for the nonnegative irreducible matrix

$$A = \begin{pmatrix} 2 & 9 & 1 & 1 \\ 2 & 0 & 6 & 0 \\ 3 & 1 & 8 & 1 \\ 2 & 0 & 6 & 0 \end{pmatrix},$$

for which an estimate using (7) is not possible. From Lemma 2, it follows that $x_3 > x_2$ and that $x_4 = x_2$. It is therefore possible to take $\underline{I} = \{2,4\}$, $\overline{I} = \{3\}$, $\overline{i} = 3$, $\underline{i} = 2$, and so $\overline{I}_0$ is empty. The definitions (2) imply $\alpha = 1/3$ and $M = 3/2$, and using these in (1) gives $\gamma_+ = (3 + \sqrt{43/3})/4 \approx 1.696$. For this matrix, it happens that $\gamma = x_3/x_2$, so $\gamma \le (3 + \sqrt{43/3})/4$. The actual value of x_3/x_2 (and of γ) is $(59 + 26\sqrt{11})/(28 + 20\sqrt{11}) \approx 1.540$ while $\rho(A) = 5 + 2\sqrt{11}$. The same procedure used to estimate x_1/x_2 gives the bound $x_1/x_2 \le (1 + \sqrt{21})/2 \approx 2.791$, while the actual value is $x_1/x_2 = 113/(28 + 20\sqrt{11}) \approx 1.198$.

Example 2. Consider the positive matrix

$$A = \begin{pmatrix} 3 & 2 & 3 & 3 & 3 \\ 2 & 1 & 2 & 2 & 2 \\ 2 & 1 & 2 & 2 & 2 \\ 2 & 1 & 2 & 2 & 2 \\ 3 & 2 & 3 & 3 & 3 \end{pmatrix}.$$

It is clear by Lemma 2 that $\{1,5\}$ and $\{2,3,4\}$ are the index sets for which the components x_i are, respectively, the largest and the smallest, so the choice $\underline{i} = 2$ and $\overline{i} = 1$ is appropriate. Minc's bound (7) then gives $\gamma \le 2$. Choosing $\underline{I} = \{2\}$ and $\overline{I} = \{1\}$ in (1) gives, using (6), $\gamma_+ = 2$, the same as Minc's bound. If however we take instead $\overline{I} = \{1,5\}$ and $\underline{I} = \{2\}$ in (2), then (6) gives $\gamma_+ = (3 + \sqrt{17})/4 \approx 1.781$, which improves the previous bound. The actual value of γ for this matrix is $(\sqrt{129} + 1)/8 \approx 1.545$. This illustrates how the bound can be improved by optimising the choice of the index sets $\overline{I}$ and $\underline{I}$ as well as the selection of indices $\overline{i}$ and $\underline{i}$ from them.

Remark 6. If two indices $\underline{i}$ and $\overline{i}$ are known for which $\chi = x_{\overline{i}}/x_{\underline{i}} = \gamma^{-1}$, then either of (1) or (7) can be used to obtain a useful lower bound for γ whenever the resulting bound for χ is less than one. However, there are other results, due primarily

to Ostrowski [5] (but see also [4]), for obtaining lower bounds for γ for any nonnegative irreducible matrix.

Acknowledgment. The author would like to thank Bob Anderssen for useful discussions that led to this work.

REFERENCES

[1] G. A. LATHAM AND R. S. ANDERSSEN, *Assessing quantification for the* EMS *algorithm*, Linear Algebra Appl., 4th special issue on linear algebra and statistics,, 210 (1994), pp. 89–122.
[2] M. MARCUS AND H. MINC, *Modern University Algebra*, Macmillan, New York, 1966.
[3] H. MINC, *On the maximal eigenvector of a positive matrix*, SIAM J. Math. Anal., 7 (1970), pp. 424–427.
[4] H. MINC, *Nonnegative matrices*, John Wiley and Sons, New York, 1988.
[5] A.M. OSTROWSKI, *On the eigenvector belonging to the maximal root of a nonnegative matrix*, Proc. Edinburgh Math. Soc., 12 (1960–1961), pp. 107–112.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 312–326, January 1995

022

ON THE USE OF CERTAIN MATRIX ALGEBRAS ASSOCIATED WITH DISCRETE TRIGONOMETRIC TRANSFORMS IN MATRIX DISPLACEMENT DECOMPOSITION*

ENRICO BOZZO† AND CARMINE DI FIORE‡

Abstract. The authors extend some recent results of Di Fiore and Zellini [*Linear Algebra Appl.*, to appear], obtaining new classes of formulas for the displacement operator-based decomposition of matrices. It is shown how an arbitrary matrix can be expressed as the sum of products of matrices belonging to matrix algebras associated with certain versions of sine and cosine transforms. Applications to the representation of the inverse of a Toeplitz and a Toeplitz plus Hankel matrix, with and without symmetry, are presented. Implications on the computation of the product of these matrices by a vector are discussed.

Key words. Toeplitz matrices, Toeplitz plus Hankel matrices, inversion formulas, displacement operators, sine transform, cosine transform

AMS subject classifications. 65F05, 65T20, 15A09

1. Introduction.

The concept of displacement rank, introduced in [18] and thereafter studied by various authors [2], [3], [5], [9], [13]–[16] has important applications in the field of computations with Toeplitz and other types of dense structured matrices both in a sequential and a parallel environment.

The main idea is as follows. Given a matrix A we look for a linear operator $\mathcal{D}$ such that we can easily recover A from its image $\mathcal{D}(A)$ in terms of simple structured matrices. Formulas of Gohberg–Semencul [16], Ammar and Gader [2], Bini and Pan [9], [5], and Gohberg and Olshevsky [14] are all examples of the application of this strategy. These formulas involve various kinds of matrices belonging to commutative algebras (Toeplitz triangulars, circulants, and others) that can be efficiently multiplied by a vector using the tool of the fast Fourier transform.

In a recent paper on this subject [12], Di Fiore and Zellini study the representation of an arbitrary square matrix as the sum of products of matrices belonging to matrix algebras generated by a Hessenberg matrix (Hessenberg algebras (HA)). This way, they are able to find a unifying approach to derive all the formulas listed above. Moreover, they propose new formulas in which an arbitrary $n \times n$ matrix is expressed as the sum of products of τ-class matrices and of matrices having a τ-class submatrix of order $n-1$ or $n-2$.

As is well known, τ-class is a commutative matrix algebra widely used in the study of spectral and computational properties of Toeplitz matrices [6], [7], [23]. Matrices belonging to τ-class are simultaneously diagonalized by the matrix of a discrete transform known as sine transform. For this reason τ-class is said to be *associated* with sine transform. The computations of the sine transform and of the Fourier transform of a real vector have approximately the same cost [19]. Besides the sine transform, other important discrete trigonometric transforms are the cosine transform [19] and the Hartley transform [8]. Different versions of these transforms exist [22] and various fast algorithms for their computation have been proposed.

* Received by the editors March 3, 1993; accepted for publication (in revised form) by R. W. Freund, November 4, 1993.

† Dipartimento di Informatica, Università di Pisa, Corso Italia 40, 56125 Pisa, Italy (bozzo@di.unipi.it).

‡ Dipartimento di Matematica, Università di Roma "La Sapienza," P. le Aldo Moro 2, 00185 Roma, Italy.

In this paper we extend some results found in [12], obtaining new classes of decomposition formulas. As an important case of the new formulas, we show how an arbitrary matrix can be expressed as the sum of products of matrices belonging to algebras associated with certain versions of sine and cosine transforms. In particular, the algebras that we consider are those generated by

$$T_{\varepsilon\varphi} = \begin{pmatrix} \varepsilon & 1 & 0 & \cdots & 0 \\ 1 & 0 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 & 1 \\ 0 & \cdots & 0 & 1 & \varphi \end{pmatrix} \tag{1}$$

for some values of $\varepsilon, \varphi \in \{0, 1, -1\}$. Obviously these algebras are HA, and for these we use the symbol $\tau_{\varepsilon\varphi}$.

The formulas presented here fit naturally in the framework that Di Fiore and Zellini set and have the computational advantage that all the transforms to be computed have the same size, say n. Actually, fast transform algorithms achieve their best efficiency when n is a power of two, and the need for computing transforms of sizes n and $n-1$ or n and $n-2$ may be a serious drawback as noted in [19]. In addition, in the particular case of an $n \times n$ real symmetric Toeplitz matrix T whose $(n-1)$-order principal minor is nonzero (for example, a real positive definite Toeplitz matrix), we are able to exploit our formulas to compute the product of T^{-1} by a vector with eight real fast Fourier transforms of order n, thus improving both [12] and [2] and matching the best result known so far given in [3].

The paper is organized as follows. In §2 we recall some results from [12]. In §3 we obtain formulas for representing a matrix as the sum of products of matrices belonging to HAs generated by persymmetric, symmetric, persymmetric tridiagonal, and symmetric-persymmetric Hessenberg matrices. In §4 we introduce $\tau_{\varepsilon\varphi}$ classes and the transforms with which they are associated. In §5 we discuss applications to Toeplitz and Toeplitz plus Hankel matrices, and in §6 we discuss the computational implications of the derived formulas.

We borrow the notations from [12]. In particular throughout the paper we indicate with $\mathbb{R}$ and $\mathbb{C}$ the real and the complex field, respectively, with $\mathbf{e}_k$ the kth vector of the canonical basis, for $k = 1, \ldots, n$, with J the reversion matrix $J = (\delta_{i\,n-j+1})$, and with Z the downshift matrix $Z = (\delta_{i\,j+1})$, $i, j = 1, \ldots, n$. Given a vector $\mathbf{x}$, we indicate $J\mathbf{x}$ with $\hat{\mathbf{x}}$. Unless otherwise stated matrices are square of order n.

2. Preliminaries.

In this section we recall some results from [12].

2.1. Hessenberg algebras. Let $A \in \mathbb{C}^{n\times n}$ and let $H_A = \{\sum_{k=0}^{n-1} a_k A^k\}$, where the a_k are complex parameters, be the algebra generated by A. Standard properties of H_A are the following [11].

PROPOSITION 2.1. 1. *Every matrix in H_A is symmetric (persymmetric, centrosymmetric, diagonalizable) if and only if A is symmetric (persymmetric, centrosymmetric, diagonalizable).*

2. *The dimension of H_A is equal to the degree of the minimal polynomial of A. In particular* $\dim H_A = n$ *if and only if A is nonderogatory.*

Now let $X \in \mathbb{C}^{n\times n}$ be a Hessenberg matrix

$$X = \begin{pmatrix} r_{11} & b_1 & \cdots & 0 \\ r_{21} & r_{22} & \ddots & \vdots \\ \vdots & & \ddots & b_{n-1} \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix}. \tag{2}$$

The algebra H_X is called Hessenberg algebra (HA). Note that if $b_i \neq 0$ for all i, then X is nonderogatory so that in this hypothesis dim $H_X = n$.

Generalizing some ideas presented in [4], Di Fiore and Zellini are able to provide a convenient basis for H_X. Let $A_k = p_{k-1}(X)$, where $p_k(\lambda)$, for $k = 1, \ldots, n$, is the characteristic polynomial of the $k \times k$ top-left submatrix of X and $p_0(\lambda) = 1$. Clearly $H_X = \{\sum_{k=1}^{n} a_k A_k\}$. Moreover, the matrices A_k have the following important property.

THEOREM 2.2 ([12]). *We have*

$$\mathbf{e}_1^T A_k = \mathbf{e}_k^T \prod_{i=1}^{k-1} b_i.$$

Thus, if $b_i \neq 0$ for all i we can set $X_k = \left(1/\prod_{i=1}^{k-1} b_i\right) A_k$. Being the A_k a basis for H_X, the X_k are a basis as well. Given a vector $\mathbf{a} = (a_k) \in \mathbb{C}^n$ we set

$$H_X(\mathbf{a}) = \sum_{k=1}^{n} a_k X_k,$$

this notation being motivated from the fact that $\mathbf{e}_1^T H_X(\mathbf{a}) = \mathbf{a}^T$. The following properties of the X_k are important.

THEOREM 2.3 ([12]). *Let $b_i \neq 0$ for all i.*

1. *We have $X_j X_k = X_k X_j = \sum_{i=1}^{n} [X_j]_{ki} X_i$ and therefore $\mathbf{e}_j^T X_k = \mathbf{e}_k^T X_j$. If X is symmetric then $X_k \mathbf{e}_j = X_j \mathbf{e}_k$. If X is persymmetric then $X_k \mathbf{e}_j = X_{n+1-j} \mathbf{e}_{n+1-k}$.*
2. *If X is symmetric X_n is nonsingular.*
3. *If X is centrosymmetric then $X_n = J$.*

2.2. Orthogonality conditions. Let us recall a result of Gader [13]. Let A be an $n \times n$ matrix with entries in an arbitrary ring $\mathfrak{R}$ with identity and let $P = Z^T + \mathbf{e}_n \mathbf{e}_1^T$. Gader considers the linear operator $\mathcal{C}(A) = A - P^T A P$ and shows that if $\mathcal{C}(A) = \sum_{m=1}^{\alpha} \mathbf{x}_m \mathbf{y}_m^T$, then for the vectors $\mathbf{x}_m$ and $\mathbf{y}_m$ the following *orthogonality conditions* hold: $\sum_{m=1}^{\alpha} (P^k \mathbf{x}_m)^T \mathbf{y}_m = 0$, for $k = 0, \ldots, n-1$.

Di Fiore and Zellini obtain a generalization of this result for linear operators of the form

$$\mathcal{D}_V(A) = AV - VA,$$

assuming only that V is a matrix with entries in $C(\mathfrak{R}) = \{r \in \mathfrak{R} | sr = rs \text{ for all } s \in \mathfrak{R}\}$. They prove what follows.

LEMMA 2.1 ([12]). *We have*

$$\sum_{i,j=1}^{n} [\mathcal{D}_V(A)]_{ij} [p(V^T)]_{ij} = \sum_{i,j=1}^{n} [p(V^T)]_{ij} [\mathcal{D}_V(A)]_{ij} = 0,$$

where p is any polynomial whose coefficients are in $\mathfrak{R}$.

Lemma 2.1 leads to the following result.

THEOREM 2.4 ([12]). *If $\mathcal{D}_V(A) = \sum_{m=1}^{\alpha} \mathbf{x}_m \mathbf{y}_m^T$ then*

$$\sum_{m=1}^{\alpha} (p(V)\mathbf{x}_m)^T \mathbf{y}_m = 0,$$

where p is any polynomial whose coefficients are in $C(\mathfrak{R})$.

The orthogonality conditions stated in Theorem 2.4 are used for the proof of a number of decomposition formulas both in [12] and in §3.

3. Displacement decomposition formulas. In this section we describe how a square matrix A with entries in an arbitrary ring $\mathfrak{R}$ with identity (so that, for example, we allow the elements of A to be matrices) can be expressed as the sum of products of matrices belonging to particular HA. Let X be the Hessenberg matrix defined in (2) with $[X]_{ij} \in \mathfrak{R}$ and consider the linear operator

$$\mathcal{D}_X(A) = AX - XA.$$

Let $C(\mathfrak{R}) = \{r \in \mathfrak{R} \mid sr = rs \text{ for all } s \in \mathfrak{R}\}$. Observe that $C(\mathfrak{R})$ is a commutative ring with identity (the identity of $\mathfrak{R}$). Throughout this section we make the following assumptions.

(i) The entries of X are in $C(\mathfrak{R})$.

(ii) b_i has inverse in $\mathfrak{R}$ for all i.

(iii) The kernel of $\mathcal{D}_X$ is $H_X = \{\sum_{k=1}^{n} a_k X_k \mid a_k \in \mathfrak{R}\}$.

By using (i) and (ii) it is possible to generalize the results in §2.1 and, in particular, Theorem 2.3 to the case where X has entries in $C(\mathfrak{R})$. Clearly, each matrix X_k is a polynomial in X with coefficients in $C(\mathfrak{R})$, i.e., $X_k \in C(\mathfrak{R})^{n\times n}$. Moreover, if X_k is invertible in $C(\mathfrak{R})^{n\times n}$ (see [10, p. 16]) then X_k^{-1} is a polynomial in X with coefficients in $C(\mathfrak{R})$. This can be easily proved by applying the Cayley–Hamiltom theorem to the matrix X_k (remember that the Cayley–Hamilton theorem is valid for matrices with entries over any commutative ring; see [10]). In particular, if X is tridiagonal and $r_{i+1\,i}$, for $i = 1, \ldots, n-1$, have inverse in $\mathfrak{R}$, then X_n is invertible in $C(\mathfrak{R})^{n\times n}$. In the case where X is symmetric, the proof of this assertion is in [12]. In our case the proof is analogous. If $\mathfrak{R} = \mathbb{R}$ or $\mathfrak{R} = \mathbb{C}$, (iii) is implied by (ii) and (i) is always verified.

In Theorem 3.1(i) of [12] a class of decomposition formulas is obtained involving the HA H_X and $H_{X'}$, with X' defined by $X = X' + (r_{n1} - \beta)\mathbf{e}_n \mathbf{e}_1^T$ and X persymmetric. These formulas include the Gohberg–Olshevsky type of formulas exploiting ϵ-circulant matrices [14].

We first generalize this result by altering a generic element in the secondary diagonal of X (Theorem 3.1). Second, we apply this technique in the principal diagonal of a symmetric (Theorem 3.2) or persymmetric-tridiagonal matrix X (Theorems 3.3 and 3.4) obtaining new classes of decomposition formulas that include, as particular cases, formulas involving algebras associated to fast discrete transforms (see §4). This way, we obtain new representations of the inverses of Toeplitz or Toeplitz plus Hankel matrices for some aspects better than the ones in [2] and [12] (§§5 and 6).

Set

$$X = X' + (r_{n+1-j\,j} - \beta)\mathbf{e}_{n+1-j}\mathbf{e}_j^T,$$

where j is any fixed number in $\{1,\ldots,\lfloor \frac{n}{2} \rfloor + 1\}$ and $\beta \in C(\mathfrak{R})$. Moreover, in the case where n is even and $j = \frac{n}{2} + 1$ ($r_{\frac{n}{2}\,\frac{n}{2}+1} = b_{\frac{n}{2}}$), β must be also invertible in $\mathfrak{R}$.

THEOREM 3.1. *If X is persymmetric, X_j and X'_j are invertible in $C(\Re)^{n\times n}$, then the equalities $\mathcal{D}_X(A) = \sum_{m=1}^{\alpha} \mathbf{x}_m \mathbf{y}_m^T$ and $\mathcal{D}_{X^T}(A) = \sum_{m=1}^{\alpha} \mathbf{x}_m \mathbf{y}_m^T$ imply, respectively,*

$$
\begin{aligned}
&(r_{n+1-j\,j} - \beta)A \\
(3)\quad &= \sum_{m=1}^{\alpha} H_{X'}(JX_j'^{-1}\mathbf{x}_m) H_X(X_j^{-T}\mathbf{y}_m) + (r_{n+1-j\,j} - \beta) H_X(X_j^{-T} A^T \mathbf{e}_j) \\
(4)\quad &= -\sum_{m=1}^{\alpha} H_X(JX_j^{-1}\mathbf{x}_m) H_{X'}(X_j'^{-T}\mathbf{y}_m) + (r_{n+1-j\,j} - \beta) H_X(JX_j^{-1} A \mathbf{e}_{n+1-j})
\end{aligned}
$$

and

$$
\begin{aligned}
&(r_{n+1-j\,j} - \beta)A \\
(5)\quad &= -\sum_{m=1}^{\alpha} H_X(X_j^{-T}\mathbf{x}_m)^T H_{X'}(JX_j'^{-1}\mathbf{y}_m)^T + (r_{n+1-j\,j} - \beta) H_X(X_j^{-T} A \mathbf{e}_j)^T \\
(6)\quad &= \sum_{m=1}^{\alpha} H_{X'}(X_j'^{-T}\mathbf{x}_m)^T H_X(JX_j^{-1}\mathbf{y}_m)^T + (r_{n+1-j\,j} - \beta) H_X(JX_j^{-1} A^T \mathbf{e}_{n+1-j})^T.
\end{aligned}
$$

Proof. For (3) using the linearity of $\mathcal{D}_X$ and the persymmetry of X and X', we have

$$
\begin{aligned}
&\mathcal{D}_X\left(\sum_{m=1}^{\alpha} H_{X'}(JX_j'^{-1}\mathbf{x}_m) H_X(X_j^{-T}\mathbf{y}_m)\right) \\
&= (r_{n+1-j\,j} - \beta) \sum_{m=1}^{\alpha} \Big(H_{X'}(JX_j'^{-1}\mathbf{x}_m)\mathbf{e}_{n+1-j}\mathbf{e}_j^T \\
&\qquad - \mathbf{e}_{n+1-j}\mathbf{e}_j^T H_{X'}(JX_j'^{-1}\mathbf{x}_m)\Big) H_X(X_j^{-T}\mathbf{y}_m) \\
&= (r_{n+1-j\,j} - \beta)\left(\sum_{m=1}^{\alpha} \mathbf{x}_m\mathbf{y}_m^T - \mathbf{e}_{n+1-j} \sum_{m=1}^{\alpha} \mathbf{x}_m^T J H_X(X_j^{-T}\mathbf{y}_m)\right) \\
&= (r_{n+1-j\,j} - \beta) \sum_{m=1}^{\alpha} \mathbf{x}_m\mathbf{y}_m^T.
\end{aligned}
$$

The last equality follows from the following relations which hold, for $i = 1, \ldots, n$, by Theorems 2.3 and 2.4

$$
\begin{aligned}
(7)\qquad &\sum_{m=1}^{\alpha} \mathbf{x}_m^T J H_X(X_j^{-T}\mathbf{y}_m)\mathbf{e}_i = \sum_{m=1}^{\alpha} \mathbf{x}_m^T J \sum_{k=1}^{n} [X_j^{-T}\mathbf{y}_m]_k X_k \mathbf{e}_i \\
&= \sum_{m=1}^{\alpha} \mathbf{x}_m^T J X_{n+1-i} J X_j^{-T}\mathbf{y}_m = \sum_{m=1}^{\alpha} \mathbf{x}_m^T X_{n+1-i}^T X_j^{-T}\mathbf{y}_m = 0.
\end{aligned}
$$

Now, assumption (iii) yields

$$
(r_{n+1-j\,j} - \beta)A - \sum_{m=1}^{\alpha} H_{X'}(JX_j'^{-1}\mathbf{x}_m) H_X(X_j^{-T}\mathbf{y}_m) = H_X(\mathbf{z})
$$

for a vector $\mathbf{z} \in \Re^n$. Since $\mathbf{e}_j^T \sum_{m=1}^{\alpha} H_{X'}(JX_j'^{-1}\mathbf{x}_m)H_X(X_j^{-T}\mathbf{y}_m) = \mathbf{0}^T$ (see (7)), $\mathbf{z}$ is defined by the equality

$$(r_{n+1-j\,j} - \beta)\mathbf{e}_j^T A = \mathbf{z}^T X_j,$$

so that we obtain (3). Regarding (4) we proceed in a similar way. Formulas (5) and (6) follow, respectively, from (3) and (4) and equality $\mathcal{D}_X(A^T) = -\mathcal{D}_{X^T}(A)^T$. □

Observe that if $j = 1$, then $X_j = X_j' = I$. Thus the case $j = 1$ [12] is the most significant.

Set

$$X = X' + (r_{ii} - \beta)\mathbf{e}_i\mathbf{e}_i^T,$$

where i is any fixed number in $\{1,\ldots,n\}$ and $\beta \in C(\Re)$.

THEOREM 3.2. *If X is symmetric, X_i and X_i' are invertible in $C(\Re)^{n\times n}$, then the equality $\mathcal{D}_X(A) = \sum_{m=1}^{\alpha} \mathbf{x}_m\mathbf{y}_m^T$ implies*

$$(8)\quad (r_{ii} - \beta)A = \sum_{m=1}^{\alpha} H_{X'}(X_i'^{-1}\mathbf{x}_m)H_X(X_i^{-1}\mathbf{y}_m) + (r_{ii} - \beta)H_X(X_i^{-1}A^T\mathbf{e}_i)$$

$$(9)\quad = -\sum_{m=1}^{\alpha} H_X(X_i^{-1}\mathbf{x}_m)H_{X'}(X_i'^{-1}\mathbf{y}_m) + (r_{ii} - \beta)H_X(X_i^{-1}A\mathbf{e}_i).$$

Proof. For (8) using the linearity of $\mathcal{D}_X$ and the symmetry of X, we have

$$\mathcal{D}_X\left(\sum_{m=1}^{\alpha} H_{X'}(X_i'^{-1}\mathbf{x}_m)H_X(X_i^{-1}\mathbf{y}_m)\right)$$
$$= (r_{ii} - \beta)\sum_{m=1}^{\alpha}\left(H_{X'}(X_i'^{-1}\mathbf{x}_m)\mathbf{e}_i\mathbf{e}_i^T - \mathbf{e}_i\mathbf{e}_i^T H_{X'}(X_i'^{-1}\mathbf{x}_m)\right)H_X(X_i^{-1}\mathbf{y}_m)$$
$$= (r_{ii} - \beta)\left(\sum_{m=1}^{\alpha}\mathbf{x}_m\mathbf{y}_m^T - \mathbf{e}_i\sum_{m=1}^{\alpha}\mathbf{x}_m^T H_X(X_i^{-1}\mathbf{y}_m)\right) = (r_{ii} - \beta)\sum_{m=1}^{\alpha}\mathbf{x}_m\mathbf{y}_m^T.$$

The last equality follows from the following relations which hold, for $j = 1,\ldots,n$, by Theorems 2.3 and 2.4

$$(10)\quad \sum_{m=1}^{\alpha}\mathbf{x}_m^T H_X(X_i^{-1}\mathbf{y}_m)\mathbf{e}_j = \sum_{m=1}^{\alpha}\mathbf{x}_m^T\sum_{k=1}^{n}[X_i^{-1}\mathbf{y}_m]_k X_k\mathbf{e}_j$$
$$= \sum_{m=1}^{\alpha}\mathbf{x}_m^T X_j X_i^{-1}\mathbf{y}_m = 0.$$

Now, assumption (iii) yields

$$(r_{ii} - \beta)A - \sum_{m=1}^{\alpha} H_{X'}(X_i'^{-1}\mathbf{x}_m)H_X(X_i^{-1}\mathbf{y}_m) = H_X(\mathbf{z})$$

for a vector $\mathbf{z} \in \Re^n$. Since $\mathbf{e}_i^T \sum_{m=1}^{\alpha} H_{X'}(X_i'^{-1}\mathbf{x}_m)H_X(X_i^{-1}\mathbf{y}_m) = \mathbf{0}^T$ (see (10)), the vector $\mathbf{z}$ is defined by the equality

$$(r_{ii} - \beta)\mathbf{e}_i^T A = \mathbf{z}^T X_i$$

so that we obtain (8). The formula (9) follows from (8) and from the equality $\mathcal{D}_X(A^T) = -\mathcal{D}_X(A)^T$. □

Observe that in the cases $i = 1$ and $i = n$, the assumptions on X_i and X_i' are satisfied.

Now set

$$X = X' + (r_{11} - \beta)(\mathbf{e}_1\mathbf{e}_1^T + \mathbf{e}_n\mathbf{e}_n^T),$$

where $\beta \in C(\Re)$.

THEOREM 3.3. *If X is persymmetric and tridiagonal and $r_{i+1\,i}$, for $i = 1, \ldots, n-1$, have inverse in $\Re$, then the equalities $\mathcal{D}_X(A) = \sum_{m=1}^{\alpha} \mathbf{x}_m\mathbf{y}_m^T$ and $\mathcal{D}_{X^T}(A) = \sum_{m=1}^{\alpha} \mathbf{x}_m\mathbf{y}_m^T$ imply, respectively,*

$$\begin{aligned}
&(r_{11} - \beta)(A + X_nAX_n^{-1}) \\
(11)\quad &= \sum_{m=1}^{\alpha} H_{X'}(\hat{\mathbf{x}}_m)H_X(X_n^{-T}\mathbf{y}_m) + (r_{11} - \beta)H_X((A + X_nAX_n^{-1})^T\mathbf{e}_1) \\
(12)\quad &= -\sum_{m=1}^{\alpha} H_X(\hat{\mathbf{x}}_m)H_{X'}(X_n'^{-T}\mathbf{y}_m) + (r_{11} - \beta)H_X(J(A + X_nAX_n^{-1})\mathbf{e}_n)
\end{aligned}$$

and

$$\begin{aligned}
&(r_{11} - \beta)(A + X_n^{-T}AX_n^T) \\
(13)\quad &= -\sum_{m=1}^{\alpha} H_X(X_n^{-T}\mathbf{x}_m)^T H_{X'}(\hat{\mathbf{y}}_m)^T + (r_{11} - \beta)H_X((A + X_n^{-T}AX_n^T)\mathbf{e}_1)^T \\
(14)\quad &= \sum_{m=1}^{\alpha} H_{X'}(X_n'^{-T}\mathbf{x}_m)^T H_X(\hat{\mathbf{y}}_m)^T + (r_{11} - \beta)H_X(J(A + X_n^{-T}AX_n^T)^T\mathbf{e}_n)^T.
\end{aligned}$$

Proof. For (11) using the linearity of $\mathcal{D}_X$, the persymmetry of X and the equality [12]

$$X_n = X_n' = \left(\delta_{i,\,n+1-j}\Big(\prod_{l=1}^{i-1} b_l\Big)^{-1} \prod_{l=n-i+1}^{n-1} r_{l+1\,l}\right),$$

we have

$$\begin{aligned}
&\mathcal{D}_X\left(\sum_{m=1}^{\alpha} H_{X'}(\hat{\mathbf{x}}_m)H_X(X_n^{-T}\mathbf{y}_m)\right) \\
&= (r_{11} - \beta)\sum_{m=1}^{\alpha}\left(H_{X'}(\hat{\mathbf{x}}_m)(\mathbf{e}_1\mathbf{e}_1^T + \mathbf{e}_n\mathbf{e}_n^T) - (\mathbf{e}_1\mathbf{e}_1^T + \mathbf{e}_n\mathbf{e}_n^T)H_{X'}(\hat{\mathbf{x}}_m)\right)H_X(X_n^{-T}\mathbf{y}_m) \\
&= (r_{11} - \beta)\left(\sum_{m=1}^{\alpha} X_n'\mathbf{x}_m\mathbf{y}_m^T X_n^{-1} + \sum_{m=1}^{\alpha}\mathbf{x}_m\mathbf{y}_m^T\right. \\
&\quad \left. - \mathbf{e}_1\sum_{m=1}^{\alpha}\mathbf{x}_m^T JH_X(X_n^{-T}\mathbf{y}_m) - \mathbf{e}_n\sum_{m=1}^{\alpha}\mathbf{x}_m^T JX_n'H_X(X_n^{-T}\mathbf{y}_m)\right) \\
&= (r_{11} - \beta)\left(\sum_{m=1}^{\alpha} X_n'\mathbf{x}_m\mathbf{y}_m^T X_n^{-1} + \sum_{m=1}^{\alpha}\mathbf{x}_m\mathbf{y}_m^T\right) = (r_{11} - \beta)\mathcal{D}_X(A + X_nAX_n^{-1}).
\end{aligned}$$

The last but one equality follows from (7), which holds also for $j = n$, and from the relation

$$\sum_{m=1}^{\alpha} \mathbf{x}_m^T J X_n' H_X(X_n^{-T}\mathbf{y}_m) = \sum_{m=1}^{\alpha} \mathbf{x}_m^T J H_X(X_n^{-T}\mathbf{y}_m) X_n' = \mathbf{0}^T X_n' = \mathbf{0}^T.$$

Now, assumption (iii) yields

$$(r_{11} - \beta)(A + X_n A X_n^{-1}) - \sum_{m=1}^{\alpha} H_{X'}(\hat{\mathbf{x}}_m) H_X(X_n^{-T}\mathbf{y}_m) = H_X(\mathbf{z})$$

for a vector $\mathbf{z} \in \Re^n$. Since $\mathbf{e}_1^T \sum_{m=1}^{\alpha} H_{X'}(\hat{\mathbf{x}}_m) H_X(X_n^{-T}\mathbf{y}_m) = \mathbf{0}^T$ (see (7) for $j = n$), the vector $\mathbf{z}$ is defined by the equality

$$(r_{11} - \beta)\mathbf{e}_1^T(A + X_n A X_n^{-1}) = \mathbf{z}^T$$

so that we obtain (11). Regarding (12), proceed in a similar way. The formulas (13) and (14) follow, respectively, from (11) and (12) and from the equality $\mathcal{D}_X(A^T) = -\mathcal{D}_{X^T}(A)^T$. □

If X is also symmetric we can state the following theorem.

THEOREM 3.4. *If X is symmetric and persymmetric and $\mathcal{D}_X(A) = \sum_{m=1}^{\alpha} \mathbf{x}_m \mathbf{y}_m^T$, then*

$$(r_{11} - \beta)(A + JAJ)$$

$$= \sum_{m=1}^{\alpha} H_{X'}(\mathbf{x}_m) H_X(\mathbf{y}_m) + (r_{11} - \beta) H_X((A + JAJ)^T \mathbf{e}_1) \tag{15}$$

$$= -\sum_{m=1}^{\alpha} H_X(\mathbf{x}_m) H_{X'}(\mathbf{y}_m) + (r_{11} - \beta) H_X((A + JAJ)\mathbf{e}_1). \tag{16}$$

Proof. Exploit Theorems 2.3 and 3.3. □

4. The algebras $\tau_{\varepsilon\varphi}$. In this section we present some matrix algebras strictly related to τ-class. As τ-class these algebras are associated with discrete trigonometric transforms computable with fast algorithms and for this reason their use in matrix displacement decomposition is attractive. At the end of this section we obtain some corollaries involving these algebras of Theorems 3.2 and 3.4.

Set

$$T_{\varepsilon\varphi} = \begin{pmatrix} \varepsilon & 1 & 0 & \cdots & 0 \\ 1 & 0 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 & 1 \\ 0 & \cdots & 0 & 1 & \varphi \end{pmatrix} \tag{17}$$

and set $H_{T_{\varepsilon\varphi}} = \tau_{\varepsilon\varphi}$. Now consider the following special cases.

Case 1. Consider $\varepsilon = 0$ and $\varphi = 0$. We have τ_{00}=τ-class. The matrix

$$M_{00} = \sqrt{\frac{2}{n+1}} \left(\sin \frac{ij\pi}{n+1}\right), \qquad i, j = 1, \ldots, n \tag{18}$$

is symmetric and orthogonal. Moreover, the following equality holds:

$$M_{00}T_{00}M_{00} = 2\,\mathrm{Diag}\left(\cos\frac{j\pi}{n+1}\right), \qquad j = 1,\dots,n. \tag{19}$$

Case 2. Consider $\varepsilon = \varphi = 1$ and $\varepsilon = \varphi = -1$. The two matrices

$$M_{11} = \sqrt{\frac{2}{n}}\left(k_j \cos\frac{(2i+1)j\pi}{2n}\right), \qquad i, j = 0,\dots,n-1, \tag{20}$$

$$M_{-1-1} = \sqrt{\frac{2}{n}}\left(k_j \sin\frac{(2i-1)j\pi}{2n}\right), \qquad i, j = 1,\dots,n, \tag{21}$$

where $k_j = 1/\sqrt{2}$ for $j = 0$ and $j = n$ and $k_j = 1$ otherwise, are orthogonal. Moreover, the two following relations hold:

$$M_{11}^T T_{11} M_{11} = 2\,\mathrm{Diag}\left(\cos\frac{j\pi}{n}\right), \qquad j = 0,\dots,n-1, \tag{22}$$

$$M_{-1-1}^T T_{-1-1} M_{-1-1} = 2\,\mathrm{Diag}\left(\cos\frac{j\pi}{n}\right), \qquad j = 1,\dots,n. \tag{23}$$

The algebras τ_{11} and τ_{-1-1} are strictly related as well as the matrices M_{11} and M_{-1-1}. In fact setting $D_{-1} = \mathrm{Diag}((-1)^j)$ with $j = 0,\dots,n-1$, we have

$$T_{-1-1} = -D_{-1}T_{11}D_{-1}, \tag{24}$$

$$M_{11}J = D_{-1}M_{-1-1}. \tag{25}$$

Case 3. Consider $\varepsilon = 1$, $\varphi = -1$ and $\varepsilon = -1$, $\varphi = 1$. We set

$$M_{1-1} = \sqrt{\frac{2}{n}}\left(\cos\frac{(2i+1)(2j+1)\pi}{4n}\right), \qquad i, j = 0,\dots,n-1, \tag{26}$$

$$M_{-11} = \sqrt{\frac{2}{n}}\left(\sin\frac{(2i+1)(2j+1)\pi}{4n}\right), \qquad i, j = 0,\dots,n-1. \tag{27}$$

The matrices M_{1-1} and M_{-11} are symmetric and orthogonal and we have

$$M_{1-1}T_{1-1}M_{1-1} = 2\,\mathrm{Diag}\left(\cos\frac{(2j+1)\pi}{2n}\right), \qquad j = 0,\dots,n-1, \tag{28}$$

$$M_{-11}T_{-11}M_{-11} = 2\,\mathrm{Diag}\left(\cos\frac{(2j+1)\pi}{2n}\right), \qquad j = 0,\dots,n-1. \tag{29}$$

Moreover, we have relations analogous to the (24) and (25):

$$T_{1-1} = -D_{-1}T_{-11}D_{-1}, \tag{30}$$

$$M_{1-1}J = D_{-1}M_{-11}. \tag{31}$$

The matrices in (18), (21), and (27) define three versions of the sine transform, while the matrices in (20), (26) define two versions of the cosine transform. All these discrete transforms are computable with fast algorithms [22]. For example, particular

attention has been devoted to the computation of the cosine transform $\mathbf{x} \to M_{11}\mathbf{x}$ and of its inverse $\mathbf{x} \to M_{11}^T\mathbf{x}$, for their importance in image and signal processing [1], [21], [17].

The following proposition will be useful.

PROPOSITION 4.1. *For* $(\varepsilon, \varphi) \in \{(0,0),(1,1),(-1,-1),(1,-1),(-1,1)\}$ *we have*

$$\tau_{\varepsilon\varphi}(\mathbf{x}) = M_{\varepsilon\varphi}\mathrm{Diag}(\mathbf{v}_{\varepsilon\varphi} \circ M_{\varepsilon\varphi}^T\mathbf{x})M_{\varepsilon\varphi}^T,$$

where $\circ$ *denotes entrywise vector product and where* $[\mathbf{v}_{\varepsilon\varphi}]_i = 1/[M_{\varepsilon\varphi}^T\mathbf{e}_1]_i$.

Proof. For $(\varepsilon, \varphi) \in \{(0,0),(1,1),(-1,-1),(1,-1),(-1,1)\}$ the matrices $M_{\varepsilon\varphi}$ are orthogonal. Moreover the vector $M_{\varepsilon\varphi}^T\mathbf{e}_1$ does not have zero entries. Thus (19), (22), (23), (28), and (29) imply that $\tau_{\varepsilon\varphi}(\mathbf{x}) = M_{\varepsilon\varphi}\mathrm{Diag}(\mathbf{y})M_{\varepsilon\varphi}^T$ for suitable $\mathbf{y}$. Thus we have $M_{\varepsilon\varphi}^T\mathbf{x}$=Diag$(\mathbf{y})M_{\varepsilon\varphi}^T\mathbf{e}_1$ and the thesis follows. □

We now rewrite the theorems in §3 choosing H_X and $H_{X'}$ to be $\tau_{\varepsilon\varphi}$ algebras. From Theorem 3.2, $i = 1$, we have the following corollary.

COROLLARY 4.1. *If* $\mathcal{D}_{T_{\varepsilon\varphi}}(A) = \sum_{m=1}^{\alpha} \mathbf{x}_m\mathbf{y}_m^T$ *then*

$$(\varepsilon - \beta)A = \sum_{m=1}^{\alpha} \tau_{\beta\varphi}(\mathbf{x}_m)\tau_{\varepsilon\varphi}(\mathbf{y}_m) + (\varepsilon - \beta)\tau_{\varepsilon\varphi}(A^T\mathbf{e}_1) \tag{32}$$

$$= -\sum_{m=1}^{\alpha} \tau_{\varepsilon\varphi}(\mathbf{x}_m)\tau_{\beta\varphi}(\mathbf{y}_m) + (\varepsilon - \beta)\tau_{\varepsilon\varphi}(A\mathbf{e}_1). \tag{33}$$

An analogous corollary can be derived from Theorem 3.2, $i = n$. We leave this task to the reader. From Theorem 3.4 we deduce the following corollary.

COROLLARY 4.2. *If* $\mathcal{D}_{T_{\varepsilon\varepsilon}}(A) = \sum_{m=1}^{\alpha} \mathbf{x}_m\mathbf{y}_m^T$ *and* $A = JAJ$, *then*

$$2(\varepsilon - \beta)A = \sum_{m=1}^{\alpha} \tau_{\beta\beta}(\mathbf{x}_m)\tau_{\varepsilon\varepsilon}(\mathbf{y}_m) + 2(\varepsilon - \beta)\tau_{\varepsilon\varepsilon}(A^T\mathbf{e}_1)$$

$$= -\sum_{m=1}^{\alpha} \tau_{\varepsilon\varepsilon}(\mathbf{x}_m)\tau_{\beta\beta}(\mathbf{y}_m) + 2(\varepsilon - \beta)\tau_{\varepsilon\varepsilon}(A\mathbf{e}_1).$$

5. Applications to inverses of Toeplitz and Toeplitz plus Hankel matrices. In this section we show how Corollaries 4.1 and 4.2 can be exploited to represent the inverse of a Toeplitz or of a Toeplitz plus Hankel matrix. Our treatment will be concise. We send the reader to [12] for an extensive discussion of the representation of the inverse of a Toeplitz or Toeplitz plus Hankel matrix by means of displacement operator-based decompositions.

5.1. Preliminaries. Let $T = (t_{i-j})$ and $H = (h_{i+j-2})$, with $i, j = 1, \ldots, n$, be a Toeplitz and a Hankel matrix. Let T and $T + H$ be nonsingular, let $S = T^{-1}$ and $W = (T + H)^{-1}$, and let $\mathbf{s}_{\cdot i}$, $\mathbf{s}_{i\cdot}$ ($\mathbf{w}_{\cdot i}$, $\mathbf{w}_{i\cdot}$) be, respectively, the ith column and row of S (W). Set

$$\mathbf{a} = (0 \quad t_{-n+1} \quad \ldots \quad t_{-1})^T,$$
$$\mathbf{b} = (t_1 \quad \ldots \quad t_{n-1} \quad 0)^T,$$
$$\mathbf{c} = (0 \quad h_0 \quad \ldots \quad h_{n-2})^T,$$
$$\mathbf{d} = (h_n \quad \ldots \quad h_{2n-2} \quad 0)^T.$$

If $\gamma = S\mathbf{a}$ and $\delta = S\mathbf{b}$, then we have [16]

$$\mathcal{D}_Z(S) = SZ - ZS = \gamma \hat{\mathbf{s}}_{\cdot 1}^T - \mathbf{s}_{\cdot 1}\hat{\gamma}^T,$$
$$\mathcal{D}_{Z^T}(S) = SZ^T - Z^T S = \delta \hat{\mathbf{s}}_{\cdot n}^T - \mathbf{s}_{\cdot n}\hat{\delta}^T,$$

and, consequently, we deduce the following proposition.

PROPOSITION 5.1. *We have*

$$\mathcal{D}_{T_{\varepsilon\varphi}}(S) = (\gamma - \varphi\mathbf{e}_n)\hat{\mathbf{s}}_{\cdot 1}^T - \mathbf{s}_{\cdot 1}(\hat{\gamma}^T - \varepsilon\mathbf{e}_1^T) + (\delta - \varepsilon\mathbf{e}_1)\hat{\mathbf{s}}_{\cdot n}^T - \mathbf{s}_{\cdot n}(\hat{\delta}^T - \varphi\mathbf{e}_n^T).$$

Proof. Exploit the equality $T_{\varepsilon\varphi} = Z + Z^T + \varepsilon\mathbf{e}_1\mathbf{e}_1^T + \varphi\mathbf{e}_n\mathbf{e}_n^T$. □

Analogously setting

$$\mathbf{x}_1 = W(\mathbf{b} + \mathbf{c}), \qquad \mathbf{x}_2 = W(\mathbf{a} + \mathbf{d}),$$
$$\mathbf{x}_3 = W^T(\hat{\mathbf{a}} + \mathbf{c}), \qquad \mathbf{x}_4 = W^T(\hat{\mathbf{b}} + \mathbf{d}),$$

it is known that [16]

$$\mathcal{D}_{T_{00}}(W) = \mathbf{x}_1\mathbf{w}_{1\cdot}^T + \mathbf{x}_2\mathbf{w}_{n\cdot}^T - \mathbf{w}_{\cdot 1}\mathbf{x}_3^T - \mathbf{w}_{\cdot n}\mathbf{x}_4^T.$$

It follows that Proposition 5.2 is true.

PROPOSITION 5.2. *We have*

$$\mathcal{D}_{T_{\varepsilon\varphi}}(W) = (\mathbf{x}_1 - \varepsilon\mathbf{e}_1)\mathbf{w}_{1\cdot}^T + (\mathbf{x}_2 - \varphi\mathbf{e}_n)\mathbf{w}_{n\cdot}^T - \mathbf{w}_{\cdot 1}(\mathbf{x}_3^T - \varepsilon\mathbf{e}_1^T) - \mathbf{w}_{\cdot n}(\mathbf{x}_4^T - \varphi\mathbf{e}_n^T).$$

5.2. Formulas for Toeplitz inverses. At this point the following result is immediate.

THEOREM 5.1. *We have*

$$(\varepsilon - \beta)S = \tau_{\beta\varphi}(\gamma - \varphi\mathbf{e}_n)\tau_{\varepsilon\varphi}(\hat{\mathbf{s}}_{\cdot 1}) - \tau_{\beta\varphi}(\mathbf{s}_{\cdot 1})\tau_{\varepsilon\varphi}(\hat{\gamma} - \varepsilon\mathbf{e}_1) \tag{34}$$
$$+ \tau_{\beta\varphi}(\delta - \beta\mathbf{e}_1)\tau_{\varepsilon\varphi}(\hat{\mathbf{s}}_{\cdot n}) - \tau_{\beta\varphi}(\mathbf{s}_{\cdot n})\tau_{\varepsilon\varphi}(\hat{\delta} - \varphi\mathbf{e}_n)$$
$$= -\tau_{\varepsilon\varphi}(\gamma - \varphi\mathbf{e}_n)\tau_{\beta\varphi}(\hat{\mathbf{s}}_{\cdot 1}) + \tau_{\varepsilon\varphi}(\mathbf{s}_{\cdot 1})\tau_{\beta\varphi}(\hat{\gamma} - \beta\mathbf{e}_1) \tag{35}$$
$$- \tau_{\varepsilon\varphi}(\delta - \varepsilon\mathbf{e}_1)\tau_{\beta\varphi}(\hat{\mathbf{s}}_{\cdot n}) + \tau_{\varepsilon\varphi}(\mathbf{s}_{\cdot n})\tau_{\beta\varphi}(\hat{\delta} - \varphi\mathbf{e}_n).$$

Proof. Use Proposition 5.1 and Corollary 4.1. □

Now let T be symmetric. In this case we have $\gamma = \hat{\delta}$ and $\hat{\mathbf{s}}_{n-i+1\cdot} = \mathbf{s}_{i\cdot}$ for all i. The above theorem can be readily rewritten in this particular case.

THEOREM 5.2. *If T is symmetric we have*

$$(\varepsilon - \beta)S$$
$$= (\tau_{\beta\varepsilon}(\gamma - \varepsilon\mathbf{e}_n)J + \tau_{\beta\varepsilon}(\hat{\gamma} - \beta\mathbf{e}_1))\tau_{\varepsilon\varepsilon}(\mathbf{s}_{\cdot 1}) - (\tau_{\beta\varepsilon}(\mathbf{s}_{\cdot 1})J + \tau_{\beta\varepsilon}(\hat{\mathbf{s}}_{\cdot 1}))\tau_{\varepsilon\varepsilon}(\gamma - \varepsilon\mathbf{e}_n)$$
$$= -\tau_{\varepsilon\varepsilon}(\gamma - \varepsilon\mathbf{e}_n)(\tau_{\beta\varepsilon}(\hat{\mathbf{s}}_{\cdot 1}) + J\tau_{\beta\varepsilon}(\mathbf{s}_{\cdot 1})) + \tau_{\varepsilon\varepsilon}(\mathbf{s}_{\cdot 1})(\tau_{\beta\varepsilon}(\hat{\gamma} - \beta\mathbf{e}_1) + J\tau_{\beta\varepsilon}(\gamma - \varepsilon\mathbf{e}_n)).$$

Proof. Set $\varphi = \varepsilon$ in Theorem 5.1 and use the fact that $J \in \tau_{\varepsilon\varepsilon}$. □

However, if T is symmetric we can more conveniently exploit Corollary 4.2.

THEOREM 5.3. *If T is symmetric we have*

$$(\varepsilon - \beta)S = \tau_{\beta\beta}(\hat{\gamma} - \beta\mathbf{e}_1)\tau_{\varepsilon\varepsilon}(\mathbf{s}_{\cdot 1}) - \tau_{\beta\beta}(\mathbf{s}_{\cdot 1})\tau_{\varepsilon\varepsilon}(\hat{\gamma} - \varepsilon\mathbf{e}_1) \tag{36}$$
$$= -\tau_{\varepsilon\varepsilon}(\hat{\gamma} - \varepsilon\mathbf{e}_1)\tau_{\beta\beta}(\mathbf{s}_{\cdot 1}) + \tau_{\varepsilon\varepsilon}(\mathbf{s}_{\cdot 1})\tau_{\beta\beta}(\hat{\gamma} - \beta\mathbf{e}_1). \tag{37}$$

Remark 1. It is well known [16] that if S is such that $s_{11} \neq 0$ then

$$\gamma = -\frac{1}{s_{11}} Z\mathbf{s}_{\cdot n}.$$

We refer the reader to [12] for a generalization of the preceding relation for S arbitrary.

5.3. Formulas for Toeplitz plus Hankel inverses. We turn now to the case of a general Toeplitz plus Hankel matrix.

THEOREM 5.4. *We have*

$$(\varepsilon-\beta)W = \tau_{\beta\varphi}(\mathbf{x}_1-\beta\mathbf{e}_1)\tau_{\varepsilon\varphi}(\mathbf{w}_1\,.) + \tau_{\beta\varphi}(\mathbf{x}_2-\varphi\mathbf{e}_n)\tau_{\varepsilon\varphi}(\mathbf{w}_n\,.) \tag{38}$$
$$- \tau_{\beta\varphi}(\mathbf{w}_{\cdot\,1})\tau_{\varepsilon\varphi}(\mathbf{x}_3-\varepsilon\mathbf{e}_1) - \tau_{\beta\varphi}(\mathbf{w}_{\cdot\,n})\tau_{\varepsilon\varphi}(\mathbf{x}_4-\varphi\mathbf{e}_n)$$
$$= -\tau_{\varepsilon\varphi}(\mathbf{x}_1-\varepsilon\mathbf{e}_1)\tau_{\beta\varphi}(\mathbf{w}_1\,.) - \tau_{\varepsilon\varphi}(\mathbf{x}_2-\varphi\mathbf{e}_n)\tau_{\beta\varphi}(\mathbf{w}_n\,.) \tag{39}$$
$$+ \tau_{\varepsilon\varphi}(\mathbf{w}_{\cdot\,1})\tau_{\beta\varphi}(\mathbf{x}_3-\beta\mathbf{e}_1) + \tau_{\varepsilon\varphi}(\mathbf{w}_{\cdot\,n})\tau_{\beta\varphi}(\mathbf{x}_4-\varphi\mathbf{e}_n).$$

Proof. We use Proposition 5.2 and Corollary 4.1. □

If $T = T^T$ then $T+H$ is symmetric and we have $\mathbf{w}_{i\,\cdot} = \mathbf{w}_{\cdot\,i}$, $\mathbf{x}_1 = \mathbf{x}_3$, and $\mathbf{x}_2 = \mathbf{x}_4$. If $JHJ = H$ then $T+H$ is persymmetric and in this case we have $\mathbf{w}_{i\,\cdot} = \hat{\mathbf{w}}_{\cdot\,n-i+1}$, $\hat{\mathbf{x}}_3 = \mathbf{x}_2$, and $\hat{\mathbf{x}}_4 = \mathbf{x}_1$. If both these conditions hold, then $T+H$ is centrosymmetric and we can rewrite the preceding theorem as follows.

THEOREM 5.5. *If $T+H$ is centrosymmetric we have*

$$\begin{aligned}(\varepsilon-\beta)W &= (\tau_{\beta\varepsilon}(\mathbf{x}_1-\beta\mathbf{e}_1) + \tau_{\beta\varepsilon}(\hat{\mathbf{x}}_1-\varepsilon\mathbf{e}_n)J)\tau_{\varepsilon\varepsilon}(\mathbf{w}_{\cdot\,1})\\ &\quad - (\tau_{\beta\varepsilon}(\mathbf{w}_{\cdot\,1}) + \tau_{\beta\varepsilon}(\hat{\mathbf{w}}_{\cdot\,1})J)\tau_{\varepsilon\varepsilon}(\mathbf{x}_1-\varepsilon\mathbf{e}_1)\\ &= -\tau_{\varepsilon\varepsilon}(\mathbf{x}_1-\varepsilon\mathbf{e}_1)(\tau_{\beta\varepsilon}(\mathbf{w}_{\cdot\,1}) + J\tau_{\beta\varepsilon}(\hat{\mathbf{w}}_{\cdot\,1}))\\ &\quad + \tau_{\varepsilon\varepsilon}(\mathbf{w}_{\cdot\,1})(\tau_{\beta\varepsilon}(\mathbf{x}_1-\beta\mathbf{e}_1) + J\tau_{\beta\varepsilon}(\hat{\mathbf{x}}_1-\varepsilon\mathbf{e}_n)).\end{aligned}$$

Proof. Set $\varphi = \varepsilon$ in Theorem 5.4, and use the fact that $J \in \tau_{\varepsilon\varepsilon}$. □

However, a more convenient expression for W can be obtained by means of Corollary 4.2.

THEOREM 5.6. *If $T+H$ is centrosymmetric we have*

$$(\varepsilon-\beta)W = \tau_{\beta\beta}(\mathbf{x}_1-\beta\mathbf{e}_1)\tau_{\varepsilon\varepsilon}(\mathbf{w}_{\cdot\,1}) - \tau_{\beta\beta}(\mathbf{w}_{\cdot\,1})\tau_{\varepsilon\varepsilon}(\mathbf{x}_1-\varepsilon\mathbf{e}_1) \tag{40}$$
$$= -\tau_{\varepsilon\varepsilon}(\mathbf{x}_1-\varepsilon\mathbf{e}_1)\tau_{\beta\beta}(\mathbf{w}_{\cdot\,1}) + \tau_{\varepsilon\varepsilon}(\mathbf{w}_{\cdot\,1})\tau_{\beta\beta}(\mathbf{x}_1-\beta\mathbf{e}_1). \tag{41}$$

Proof. Use Proposition 5.2 and Corollary 4.2. □

6. Applications to the solution of real Toeplitz and Toeplitz plus Hankel systems of equations. As is well known, many algorithms for the solution of a Toeplitz system of equations $T\mathbf{z} = \mathbf{b}$ have the following two-stage structure.

Stage 1. Given T, compute the information relevant to obtain a displacement representation for $S = T^{-1}$. For example, if T is symmetric and $s_{11} \neq 0$, only $\mathbf{s}_{\cdot\,1}$ must be computed by virtue of Remark 1.

Stage 2. Compute $\mathbf{z} = S\mathbf{b}$, exploiting the properties of the matrices involved in the displacement representation of S.

Obviously the same kind of algorithm can be used for the solution of a Toeplitz plus Hankel system of equations $(T+H)\mathbf{z} = \mathbf{b}$.

In this section we look for an efficient implementation of stage 2 using the formulas of §5 in the case of real systems of equations, making, when possible, a comparison with the literature on the subject [15], [2], [3], [12].

As is customary, we will use the number of fast Fourier transforms (FFT) that must be executed as a measure of cost for our algorithms. More precisely with $\mathrm{FFT}_{\mathbb{R}}(n)$ we denote the cost of the FFT of a real vector of order n. Remember that $\mathrm{FFT}_{\mathbb{R}}(2n) \simeq 2\mathrm{FFT}_{\mathbb{R}}(n)$. Moreover, the computations of the trigonometric transforms $\mathbf{x} \to M_{11}^T\mathbf{x}$, $\mathbf{x} \to M_{11}\mathbf{x}$, $\mathbf{x} \to M_{-11}^T\mathbf{x}$, and $\mathbf{x} \to M_{-11}\mathbf{x}$, where $\mathbf{x}$ is a real vector, also have a cost $\simeq \mathrm{FFT}_{\mathbb{R}}(n)$ [21], [22], [17].

It is worthwhile to distinguish between computations that involve the vector $\mathbf{b}$ and computations that involve only entries of S, which can be embodied in the *preconditioning* stage [15] if any is performed. So we use the notation $x\mathrm{FFT}_{\mathbb{R}}(n)+y\mathrm{FFT}_{\mathbb{R}}(n)$ to indicate a cost of $x+y$ real FFTs, y of which can be performed only once if the system must be solved for many different known term vectors.

6.1. Toeplitz and Toeplitz plus Hankel general systems. Using formulas (34)–(35) and (38)–(39) it is possible to compute $S\mathbf{b}$ or $W\mathbf{b}$ with a cost of $10\mathrm{FFT}_{\mathbb{R}}(n) + 8\mathrm{FFT}_{\mathbb{R}}(n) + O(n)$ arithmetic operations. To prove this, consider, for example, formula (39) and set $\varepsilon = \varphi = 1$, $\beta = -1$. By virtue of Proposition 4.1, this implies

$$\begin{aligned} W = \frac{1}{2} M_{11} \{ & -\Lambda_{11}(M_{11}^T(\mathbf{x}_1 - \mathbf{e}_1)) M_{11}^T M_{-11} \Lambda_{-11}(M_{-11}^T \mathbf{w}_{1\cdot}) \\ & - \Lambda_{11}(M_{11}^T(\mathbf{x}_2 - \mathbf{e}_n)) M_{11}^T M_{-11} \Lambda_{-11}(M_{-11}^T \mathbf{w}_{n\cdot}) \\ & + \Lambda_{11}(M_{11}^T \mathbf{w}_{\cdot 1}) M_{11}^T M_{-11} \Lambda_{-11}(M_{-11}^T(\mathbf{x}_3 + \mathbf{e}_1)) \\ & + \Lambda_{11}(M_{11}^T \mathbf{w}_{\cdot n}) M_{11}^T M_{-11} \Lambda_{-11}(M_{-11}^T(\mathbf{x}_4 - \mathbf{e}_n)) \} M_{-11}^T, \end{aligned}$$

where $\Lambda_{\varepsilon\varphi}(\mathbf{x}) = \mathrm{Diag}(\mathbf{v}_{\varepsilon\varphi} \circ \mathbf{x})$. (See Proposition 4.1.)

6.2. Toeplitz symmetric and Toeplitz plus Hankel symmetric and persymmetric systems. If T is symmetric or $T + H$ is symmetric and persymmetric using formulas (36)–(37) and (40)–(41), respectively, it is possible to compute $S\mathbf{b}$ and $W\mathbf{b}$ with the cost of $6\mathrm{FFT}_{\mathbb{R}}(n) + 4\mathrm{FFT}_{\mathbb{R}}(n) + O(n)$. In fact, consider, for example, formula (37) and set $\varepsilon = 1$ and $\beta = -1$. Observe that, by Proposition 4.1, we have

$$\tau_{-1-1}(\mathbf{x}) = M_{-1-1}\mathrm{Diag}(\mathbf{v}_{-1-1} \circ M_{-1-1}^T \mathbf{x}) M_{-1-1}^T,$$

whence, using relation (25) we have

$$\begin{aligned} \tau_{-1-1}(\mathbf{x}) &= D_{-1} M_{11} \mathrm{Diag}(\mathbf{v}_{11} \circ M_{11}^T D_{-1} \mathbf{x}) M_{11}^T D_{-1} \\ &= D_{-1} M_{11} \Lambda_{11}(M_{11}^T D_{-1} \mathbf{x}) M_{11}^T D_{-1}. \end{aligned}$$

Formula (37) becomes

$$\begin{aligned} (42) \quad S = \frac{1}{2} M_{11} \{ & -\Lambda_{11}(M_{11}^T(\hat{\gamma} - \mathbf{e}_1)) M_{11}^T D_{-1} M_{11} \Lambda_{11}(M_{11}^T D_{-1} \mathbf{s}_{\cdot 1}) \\ & + \Lambda_{11}(M_{11}^T \mathbf{s}_{\cdot 1}) M_{11}^T D_{-1} M_{11} \Lambda_{11}(M_{11}^T D_{-1}(\hat{\gamma} + \mathbf{e}_1)) \} M_{11}^T D_{-1}. \end{aligned}$$

Setting $\varepsilon = 1$ and $\beta = -1$ in (41) we have an analogous formula for W. It can be simply obtained from (42) by replacing $\hat{\gamma}$ and $\mathbf{s}_{\cdot 1}$ with $\mathbf{x}_1$ and $\mathbf{w}_{\cdot 1}$, respectively.

The algorithm that can be obtained from the preceding representation of S (W) has substantially the same cost of that suggested in [12]. However in some situations the algorithm has the practical advantage that the transforms to be computed have the same size, say n. This is particularly favourable when n is a power of two. In the algorithm proposed in [12] it is necessary to compute sine transforms of size n and $n-2$ and, if n is a power of two, the computation of a sine transform of order $n-2$ may be more expensive with respect to that of a sine transform of order n.

6.2.1. Toeplitz positive definite systems. Let T be real and symmetric and let $s_{11} \neq 0$. Note that this includes the important special case where T is positive definite. It is possible to compute $S\mathbf{b}$ with a cost of $6\mathrm{FFT}_{\mathbb{R}}(n) + 2\mathrm{FFT}_{\mathbb{R}}(n) + O(n)$.

For this purpose, consider again the representation (42) and take into account that $\hat{\gamma} = -\frac{1}{s_{11}} Z^T \mathbf{s}_{\cdot 1}$. (See Remark 1.) Thus,

$$S = \frac{1}{2s_{11}} M_{11} \left\{ \Lambda_{11}(M_{11}^T(Z^T \mathbf{s}_{\cdot 1} + s_{11}\mathbf{e}_1)) M_{11}^T D_{-1} M_{11} \Lambda_{11}(M_{11}^T D_{-1} \mathbf{s}_{\cdot 1}) \right.$$
$$\left. - \Lambda_{11}(M_{11}^T \mathbf{s}_{\cdot 1}) M_{11}^T D_{-1} M_{11} \Lambda_{11}(M_{11}^T D_{-1}(Z^T \mathbf{s}_{\cdot 1} - s_{11}\mathbf{e}_1)) \right\} M_{11}^T D_{-1}.$$

Let us consider the four transforms

$$\begin{matrix} M_{11}^T \mathbf{s}_{\cdot 1}, & M_{11}^T D_{-1} \mathbf{s}_{\cdot 1}, \\ M_{11}^T Z^T \mathbf{s}_{\cdot 1}, & M_{11}^T D_{-1} Z^T \mathbf{s}_{\cdot 1}. \end{matrix} \tag{43}$$

We note that the following relation holds [1]:

$$M_{11}^T = \sqrt{\frac{2}{n}} \mathrm{Re}\left(\mathrm{Diag}(k_i \rho_{2n}^i)(\rho_n^{ij})\right), \qquad i, j = 0, \ldots, n-1,$$

where $\rho_n = e^{-\mathbf{i}\frac{\pi}{n}}$ being $\mathbf{i}$ the complex unit. The matrix $R = (\rho_n^{ij})$, with $i, j = 0, \ldots, n-1$, is the $n \times n$ left upper corner of the Fourier matrix of order $2n$ defined as $F_{2n} = (\rho_{2n}^{2ij})$, for $i, j = 0, \ldots, 2n-1$. Observe that F_{2n} has the form

$$F_{2n} = \begin{pmatrix} R & D_{-1}R \\ RD_{-1} & (-1)^n D_{-1} R D_{-1} \end{pmatrix}.$$

Thus, if we compute

$$F_{2n} \begin{pmatrix} \mathbf{s}_{\cdot 1} \\ 0 \end{pmatrix} \quad \text{and} \quad F_{2n} Z^T \begin{pmatrix} \mathbf{s}_{\cdot 1} \\ 0 \end{pmatrix} \tag{44}$$

we can recover the transforms in (43) with $O(n)$ arithmetic operations. Moreover

$$F_{2n} Z^T = F_{2n}(P - \mathbf{e}_{2n}\mathbf{e}_1^T),$$

where $P = Z^T + \mathbf{e}_{2n}\mathbf{e}_1^T$ is the unit circulant matrix [11] and is such that $F_{2n}P = \mathrm{Diag}(\rho_n^{-i}) F_{2n}$, $i = 0, \ldots, 2n-1$. Thus

$$F_{2n} Z^T = -F_{2n}\mathbf{e}_{2n}\mathbf{e}_1^T + \mathrm{Diag}(\rho_n^{-i}) F_{2n}.$$

We deduce that once

$$F_{2n} \begin{pmatrix} \mathbf{s}_{\cdot 1} \\ 0 \end{pmatrix}$$

has been computed, the other transform in (44) can be recovered with $O(n)$ arithmetic operations.

The present result improves both the bound $6\mathrm{FFT}_{\mathbb{R}}(n) + 4\mathrm{FFT}_{\mathbb{R}}(n) + O(n)$ given in [12] and the bound $7\mathrm{FFT}_{\mathbb{R}}(n) + 2\mathrm{FFT}_{\mathbb{R}}(n) + O(n)$ given in [2], matching the best result known so far, given in [3].

Acknowledgments. We would like to thank Professors Dario Bini and Paolo Zellini for proposing the problem and for their suggestions.

REFERENCES

[1] N. AHMED, T. NATARAJAN, AND K. RAO, *Discrete cosine transform*, IEEE Trans. Comput., 23 (1974), pp. 90–93.

[2] G. AMMAR AND P. GADER, *A variant of the Gohberg–Semencul formula involving circulant matrices*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 534–540.

[3] ———, *New decompositions of the inverse of a Toeplitz matrix*, in Signal Processing, Scattering and Operator Theory and Numerical Methods, Proc. Internat. Symp. MTNS-89, Vol. 3, Birkäuser, Boston, 1990, pp. 421–428.

[4] R. B. BAPAT AND V. S. SUNDER, *On hypergroups of matrices*, Linear Multilinear Algebra, 29 (1991), pp. 125–140.

[5] D. BINI, *On a class of matrices related to Toeplitz matrices*, Tech. Report TR83-5, Computer Science Dept., State University of New York at Albany, 1983.

[6] D. BINI AND M. CAPOVANI, *Spectral and computational properties of band symmetric Toeplitz matrices*, Linear Algebra Appl., 52/53 (1983), pp. 99–126.

[7] ———, *Tensor rank and border rank of band Toeplitz matrices*, SIAM J. Comput., 16 (1987), pp. 252–258.

[8] D. BINI AND P. FAVATI, *On a matrix algebra related to discrete Hartley transform*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 500–507.

[9] D. BINI AND V. PAN, *Numerical and Algebraic Computations with Matrices and Polynomials*, Birkäuser, Boston, MA, to appear.

[10] W. C. BROWN, *Matrices over Commutative Rings*, Marcel Dekker, New York, 1993.

[11] P. J. DAVIS, *Circulant Matrices*, Wiley, New York, 1979.

[12] C. DI FIORE AND P. ZELLINI, *Matrix decompositions using displacement rank and classes of commutative matrix algebras*, Linear Algebra Appl., to appear.

[13] P. GADER, *Displacement operator based decompositions of matrices using circulants or other group matrices*, Linear Algebra Appl., 139 (1990), pp. 111–131.

[14] I. GOHBERG AND V. OLSHEVSKY, *Circulants, displacements and decompositions of matrices*, Integral Equations Operator Theory, 15 (1992), pp. 730–743.

[15] ———, *Fast algorithms with preprocessing for multiplication of transpose Vandermonde matrix and Cauchy matrix with vectors*, preprint.

[16] G. HEINIG AND K. ROST, *Algebraic Methods for Toeplitz-like Matrices and Operators*, Birkhäuser, Boston, MA, 1984.

[17] H. HOU, *A fast recursive algorithm for computing the discrete cosine transform*, IEEE Trans. Acoust. Speech Signal Process., 35 (1987), pp. 1455–1461.

[18] T. KAILATH, S. KUNG, AND M. MORF, *Displacement rank of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979), pp. 395–407.

[19] W. PRESS, B. FLANNERY, S. TEUKOLSKY, AND W. VETTERLING, *Numerical Recipes. The Art of Scientific Computing*, Cambridge University Press, Cambridge, 1986.

[20] H. SORENSEN, D. JONES, M. HEIDEMAN, AND C. BURRUS, *Real-valued fast Fourier transform algorithms*, IEEE Trans. Acoust. Speech Signal Process., 35 (1987), pp. 849–863.

[21] M. VETTERLI AND H. NUSSBAUMER, *Simple* FFT *and* DCT *algorithms with reduced number of operations*, Signal Process., 6 (1984), pp. 267–278.

[22] Z. WANG, *Fast algorithms for the discrete W transforms and the discrete Fourier transform*, IEEE Trans. Acoust. Speech Signal Process., 32 (1984), pp. 803–816.

[23] P. ZELLINI, *On the optimal computation of a set of symmetric and persymmetric bilinear forms*, Linear Algebra Appl., 23 (1979), pp. 101–119.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 327–332, January 1995

023

NEW PERTURBATION BOUNDS FOR THE UNITARY POLAR FACTOR*

REN-CANG LI†

Abstract. Let A be an $m \times n$ ($m \geq n$) complex matrix. It is known that there is a unique *polar decomposition* $A = QH$, where $Q^*Q = I$, the $n \times n$ identity matrix, and H is positive definite, provided A has full column rank. This note addresses the following question: How much may Q change if A is perturbed? For the square case $m = n$ our bound, which is valid for any unitarily invariant norm, is sharper and simpler than that of Mathias [*SIAM J. Matrix Anal. Appl.*, 14 (1993), pp. 588–597]. For the nonsquare case, a bound is also established for unitarily invariant norm, which has not been done in the literature.

Key words. polar decomposition, perturbation bound, unitarily invariant norm

AMS subject classifications. 15A12, 15A18, 15A23, 15A45

Let A be an $m \times n$ ($m \geq n$) complex matrix. It is known that there are Q with orthonormal column vectors, i.e., $Q^*Q = I$, and a unique positive semidefinite H such that

$$A = QH. \tag{1}$$

Hereafter I denotes an identity matrix with appropriate dimensions that are either specified or that are clear from the context. The decomposition (1) is called the *polar decomposition* of A. If, in addition, A has full column rank, then Q is uniquely determined also. In fact,

$$H = (A^*A)^{1/2}, \qquad Q = A(A^*A)^{-1/2}, \tag{2}$$

where the superscript * denotes conjugate transpose. The decomposition (1) can also be computed from the *singular value decomposition* (SVD) $A = U\Sigma V^*$ by

$$H = V\Sigma_1 V^*, \qquad Q = U_1 V^*, \tag{3}$$

where $U = (U_1, U_2)$ and V are unitary, U_1 is $m \times n$, $\Sigma = \left(\begin{smallmatrix}\Sigma_1\\0\end{smallmatrix}\right)$ and $\Sigma_1 = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$ is nonnegative.

There are published bounds stating how much the two factor matrices Q and H may change if entries of A are perturbed. Among the papers written on this subject are [1], [3], [4], [6]–[10], the perturbation bounds for Q when $m = n$ proved by Mathias [9], cover every unitarily invariant norm, while others cover the Frobenius norm only. Chen and Sun [6], [3] and Li [8] also deal with the case $m \geq n$ as we do here. A surprise is how heavily the sensitivity of the Q factor depends upon whether the working number field is real or complex [1], [7], [9].

In this paper, we obtain some bounds for the perturbations of Q, assuming A is complex. Our bound for the case $m = n$ is achievable and improves on that of Mathias slightly for small perturbations and significantly for big ones.

For the sake of convenience in our presentation, we use A and $\widetilde{A}$ for two matrices having full column rank, one of which is a perturbation of the other. Let

$$A = QH, \qquad \widetilde{A} = \widetilde{Q}\widetilde{H} \tag{4}$$

* Received by the editors September 7, 1993; accepted for publication (in revised form) by N. J. Higham, November 9, 1993.

† Department of Mathematics, University of California at Berkeley, Berkeley, California 94720 (li@math.berkeley.edu).

be the polar decompositions of A and $\widetilde{A}$, respectively, and let

$$A = U\Sigma V^*, \qquad \widetilde{A} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^* \tag{5}$$

be the SVDs of A and $\widetilde{A}$, respectively, where $\widetilde{U} = (\widetilde{U}_1, \widetilde{U}_2)$, $\widetilde{U}_1$ is $m \times n$, and $\widetilde{\Sigma} = \begin{pmatrix} \widetilde{\Sigma}_1 \\ 0 \end{pmatrix}$ and $\widetilde{\Sigma}_1 = \operatorname{diag}(\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_n)$. Assume as usual that

$$\sigma_1 \geq \cdots \geq \sigma_n > 0 \quad \text{and} \quad \widetilde{\sigma}_1 \geq \cdots \geq \widetilde{\sigma}_n > 0. \tag{6}$$

It follows from (2) and (5) that

$$Q = U_1 V^*, \qquad \widetilde{Q} = \widetilde{U}_1 \widetilde{V}^*.$$

In what follows, $\|X\|_2$ denotes the spectral norm which is the biggest singular value of X and $\|X\|_{\mathrm{F}}$ the Frobenius norm which is the square root of the trace of X^*X. We shall use $|||\cdot|||$ to denote a general unitarily invariant norm [5], [11]. Two particular ones are $\|\cdot\|_2$ and $\|\cdot\|_{\mathrm{F}}$. Consider

$$\left|\left|\left|A - \widetilde{A}\right|\right|\right| = \left|\left|\left|U^*(A - \widetilde{A})\widetilde{V}\right|\right|\right| = \left|\left|\left|\Sigma V^*\widetilde{V} - U^*\widetilde{U}\widetilde{\Sigma}\right|\right|\right| \tag{7}$$

$$= \left|\left|\left|\widetilde{U}^*(\widetilde{A} - A)V\right|\right|\right| = \left|\left|\left|\widetilde{\Sigma}\widetilde{V}^*V - \widetilde{U}^*U\Sigma\right|\right|\right|. \tag{8}$$

Define

$$E \stackrel{\text{def}}{=} \Sigma V^*\widetilde{V} - U^*\widetilde{U}\widetilde{\Sigma} \quad \text{and} \tag{9}$$

$$\widetilde{E} \stackrel{\text{def}}{=} \widetilde{\Sigma}\widetilde{V}^*V - \widetilde{U}^*U\Sigma \tag{10}$$

to infer from (7) and (8) that

$$|||E||| = \left|\left|\left|\widetilde{E}\right|\right|\right| = \left|\left|\left|A - \widetilde{A}\right|\right|\right|. \tag{11}$$

Notice that by (9) and (10)

$$(I, 0)E = \Sigma_1 V^*\widetilde{V} - U_1^*\widetilde{U}_1\widetilde{\Sigma}_1 \quad \text{and}$$
$$(I, 0)\widetilde{E} = \widetilde{\Sigma}_1\widetilde{V}^*V - \widetilde{U}_1^*U_1\Sigma_1,$$

where I is $n \times n$. Adding the conjugate transpose of the second equation to the first yields

$$\Sigma_1(V^*\widetilde{V} - U_1^*\widetilde{U}_1) + (V^*\widetilde{V} - U_1^*\widetilde{U}_1)\widetilde{\Sigma}_1 = (I, 0)E + \widetilde{E}^*\begin{pmatrix} I \\ 0 \end{pmatrix}. \tag{12}$$

This is our perturbation equation to derive our perturbation bounds for Q because for any unitarily invariant norm $|||\cdot|||$,

$$\left|\left|\left|V^*\widetilde{V} - U_1^*\widetilde{U}_1\right|\right|\right| = \left|\left|\left|I - VU_1^*\widetilde{U}_1\widetilde{V}^*\right|\right|\right| = \left|\left|\left|I - Q^*\widetilde{Q}\right|\right|\right|. \tag{13}$$

We shall use Lemma 1, which is a special case of Davis and Kahan [2, Thm. 5.2].

LEMMA 1. *Let M and N be two Hermitian matrices and let S be a complex matrix with suitable dimensions. Suppose there are two disjoint intervals separated by a gap of width at least η, one of which contains the spectrum of M and the other*

contains that of N. If $\eta > 0$, then there is a unique solution X to the matrix equation $MX - XN = S$, and moreover $|||X||| \le \frac{1}{\eta} |||S|||$ for every unitarily invariant norm $|||\cdot|||$.

Applying this lemma to (11), (12), and (13) with $M = \Sigma_1$, $N = -\widetilde{\Sigma}_1$ and $X = V^*\widetilde{V} - U_1^*\widetilde{U}_1$ yields Lemma 2.

LEMMA 2. *It holds that*

$$|||I - Q^*\widetilde{Q}||| \le \frac{2}{\sigma_n + \widetilde{\sigma}_n} |||A - \widetilde{A}|||. \tag{14}$$

When $m = n$, both Q and $\widetilde{Q}$ are unitary. Thus $|||I - Q^*\widetilde{Q}||| = |||Q - \widetilde{Q}|||$, and Lemma 2 yields the following theorem.[1]

THEOREM 1. *Let A and $\widetilde{A}$ be two $n \times n$ nonsingular complex matrices whose polar decompositions are given by (4), and let σ_n and $\widetilde{\sigma}_n$ be the smallest singular values of A and $\widetilde{A}$, respectively. Then*

$$|||Q - \widetilde{Q}||| \le \frac{2}{\sigma_n + \widetilde{\sigma}_n} |||A - \widetilde{A}|||. \tag{15}$$

If, however, $m > n$, then it follows from (9) and (10) that

$$(0, I)E = -U_2^*\widetilde{U}_1\widetilde{\Sigma}_1 \quad \text{and}$$
$$(0, I)\widetilde{E} = -\widetilde{U}_2^*U_1\Sigma_1,$$

where I is $(m-n) \times (m-n)$. Therefore

$$|||U_2^*\widetilde{U}_1||| \le |||-U_2^*\widetilde{U}_1\widetilde{\Sigma}_1||| \left\|\widetilde{\Sigma}_1^{-1}\right\|_2 \le \frac{|||(0, I)E|||}{\widetilde{\sigma}_n} \le \frac{|||A - \widetilde{A}|||}{\widetilde{\sigma}_n}.$$

Similarly,

$$|||\widetilde{U}_2^*U_1||| \le \frac{|||(0, I)\widetilde{E}|||}{\sigma_n} \le \frac{|||A - \widetilde{A}|||}{\sigma_n}.$$

Notice that $(U_1V^*, U_2) = (Q, U_2)$ and $(\widetilde{U}_1\widetilde{V}^*, \widetilde{U}_2) = (\widetilde{Q}, \widetilde{U}_2)$ are unitary. Hence $U_2^*Q = 0$ and

$$\begin{aligned} |||Q - \widetilde{Q}||| = |||(Q, U_2)^*(Q - \widetilde{Q})||| &= \left|\left|\left| \begin{pmatrix} I - Q^*\widetilde{Q} \\ -U_2^*\widetilde{Q} \end{pmatrix} \right|\right| \\ &\le |||I - Q^*\widetilde{Q}||| + |||-U_2^*\widetilde{U}_1\widetilde{V}^*||| \\ &= |||I - Q^*\widetilde{Q}||| + |||U_2^*\widetilde{U}_1||| \\ &\le \left(\frac{2}{\sigma_n + \widetilde{\sigma}_n} + \frac{1}{\widetilde{\sigma}_n}\right) |||A - \widetilde{A}||| . \end{aligned} \tag{16}$$

[1] Professor R. Bhatia kindly pointed out to me that Theorem 1 would be true in infinite dimensions. That is because of the infinite dimensional version of Lemma 1 in [2]. In the infinite dimensional version of the inequality (15), σ_n and $\widetilde{\sigma}_n$ should be replaced by $\|A^{-1}\|^{-1}$ and $\|\widetilde{A}^{-1}\|^{-1}$, respectively, where $\|\cdot\|$ is the operator norm in the Hilbert space where A and $\widetilde{A}$ live.

Similarly, we can prove

$$|||Q-\widetilde{Q}||| \le \left(\frac{2}{\sigma_n+\widetilde{\sigma}_n}+\frac{1}{\sigma_n}\right)|||A-\widetilde{A}|||. \tag{17}$$

Therefore, generally, we have Theorem 2.

THEOREM 2. *Let A and $\widetilde{A}$ be two $m\times n$ $(m>n)$ complex matrices having full column rank and with the polar decompositions* (4), *and let σ_n and $\widetilde{\sigma}_n$ be the smallest singular values of A and $\widetilde{A}$, respectively. Then*

$$|||Q-\widetilde{Q}||| \le \left(\frac{2}{\sigma_n+\widetilde{\sigma}_n}+\frac{1}{\max\{\sigma_n,\widetilde{\sigma}_n\}}\right)|||A-\widetilde{A}||| . \tag{18}$$

Estimates (16) and (17) can be sharpened a little bit when $|||\cdot||| = \|\cdot\|_F$. As a matter of fact, we shall have

$$\begin{aligned}\|Q-\widetilde{Q}\|_F &= \sqrt{\|I-Q^*\widetilde{Q}\|_F^2+\|U_2^*\widetilde{U}_1\|_F^2}\\ &\le \sqrt{\left(\frac{2}{\sigma_n+\widetilde{\sigma}_n}\right)^2+\frac{1}{\widetilde{\sigma}_n^2}}\,\|A-\widetilde{A}\|_F \quad \text{and}\\ \|Q-\widetilde{Q}\|_F &\le \sqrt{\left(\frac{2}{\sigma_n+\widetilde{\sigma}_n}\right)^2+\frac{1}{\sigma_n^2}}\,\|A-\widetilde{A}\|_F .\end{aligned}$$

A consequence of these two inequalities is the following theorem.

THEOREM 3. *Under the conditions of Theorem* 2,

$$\|Q-\widetilde{Q}\|_F \le \sqrt{\left(\frac{2}{\sigma_n+\widetilde{\sigma}_n}\right)^2+\left(\frac{1}{\max\{\sigma_n,\widetilde{\sigma}_n\}}\right)^2}\,\|A-\widetilde{A}\|_F . \tag{19}$$

We conclude this paper with a few remarks.

Remark 1. The bound in (15) is the best possible, in the sense that the equality can be achieved. Take the following case for an example: Both A and $\widetilde{A}$ are $n\times n$ unitary matrices. Thus $\sigma_n=\widetilde{\sigma}_n=1$, $Q=A$, $\widetilde{Q}=\widetilde{A}$, and

$$|||Q-\widetilde{Q}||| = \frac{2}{\sigma_n+\widetilde{\sigma}_n}|||A-\widetilde{A}|||.$$

It is even achievable in the real number field by taking A and $\widetilde{A}$ to be two $n\times n$ orthogonal matrices although, as we know, Q behaves quite differently in the real number field (Remark 5). All previously published bounds do not achieve this!

Remark 2. Bounds (15), (18), and (19) involve both σ_n and $\widetilde{\sigma}_n$. To obtain bounds involving σ_n alone, one can weaken them by utilizing the following fact:

$$\|A-\widetilde{A}\|_2 \ge |\sigma_n-\widetilde{\sigma}_n|.$$

For example, (15) yields

$$|||Q-\widetilde{Q}||| \le \frac{2}{2\sigma_n-\|A-\widetilde{A}\|_2}|||A-\widetilde{A}|||, \tag{20}$$

provided $\left\|A-\widetilde{A}\right\|_2 < 2\sigma_n$.

Remark 3. Mathias [9] proved that for $m = n$ if $\left\|A-\widetilde{A}\right\|_2 < \sigma_n$, then

$$\left|\!\left|\!\left|Q-\widetilde{Q}\right|\!\right|\!\right| \le -\frac{\left|\!\left|\!\left|A-\widetilde{A}\right|\!\right|\!\right|}{\left\|A-\widetilde{A}\right\|_2} \times \ln\left(1-\frac{\left\|A-\widetilde{A}\right\|_2}{\sigma_n}\right). \tag{21}$$

Although his bound uses slightly different information than ours, it is always a bigger, sometimes much bigger, bound than (15) (since the left-hand side of (21) could blow up). To see why this is true, we claim that even (20), the weakened form of (15), is still no weaker than that of Mathias because their ratio (his/ours) is

$$-\frac{\ln(1-x)}{x}\cdot\left(1-\frac{x}{2}\right) = 1+\sum_{j=2}^{\infty}\left(\frac{1}{j+1}-\frac{1}{2j}\right)x^j > 1$$

for $0 < x = \left\|A-\widetilde{A}\right\|_2 \Big/ \sigma_n < 1$.

Remark 4. Chen and Sun [6] studied the case $m > n$, also. But only the Frobenius norm was considered. They proved

$$\left\|Q-\widetilde{Q}\right\|_{\mathrm{F}} \le \frac{2}{\sigma_n}\left\|A-\widetilde{A}\right\|_{\mathrm{F}}. \tag{22}$$

Without loss of generality, assume $\widetilde{\sigma}_n \le \sigma_n$. Then it is easy to see our bound (19) is sharper than (22) when

$$\widetilde{\sigma}_n \le \sigma_n \le \frac{\sqrt{3}}{2-\sqrt{3}}\widetilde{\sigma}_n \approx 6.5\widetilde{\sigma}_n;$$

otherwise (22) is a little sharper because

$$\sqrt{\left(\frac{2}{\sigma_n+\widetilde{\sigma}_n}\right)^2+\left(\frac{1}{\sigma_n}\right)^2} \le \frac{\sqrt{5}}{\sigma_n} \approx \frac{2.2}{\sigma_n}$$

always. More generally, Sun and Chen [3] and Li [8] treated the cases when A and $\widetilde{A}$ do not necessarily have full column rank. Applied to our full column rank case here, the perturbation bound for the polar factor in [3] reads exactly the same as (22), and in [8] reads

$$\left\|Q-\widetilde{Q}\right\|_{\mathrm{F}} \le \frac{1}{\min\{\sigma_n,\widetilde{\sigma}_n\}}\left\|A-\widetilde{A}\right\|_{\mathrm{F}},$$

which is clearly sharper than (19) and (22) when $\sigma_n \approx \widetilde{\sigma}_n$. However, it may be very bad if one of σ_n and $\widetilde{\sigma}_n$ is much smaller than the other.

Remark 5. Perturbation bounds for the Q factor in polar decomposition illustrate that the change in Q is proportional to the reciprocal of the smallest singular value of A when $m = n$ and when the working number field is complex. However, it was discovered by Barrlund [1], Kenney and Laub [7], and Mathias [9] that for the real case the change in Q is proportional to the reciprocal of the sum of the two smallest singular values of A if $m = n$, which means Q is (much) less sensitive to perturbations in A

in the real case than in the complex case. The above derivation of the perturbation bound (15) for the complex case is very elementary while giving the best among the derivations that have been published. However the author was unable to extend this derivation to perform for the real case. It is worth stating (as pointed out by one of the anonymous referees) that even in the real number field when $m > n$, the change in Q is not proportional to $1/(\sigma_{n-1} + \sigma_n)$ instead of $1/\sigma_n$. The following example offered by the referee makes this point very clear:

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0.8 \times 10^{-6} \\ 0 & 0 \end{pmatrix}, \qquad Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix};$$

$$\widetilde{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0.8 \times 10^{-6} \\ 0 & 0.6 \times 10^{-6} \end{pmatrix}, \qquad \widetilde{Q} = \begin{pmatrix} 1 & 0 \\ 0 & 0.8 \\ 0 & 0.6 \end{pmatrix}.$$

Acknowledgments. The author is grateful for the encouragement of Professors W. Kahan and J. Demmel. He thanks Dr. N. J. Higham for his valuable comments concerning the manuscript. He is indebted to the referees for their many helpful suggestions, especially the last part of Remark 5 which he had not previously addressed.

REFERENCES

[1] A. BARRLUND, *Perturbation bounds on the polar decomposition*, BIT, 31 (1990), pp. 101–113.
[2] C. DAVIS AND W. M. KAHAN, *The rotation of eigenvectors by a perturbation. iii*, SIAM J. Numer. Anal., 1 (1970), pp. 1–46.
[3] J.-G. SUN AND C. HUI CHEN, *Generalized polar decomposition*, Math. Numer. Sinica, 11 (1989), pp. 262–273. (In Chinese.)
[4] N. J. HIGHAM, *Computing the polar decomposition—with applications*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1160–1174.
[5] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.
[6] C. HUI CHEN AND J. GUANG SUN, *Perturbation bounds for the polar factors*, J. Comput. Math., 7 (1989), pp. 397–401.
[7] C. KENNEY AND A. J. LAUB, *Polar decompostion and matrix sign function condition estimates*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 488–504.
[8] R.-C. LI, *A perturbation bound for the generalized polar decomposition*, BIT, 34 (1993), pp. 304–308.
[9] R. MATHIAS, *Perturbation bounds for the polar decomposition*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 588–597.
[10] J. QIN MAO, *The perturbation analysis of the product of singular vector matrices* uv^h, J. Comput. Math., 4 (1986), pp. 245–248.
[11] G. W. STEWART AND J.-G. SUN, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 1, pp. 333–340, January 1995

024

SINGULAR VALUES OF COMPANION MATRICES AND BOUNDS ON ZEROS OF POLYNOMIALS *

FUAD KITTANEH†

Abstract. Let $C(p)$ denote the companion matrix of a monic polynomial p with complex coefficients. Then the zeros of p are exactly the eigenvalues of $C(p)$. In this paper, the singular values of $C(p)$ are computed. Applying some basic eigenvalue-singular value majorization relations, several sharp estimates are obtained for the zeros of p in terms of its coefficients. These estimates improve some classical bounds on zeros of polynomials.

Key words. zeros of a polynomial, eigenvalue, singular value, companion matrix

AMS subject classifications. Primary 15A18, 15A42, 30C15

1. Introduction.

In this paper we are concerned with some estimates for the zeros of a monic polynomial p of the form

$$p(z) = z^n + a_n z^{n-1} + \cdots + a_2 z + a_1, \tag{1}$$

where $n \geq 2$ and the coefficients $a_1, a_2, \ldots, a_n$ are complex numbers with $a_1 \neq 0$.

The problem of locating the zeros of a polynomial of the form given in (1) has attracted the attention of many mathematicians in the past and it is still a fascinating topic to both complex and numerical analysts. An excellent approach to this problem using matrix analysis has been demonstrated in [4], [5], [7, pp. 316–319], and [9, pp. 139–146]. The related problem of estimating the distance between the zeros of two monic polynomials and its connection with the general spectral variation problem has also been dealt with in the literature (see, e.g., [2, Chap. 5], [3], and [12, Appendices A, B, and K]). For comprehensive accounts on polynomial inequalities, the reader is referred to [9, Chap. VII and VIII], [11, pp. 217–235], [13, Part Three], and references therein.

Recall that the companion matrix of the polynomial p given in (1) is defined by

$$C(p) = \begin{bmatrix} -a_n & -a_{n-1} & \cdots & -a_2 & -a_1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}. \tag{2}$$

Since the characteristic polynomial of $C(p)$ is p, it follows that the zeros of p are exactly the eigenvalues of $C(p)$ (see, e.g., [7, p. 147]).

Throughout the paper we let $z_1, z_2, \ldots, z_n$ denote the zeros of p (or the eigenvalues of $C(p)$) enumerated as $|z_1| \geq |z_2| \geq \cdots \geq |z_n|$ with multiplicity counted in this enumeration. Employing the fact that the spectral radius of $C(p)$ is dominated by any matrix norm of $C(p)$, many of the classical bounds on $z_1, z_2, \ldots, z_n$ in terms of

* Received by the editors March 28, 1993; accepted for publication (in revised form) by R. A. Horn, November 24, 1993.

† Department of Mathematics, University of Jordan, Amman, Jordan.

the coefficients $a_1, a_2, \ldots, a_n$ have been given very elegant proofs (see [4] and [7, pp. 316–319]). The following bounds are examples of what we have in mind.

$$\text{(3)} \qquad \text{(i)} \quad \text{(Cauchy bound)} \ |z_1| \leq 1 + \max_{1 \leq j \leq n} |a_j|.$$

$$\text{(4)} \qquad \text{(ii)} \quad \text{(Carmichael–Mason bound)} \ |z_1| \leq \left(1 + \sum_{j=1}^{n} |a_j|^2 \right)^{1/2}.$$

The proofs of the Carmichael–Mason bound given in [4] and [7, Problem 28, p. 317] are based on the fact that the usual operator (spectral) norm of $C(p)$ is not greater than

$$\left(1 + \sum_{j=1}^{n} |a_j|^2 \right)^{1/2}.$$

In this paper we improve this bound by computing the exact value of the usual operator norm (the largest singular value) of $C(p)$. In view of this computation, some basic majorization relations between the eigenvalues and singular values of a matrix enable us to establish several sharp inequalities involving the zeros and coefficients of the polynomial p.

Let $M_n(\mathbf{C})$ denote the algebra of all $n \times n$ complex matrices. For $A \in M_n(\mathbf{C})$ let $\lambda_1(A), \lambda_2(A), \ldots, \lambda_n(A)$ be the eigenvalues of A and let $s_1(A), s_2(A), \ldots, s_n(A)$ be the singular values of A (the eigenvalues of the positive semidefinite matrix $|A| = (A^*A)^{1/2}$). We enumerate these numbers as $|\lambda_1(A)| \geq |\lambda_2(A)| \geq \cdots \geq |\lambda_n(A)|$ and $s_1(A) \geq s_2(A) \geq \cdots \geq s_n(A)$, where multiplicity is counted in these enumerations. It is known that if $A \in M_n(\mathbf{C})$, then

$$\text{(5)} \qquad s_j^2(A) = \lambda_j(A^*A) = \lambda_j(AA^*) \quad \text{for } j = 1, 2, \ldots, n.$$

Recall also that $s_1(A)$ is the usual operator norm of A (see, e.g., [7, p. 437] and [8, p. 146]).

2. Preliminary results. To achieve our goal we need the following inequalities between the eigenvalues and singular values of a matrix.

LEMMA 1. *If $A \in M_n(\mathbf{C})$, then*

$$\text{(6)} \qquad \prod_{j=1}^{k} |\lambda_j(A)| \leq \prod_{j=1}^{k} s_j(A) \quad \textit{for } k = 1, 2, \ldots, n,$$

with equality for $k = n$; and

$$\text{(7)} \qquad \prod_{j=k}^{n} |\lambda_j(A)| \geq \prod_{j=k}^{n} s_j(A) \quad \textit{for } k = 1, 2, \ldots, n,$$

with equality for $k = 1$. In particular we have

$$\text{(8)} \qquad s_n(A) \leq |\lambda_j(A)| \leq s_1(A) \quad \textit{for } j = 1, 2, \ldots, n.$$

It should be mentioned here that equality holds simultaneously in all the relations (6) (or (7)) if and only if A is normal (see [6, p. 36]).

LEMMA 2. *If $A \in M_n(\mathbf{C})$ and r is a positive real number, then*

$$\sum_{j=1}^{k} |\lambda_j(A)|^r \leq \sum_{j=1}^{k} s_j^r(A) \quad \text{for } k = 1, 2, \ldots, n. \tag{9}$$

When $k = n$, equality in (9) *holds if and only if A is normal.*

LEMMA 3. *If $A \in M_n(\mathbf{C})$ is invertible and r is a nonzero real number, then*

$$\sum_{j=1}^{n} |\lambda_j(A)|^r \leq \sum_{j=1}^{n} s_j^r(A). \tag{10}$$

Moreover, equality in (10) *holds if and only if A is normal.*

LEMMA 4. *If $A \in M_n(\mathbf{C})$ and r is a positive real number, then*

$$\prod_{j=1}^{k} (1 + r|\lambda_j(A)|) \leq \prod_{j=1}^{k} (1 + r s_j(A)) \quad \text{for } k = 1, 2, \ldots, n. \tag{11}$$

When $k = n$, equality in (11) *holds if and only if A is normal.*

The famous majorization relations of Lemmas 1, 2, 3, and 4 are due to Weyl (see, e.g., [6, pp. 35–41] and [8, Chap. 3]). We remark that the inequalities (6) and (7) of Lemma 1 are equivalent when A is invertible. It should be mentioned here that the inequality (6) is an essential ingredient in the proofs of more general eigenvalue-singular value majorization results including (9), (10), and (11) as special cases (see [6, pp. 39–45] and [8, p. 176 and pp. 182–185]).

Our main results are immediate consequences of Lemmas 1, 2, 3, and 4 applied to the matrix $C(p)$. To compute the singular values of $C(p)$ we need to recall the following formula concerning determinants of partitioned matrices (see [7, Prob. 15, p. 175]).

LEMMA 5. *Let $A = [a_{ij}] \in M_n(\mathbf{C})$ be written in partitioned form as*

$$A = \begin{bmatrix} \tilde{A} & x \\ x^* & a_{nn} \end{bmatrix},$$

where $x \in \mathbf{C}^{n-1}$ and $\tilde{A} \in M_{n-1}(\mathbf{C})$. Then

$$\det A = a_{nn} \det \tilde{A} - x^* (\operatorname{adj} \tilde{A}) x, \tag{12}$$

where $\operatorname{adj}\tilde{A}$ is the adjugate (classical adjoint) of $\tilde{A}$.

Note that the characteristic polynomial of $C(p)C(p)^*$ is the determinant of the partitioned matrix

$$tI - C(p)C(p)* = \left[\begin{array}{ccccc|c} t-\alpha & a_n & a_{n-1} & \ldots & a_3 & a_2 \\ \bar{a}_n & t-1 & 0 & \ldots & 0 & 0 \\ \bar{a}_{n-1} & 0 & t-1 & & 0 & 0 \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ \bar{a}_3 & 0 & 0 & \ldots & t-1 & 0 \\ \hline \bar{a}_2 & 0 & 0 & \cdots & 0 & t-1 \end{array}\right], \tag{13}$$

where

$$\alpha = \sum_{j=1}^{n} |a_j|^2.$$

Now using induction and Lemma 5, one can prove that the characteristic polynomial of $C(p)C(p)^*$ is given by

$$\det(tI - C(p)C(p)^*) = (t-1)^{n-2}(t^2 - (\alpha+1)t + |a_1|^2). \tag{14}$$

As an immediate consequence of (5) and (14), we have

$$s_1(C(p)) = \left(\frac{\alpha + 1 + ((\alpha+1)^2 - 4|a_1|^2)^{1/2}}{2}\right)^{1/2}, \tag{15}$$

$$s_n(C(p)) = \left(\frac{\alpha + 1 - ((\alpha+1)^2 - 4|a_1|^2)^{1/2}}{2}\right)^{1/2}, \tag{16}$$

and

$$s_j(C(p)) = 1 \quad \text{for } j = 2, \ldots, n-1. \tag{17}$$

The following alternative method of computing the eigenvalues of $C(p)C(p)^*$ has been suggested by Roger A. Horn. Using (13), one can easily see that the Hermitian matrix $I - C(p)C(p)^*$ has at most two linearly independent columns, so its rank is at most two, and hence (at least) $n-2$ of the eigenvalues of $C(p)C(p)^*$ are equal to 1. The two remaining eigenvalues (call them λ and μ) are easily determined from the relations

$$\lambda + \mu + \underbrace{1 + \cdots + 1}_{n-2} = \lambda + \mu + n - 2 = \operatorname{tr} C(p)C(p)^* = \alpha + n - 1,$$

and

$$\lambda\mu \underbrace{1 \cdots 1}_{n-2} = \lambda\mu = \det C(p)C(p)^* = |a_1|^2.$$

Since $\lambda + \mu = \alpha + 1$ and $\lambda\mu = |a_1|^2$, it follows that λ and μ are the roots of the quadratic equation $t^2 - (\alpha+1)t + |a_1|^2 = 0$.

For full implementation of Lemmas 1, 2, 3, and 4 we need to know the condition on p that is equivalent to the normality of $C(p)$. A simple computation shows that $C(p)$ is normal if and only if $|a_1| = 1$ and $a_j = 0$ for $j = 2, \ldots, n$. Thus, $C(p)$ is normal if and only if $p(z) = z^n + a_1$ with $|a_1| = 1$. Note that $C(p)$ is normal if and only if $C(p)$ is unitary.

3. Main results. Having found the singular values of $C(p)$, we are now in a position to present our main estimates for the zeros of p. In what follows we let

$$\alpha = \sum_{j=1}^{n} |a_j|^2$$

and we simply write s_j for $s_j(C(p))$.

It has been shown in [7, pp. 316–319] that all the zeros of p lie in the annulus described by

$$\frac{|a_1|}{(1+\alpha)^{1/2}} \leq |z| \leq (1+\alpha)^{1/2}. \tag{18}$$

The second inequality of (18) is the Carmichael–Mason bound and the first inequality is the lower bound counterpart of the Carmichael–Mason bound (obtained by applying the upper bound to the polynomial $z^n p(z^{-1})/a_1$).

Our first result concerning the location of the zeros of p is a considerable improvement of (18). It also improves some classical estimates of Landau and Specht (see [11, p. 224] and references therein). Furthermore, the result tells us that all the zeros of p lie in the annulus given by

$$s_n \leq |z| \leq s_1, \tag{19}$$

which is included in the annulus given in (18).

Since $s_j = 1$ for $j = 2, \ldots, n-1$, the proof of the following theorem follows immediately from Lemma 1.

THEOREM 1. *We have*

$$\prod_{j=1}^{k} |z_j| \leq s_1 \quad \text{for } k = 1, \ldots, n-1, \tag{20}$$

and

$$\prod_{j=k}^{n} |z_j| \geq s_n \quad \text{for } k = 2, \ldots, n. \tag{21}$$

Complementing Theorem 1, it should be noted that

$$\prod_{j=1}^{n} |z_j| = s_1 s_n = |a_1|. \tag{22}$$

Moreover, equality holds simultaneously in all the relations (20) (or (21)) if and only if $p(z) = z^n + a_1$ with $|a_1| = 1$.

As an application of Lemma 2 to our investigation we can easily prove the following result.

THEOREM 2. *If r is a positive real number, then*

$$\sum_{j=1}^{k} |z_j|^r \leq s_1^r + k - 1 \quad \text{for } k = 1, \ldots, n-1. \tag{23}$$

Since $a_1 \neq 0$, it follows that the matrix $C(p)$ is invertible. Thus by Lemma 3 applied to $C(p)$ we have Theorem 3.

THEOREM 3. *If r is a nonzero real number, then*

$$\sum_{j=1}^{n} |z_j|^r \leq s_1^r + s_n^r + n - 2. \tag{24}$$

Inequality (24) *is an equality if and only if* $p(z) = z^n + a_1$ *with* $|a_1| = 1$.

A useful inequality that is essentially due to Schur (see [14, p. 70]) asserts that

$$\prod_{j=1}^{n} (1 + |z_j|) \leq 2^n (1 + \alpha)^{1/2}. \tag{25}$$

The classical complex analysis proof of (25) given in [14] shows how it is possible to connect (25) with the Mahler measure of p (see [11, pp. 232–233] and references therein).

A considerable improvement of (25) that is based on Lemma 4 can be stated as follows.

THEOREM 4. *If* r *is a positive real number, then*

$$\prod_{j=1}^{k} (1 + r|z_j|) \leq (1 + r)^{k-1}(1 + rs_1) \quad \text{for } k = 1, \ldots, n-1; \tag{26}$$

and

$$\prod_{j=1}^{n} (1 + r|z_j|) \leq (1 + r)^{n-2}(1 + rs_1)(1 + rs_n). \tag{27}$$

In particular (*when* $r = 1$), *we have*

$$\prod_{j=1}^{n} (1 + |z_j|) \leq 2^{n-2}(1 + s_1)(1 + s_n). \tag{28}$$

The inequality (27) *becomes an equality if and only if*

$$p(z) = z^n + a_1 \text{ with } |a_1| = 1.$$

4. Related inequalities. Using the fact that two similar matrices have the same spectral radius, some weighted inequalities generalizing the Cauchy bound (3) have been given in [7, p. 319]. Following the techniques in [7, p. 319] and invoking the fact that two similar matrices have the same eigenvalues, we establish a weighted inequality related to (24) without using the singular values of $C(p)$.

If $r_1, r_2, \ldots, r_n$ are positive real numbers and $Q = \operatorname{diag}(r_1, r_2, \ldots, r_n)$, then

$$Q^{-1}C(p)Q = \begin{bmatrix} -a_n & -\frac{r_2}{r_1}a_{n-1} & \cdots & -\frac{r_{n-1}}{r_1}a_2 & -\frac{r_n}{r_1}a_1 \\ \frac{r_1}{r_2} & 0 & \cdots & 0 & 0 \\ 0 & \frac{r_2}{r_3} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{r_{n-1}}{r_n} & 0 \end{bmatrix}. \tag{29}$$

It is known that if $A = [a_{ij}] \in M_n(\mathbf{C})$, then

$$\sum_{j=1}^{n} s_j^2(A) = \sum_{i,j=1}^{n} |a_{ij}|^2 \tag{30}$$

and

$$\sum_{j=1}^{n} |\lambda_j(A)|^2 \leq \sum_{i,j=1}^{n} |a_{ij}|^2 \tag{31}$$

with equality if and only if A is normal (see [7, pp. 316, 421], [8, p. 156], or [15, p. 11]).

By (31) applied to $C(p)$ we have

$$\sum_{j=1}^{n} |z_j|^2 \leq \alpha + n - 1, \tag{32}$$

which is a special case of (24). It should be mentioned that (32) can be also concluded from a closely related result of Ostrowski (see [11, p. 224]).

Applying (31) to the matrix $Q^{-1}C(p)Q$ and using the fact that $z_j = \lambda_j(Q^{-1}C(p)Q)$ for $j = 1, 2, \ldots, n$, we have

$$\sum_{j=1}^{n} |z_j|^2 \leq \sum_{j=1}^{n-1} \left(\frac{r_j}{r_{j+1}}\right)^2 + \frac{1}{r_1^2} \sum_{j=1}^{n} |a_j|^2 r_{n-j+1}^2. \tag{33}$$

Note that if we let $r_j = 1$ for $j = 1, 2, \ldots, n$, then we obtain (32). If we let $r_j = r^j$ for some $r > 0$ and for $j = 1, 2, \ldots, n$, then we have the estimate

$$\sum_{j=1}^{n} |z_j|^2 \leq \frac{n-1}{r^2} + r^{2n} \sum_{j=1}^{n} \frac{|a_j|^2}{r^{2j}}. \tag{34}$$

If all the coefficients a_j are nonzero, then by choosing

$$r_j = \frac{r_1}{|a_{n-j+1}|} \quad \text{for } j = 2, \ldots, n,$$

we obtain

$$\sum_{j=1}^{n} |z_j|^2 \leq n - 1 + |a_{n-1}|^2 + |a_n|^2 + \sum_{j=1}^{n-2} \left|\frac{a_j}{a_{j+1}}\right|^2. \tag{35}$$

On comparing the estimates (32) and (35), it can be verified that (32) is better than (35) if $|a_j| \leq 1$ for $j = 2, \ldots, n-1$ with strict inequality for at least one value of j. On the other hand, (35) is better than (32) if $|a_j| \geq 1$ for $j = 2, \ldots, n-1$ with strict inequality for at least one value of j.

5. Remarks. We conclude the paper with the following remarks concerning our main results.

Remark 1. Observe that while the Cauchy bound (3) relates the max norms (or l_∞ norms) of $(z_1, z_2, \ldots, z_n)$ and $(a_1, a_2, \ldots, a_n)$ as vectors in $\mathbf{C}^n$, the estimate (32) relates the Euclidean norms (or l_2 norms) of these vectors.

Remark 2. Since we have explicitly found all the singular values of the companion matrix $C(p)$, we can establish other estimates for the zeros of p. In fact, for every inequality relating the eigenvalues and singular values of matrices, there is a corresponding inequality relating the zeros and coefficients of polynomials. For a host of

eigenvalue-singular value inequalities involving different classes of Schur-convex (or isotone) functions, the reader is referred to [1], [6, Chap. 2], [8, Chap. 3], and [10, Chap. 9], and is then invited to formulate the corresponding polynomial inequalities.

Remark 3. Our explicit evaluation of s_1 and s_n gives an explicit formula for the spectral condition number of a companion matrix. Recall that if $A \in M_n(\mathbf{C})$ is invertible, then its spectral condition number is given by

$$k(A) = \frac{s_1(A)}{s_n(A)}$$

(see [7, p. 442] or [8, p. 158]).

Acknowledgment. The author is grateful to Professor Roger A. Horn and the referees for their careful reading of this paper and for their useful comments and suggestions.

REFERENCES

[1] T. ANDO, *Majorization, doubly stochastic matrices and comparison of eigenvalues*, Linear Algebra Appl., 118 (1989), pp. 163–248.

[2] R. BHATIA, *Perturbation bounds for matrix eigenvalues*, Pitman Research Notes in Mathematics, No. 162, Longman Scientific and Technical, Essex, U.K., 1987.

[3] R. BHATIA, L. ELSNER, AND G. KRAUSE, *Bounds for the variation of the roots of a polynomial and the eigenvalues of a matrix*, Linear Algebra Appl., 142 (1990), pp. 195–210.

[4] M. FUJII AND F. KUBO, *Operator norms as bounds for roots of algebraic equations*, Proc. Japan Acad., 49 (1973), pp. 805–808.

[5] M. FUJII AND F. KUBO, *Buzano's inequality and bounds for roots of algebraic equations*, Proc. Amer. Math. Soc., 117 (1993), 359–361.

[6] I. C. GOHBERG AND M. G. KREIN, *Introduction to the theory of linear nonselfadjoint operators*, Transl. Math. Monographs, Vol. 18, Amer. Math. Soc., Providence, RI, 1969.

[7] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, 1987.

[8] ———, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1991.

[9] M. MARDEN, *Geometry of Polynomials*, Amer. Math. Soc. Surveys, Vol. 3, 2nd ed., Providence, RI, 1966.

[10] A. W. MARSHALL AND I. OLKIN, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.

[11] D. S. MITRINOVIC, *Analytic Inequalities*, Springer-Verlag, Berlin, 1970.

[12] A. M. OSTROWSKI, *Solution of Equations in Euclidean and Banach Spaces*, 3rd ed., Academic Press, New York, 1973.

[13] G. PÓLYA AND G. SZEGÖ, *Problems and Theorems in Analysis*, Vol. I, Springer-Verlag, Berlin, 1972.

[14] ———, *Problems and Theorems in Analysis*, Vol. II, Springer-Verlag, New York, 1976.

[15] B. SIMON, *Trace Ideals and Their Applications*, Cambridge University Press, Cambridge, 1979.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 343–358, April 1995

MATRIX POWERS IN FINITE PRECISION ARITHMETIC*

NICHOLAS J. HIGHAM† AND PHILIP A. KNIGHT‡

Abstract. If A is a square matrix with spectral radius less than 1 then $A^k \to 0$ as $k \to \infty$, but the powers computed in finite precision arithmetic may or may not converge. We derive a sufficient condition for $fl(A^k) \to 0$ as $k \to \infty$ and a bound on $\|fl(A^k)\|$, both expressed in terms of the Jordan canonical form of A. Examples show that the results can be sharp. We show that the sufficient condition can be rephrased in terms of a pseudospectrum of A when A is diagonalizable, under certain assumptions. Our analysis leads to the rule of thumb that convergence or divergence of the computed powers of A can be expected according as the spectral radius computed by any backward stable algorithm is less than or greater than 1.

Key words. matrix powers, rounding errors, Jordan canonical form, nonnormal matrices, pseudospectrum

AMS subject classifications. primary 65F99, 65G05

1. Introduction.

Many numerical processes depend for their success upon the powers of a matrix tending to zero. A fundamental example is stationary iteration for solving a linear system $Ax = b$, in which a sequence of vectors is defined by $Mx_{k+1} = Nx_k + b$, where $A = M - N$ and M is nonsingular. The errors $e_k = x - x_k$ satisfy $e_k = (M^{-1}N)^k e_0$, so the iteration converges for all x_0 if $(M^{-1}N)^k \to 0$ as $k \to \infty$. Many theorems are available about the convergence of stationary iteration, but virtually all of them are concerned with exact arithmetic (for exceptions see [12], [13] and the references therein). While the errors in stationary iteration are not precisely modelled by the errors in matrix powering, as matrix powers are not formed explicitly, the behaviour of the computed powers $fl((M^{-1}N)^k)$ can be expected to give some insight into the behaviour of stationary iteration (indeed, the basic error recurrences in [12] and [13] involve powers of $M^{-1}N$ acting on vectors of rounding errors).

In [18, Chap. 20], Ostrowski proves a theorem about a product of perturbed matrices $A + \Delta A_i$ that he states "assures the theoretical *stability of the convergence* of A^μ to 0 with respect to rounding off" as $\mu \to \infty$ for any matrix A with spectral radius $\rho(A) < 1$. Although Ostrowski's theorem is correct, its interpretation with respect to computed powers is not as simple as this statement implies, because for any finite precision arithmetic, no matter how accurate, there are matrices that are sensitive enough to perturbations to cause the theoretically convergent sequence of powers to diverge. To illustrate this point, Fig. 1.1[1] plots the 2-norms of the first 200 powers of a 14×14 nilpotent matrix C_{14} discussed by Trefethen and Trummer [23] (see §3 for details). The plot confirms the statement of these authors that the matrix is not power-bounded in floating point arithmetic, even though its 14th power should be zero. The powers for our plot were computed in MATLAB, which has unit

* Received by the editors September 22, 1993; accepted for publication (in revised form) by Charles Van Loan, March 25, 1994.

† Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (na.nhigham@na-net.ornl.gov). The work of this author was supported by Science and Engineering Research Council grant GR/H52139.

‡ Department of Mathematics, University of Manchester, Manchester, M13 9PL, England. Present address: Department of Mathematics, University of Strathclyde, Glasgow, G1 1XH, Scotland (p.a.knight@strath.ac.uk). This author was supported by a Science and Engineering Research Council Research Studentship.

[1] As in all our plots of norms of powers, k on the x-axis is plotted against $\|fl(A^k)\|_2$ on the y-axis.

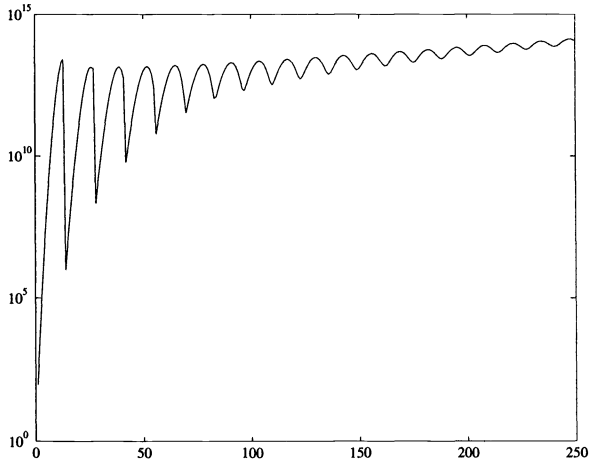


FIG. 1.1. *Diverging powers of a nilpotent matrix,* C_{14}.

roundoff $u = 2^{-53} \approx 1.1 \times 10^{-16}$. Reichel and Trefethen [19] also give an example of a matrix that is nilpotent in theory but not power-bounded in practice. In this paper we determine conditions on a matrix A that ensure that the computed powers converge to zero.

In §2 we examine the behaviour of matrix powers in exact arithmetic. In particular, we review a number of bounds on the norms of powers. In §3 we use the Jordan canonical form of A to bound $\|fl(A^k)\|$ and to determine a sufficient condition for $fl(A^k) \to 0$ as $k \to \infty$. We also show that for certain matrices our bounds are tight. Finally, in §4 we rephrase our sufficient condition in terms of a pseudospectrum of A, under certain assumptions, including that A is diagonalizable; the modified result is not any sharper than the original, but offers an alternative viewpoint that is intuitively attractive.

In our analysis we use the standard model for floating point arithmetic:

$$fl(x \pm y) = x(1+\alpha) \pm y(1+\beta), \qquad |\alpha|, |\beta| \le u,$$
$$fl(x \operatorname{op} y) = (x \operatorname{op} y)(1+\delta), \qquad |\delta| \le u, \quad \operatorname{op} = *, /,$$

where u is the unit roundoff. This model is valid for machines that do not use a guard digit in addition and subtraction.

We will use the Frobenius norm, $\|A\|_F = (\sum_{i,j} |a_{ij}|^2)^{1/2}$, and the p-norms $\|A\|_p = \max_{x \neq 0} \|Ax\|_p / \|x\|_p$, where $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$ and $1 \le p \le \infty$. From §3 onwards we will drop the subscripts on $\|\cdot\|_p$, since all the norms from that point on are p-norms.

2. Matrix powers in exact arithmetic. We begin by discussing the behaviour of matrix powers in the absence of rounding errors. In exact arithmetic the limiting behaviour of the powers of $A \in \mathbf{C}^{n \times n}$ is determined by A's eigenvalues. If the spectral radius $\rho(A) < 1$ then $A^k \to 0$ as $k \to \infty$; if $\rho(A) > 1$, $A^k \to \infty$ as $k \to \infty$. If $\rho(A) = 1$ then $\|A^k\| \to \infty$ if A has a defective eigenvalue λ such that $|\lambda| = 1$; A^k does not converge if A has a nondefective eigenvalue $\lambda \neq 1$ such that $|\lambda| = 1$ (although the norms of the powers may converge); otherwise, the only eigenvalue of modulus 1 is the

nondefective eigenvalue 1, and A^k converges to a nonzero matrix. These statements are easily proved using the Jordan canonical form

$$A = XJX^{-1} \in \mathbf{C}^{n\times n}, \tag{2.1a}$$

where X is nonsingular and

$$J = \operatorname{diag}(J_1, J_2, \ldots, J_s), \qquad J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \lambda_i & 1 \\ & & & \lambda_i \end{bmatrix} \in \mathbf{C}^{n_i\times n_i}, \tag{2.1b}$$

where $n_1 + n_2 + \cdots + n_s = n$. We will call a matrix for which $A^k \to 0$ as $k \to \infty$ (or equivalently, $\rho(A) < 1$) a *convergent matrix.*

The norm of a convergent matrix can be arbitrarily large, as is shown trivially by the example

$$A_2(\alpha) = \begin{bmatrix} \lambda & \alpha \\ 0 & \lambda \end{bmatrix}, \qquad |\lambda| < 1, \quad \alpha \gg 1. \tag{2.2}$$

While the spectral radius determines the asymptotic rate of growth of matrix powers, the norm influences the initial behaviour of the powers. The interesting result that $\rho(A) = \lim_{k\to\infty} \|A^k\|^{1/k}$ for any norm (see [14, p. 299], for example) confirms the asymptotic role of the spectral radius. An important quantity is the "hump" $\max_k \|A^k\|/\|A\|$, which can be arbitrarily large for a convergent matrix, as can be seen from $A_3(\alpha)$, the 3×3 analogue of the matrix in (2.2), for which $\|A_3(\alpha)^2\|/\|A_3(\alpha)\| = O(\alpha)$. Figure 2.1 shows an example of the hump phenomenon: the plot is for $A_3(2)$ with $\lambda = 3/4$; here, $\|A_3(2)\|_2 = 3.57$. The shape of the plot is typical of that for a convergent matrix with norm bigger than 1. Note that if A is normal (so that in (2.1a) J is diagonal and X can be taken to be unitary) we have $\|A^k\|_2 = \|\operatorname{diag}(\lambda_i^k)\|_2 = \|A\|_2^k = \rho(A)^k$, so the problem of bounding $\|A^k\|$ is of interest only for nonnormal matrices. The hump phenomenon arises in various areas of numerical analysis. For example, it is discussed for matrix powers in the context of stiff differential equations by D. J. Higham and Trefethen [8], and by Moler and Van Loan [17] for the matrix exponential e^{At} with $t \to \infty$.

In the rest of this section we briefly survey bounds for $\|A^k\|$. First, however, we comment on the condition number $\kappa(X) = \|X\|\|X^{-1}\|$ that appears in various bounds in this paper. The matrix X in the Jordan form (2.1a) is by no means unique [3, pp. 220–221], [6]: if A has distinct eigenvalues (hence J is diagonal) then X can be replaced by XD, for any nonsingular diagonal D, while if A has repeated eigenvalues then X can be replaced by XT, where T is a block matrix with block structure conformal with that of J and which contains some arbitrary upper trapezoidal Toeplitz blocks. We adopt the convention that $\kappa(X)$ denotes the minimum possible value of $\kappa(X)$ over all possible choices of X. In general it is difficult to determine this optimal value. However, for any nonsingular X we have the bound

$$\kappa_F(X) \geq \sum_i \|x_i\|_2 \|y_i\|_2,$$

where $X = [x_1, \ldots, x_n]$ and $X^{-1} = [y_1, \ldots, y_n]^H$, with equality if there is a nonzero α such that $\|x_i\|_2 = \alpha\|y_i\|_2$ for all i [21, Thm. 4.3.5]. If A has distinct eigenvalues then

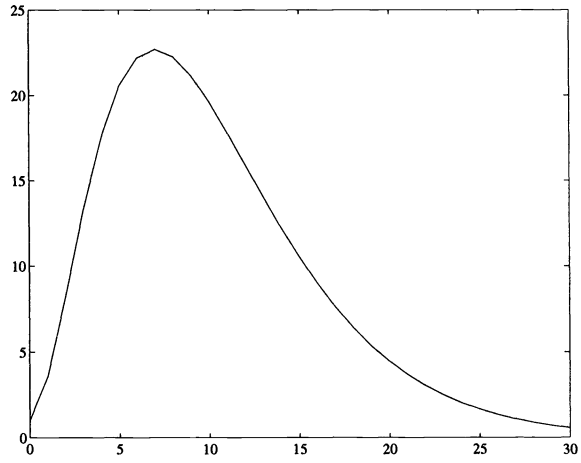


FIG. 2.1. *A typical hump for a convergent, nonnormal matrix.*

this lower bound is the same for all X in the Jordan form and is the minimum value of $\kappa_F(X)$. An alternative approach for matrices with distinct eigenvalues is to insist that the columns of X have unit 2-norm, for this gives a 2-norm condition number within a factor $n^{1/2}$ of the minimum, in view of a result by van der Sluis on diagonal scalings [24, Thm. 3.5]. However, we will see in §3 that to appreciate fully the various instability phenomena, we must consider defective problems.

If A is diagonalizable then, from (2.1a), we have the bound

$$\|A^k\|_p \leq \kappa_p(X)\rho(A)^k, \tag{2.3}$$

for any p-norm. (Since $\rho(A) \leq \|A\|$ for any norm, we also have the lower bound $\rho(A)^k \leq \|A^k\|_p$.) This bound is unsatisfactory for two reasons. First, by choosing A to have well-conditioned large eigenvalues and ill-conditioned small eigenvalues we can make the bound arbitrarily pessimistic. Second, it models norms of powers of convergent matrices as monotonically decreasing sequences, which is qualitatively incorrect if there is a large hump.

The Jordan canonical form can also be used to bound the norms of the powers of a defective matrix. If XJX^{-1} is the Jordan canonical form of $\delta^{-1}A$ then

$$\|A^k\|_p \leq \kappa_p(X)(\rho(A) + \delta)^k, \tag{2.4}$$

for all $\delta > 0$. This is a special case of a result of Ostrowski [18, Thm. 20.1], and a proof is straightforward: We can write $\delta^{-1}A = X(\delta^{-1}D + M)X^{-1}$, where $D = \operatorname{diag}(\lambda_i)$ and M is the off-diagonal part of the Jordan form. Then $A = X(D + \delta M)X^{-1}$, and (2.4) follows by taking norms. An alternative way of writing this bound is

$$\|A^k\|_p \leq \kappa_p(X)\kappa_p(D)(\rho(A) + \delta)^k,$$

where $A = XJX^{-1}$ and $D = \operatorname{diag}(\delta^{n-1}, \delta^{n-2}, \ldots, 1)$. Note that this is not the same X as in (2.4): multiplying A by a scalar changes $\kappa(X)$ when A is not diagonalizable. Both bounds suffer from the same problems as the bound (2.3) for diagonalizable matrices.

Another bound in terms of the Jordan canonical form (2.1) of A is given by Gautschi [4]. For convergent matrices, it can be written in the form

$$\|A^k\|_F \le c\, k^{p-1} \rho(A)^k, \tag{2.5}$$

where $p = \max\{n_i : \lambda_i \ne 0\}$ and c is a constant depending only on A (c is not defined explicitly in [4]). The factor k^{p-1} makes this bound somewhat more effective at predicting the shapes of the actual curve than (2.4), but again c can be unsuitably large.

Another way to estimate $\|A^k\|$ is to introduce a measure of nonnormality. Consider the Schur decomposition $Q^H A Q = D + N$, where N is strictly upper triangular, and let S represent the set of all such N. The nonnormality of A can be measured by Henrici's departure from normality [7]

$$\Delta(A, \|\cdot\|) \equiv \Delta(A) = \min_{N \in S} \|N\|.$$

For the Frobenius norm, Henrici shows that $\|N\|_F$ is independent of the particular Schur form and that

$$\Delta_F(A) = \left(\|A\|_F^2 - \sum_i |\lambda_i|^2\right)^{1/2} \le \left(\frac{n^3 - n}{12}\right)^{1/4} \|A^H A - A A^H\|_F^{1/2}.$$

László [15] has recently shown that $\Delta_F(A)$ is within a constant factor of the distance from A to the nearest normal matrix:

$$\Delta_F(A)/\sqrt{n} \le \nu(A) \le \Delta_F(A),$$

where $\nu(A) = \min\{\|E\|_F : A + E \text{ is normal}\}$. Henrici uses the departure from normality to derive the 2-norm bounds

$$\|A^k\|_2 \le \begin{cases} \displaystyle\sum_{i=0}^{n-1} \binom{k}{i} \rho(A)^{k-i} \Delta_2(A)^i, & \rho(A) > 0, \\ \Delta_2(A)^k, & \rho(A) = 0 \text{ and } k < n. \end{cases} \tag{2.6}$$

Empirical evidence suggests that the first bound in (2.6) can be very pessimistic. However, for normal matrices both the bounds are equalities. A bound of the same form as the first bound in (2.6), but with $\|A\|_2$ replacing $\Delta_2(A)$ and with an extra factor $2^{(n-1)/2}$, is obtained from a bound of Stafney in [20, Thm. 2.1] for $\|p(A)\|$, where p is a polynomial.

Another bound involving nonnormality is given by Golub and Van Loan [5, Lem. 7.3.2]. They show that, in the above notation,

$$\|A^k\|_2 \le (1+\theta)^{n-1} \left(\rho(A) + \frac{\Delta_F(A)}{1+\theta}\right)^k$$

for any $\theta \ge 0$. This bound is an analogue of (2.4) with the Schur form replacing the Jordan form. Again, there is equality when A is normal (if we set $\theta = 0$).

To compare bounds based on the Schur form with ones based on the Jordan form we need to compare $\Delta(A)$ with $\kappa(X)$. If A is diagonalizable then [16, Thm. 4]

$$\kappa_2(X) \ge \left(1 + \frac{\Delta_F(A)^2}{\|A\|_F^2}\right)^{1/2},$$

and it can be shown by a 2×2 example that $\min_X \kappa_2(X)$ can exceed $\Delta_F(A)/\|A\|_F$ by an arbitrary factor [2, §8.1.2], [1, §4.2.7].

Another tool that can be used to bound the norms of powers is the pseudospectrum of a matrix [22]. The ϵ-pseudospectrum of $A \in \mathbf{C}^{n\times n}$ is defined for a given $\epsilon > 0$ to be the set

$$\Lambda_\epsilon(A) = \{ z : z \text{ is an eigenvalue of } A + E \text{ for some } E \text{ with } \|E\|_2 \le \epsilon \},$$

and it can also be represented, in terms of the resolvent $(zI - A)^{-1}$, as

$$\Lambda_\epsilon(A) = \{ z : \|(zI - A)^{-1}\|_2 \ge \epsilon^{-1} \}.$$

As Trefethen notes [22], by using the Cauchy integral representation of A^k (which involves a contour integral of the resolvent) one can show that

$$\|A^k\|_2 \le \epsilon^{-1} \rho_\epsilon(A)^{k+1}, \tag{2.7}$$

where the ϵ-pseudospectral radius

$$\rho_\epsilon(A) = \max\{ |z| : z \in \Lambda_\epsilon(A) \}. \tag{2.8}$$

This bound is very similar in flavour to (2.4). The difficulty is transferred from estimating $\kappa(X)$ to choosing ϵ and estimating $\rho_\epsilon(A)$.

Finally, we mention that the Kreiss matrix theorem provides a good estimate of $\sup_{k\ge 0} \|A^k\|$ for a general $A \in \mathbf{C}^{n\times n}$, albeit in terms of an expression that involves the resolvent and is not easy to compute:

$$r(A) \le \sup_{k\ge 0} \|A^k\|_2 \le n\, e\, r(A),$$

where $r(A) = \sup\{ (|z| - 1)\|(zI - A)^{-1}\|_2 : |z| > 1 \}$ and $e = \exp(1)$. Details and references are given by Wegert and Trefethen [25].

3. Bounds for finite precision arithmetic. The formulae $A \cdot A^k$ or $A^k \cdot A$ can be implemented in several ways, corresponding to different loop orderings in each individual product, but as long as each product is formed using the standard formula $(AB)_{ij} = \sum_k a_{ij} b_{kj}$, all these variations satisfy the same rounding error bounds. We do not analyse here the use of the binary powering technique, where, for example, A^9 is formed as $A((A^2)^2)^2$, alternate multiplication on the left and right: $fl(A^k) = fl(Afl(A^{k-2})A)$, or the use of fast matrix multiplication techniques such as Strassen's method, since none of these methods is equivalent to repeated multiplication in finite precision arithmetic.

We suppose, without loss of generality, that the columns of A^m are computed one at a time, the jth as $fl(A(A(\ldots(Ae_j)\ldots)))$, where e_j is the jth unit vector. Standard error analysis shows that the jth computed column of A^m satisfies

$$fl(A^m e_j) = (A + \Delta A_1)(A + \Delta A_2)\ldots(A + \Delta A_m)e_j, \tag{3.1}$$

where

$$|\Delta A_i| \le c_n u |A|, \tag{3.2}$$

with c_n a constant of order n. (The inequality and absolute value are taken componentwise.) This bound holds for both real and complex matrices. It follows that

$$|fl(A^m e_j)| \le (1 + c_n u)^m |A|^m e_j,$$

and so a sufficient condition for convergence of the computed powers is that

$$\rho(|A|) < \frac{1}{1+c_n u}.$$

This result is useful in certain special cases: $\rho(|A|) = \rho(A)$ if A is triangular or has a checkerboard sign pattern (since then $|A| = DAD^{-1}$ where $D = \mathrm{diag}(\pm 1)$); if A is normal then $\rho(|A|) \leq \sqrt{n}\rho(A)$ (this bound being attained for a Hadamard matrix); and in Markov processes, where the a_{ij} are transition probabilities, $|A| = A$. However, in general $\rho(|A|)$ can exceed $\rho(A)$ by an arbitrary factor.

To obtain sharper and more informative results it is necessary to use more information about the matrix. Although the Jordan form is usually avoided by numerical analysts because of its sensitivity to perturbations, it is convenient to work with in this application and leads to informative results.

We point out that, because the analysis below is based on (3.1), our proofs of sufficient conditions for $fl(A^m) \to 0$ yield, with only trivial changes, sufficient conditions for $fl(A^m b) \to 0$, for any vector b. These conditions do not, however, exploit any special relations between A and b (such, as for example, b being an eigenvector of A).

3.1. Nilpotent matrices. We begin by considering nilpotent matrices, that is, those whose spectral radius is zero. The fact that nth power of an $n \times n$ nilpotent matrix is zero simplifies the analysis. The following theorem gives a bound on the norm of a computed power, together with a condition for the limit of the powers to be zero.

THEOREM 3.1. *Let $A \in \mathbf{C}^{n\times n}$ be a nilpotent matrix with the Jordan canonical form (2.1). A sufficient condition for $fl(A^m) \to 0$ as $m \to \infty$ is*

$$d_n u \kappa(X)\|A\| < 1 \tag{3.3}$$

for some p-norm, where u is the unit roundoff and d_n is a modest constant that depends only on n. Furthermore, if, for some $k \geq 1$ and $\theta > 1$,

$$\theta d_n u \kappa(X)^{\frac{k+1}{k}}\|A\| < 1, \tag{3.4}$$

then, with $t = \max_i n_i$,

$$\|fl(A^{rt})\| \leq n^{1-\frac{1}{p}} \frac{\kappa(X)^{1-\frac{r}{k}}\theta^{-r}}{1-\theta^{-1}\kappa(X)^{-\frac{1}{k}}} = O(\theta^{-r}), \quad r \geq k. \tag{3.5}$$

Proof. Taking norms in (3.1) we have

$$\|fl(A^m e_j)\| \leq \|(A+\Delta A_1)(A+\Delta A_2)\dots(A+\Delta A_m)\|.$$

Using the inequality

$$\|A\| \leq n^{1-\frac{1}{p}} \max_j \|Ae_j\| \tag{3.6}$$

from [10], we have

$$\|fl(A^m)\| \leq n^{1-\frac{1}{p}}\|(A+\Delta A_1)(A+\Delta A_2)\dots(A+\Delta A_m)\|.$$

Expanding this product and collecting together terms with the same number of ΔA_i factors we obtain the bound

$$\begin{aligned} n^{\frac{1}{p}-1}\|fl(A^m)\| \le \|A^m\| &+ \sum_{i=1}^{m} \|A^{i-1}\Delta A_i A^{m-i}\| \\ &+ \sum_{i=1}^{m-1}\sum_{j=1}^{m-i} \|A^{i-1}\Delta A_i A^{j-1}\Delta A_{i+j} A^{m-i-j}\| + \cdots \\ &+ \|\Delta A_1 \Delta A_2 \ldots \Delta A_m\|. \end{aligned}$$

From (3.2) we find, using (3.6) and an analogous result involving duality, that $\|\Delta A_i\| \le c_n' u\|A\|$, where $c_n' = n^{\min(1/p,1-1/p)}c_n$. Since $A^l = X^{-1}J^lX$ we have

$$\begin{aligned} n^{\frac{1}{p}-1}\|fl(A^m)\| \le \|A^m\| &+ \kappa(X)^2 c_n' u\|A\| \sum_{i=1}^{m} \|J^{i-1}\|\|J^{m-i}\| \\ &+ \kappa(X)^3 (c_n' u\|A\|)^2 \sum_{i=1}^{m-1}\sum_{j=1}^{m-i} \|J^{i-1}\|\|J^{j-1}\|\|J^{m-i-j}\| + \cdots \\ &+ (c_n' u\|A\|)^m. \end{aligned} \tag{3.7}$$

Now let $m = rt$, where $r \ge 1$. Since A is nilpotent, $A^t = J^t = 0$, and since every term in the first $r-1$ summations in (3.7) contains a factor $\|J^i\|$ with $i \ge t$, all these terms disappear. Furthermore, in the remaining summations we need only count terms in which all the exponents of J are less than t (again, the other terms disappear). Overall, we have, using the fact that $\|J^i\| = 1$ $(0 \le i < t)$,

$$n^{\frac{1}{p}-1}\|fl(A^{rt})\| \le \kappa(X) \sum_{j=r}^{rt} \left(tc_n' u\kappa(X)\|A\|\right)^j.$$

Now suppose that (3.4) holds with $d_n = tc_n'$, for some $\theta > 1$. Then, for $r \ge k$,

$$\begin{aligned} n^{\frac{1}{p}-1}\|fl(A^{rt})\| &\le \kappa(X) \sum_{j=r}^{rt} \left(\theta\kappa(X)^{\frac{1}{k}}\right)^{-j} \\ &\le \kappa(X)^{1-\frac{r}{k}}\theta^{-r} \sum_{j=0}^{rt-r} \left(\theta\kappa(X)^{\frac{1}{k}}\right)^{-j} \\ &\le \frac{\kappa(X)^{1-\frac{r}{k}}\theta^{-r}}{1-\theta^{-1}\kappa(X)^{-\frac{1}{k}}}. \end{aligned}$$

This gives the second part of the theorem. The first part follows immediately by choosing $\theta = 1+\epsilon$, with ϵ an arbitrarily small positive number, and taking the limit as $k \to \infty$. □

In practice we may have a computed matrix $\widehat{A} \approx A$ that is not exactly nilpotent. As long as $\|A - \widehat{A}\| \le c_n u\|A\|$, we can absorb the error $A - \widehat{A}$ into the terms ΔA_i in the proof, and so by applying the theorem to A we will obtain conclusions valid for $\widehat{A}$.

To exhibit the sharpness of the bounds we give the following example, using the Chebyshev spectral differentiation matrix $C_n \in \mathbb{R}^{n\times n}$ described in [23]. The matrix

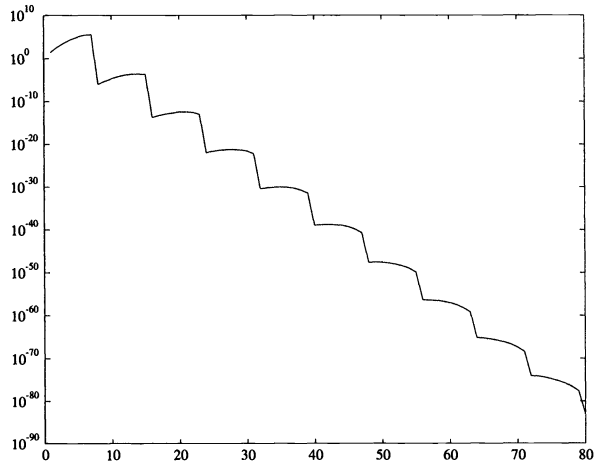


FIG. 3.1. *Converging powers of the nilpotent matrix* C_8.

C_n arises from degree $n-1$ polynomial interpolation of n arbitrary data values at n Chebyshev points, including a boundary condition at $x_0 = 1$. It is nilpotent and is similar to a single Jordan block of dimension n. We generate C_n in MATLAB using the routine **chebspec** from the test collection of Higham [9], [11].

Figure 3.1 shows the 2-norms of the computed powers of C_8 and Fig. 1.1 those of C_{14}. The powers of C_8 converge to zero, while the powers of C_{14} diverge.

To check the sharpness of the bounds in Theorem 3.1 we need an estimate of the condition number of the matrix X_n in the Jordan canonical form of C_n. We outline our approach in Appendix A. Our estimate for $\kappa_2(X_8)$ is 3.42×10^5, and $\|C_8\|_2 = 28.56$. Table 3.1 gives the order of the bound (3.5) for a number of powers, with $r = k$ and θ chosen as large as possible so that (3.4) is satisfied (we take $d_n = n$, instead of the actual value $d_n \approx n^{5/2}$ for this example, to allow for the inevitable overestimation of errors inherent in a strict rounding error bound of this type). The actual computed order is given for comparison and clearly there is reasonable agreement. According to (3.3), we require $d_n \kappa(X)\|A\|u < 1$ to guarantee that the computed powers of A converge to zero. For C_{14} we have $u\kappa_2(X)\|C_{14}\|_2 \approx 0.28$ so, allowing for d_n, (3.3) correctly does not predict convergence of the computed powers.

To emphasize that the behaviour of the computed powers is scale-dependent, we mention that the computed powers of $15C_8$ diverge. Again, this is in accord with Theorem 3.1 because $u\kappa_2(X)\|15C_8\|_2 \approx 2.7$. Finally, we note that for C_{12}, Theorem 3.1 again correctly predicts convergence of the computed powers, but the powers computed by alternate left and right multiplication and by binary powering diverge; this confirms that our analysis is not applicable to these forms of multiplication.

3.2. General matrices. Now we turn to general convergent matrices. In contrast to the theory we have developed for nilpotent matrices, we need separate theorems to describe the limiting behaviour of the matrix powers and to bound the norm for a finite exponent. In the following theorem we give a sufficient condition, based on the Jordan canonical form, for the computed powers of a matrix to converge to zero.

TABLE 3.1
Expected and actual orders of $\|fl(C_8^m)\|_2$.

Power	Bound	Actual
$m = 8$	10^{-2}	10^{-6}
$m = 16$	10^{-11}	10^{-14}
$m = 32$	10^{-27}	10^{-31}
$m = 64$	10^{-59}	10^{-66}

THEOREM 3.2. *Let* $A \in \mathbf{C}^{n\times n}$ *with the Jordan form* (2.1) *have spectral radius* $\rho(A) < 1$. *A sufficient condition for* $fl(A^m) \to 0$ *as* $m \to \infty$ *is*

$$d_n u\kappa(X)\|A\| < (1-\rho(A))^t \tag{3.8}$$

for some p-norm, where $t = \max_i n_i$ *and* d_n *is a modest constant depending only on* n.

Proof. Since any two p-norms differ by a factor at most n, we need only show convergence for one particular norm. We choose the ∞-norm.

It is easy to see that if we can find a nonsingular matrix S such that

$$\|S^{-1}AS\|_\infty + \kappa_\infty(S)\|\Delta A_i\|_\infty < 1 \tag{3.9}$$

for all i, then the product $(A+\Delta A_1)\dots(A+\Delta A_m) = S(S^{-1}AS+S^{-1}\Delta A_1 S)\dots(S^{-1}AS+S^{-1}\Delta A_m S)S^{-1} \to 0$ as $m \to \infty$. In the rest of the proof we construct such a matrix S for the ΔA_i in (3.1).

Let $P(\epsilon) = \operatorname{diag}(P_1(\epsilon),\dots,P_s(\epsilon))$ where $0 < \epsilon < 1-\rho(A)$ and

$$P_i(\epsilon) = \operatorname{diag}((1-|\lambda_i|-\epsilon)^{1-n_i}, (1-|\lambda_i|-\epsilon)^{2-n_i},\dots,1) \in \mathbb{R}^{n_i\times n_i}.$$

Now consider the matrix $P(\epsilon)^{-1}JP(\epsilon)$. Its ith diagonal block is of the form $\lambda_i I + (1-|\lambda_i|-\epsilon)N$, where the only nonzeros in N are 1s on the first superdiagonal, and so

$$\|P(\epsilon)^{-1}X^{-1}AXP(\epsilon)\|_\infty = \|P(\epsilon)^{-1}JP(\epsilon)\|_\infty \le \max_i(|\lambda_i| + 1 - |\lambda_i| - \epsilon) = 1-\epsilon.$$

Defining $S = XP(\epsilon)$, we have $\|S^{-1}AS\|_\infty \le 1-\epsilon$ and

$$\kappa_\infty(S) \le \kappa_\infty(P(\epsilon))\kappa_\infty(X) \le (1-\rho(A)-\epsilon)^{1-t}\kappa_\infty(X). \tag{3.10}$$

Now we set $\epsilon = \theta(1-\rho(A))$ where $0 < \theta < 1$ and we determine θ so that (3.9) is satisfied, that is, so that $\kappa_\infty(S)\|\Delta A_i\|_\infty < \epsilon$ for all i. From (3.2) and (3.10) we have

$$\kappa_\infty(S)\|\Delta A_i\|_\infty \le c_n u(1-\theta)^{1-t}(1-\rho(A))^{1-t}\kappa_\infty(X)\|A\|_\infty.$$

Therefore (3.9) is satisfied if

$$c_n u(1-\theta)^{1-t}(1-\rho(A))^{1-t}\kappa_\infty(X)\|A\|_\infty < \theta(1-\rho(A)),$$

that is, if

$$c_n u\kappa_\infty(X)\|A\|_\infty < (1-\theta)^{t-1}\theta(1-\rho(A))^t.$$

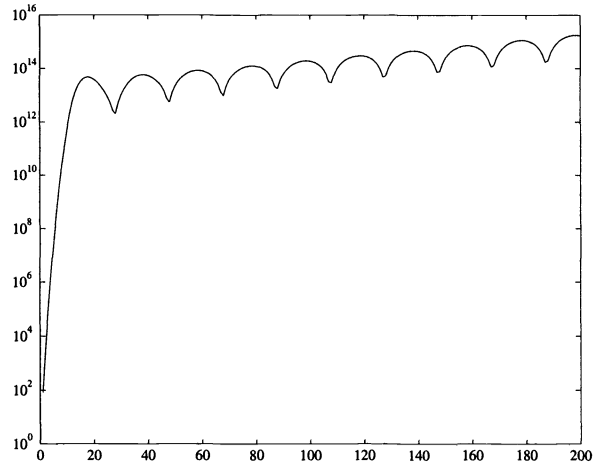


FIG. 3.2. *Diverging powers of* $C_{13} + 0.36I$.

If the integer t is greater than 1 then the function $f(\theta) = (1-\theta)^{t-1}\theta$ has a maximum on $[0,1]$ at $\theta_* = t^{-1}$ and $f(\theta_*) = (t-1)^{-1}(1-t^{-1})^t$ satisfies $(4(t-1))^{-1} \leq f(\theta_*) < e^{-1}$. We conclude that for all integers $1 \leq t \leq n$,

$$c_n u \kappa_\infty(X)\|A\|_\infty < \frac{1}{4t}(1-\rho(A))^t$$

is sufficient to ensure that (3.9) holds. The theorem is proved with $d_n = 4tc_n$. □

If A is normal then $\|A\|_2 = \rho(A) < 1$, $t = 1$, and $\kappa_2(X) = 1$, so (3.8) takes the form

$$\rho(A) < \frac{1}{1+d_n u}.$$

This condition is also easily derived by taking 2-norms in (3.1) and (3.2).

Again, we can show the sharpness of this bound by using the Chebyshev spectral differentiation matrix C_n, this time adding multiples of the identity matrix.

Figure 3.2 shows the nonconverging 2-norms of the first 200 computed powers of $A = C_{13} + 0.36I$. Since the same matrix X takes both C_{13} and A to Jordan form, we can use the same routines as for our nilpotent examples to estimate $\kappa(X)$. Our estimate for $\kappa_2(X)\|A\|_2 u/(1-\rho(A))^{13}$ in this case is 3.05. On the other hand, the computed powers of $A = C_{13}+0.01I$ converge to zero, and $\kappa_2(X)\|A\|_2 u/(1-\rho(A))^{13} \approx 0.01$. Thus our bound (3.8) is reasonably sharp.

Figure 3.2 reveals an interesting scalloping pattern in the curve of the norms. In Figs. 1.1 and 3.1 for nilpotent matrices the norm dips whenever the power is a multiple of the dimension of the matrix. Here the norm first dips for $\|fl(A^{28})\|_2$ and then regularly after every further 20 powers, but the point of first dip and the dipping intervals can be altered by adding different multiples of the identity matrix. The reason for this behaviour is not clear.

A difficulty we have when attempting to bound $\|fl(A^m)\|$ for a finite m is that, as explained in §2, we do not have a good estimate of the true value $\|A^m\|$. If we do

have such an estimate we can prove results similar to (3.4) and (3.5), although we are not able to determine a precise bound for $\|fl(A^m)\|$ in simple form.

THEOREM 3.3. *Let $A \in \mathbf{C}^{n\times n}$ with the Jordan form (2.1) have spectral radius $\rho(A) < 1$. Let q be such that $\|A^q\| = cu$ where $c = O(1)$ and suppose that, for some $k \geq 1$ and $\theta > 1$,*

$$\theta d_n q u \mu^{\frac{k+1}{k}} \|A\| < 1,$$

where $\mu = \kappa(X)/(1-\rho(A))^{t-1}$ and $t = \max_i n_i$. Then

$$\|fl(A^{rq})\| = O(\theta^{-r}), \quad r \geq k.$$

Proof. We omit the proof of the theorem, which is very similar to the proof of Theorem 3.1. □

We conclude this section by commenting that the proof of Theorem 3.2 can be adapted to use the Schur decomposition in place of the Jordan canonical form. The modified analysis leads to the sufficient condition

$$d_n' u \|N\|_F^{n-1} \|A\|_2 < (1-\rho(A))^n \tag{3.11}$$

for $fl(A^m) \to 0$ as $m \to \infty$, where N is the strictly upper triangular part of the Schur form. This condition is weaker than (3.8) in two respects. First, it takes no account of the defectiveness of A, because it contains a power n on the right-hand side instead of $t = \max_i n_i \leq n$. Second, under the scaling $A \leftarrow \alpha A$ the left-hand side of (3.11) scales by $|\alpha|^n$, which tends to make the left-hand side of (3.11) much larger than that of (3.8) when $\|A\|_F > 1$. It is an open question how to obtain a sharp sufficient condition for convergence in terms of the Schur decomposition.

4. A pseudospectral approach. In this section we show how the pseudospectrum can be used to predict the limiting behaviour of a computed sequence of powers. Figure 4.1 shows approximations to pseudospectra for the matrices in the examples of Figs. 1.1, 3.1, and 3.2; we have approximated $\Lambda_\epsilon(A)$ with $\epsilon = c_n\|A\|_2 u$, taking $c_n = n$ for simplicity. The inner ring is an approximation to the pseudospectrum of C_8, that of C_{14} is marked by + and that of $C_{13} + 0.36I$ is marked by o. The solid curve is the unit disc.

A heuristic argument based on (3.1) and (3.2) suggests that, if for randomly chosen perturbations ΔA_i with $\|\Delta A_i\| \leq c_n u \|A\|$, most of the eigenvalues of the perturbed matrices lie outside the unit disc, then we can expect a high percentage of the terms $A + \Delta A_i$ in (3.1) to have spectral radius bigger than one and hence we can expect the product to diverge. On the other hand, if the $c_n u\|A\|$-pseudospectrum is wholly contained within the unit disc, each $A+\Delta A_i$ will have spectral radius less than one and the product can be expected to converge. (Note, however, that if $\rho(A) < 1$ and $\rho(B) < 1$ it is not necessarily the case that $\rho(AB) < 1$.) To make this heuristic precise, we need an analogue of Theorem 3.2 phrased in terms of the pseudospectrum rather than the Jordan form.

To obtain such an analogue directly from Theorem 3.2 we need to relate $\kappa(X)$ to the pseudospectral radius $\rho_\epsilon(A)$ (see (2.8)). If we can show that

$$\rho_\epsilon(A) \geq \rho(A) + (c_n \epsilon \kappa(X))^{1/t} \tag{4.1}$$

for a particular ϵ, then $\rho_\epsilon(A) < 1$ implies $c_n\epsilon\kappa(X) < (1-\rho(A))^t$, which is a condition of the same form as (3.8). In Theorem 4.2 we show that (4.1) holds for diagonalizable

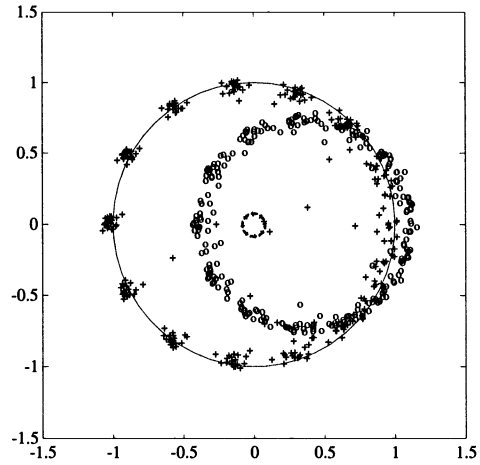


FIG. 4.1. *Pseudospectra of the three example matrices.*

matrices to first order, under a certain assumption. We need the following standard result (see, for example, [21, pp. 183–184]).

THEOREM 4.1. *Let λ be a simple eigenvalue of the matrix A, with right and left eigenvectors x and y, and let $\widetilde{A} = A + E$ be a perturbation of A. Then there is an eigenvalue $\widetilde{\lambda}$ of $\widetilde{A}$ such that*

$$\widetilde{\lambda} = \lambda + \frac{y^H E x}{y^H x} + O(\|E\|^2).$$

We can now prove the following theorem.

THEOREM 4.2. *Let $A \in \mathbf{C}^{n\times n}$ have the Jordan canonical form (2.1), with $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$. Suppose that $\|X\|_1 = \sum_{i=1}^n |x_{i1}|$ and $\|X^{-1}\|_\infty = \sum_{j=1}^n |y_{1j}|$, where $X^{-1} = (y_{ij})$. Then there is a perturbation $\widetilde{A} = A + E$ of A, with $\|E\| = \epsilon$ for all p-norms, such that*

$$\rho(\widetilde{A}) \geq \rho(A) + \frac{\kappa(X)\epsilon}{n^2} + O(\epsilon^2). \tag{4.2}$$

Proof. By assumption, λ_1 is a simple eigenvalue, so $x_1 = Xe_1$ and $y_1 = (e_1^T X^{-1})^H$ are the right eigenvector and the left eigenvector corresponding to λ_1. From Theorem 4.1 we know that any perturbation $\widetilde{A}$ will have an eigenvalue

$$\widetilde{\lambda} = \lambda_1 + y_1^H E x_1 + O(\|E\|^2)$$

(since $y_1^H x_1 = 1$). Define E by $|e_{ij}| = \epsilon/n$ and $\arg(e_{ij}) = \arg(y_{1i}) - \arg(x_{j1}) + \arg(\lambda_1)$. Then E has rank 1, $\|E\| = \epsilon$ for all p, and

$$\widetilde{\lambda} = e^{i\arg(\lambda_1)}(|\lambda_1| + \frac{\epsilon}{n}\|y_1^H\|_1\|x_1\|_1) + O(\epsilon^2). \tag{4.3}$$

Now for an $n \times n$ matrix B and any $1 \leq p, q \leq \infty$ [10],

$$\|B\|_p \leq n^{\left(\frac{1}{\min(p,q)} - \frac{1}{\max(p,q)}\right)}\|B\|_q,$$

and together with the conditions of the theorem this gives

$$\|y_1^H\|_1\|x_1\|_1 = \|X^{-1}\|_\infty\|X\|_1 \geq \frac{\kappa(X)}{n}.$$

The proof is completed by taking absolute values in (4.3). □

Because of the assumptions on $\|X\|_1$ and $\|X^{-1}\|_\infty$ we do not have the freedom to choose X to minimize $\kappa(X)$ in Theorem 4.2. We note, however, that if A is diagonalizable and X has columns all of the same norm, then the condition in Theorem 4.2 on the rows and columns of X and X^{-1} reduces to the requirement that the eigenvalue of largest modulus be the most ill conditioned.

Theorem 4.2 enables us to obtain the following corollary of Theorem 3.2.

COROLLARY 4.3. *Suppose that $A \in \mathbf{C}^{n\times n}$ is diagonalizable and satisfies the conditions of Theorem 4.2, and suppose that the $O(\epsilon^2)$ term in (4.2) is negligible. If $\rho_\epsilon(A) < 1$ for $\epsilon = c_n\|A\|u$, where c_n is a modest constant depending only on n, then $\lim_{m\to\infty} fl(A^m) = 0$.*

Proof. By Theorem 4.2, if $\rho_\epsilon(A) < 1$ then

$$\rho(A) + \epsilon\kappa(X)/n^2 < 1.$$

Rearranging gives

$$c_n u\kappa(X)\|A\|/n^2 < 1 - \rho(A).$$

Using Theorem 3.2 we have the required result for $c_n = n^2 d_n$, since $t = 1$. □

Suppose we compute the eigenvalues of A by a backward stable algorithm, that is, one that yields the exact eigenvalues of $A + E$, where $\|E\|_2 \leq c_n u\|A\|_2$, with c_n a modest constant. (An example of such an algorithm is the QR algorithm [5, §7.5]). Then the computed spectral radius $\widehat{\rho}$ satisfies $\widehat{\rho} \leq \rho_{c_n u\|A\|_2}(A)$. In view of Corollary 4.3 we can formulate a rule of thumb:

> The computed powers of A can be expected to converge to zero if the spectral radius computed via a backward stable eigensolver is less than 1.

This rule of thumb has also been discussed by Trefethen and Trummer [23] and Reichel and Trefethen [19]. In our experience the rule of thumb is fairly reliable when $\widehat{\rho}$ is not too close to 1. For the matrices used in our examples we have

$$\widehat{\rho}(C_8) = 0.073, \quad \widehat{\rho}(15C_8) = 2.7, \quad \widehat{\rho}(C_{14}) = 1.005,$$
$$\widehat{\rho}(C_{13} + 0.01I) = 0.70, \quad \widehat{\rho}(C_{13} + 0.36I) = 1.05,$$

and we observed convergence of the computed powers for C_8 and $C_{13} + 0.01I$ and divergence for the other matrices.

Appendix A. Approximating X in the jordan form of C_n. In §3 we needed an estimate of $\kappa(X)$ for the Chebyshev differentiation matrix, C, where $C = XJX^{-1}$ is the Jordan form. In this appendix we outline our approach for computing an estimate of $\kappa(X)$.

Recall that

$$C = XJX^{-1}, \tag{A.1}$$

where J is a single Jordan block whose diagonal is zero. Suppose we decompose C into Schur form via the orthogonal matrix Q (which is real since C's spectrum is real), that is,

$$C = QTQ^T,$$

where T is upper triangular with zero diagonal. If we can find an upper triangular matrix R such that $T = RJR^{-1}$ then $X = QR$ and $\kappa_2(X) = \kappa_2(R)$. We require $TR = RJ$, that is, $Tr_j = r_{j-1}$, $2 \le j \le n$, and $Tr_n = 0$, where r_j is the jth column of R.

We choose the arbitrary last column of R to be the last column of the identity matrix. The following algorithm computes R (here, we use the MATLAB colon notation).

$R(:,n) = e_n$
for $j = n-1:-1:1$
  $R(1:j,j) = T(1:j,1:j+1)R(1:j+1,j+1)$
end

It remains to compute the Schur form of C. We do not use the QR algorithm to compute the Schur form, as for nilpotent matrices it can lead to triangular matrices with elements of order 1 on the diagonal. We use the following algorithm described by Golub and Wilkinson in [6, §10], which, although computationally expensive, has good error properties.

Compute the SVD of $C_1 = C = U_1\Sigma_1 V_1^T$.
for $i = 1:n-2$
  $C_{i+1} = V_i^T C_i V_i$
  Compute the SVD $C_{i+1}(1:n-i, 1:n-i) = U_{i+1}\Sigma_{i+1}W_{i+1}^T$.
  $V_{i+1} = \mathrm{diag}(W_{i+1}, I_i)$
end
$L = V_{n-1}^T C_{n-1} V_{n-1}$
$Q = V_1 V_2 \ldots V_{n-1}$

Upon completion of the algorithm we have $C = QLQ^T$ with L lower triangular, and so we apply our first algorithm to L^T to estimate $\kappa_2(X)$ (note that the Jordan matrix for A^T is a permutation of the one for A [14, §3.2.3]).

Acknowledgments. We thank Des Higham and Nick Trefethen for their comments on the manuscript.

REFERENCES

[1] F. CHATELIN, *Eigenvalues of Matrices*, John Wiley, Chichester, 1993.
[2] F. CHATELIN AND V. FRAYSSÉ, *Qualitative computing: Elements of a theory for finite precision computation*, lecture notes, CERFACS, Toulouse, France and THOMSON-CSF, Orsay, France, June 1993.
[3] F. R. GANTMACHER, *The Theory of Matrices*, Vol. 1, Chelsea, New York, 1959.
[4] W. GAUTSCHI, *The asymptotic behaviour of powers of matrices*, Duke Math. J., 20 (1953), pp. 127–140.
[5] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, MD, 1989.
[6] G. H. GOLUB AND J. H. WILKINSON, *Ill-conditioned eigensystems and the computation of the Jordan canonical form*, SIAM Rev., 18 (1976), pp. 578–619.
[7] P. HENRICI, *Bounds for iterates, inverses, spectral variation and fields of values of non-normal matrices*, Numer. Math., 4 (1962), pp. 24–40.

[8] D. J. HIGHAM AND L. N. TREFETHEN, *Stiffness of ODEs*, BIT, 33 (1993), pp. 285–303.
[9] N. J. HIGHAM, *Algorithm* 694: *A collection of test matrices in MATLAB*, ACM Trans. Math. Software, 17 (1991), pp. 289–305.
[10] ———, *Estimating the matrix p-norm*, Numer. Math., 62 (1992), pp. 539–555.
[11] ———, *The Test Matrix Toolbox for Matlab*, Numerical Analysis Report No. 237, University of Manchester, England, Dec. 1993.
[12] N. J. HIGHAM AND P. A. KNIGHT, *Componentwise error analysis for stationary iterative methods*, in Linear Algebra, Markov Chains, and Queueing Models, C. D. Meyer and R. J. Plemmons, eds., Vol. 48, IMA Volumes in Mathematics and its Applications, Springer-Verlag, New York, 1993, pp. 29–46.
[13] ———, *Finite precision behavior of stationary iteration for solving singular systems*, Linear Algebra Appl., 192 (1993), pp. 165–186.
[14] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, 1985.
[15] L. LÁSZLÓ, *An attainable lower bound for the best normal approximation*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1035–1043.
[16] G. LOIZOU, *Nonnormality and Jordan condition numbers of matrices*, J. Assoc. Comput. Mach., 16 (1969), pp. 580–584.
[17] C. B. MOLER AND C. F. VAN LOAN, *Nineteen dubious ways to compute the exponential of a matrix*, SIAM Rev., 20 (1978), pp. 801–836.
[18] A. M. OSTROWSKI, *Solution of Equations in Euclidean and Banach Spaces*, Academic Press, New York, 1973; Solution of Equations and Systems of Equations, 3rd ed.
[19] L. REICHEL AND L. N. TREFETHEN, *Eigenvalues and pseudo-eigenvalues of Toeplitz matrices*, Linear Algebra Appl., 162–164 (1992), pp. 153–185.
[20] J. D. STAFNEY, *Functions of a matrix and their norms*, Linear Algebra Appl., 20 (1978), pp. 87–94. Correction in Linear Algebra Appl., 39 (1981), pp. 259–260.
[21] G. W. STEWART AND J. SUN, *Matrix Perturbation Theory*, Academic Press, London, 1990.
[22] L. N. TREFETHEN, *Pseudospectra of matrices*, in Numerical Analysis 1991, Proc. 14th Dundee Conference, D. F. Griffiths and G. A. Watson, eds., Vol. 260, Pitman Research Notes in Mathematics, Longman Scientific and Technical, Essex, UK, 1992, pp. 234–266.
[23] L. N. TREFETHEN AND M. R. TRUMMER, *An instability phenomenon in spectral methods*, SIAM J. Numer. Anal., 24 (1987), pp. 1008–1023.
[24] A. VAN DER SLUIS, *Condition numbers and equilibration of matrices*, Numer. Math., 14 (1969), pp. 14–23.
[25] E. WEGERT AND L. N. TREFETHEN, *From the Buffon needle problem to the Kreiss matrix theorem*, Amer. Math. Monthly, 101 (1994), pp. 132–139.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 359–368, January 1995

002

THE EXTENDED LINEAR COMPLEMENTARITY PROBLEM*

O. L. MANGASARIAN† AND J. S. PANG‡

Abstract. We consider an extension of the horizontal linear complementarity problem, which we call the extended linear complementarity problem (XLCP). With the aid of a natural bilinear program, we establish various properties of this extended complementarity problem; these include the convexity of the bilinear objective function under a monotonicity assumption, the polyhedrality of the solution set of a monotone XLCP, and an error bound result for a nondegenerate XLCP. We also present a finite, sequential linear programming algorithm for solving the nonmonotone XLCP.

Key words. complementarity problems, monotonicity, error bound, bilinear program

AMS subject classifications. 90C30, 90C33

1. Introduction.

In the past couple of years, the horizontal linear complementarity problem (HLCP) has received an increasing amount of attention among researchers interested in the family of interior-point methods for solving linear programs and complementarity problems. This surge of interest originates from an article by Zhang [17] who used the HLCP as a unifying framework for the convergence analysis of a class of so-called "infeasible-interior-point algorithms." Subsequent work in this area includes [2], [7], [12], [14]. Independently, Sznajder and Gowda [13] have studied some matrix-theoretic properties and their roles in the horizontal and vertical LCPs. Inspired by this flurry of activities and other applications (like the one described in [4], [15]), we became interested in undertaking a further study of the HLCP. In particular, our goal in this paper is twofold: one, to derive some basic results of the HLCP along the line of the classical LCP [3]; and, two, to present an alternative solution method for the HLCP (particularly, for the "nonmonotone" problems).

The problem we study in this paper is defined as follows. Let M and N be two real matrices of order $m \times n$, and let C be a polyhedral set in R^m. The extended linear complementarity problem, which we denote XLCP (M, N, C), is to find a pair of vectors $(x, y) \in R^{2n}_+$ such that

$$Mx - Ny \in C, \qquad x \perp y,$$

where the notation $x \perp y$ means that x is orthogonal to y, i.e., $x^T y = 0$. When $m = n$ and C consists of the single vector $p \in R^n$, this problem reduces to the HLCP that has motivated our work. In general, when

$$C = \{Lz + q : z \in R^\ell\}$$

for some matrix $L \in R^{m \times \ell}$ and vector $q \in R^m$, the XLCP (M, N, C) becomes the "general linear complementarity problem" studied by Ye [16]. However, Gowda [6]

* Received by the editors November 19, 1993; accepted for publication (in revised form) by R. Cottle, April 8, 1994. The work of the first author was based on research supported by Air Force Office of Scientific Research grant AFOSR-F49620-94-1-0036 and National Science Foundation grant CCR-9101801. The work of the second author was based on research supported by National Science Foundation under grants DDM-9104078 and CCR-92137389 and by Office of Naval Research grant N00014-93-1-0228.

† Department of Computer Sciences, University of Wisconsin, Madison, Wisconsin 53706 (olvi@cs.wisc.edu).

‡ Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, Maryland 21218-2689 (jsp@vicp1.mts.jhu.edu).

pointed out that the XLCP can also be obtained from Ye's general linear complementarity problem, and hence the two are equivalent. Ye proposed an interior point approach for this problem, whereas our approach is based on finite bilinear programming.

The feasible region of XLCP (M, N, C) is denoted $\mathrm{FEA}(M, N, C)$; it is defined to be the set

$$\mathrm{FEA}(M, N, C) \equiv \{(x, y) \in R^{2n}_{+} : Mx - Ny \in C\},$$

which is a polyhedral subset of R^{2n}_{+}. We say that XLCP (M, N, C) is *feasible* if $\mathrm{FEA}(M, N, C)$ is nonempty. The set of complementary solutions of the XLCP (M, N, C) is given by

$$\mathrm{SOL}(M, N, C) \equiv \{(x, y) \in \mathrm{FEA}(M, N, C) : x \perp y\}.$$

2. The equivalent bilinear program. Associated with the XLCP (M, N, C) is a natural bilinear program defined on the same feasible region:

$$\begin{aligned} &\text{minimize} && x^T y \\ &\text{subject to} && (x, y) \in \mathrm{FEA}(M, N, C). \end{aligned}$$

We denote this problem by BLP (M, N, C). The BLP (M, N, C) should be contrasted with the "natural" quadratic program that one associates with the standard LCP (q, M), which corresponds to the special case of the XLCP (M, N, C) with $m = n$, $N = I$, and $C = \{-q\}$. The latter quadratic program is [3]

$$\begin{aligned} &\text{minimize} && x^T(q + Mx) \\ &\text{subject to} && x \geq 0, \quad q + Mx \geq 0. \end{aligned} \tag{1}$$

One important distinction between the BLP (M, N, C) and the quadratic program (1) is that the latter is defined by the variable x only, whereas the former involves the pair (x, y). We shall see shortly that the BLP (M, N, C) plays a similar role in the study of the XLCP (M, N, C) as (1) in the LCP (q, M).

Since the objective function of BLP (M, N, C) is nonnegative on $\mathrm{FEA}(M, N, C)$, the XLCP (M, N, C) is equivalent to the BLP (M, N, C) in the sense that a pair of vectors (x, y) solves the former problem if and only if (x, y) is a globally optimal solution of the latter problem with a zero objective value. Moreover, by the well-known Frank–Wolfe Theorem of quadratic programming [5], the BLP (M, N, C) always has an optimal solution provided that it is feasible. Of course, it is generally not necessary for an optimal solution of the BLP (M, N, C) to have zero objective value. In what follows, we establish several results that pertain to the relationship between the XLCP and the associated BLP.

PROPOSITION 2.1. *Let M and N be $m \times n$ matrices and C a polyhedral set in R^m. The bilinear function $f(x, y) \equiv x^T y$ is convex on the set $\mathrm{FEA}(M, N, C)$ if and only if the following implication holds:*

$$\left[(x^i, y^i) \in \mathrm{FEA}(M, N, C), i = 1, 2 \right] \Rightarrow (x^1 - x^2)^T (y^1 - y^2) \geq 0. \tag{2}$$

Proof. By an easy calculation, it can be verified that the following identity holds for any two pairs of vectors $(x^i, y^i) \in R^{2n}$ and any scalar τ,

$$\tau (x^1)^T y^1 + (1 - \tau)(x^2)^T y^2 - x(\tau)^T y(\tau) = \tau(1 - \tau)(x^1 - x^2)^T (y^1 - y^2),$$

where

$$\begin{pmatrix} x(\tau) \\ y(\tau) \end{pmatrix} = \tau \begin{pmatrix} x^1 \\ y^1 \end{pmatrix} + (1-\tau) \begin{pmatrix} x^2 \\ y^2 \end{pmatrix}.$$

Thus, the claimed equivalence follows easily. □

With the somewhat notorious reputation of the bilinear function, the above proposition is a pleasant surprise in that it exhibits an important instance in which the BLP (M, N, C) is actually a "convex program" (in the sense that it has a convex objective function on the feasible set). Indeed, when one specializes this result to the case of the standard LCP (q, M), one may conclude that if M is a positive semidefinite matrix, then the bilinear form $x^T y$ is a convex function on the set $\{(x, y) \in R^{2n}_+ : Mx - y = q\}$. This fact, though trivial to prove, seems to have been completely overlooked in the LCP literature.

To state the next result, which gives a sufficient condition for every Karush–Kuhn–Tucker (KKT) vector of the (general) BLP (M, N, C) to be a solution of the XLCP (M, N, C), we recall that a matrix $L \in R^{m \times m}$ is copositive on a cone $K \subseteq R^m$ if $u^T L u \geq 0$ for every $u \in K$. Also, we denote the recession cone of the set C by 0^+C; finally, the dual cone of a set $S \subseteq R^m$ is denoted S^*.

PROPOSITION 2.2. *Let M and N be $m \times n$ matrices and C a polyhedral set in R^m. If the matrix $MN^T \in R^{m \times m}$ is copositive on $(0^+C)^*$, then every* KKT *vector of the* BLP (M, N, C), *if it exists, solves the* XLCP (M, N, C). *Thus, if in addition* $\mathrm{FEA}(M, N, C) \neq \emptyset$, *then* $\mathrm{SOL}(M, N, C) \neq \emptyset$.

Proof. Without loss of generality, we represent the set C in the following form:

$$C = \{u \in R^m : Au \geq b\},$$

for some matrix $A \in R^{\ell \times m}$ and vector $b \in R^\ell$. Then we have

$$(0^+C)^* = \{v \in R^m : v = A^T \lambda \text{ for some } \lambda \in R^\ell_+\},$$

and the BLP (M, N, C) becomes

$$\begin{aligned} \text{minimize} \quad & x^T y \\ \text{subject to} \quad & AMx - ANy \geq b, \\ & (x, y) \geq 0. \end{aligned}$$

Now, if (x, y) is a KKT vector of the BLP (M, N, C), then there exist nonnegative vectors $\lambda \in R^\ell$, and $(r, s) \in R^{2n}$ such that

$$\begin{aligned} y &= M^T A^T \lambda + r, \qquad x = -N^T A^T \lambda + s, \\ x^T r &= y^T s = \lambda^T (AMx - ANy - b) = 0. \end{aligned}$$

Clearly, we have

$$\begin{aligned} x^T y &= x^T y - x^T r - s^T y + s^T r - s^T r \\ &= (x - s)^T (y - r) - s^T r = -\left((\lambda^T A) M N^T (A^T \lambda) + r^T s\right) \leq 0, \end{aligned}$$

where the last inequality follows from the copositivity of MN^T on $(0^+C)^*$. Since $x^T y$ is nonnegative, it follows that $(x, y) \in \mathrm{SOL}(M, N, C)$.

The last assertion of the proposition holds because the BLP (M, N, C) must have an optimal solution if it is feasible, and such a minimum solution must also be a complementary solution by what has just been proved. □

In order to combine the above two propositions, we establish a lemma that gives a sufficient condition for the matrix MN^T to be positive semidefinite (hence copositive on any cone).

LEMMA 2.3. *Let C be a polyhedron in R^n and let M and N be square matrices of order n. If* $\mathrm{FEA}(M, N, C) \neq \emptyset$ *and the pair (M, N) satisfies the condition*

$$[\, Mx^i - Ny^i \in C, i = 1, 2\,] \Rightarrow (x^1 - x^2)^T(y^1 - y^2) \geq 0, \tag{3}$$

then MN^T is positive semidefinite.

Proof. We first show that the following implication holds:

$$Mx - Ny \in 0^+C \Rightarrow x^T y \geq 0. \tag{4}$$

Indeed, let (x, y) be a pair of vectors satisfying $Mx - Ny \in 0^+C$. For an arbitrary pair of vectors $(\bar{x}, \bar{y}) \in \mathrm{FEA}(M, N, C)$, we have $Mx_\tau - Ny_\tau \in C$ for every scalar $\tau \geq 0$ where $(x_\tau, y_\tau) \equiv (\bar{x}, \bar{y}) + \tau(x, y)$. By the implication (3), it follows easily that $x^T y \geq 0$. Since the origin is always an element in the recession cone, it follows that

$$Mx - Ny = 0 \Rightarrow x^T y \geq 0.$$

Hence, (M, N) is a column monotone pair in the sense defined in [13]. In particular, by Theorem 11 in this reference, it follows that MN^T is positive semidefinite. □

Remark. We need to assume that M and N are square in order to apply the result in [13] to deduce the positive semidefiniteness of MN^T from the column monotonicity of (M, N).

When $m = n$ and C is a singleton, condition (3) is equivalent to the column monotonicity of the pair (M, N) in the following sense. If (M, N) is column monotone, then (3) holds for $C = \{q\}$ for every n-vector q; conversely, by the proof of Lemma 2.3, if (3) holds for $C = \{q\}$ for some n-vector q belonging to the column space of (M, N), then the pair (M, N) must be column monotone. According to [13, Theorem 11], the column monotonicity of (M, N) is in turn equivalent to two conditions: (i) $M + N$ is nonsingular, and (ii) MN^T is positive semidefinite. By this characterization, it is easy to construct pairs of matrices (M, N) for which MN^T is positive semidefinite but (M, N) is not column monotone. If (M, N) is such a pair of matrices, then with an appropriately defined vector q, the BLP $(M, N, \{q\})$ will have the property that every one of its KKT points is a solution of the XLCP $(M, N, \{q\})$ but the BLP itself is not a convex program. A pair of matrices (M, N) with the property that MN^T is positive semidefinite will be called a *monotone product pair*. Unlike a column monotone pair, a monotone product pair (M, N) need not contain any nonsingular column representative matrix (as defined in [13]). Incidentally, Ye [16] showed that if (M, N) is a monotone product pair, then his potential reduction algorithm will compute a solution of the XLCP (M, N, C) in polynomial time. His results also provided a proof that in this case, the feasibility of the XLCP implies its solvability. This conclusion is a special case of Proposition 2.2.

3. Monotone problems. We say that a pair of $n \times n$ matrices (M, N) is *monotone* with respect to the polyhedral set $C \subseteq R^n$ or, in short, (M, N, C) is a *monotone triple*, if the implication (3) holds. (Note that this definition requires that M and N

be square.) Summarizing the discussion in the last section, we may state the following result for an XLCP with a monotone triple (M, N, C).

THEOREM 3.1. *Let C be a nonempty polyhedron in R^n and let M and N be square matrices of order n. Suppose that (M, N) is monotone with respect to C and that* $\mathrm{FEA}(M, N, C) \neq \emptyset$. *Then the following statements hold:*

(a) *the bilinear function $x^T y$ is convex on* $\mathrm{FEA}(M, N, C)$;

(b) $\mathrm{SOL}(M, N, C) \neq \emptyset$ *and* $\mathrm{SOL}(M, N, C)$ *is a polyhedron.*

Proof. Only the polyhedrality of $\mathrm{SOL}(M, N, C)$ requires a proof. We observe that $\mathrm{SOL}(M, N, C)$ is a convex set by (a). Since the BLP (M, N, C) is a quadratic program and the set of optimal solutions of any quadratic program is equal to the union of a finite number of convex polyhedra [10], the convexity of $\mathrm{SOL}(M, N, C)$ must imply its polyhedrality. □

Under the assumptions of Theorem 3.1, it is possible to give an explicit (polyhedral) representation for $\mathrm{SOL}(M, N, C)$. Instead of presenting such an expression in its fullest generality, we devote the remainder of this section to a discusion of HLCP that has $C = \{-q\}$. For this case, we first introduce a special set associated with a column monotone pair. (*Remark.* Although the next three results can be proved by invoking the close connection between a column monotone pair and a positive semidefinite matrix, our derivation is more direct and reveals some interesting features of the HLCP.)

PROPOSITION 3.2. *Let (M, N) be a column monotone pair of $n \times n$ matrices. Let*

$$\mathcal{K}(M, N) \equiv \{(u, v) \in R^{2n} : Mu - Nv = 0, u \perp v\}.$$

Then $(u, v) \in \mathcal{K}(M, N)$ if and only if there exists a vector λ in the null space of $MN^T + NM^T$ such that

$$u = -N^T\lambda \quad \text{and} \quad v = M^T\lambda. \tag{5}$$

Thus, $\mathcal{K}(M, N)$ is a linear subspace of R^{2n}.

Proof. The column monotonicity of (M, N) implies that $(\bar{u}, \bar{v}) \in \mathcal{K}(M, N)$ if and only if $(\bar{u}, \bar{v})$ is an optimal solution of the (equality constrained) quadratic program:

$$\begin{aligned} \text{minimize} \quad & u^T v \\ \text{subject to} \quad & Mu - Nv = 0, \end{aligned}$$

and $\bar{u}^T\bar{v} = 0$. Thus, if $(\bar{u}, \bar{v}) \in \mathcal{K}(M, N)$, then there must exist a vector λ such that $\bar{u} = -N^T\lambda$ and $\bar{v} = M^T\lambda$. Moreover, we must have

$$\lambda^T MN^T\lambda = -\bar{u}^T\bar{v} = 0.$$

Since MN^T is positive semidefinite, it follows that $(MN^T + NM^T)\lambda = 0$. The converse is easily proved. From this characterization of the set $\mathcal{K}(M, N)$, it follows trivially that this set must be a linear subspace. □

In the next result, we give two representations of the solution set of the "monotone" HLCP:

$$\begin{aligned} & Mx - Ny + q = 0, \\ & (x, y) \geq 0, \quad x \perp y, \end{aligned} \tag{6}$$

where (M, N) is a column monotone pair. One representation is valid in general and the other is valid in the case when the problem is *nondegenerate*, i.e., when it has a

solution $(\bar{x}, \bar{y})$ satisfying $\bar{x} + \bar{y} > 0$. Throughout the remainder of the paper, we write (M, N, q) for $(M, N, \{-q\})$.

PROPOSITION 3.3. *Let (M, N) be a column monotone pair of $n \times n$ matrices and let $(x^0, y^0) \in \mathrm{SOL}(M, N, q)$ be arbitrary. Then*

$$\mathrm{SOL}(M, N, q) = \{ (x, y) \in \mathrm{FEA}(M, N, q) : x^T y^0 + y^T x^0 = 0, (x, y) \in (x^0, y^0) + \mathcal{K}(M, N) \}. \tag{7}$$

If the HLCP *(M, N, q) is nondegenerate, then*

$$\mathrm{SOL}(M, N, q) = \{(x, y) \in \mathrm{FEA}(M, N, q) : x^T y^0 + y^T x^0 = 0\}. \tag{8}$$

Proof. Since $(x^0)^T y^0 = 0$, we may write

$$x^T y = x^T y^0 + y^T x^0 + (x - x^0)^T (y - y^0).$$

By the column monotonicity of (M, N), it follows that $(x, y) \in \mathrm{SOL}(M, N, q)$ if and only if $(x, y) \in \mathrm{FEA}(M, N, q)$, $x^T y^0 + y^T x^0 = 0$, and $(x - x^0)^T (y - y^0) = 0$, or by Proposition 3.2, if and only if (x, y) belongs to the right-hand set in (7).

Suppose that the HLCP (M, N, q) is nondegenerate. It suffices to verify that the right-hand set in (8) is contained in $\mathrm{SOL}(M, N, q)$. Take any vector (x, y) belonging to this right-hand set. Let $(\bar{x}, \bar{y})$ be a nondegenerate solution of the HLCP (M, N, q); then $(x^0, y^0) \in (\bar{x}, \bar{y}) + \mathcal{K}(M, N)$. Since $(x, y) \in \mathrm{FEA}(M, N, q)$, we can verify, by the characterization of the set $\mathcal{K}(M, N)$ in Proposition 3.2, that

$$x^T y^0 + y^T x^0 = x^T \bar{y} + y^T \bar{x}. \tag{9}$$

Indeed, for some vector λ, we have

$$x^0 = \bar{x} - N^T \lambda,$$
$$y^0 = \bar{y} + M^T \lambda.$$

Multiplying the first equation by $(y - y^0)^T$ and the second equation by $(x - x^0)^T$, adding the resulting equations, and using the fact that $(x^0)^T y^0 = \bar{x}^T y^0 + \bar{y}^T x^0 = 0$ and $M(x - x^0) - N(y - y^0) = 0$, we immediately deduce the desired equation (9). Consequently, we have $x^T \bar{y} + y^T \bar{x} = 0$, which easily implies $x^T y = 0$ by the nondegeneracy of the solution $(\bar{x}, \bar{y})$. □

The polyhedral representations (7) and (8) allow us to obtain some error bounds for the monotone HLCP. (The polyhedrality of (7) follows from Proposition 3.2.) Although some such bounds have been obtained in [9] for the general HLCP, they are valid only for test vectors that lie in a compact set. In what follows, we use (8) to obtain a sharpened error bound for the nondegenerate, monotone, HLCP.

COROLLARY 3.4. *Let (M, N) be a column monotone pair of $n \times n$ matrices. If the* HLCP *(M, N, q) has a nondegenerate solution, then there exists a constant $\sigma > 0$, dependent on (M, N, q), such that for all $(x, y) \in \mathrm{FEA}(M, N, q)$,*

$$\mathrm{dist}((x, y), \mathrm{SOL}(M, N, q)) \leq \sigma x^T y,$$

where dist *denotes the distance (measured by any norm) from a vector to a set.*

Proof. It suffices to apply the well-known error bound for polyhedra [8],[11] to the representation (8) and to note that for any solution $(x^0, y^0) \in \mathrm{SOL}(M, N, q)$ and feasible vector $(x, y) \in \mathrm{FEA}(M, N, q)$, we have

$$x^T y^0 + y^T x^0 = x^T y - (x - x^0)^T (y - y^0) \leq x^T y.$$

This establishes the corollary. □

4. A finite SLP algorithm. We now return to the general XLCP (M, N, C). The bilinear programming formulation of this problem allows us to compute a solution by solving a finite sequence of linear programs (SLP) when the triple (M, N, C) satisfies the assumptions of Proposition 2.2. Since these assumptions are considerably more general than the column monotonicity property (for one thing, M and N need not be square matrices), the SLP procedure is applicable to a broader class of XLCPs than the (square) monotone class.

The algorithm described below was formulated in [1] and its finite termination was established for bilinear programs, not necessarily convex. We rephrase the algorithm for the BLP (M, N, C) and use the convergence results from the reference to establish its finite termination. In essence, this algorithm is a modified Frank–Wolfe algorithm for solving the BLP (M, N, C) as a quadratic program, whose convergence was originally proved for convex functions [5].

AN SLP ALGORITHM. Start with any feasible $(x^0, y^0) \in \text{FEA}(M, N, C)$. In general, determine (x^{i+1}, y^{i+1}) from (x^i, y^i) as follows:

- Let (u^i, v^i) be a vertex optimal solution of the linear program:

$$\begin{aligned} &\text{minimize} && x^T y^i + y^T x^i \\ &\text{subject to} && (x, y) \in \text{FEA}(M, N, C). \end{aligned}$$

- Stop if $(u^i)^T y^i + (v^i)^T x^i = 2(x^i)^T y^i$.
- Otherwise, let

$$\begin{pmatrix} x^{i+1} \\ y^{i+1} \end{pmatrix} = (1 - \tau_i) \begin{pmatrix} x^i \\ y^i \end{pmatrix} + \tau_i \begin{pmatrix} u^i \\ v^i \end{pmatrix},$$

where

$$\tau_i \in \operatorname{argmin}_{\tau \in [0,1]} (x^i + \tau(u^i - x^i))^T (y^i + \tau(v^i - y^i)).$$

THEOREM 4.1. *Let M and N be $m \times n$ matrices and C a polyhedral set in R^m. Suppose* $\text{FEA}(M, N, C) \neq \emptyset$. *If the* BLP (M, N, C) *has the property that every one of its* KKT *points solves the* XLCP (M, N, C), *then in a finite number of iterations, the above algorithm will produce a vertex $(u^i, v^i) \in \text{FEA}(M, N, C)$ satisfying $(u^i)^T v^i = 0$.*

Proof. Note that the sequence $\{(x^i, y^i)\}$ generated by the SLP algorithm is bounded because it lies in the convex hull of the vertices of $\text{FEA}(M, N, C)$ and (x^0, y^0). Hence $\{(x^i, y^i)\}$ has at least one accumulation point $(\bar{x}, \bar{y})$ that must satisfy the minimum principle necessary optimality condition [1, Theorem A.1], and hence the KKT conditions for the BLP (M, N, C). By assumption, it follows that $(\bar{x}, \bar{y})$ solves the XLCP (M, N, C). Consequently, $\bar{x}^T \bar{y} = 0$. By [1, Theorem A.2], a vertex (u^i, v^i) generated by the SLP algorithm solves the BLP (M, N, C) with zero minimum. Hence this vertex also solves the XLCP (M, N, C). □

4.1. Sufficient pairs of matrices. Specializing Theorem 4.1 to the HLCP (M, N, q), we obtain the following corollary.

COROLLARY 4.2. *Let (M, N) be a monotone product pair of $n \times n$ matrices. If the* HLCP (M, N, q) *is feasible, then in a finite number of iterations, the* SLP *algorithm will produce a vertex solution of this* HLCP.

Inspired by the class of (row/column) sufficient matrices [3, §3.5], we can broaden the class of matrix pairs (M, N) for which the above corollary is valid. Specifically, we

say that a pair of $n \times n$ matrices (M, N) is row sufficient if the following implication holds: with $A \equiv (M, N) \in R^{n \times 2n}$,

$$\left[\begin{pmatrix} u \\ v \end{pmatrix} \in \text{range } A^T,\ u \circ v \leq 0 \right] \Rightarrow u \circ v = 0,$$

where range denotes the column space of a matrix and $\circ$ denotes the Hadamard product of two vectors; i.e., $x \circ y$ is the vector whose components are the products of the corresponding components of x and y. Similarly, (M, N) is said to be column sufficient if the following implication holds: with $\tilde{A} \equiv (M, -N) \in R^{n \times 2n}$,

$$\left[\begin{pmatrix} u \\ v \end{pmatrix} \in \text{null } \tilde{A},\ u \circ v \leq 0 \right] \Rightarrow u \circ v = 0,$$

where null denotes the null space of a matrix. Finally, the pair (M, N) is said to be sufficient if it is both row and column sufficient.

While a monotone product pair must be row sufficient but not necessarily column sufficient, a column monotone pair must be (both row and column) sufficient. The role played by the (row/column) sufficient pairs in the HLCP is similar to that by the (row/column) sufficient matrices in the standard LCP. For the sake of completeness, we state the following characterization result for the HLCP.

THEOREM 4.3. *Let (M, N) be a pair of $n \times n$ matrices.*

(a) *The pair (M, N) is row sufficient if and only if for every vector $q \in R^n$ for which the* HLCP *(M, N, q) is feasible, every* KKT *vector of the* BLP *(M, N, q) solves the* HLCP *(M, N, q).*

(b) *The pair (M, N) is column sufficient if and only if for every vector $q \in R^n$, the solution set of the* HLCP *(M, N, q), if nonempty, is convex.*

Proof. Assume that (M, N) is a row sufficient pair. Suppose that (x, y) is a KKT vector of the BLP (M, N, q). Then there exist vectors $\lambda \in R^n$, and $(r, s) \in R^{2n}_+$ such that

$$y = M^T\lambda + r, \qquad x = -N^T\lambda + s,$$
$$x^T r = y^T s = 0.$$

By a similar derivation as in the proof of Proposition 2.2, we can show that

$$x \circ y = -((M^T\lambda) \circ (N^T\lambda) + r \circ s).$$

Thus, $(M^T\lambda) \circ (N^T\lambda) \leq 0$. The row sufficiency of (M, N) therefore implies that $(M^T\lambda) \circ (N^T\lambda) = 0$ which in turn yields $x \circ y = 0$.

To prove the converse in (a), suppose that the pair (M, N) is not row sufficient. Then, for some vector $\lambda \in R^n$, we have $(M^T\lambda) \circ (N^T\lambda) \leq 0$ and $(M^T\lambda)_i (N^T\lambda)_i < 0$ for at least one component i. Without loss of generality, we may assume that $(M^T\lambda)_i > 0$ and $(N^T\lambda)_i < 0$. Let

$$y \equiv (M^T\lambda)^+, \qquad r \equiv (M^T\lambda)^-,$$
$$s \equiv (N^T\lambda)^+, \qquad x \equiv (N^T\lambda)^-,$$

where v^+ and v^- denote, respectively, the nonnegative and nonpositive part of a vector v. Also let $q = Ny - Mx$. It is then easy to verify that (x, y) is a KKT vector

of the BLP (M, N, q) with (r, s) as the corresponding multipliers; nevertheless, x is not complementary to y. Thus (a) holds.

To prove (b), suppose the pair (M, N) is column sufficient. Let (x^i, y^i) for $i = 1, 2$ be two solutions of the HLCP (M, N, q). It is then easy to verify for all components $k = 1, \ldots, n$, we have

$$(x^1 - x^2)_k (y^1 - y^2)_k = -(x_k^1 y_k^2 + x_k^2 y_k^1) \leq 0.$$

Since we also have $M(x^1 - x^2) - N(y^1 - y^2) = 0$, it follows that $x_k^1 y_k^2 = x_k^2 y_k^1 = 0$ for all k. In turn, this easily implies that

$$(\tau x^1 + (1 - \tau) x^2)^T (\tau y^1 + (1 - \tau) y^2) = 0$$

for all $\tau \in [0, 1]$. Thus, the convexity of $\mathrm{SOL}(M, N, q)$ follows. Conversely, suppose that (M, N) is not column monotone. Then there exists a vector $(x, y) \in R^{2n}$ satisfying $Mx - Ny = 0, x \circ y \leq 0$, and $x_i y_i < 0$ for at least one index i. Let

$$-q \equiv Mx^+ - Ny^+ = Mx^- - Ny^-.$$

It is then easy to verify that (x^+, y^+) and (x^-, y^-) are solutions of the HLCP (M, N, q) but that these solutions are not "cross complementary," i.e., either $(x^+)^T y^- > 0$ or $(x^-)^T y^+ > 0$. The latter cross complementarity property is easily seen to be both necessary and sufficient for the solution set of any HLCP to be convex. □

It follows immediately from Corollary 4.2 and Theorem 4.3 that if (M, N) is a row sufficient pair, then the SLP algorithm will compute a solution to the HLCP (M, N, q) for every q for which $\mathrm{FEA}(M, N, q) \neq \emptyset$.

In [13], a pair of square matrices (M, N) was defined to be *row monotone* if (M^T, N^T) is column monotone. We have previously mentioned that a column monotone pair must be (column and row) sufficient. Nevertheless, a row monotone pair need not be either column or row sufficient. Indeed, borrowing from [13, Example 10], let us consider the pair

$$M = \begin{bmatrix} 3/2 & 1/2 \\ -1/2 & -1/2 \end{bmatrix}, \qquad N = \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix},$$

which is obtained by transposing, respectively, the matrices C and D in the cited example. The pair (M, N) is row monotone because as shown in the reference (C, D) is column monotone. But the pair (M, N) is neither column nor row sufficient. Column sufficiency is violated with

$$u = (2, 0), \qquad v = (-1, 5);$$

whereas row sufficiency is violated with

$$u = (-\frac{1}{2}, -\frac{1}{2}), \qquad v = (1, 0).$$

The reason for this dichotomy is that the definition of row monotonicity in [13] was relevant for the vertical LCP and was not shown to have any relation to the HLCP. On the other hand, the column and row sufficiency defined herein have direct implications for the HLCP. Thus, it is not surprising that these (column/row) sufficiency and monotonicity concepts for matrix pairs would be quite different. The reader is referred to [6] for more discussion on these concepts.

5. Acknowledgments. The authors are grateful to Professors M. S. Gowda and Y. Ye for pointing out the connection between the XLCP and Ye's general linear complementarity problem.

REFERENCES

[1] K. P. BENNETT AND O. L. MANGASARIAN, *Bilinear separation of two sets in n-space*, Comput. Optim. Appl., 2 (1993), pp. 207–227.

[2] J. F. BONNANS AND C. C. GONZAGA, *Convergence of interior point algorithms for the monotone linear complementarity problem*, manuscript, INRIA, Rocquencourt, October 1993.

[3] R. W. COTTLE, J. S. PANG, AND R. E. STONE, *The Linear Complementarity Problem*, Academic Press, Boston, 1992.

[4] B. DE MOOR, *Mathematical Concepts and Techniques for Modelling of Static and Dynamic Systems*, Katholieke Universiteit Leuven, Fakulteit der Toegepaste Wetenschappen, Departement Elektrotechniek, Leuven, The Netherlands, 1988.

[5] M. FRANK AND P. WOLFE, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110.

[6] M.S. GOWDA, *On the extended linear complementarity problem*, manuscript, Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, January 1994.

[7] O. GÜLER, *Generalized linear complementarity problems*, Math. Oper. Res., to appear.

[8] A. J. HOFFMAN, *On approximate solutions of systems of linear inequalities*, J. Res. National Bureau of Standards, 49 (1952), pp. 263–265.

[9] Z.Q. LUO AND J.S. PANG, *Error bounds for analytic systems and their applications*, Math. Progr., 67 (1994), pp. 1–28.

[10] Z.Q. LUO AND P. TSENG, *Error bound and the convergence analysis of matrix splitting algorithms for the affine variational inequality problem*, SIAM J. Optim., 2 (1992), pp. 43-54.

[11] O.L. MANGASARIAN, *A condition number for linear inequalities and linear programs*, in Proc. 6th Symp. über Operations Research, G. Bamberg, and O. Opitz, eds., Augsburg, 7–9 September 1981, Konigstein, Verlagsgruppe Athenaum, Hain, Scriptor, Hanstein, 1981, pp. 3–15.

[12] R. D. C. MONTEIRO AND T. TSUCHIYA, *Limiting behavior of the derivatives of certain trajectory associated with a monotone horizontal linear complementarity problem*, manuscript, Department of Systems and Industrial Engineering, University of Arizona, Tucson, December 1992.

[13] R. SZNAJDER AND M. S. GOWDA, *Generalizations of P_0- and P-properties; Extended vertical and horizontal* LCP*'s*, Linear Algebra Appl., to appear.

[14] R. H. TÜTÜNCÜ AND M. J. TODD, *Reducing horizontal linear complementarity problems*, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, October 1993, manuscript.

[15] L. VANDENBERGHE, B. DE MOOR, AND J. VANDEWALLE, *The generalized linear complementarity problem applied to the complete analysis of resistive piecewise-linear circuits*, IEEE Trans. Circuits and Systems, 36 (1989), pp. 1382–1391.

[16] Y. YE, *A fully polynomial-time approximation algorithm for computing a stationary point of the general linear complementarity problem*, Math. Oper. Res., 18 (1993), pp. 334–345.

[17] Y. ZHANG, *On the convergence of a class of infeasible interior-point algorithm for the horizontal linear complementarity problem*, SIAM J. Optim., 4 (1994), pp. 208–227.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 369–376, April 1995

003

MAXIMUM ENTROPY ELEMENTS IN THE INTERSECTION OF AN AFFINE SPACE AND THE CONE OF POSITIVE DEFINITE MATRICES*

MIHÁLY BAKONYI† AND HUGO J. WOERDEMAN‡

Abstract. It is shown that for given positive definite A and B and a linear subspace $\mathcal{W}$ consisting of $n \times n$ indefinite (or trivial) Hermitian matrices, there exists a unique positive definite matrix F in $A + \mathcal{W}$ such that $F^{-1} - B \in \mathcal{W}^{\perp}$. This matrix F appears as the maximizer of a certain entropy function. The theorem generalizes a result on Gaussian measures with prescribed margins. Several special cases are presented, yielding new results and recovering known matrix completion results. In case $\mathcal{W}$ is a coordinate subspace, algorithms to find the optimal F are described and numerical results are presented.

Key words. positive definite completions, maximum entropy, Toeplitz matrix completion

AMS subject classifications. Primary 15A15, 15A48; Secondary 62F99

1. Introduction.

Let $\mathcal{M}$ denote the real vector space of all $n \times n$ selfadjoint complex matrices endowed with the scalar product $(C, D) = \operatorname{tr}(CD^*)$, and let $\mathbf{PSD} = \{A \in \mathcal{M} : A \geq 0\}$ and $\mathbf{PD} = \{A \in \mathcal{M} : A > 0\}$ denote the cone of positive semidefinite and positive definite matrices, respectively.

We prove the following result.

THEOREM 1.1. *Let $\mathcal{W} \subset \mathcal{M}$ be a linear subspace such that $\mathcal{W} \cap \mathbf{PSD} = \{0\}$, $A \in \mathbf{PD}$ and $B \in \mathcal{M}$. Then there is a unique $F \in (A + \mathcal{W}) \cap \mathbf{PD}$ such that $F^{-1} - B \perp \mathcal{W}$. Moreover, F is the unique maximizer of the function*

$$f(X) = \log\det X - \operatorname{tr}(BX), \qquad X \in (A + \mathcal{W}) \cap \mathbf{PD}. \tag{1}$$

If A and B are real matrices then so is F.

This result may be viewed as a generalization of Theorem 1 in [19], which solves the problem of finding Gaussian measures with prescribed margins. We cite this result of [19] below in Corollary 1.2. For the particular case that $B = 0$ and $\mathcal{W} \subseteq \{W \in \mathcal{W} : W_{ii} = 0, i = 1, \ldots, n\}$ the result appeared earlier in [17] (in different terminology). The function (1) may be viewed as an entropy function. In optimization context this function is used as a barrier function, and the above problem corresponds to the log barrier problem with parameter equal to 1 (see, e.g., [2], [3]). When $B = I$ the function (1) appears in [10] (see also [4]) where the author shows that the Broyden–Fletcher, Goldfarb, Shannon (BFGS) and Davidon–Fletcher–Powell (DFP) updates of quasi-Newton methods can be derived using this entropy function. In addition there are connections with the sum decomposition results in [12] and [13].

Using the Hahn–Banach separation theorem one can easily show that $\mathcal{W} \cap \mathbf{PSD} = \{0\}$ implies that for every $B \in \mathcal{M}$ we have that $(B + \mathcal{W}^{\perp}) \cap \mathbf{PD} \neq \emptyset$. Indeed, if the latter intersection is empty then the separation theorem yields the existence of a $\Phi \in \mathcal{M}$ and a real number α such that $\operatorname{tr}(P\Phi) > \alpha$ for all $P \in \mathbf{PD}$ and $\operatorname{tr}(X\Phi) \leq \alpha$ for all $X \in B + \mathcal{W}^{\perp}$. Since $\mathbf{PD}$ is a cone, it is easy to see that we must have $\alpha \leq 0$,

* Received by the editors January 27, 1993; accepted for publication (in revised form) by P. Lancaster, April 12, 1994.

† Department of Mathematics, Georgia State University, Atlanta, Georgia 30303 (matmmb@gsusgi2.gsu.edu).

‡ Department of Mathematics, The College of William and Mary, Williamsburg, Virginia 23187-8795 (hugo@cs.wm.edu).

and since $\mathcal{W}^{\perp}$ is a subspace it is easy to see that we must have $\Phi \in \mathcal{W}$. But then it follows that $0 \neq \Phi \in \mathcal{W} \cap \mathbf{PSD}$. Thus the content of the theorem is not weakened when one restricts oneself to $B \in \mathbf{PD}$.

COROLLARY 1.2. *Given is a positive definite matrix* $A = (A_{ij})_{i,j=1}^{n}$, *a Hermitian matrix* $B = (B_{ij})_{i,j=1}^{n}$, *and a set* $J \subseteq \{(i,j) : 1 \leq i < j \leq n\}$. *Let* $\hat{J} = \{(i,j) : (i,j) \in J \text{ or } (j,i) \in J\}$. *Then there exists a unique positive definite matrix* $F = (F_{ij})_{i,j=1}^{n}$ *such that* $F_{ij} = A_{ij}$ *for* $(i,j) \notin \hat{J}$ *and* $(F^{-1})_{ij} = B_{ij}$ *for* $(i,j) \in \hat{J}$. *Moreover,* F *maximizes the function* $f(X) = \log\det X - \operatorname{tr}(BX)$ *over the set of all positive definite matrices* X *whose entries* $(i,j) \notin \hat{J}$ *coincide with those in* A. *In case* A *and* B *are real,* F *is also real.*

The first part of Corollary 1.2 appears in [19]. In [9] this result is described as one that follows easily from an estimate derived in [7], a paper that we unfortunately were not able to retrieve. However, the case when $B = 0$ appears in [8] and was also independently obtained in [5] and [11]. For a construction of the solution in the special case of band matrices and $B = 0$, see [6], and in the special case of a chordal pattern and $B = 0$, see [15] and [1]. In the latter four references the problem is formulated as a matrix completion problem. From this viewpoint one may view Corollary 1.2 as the solution to the completion problem in which a partial matrix is given whose entries in positions $(i,j) \notin \hat{J}$ are prescribed and whose inverse is prescribed in positions $(i,j) \in \hat{J}$. In §3 we present two algorithms to obtain the optimal F in Corollary 1.2, and in §4 we present some test results.

Another corollary is the following.

COROLLARY 1.3. *Let* $A = (A_{ij})_{i,j=1}^{n}$ *be positive definite. There exists a unique positive definite matrix* $F = (F_{ij})_{i,j=1}^{n}$ *such that* $F_{ij} = A_{ij}$, $i \neq j$, $\operatorname{tr} F = \operatorname{tr} A$ *and the diagonal entries of the inverse of* F *are all the same. In case* A *is real,* F *is also real.*

For Toeplitz matrices we obtain the following result.

COROLLARY 1.4. *Let* $A = (A_{j-i})_{i,j=1}^{n}$ *be a positive definite Toeplitz matrix,* $0 < p_1 < p_2 < \cdots < p_r < n$ *and* $\alpha_1, \ldots, \alpha_r \in \mathbf{C}$. *Then there exists a unique Toeplitz matrix* $F = (F_{j-i})_{i,j=1}^{n}$ *with* $F_q = A_q$, $|q| \neq p_1, \ldots, p_r$, *and*

$$\sum_{j-i=p_k} (F^{-1})_{ij} = \alpha_k, \qquad k = 1, \ldots, r. \tag{2}$$

In case A *and* $\alpha_1, \ldots, \alpha_r$ *are real, the matrix* F *is also real.*

One may view Corollary 1.4 as the answer to a Toeplitz matrix completion problem. In the case when $\alpha_k = 0$ for all k, the result appears in [17]. In Corollaries 1.3 and 1.4 the solution F appears as the unique maximizer of the function (1) with suitable choices of $\mathcal{W}$, A, and B.

2. The proofs.

Proof of Theorem 1.1. Since $\mathcal{W} \cap \mathbf{PSD} = \{0\}$, $(A + \mathcal{W}) \cap \mathbf{PSD}$ is a bounded set (we are in a finite dimensional space). The set $(A + \mathcal{W}) \cap \mathbf{PSD}$ is convex. It is known that log det is strictly concave on **PSD** (see, e.g., Theorem 7.6.7 in [14]). Since tr (BX) is linear in X, $f(X)$ is strictly concave and thus has a unique maximum on $(A + \mathcal{W}) \cap \mathbf{PSD}$ denoted by F. Since near the boundary f tends to $-\infty$, F is a point of $(A + \mathcal{W}) \cap \mathbf{PD}$.

Fix an arbitrary $W \in \mathcal{W}$. Consider the function $f_{F,W}(x) = \log\det(F + xW) - \operatorname{tr}(B(F + xW))$ defined in a neighborhood of 0 in $\mathbf{C}$. Then $f'_{F,W}(0) = 0$ (since f has

its maximum at F). It is easy to see that

$$f'_{F,W}(0) = \frac{(\det(I + xF^{-1}W))'}{\det(I + xF^{-1}W)}\Big|_{x=0} - (\operatorname{tr}(B(F + xW)))'|_{x=0}$$
$$= \operatorname{tr}(F^{-1}W) - \operatorname{tr}(BW) = \operatorname{tr}((F^{-1} - B)W) = 0.$$

Since W is an arbitrary element of $\mathcal{W}$ we have that $F^{-1} - B \perp \mathcal{W}$. Assume that $G \in (A + \mathcal{W}) \cap \mathbf{PD}$ and $G^{-1} - A \perp \mathcal{W}$. Then, $f'_{G,W}(0) = 0$ for any $W \in \mathcal{W}$ and since f is strictly concave it follows that $G = F$. This proves the uniqueness of F.

In case A and B are real matrices, one can restrict the attention to real matrices, and repeat the above argument. The resulting matrix F will also be real. □

Proof of Corollary 1.2. Introduce

$$\mathcal{W} = \{W \in \mathcal{M} : W_{kj} = 0 \text{ whenever } (k,j) \notin \hat{J}\}. \tag{3}$$

Then $\mathcal{W} \cap \mathbf{PSD} = \{0\}$. By Theorem 1.1 there exists a unique $F \in (A + \mathcal{W}) \cap \mathbf{PD}$ such that $F^{-1} - B \perp \mathcal{W}$. For any $(k,j) \in J$, consider the matrix $W_R^{(k,j)} \in \mathcal{W}$ having all its entries 0 except those on the positions (k,j) and (j,k) which equal 1, respectively, the matrix $W_I^{(k,j)}$ having i on the position (k,j), $-i$ on the position (j,k) and 0 elsewhere. The conditions $\operatorname{tr}((F^{-1} - B)W_R^{(k,j)}) = \operatorname{tr}((F^{-1} - B)W_I^{(k,j)}) = 0$ imply that $(F^{-1})_{kj} = B_{kj}$ for any $(k,j) \in \hat{J}$. □

A subspace $\mathcal{W}$ of type (3) is called a *coordinate subspace.*

Proof of Corollary 1.3. Let $B = 0$ and $\mathcal{W} = \{W \in \mathcal{M} : W_{ij} = 0, i \neq j \text{ and } \operatorname{tr} W = 0\}$. Since

$$\mathcal{W}^{\perp} = \{W \in \mathcal{M} : W_{ii} = W_{jj},\ i,j = 1,\dots,n\}$$

the result follows immediately from Theorem 1.1. □

Proof of Corollary 1.4. Let $B = (B_{ij})$ with $B_{1p_k} = B^*_{p_k 1} = \alpha_k$, $k = 1,\dots,r$ and $B_{ij} = 0$ otherwise. Furthermore, let $\mathcal{W}$ be the span of Toeplitz matrices of the form

$$W_1^{(j)} = \begin{pmatrix} 0 & 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & \dots & 0 & 0 & \dots & 1 \\ 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 & \dots & 0 \end{pmatrix},$$

$$W_2^{(j)} = \begin{pmatrix} 0 & 0 & \dots & i & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & i & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ -i & 0 & \dots & 0 & 0 & \dots & i \\ 0 & -i & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & -i & 0 & \dots & 0 \end{pmatrix}$$

supported on the jth diagonal for $j \in \{p_1,\dots,p_r\}$. Then $(M - B, W_k^{(j)}) = 0$, $k = 1,2$, if and only if the sum of the elements on the jth diagonal of $M - B$ is equal to 0. The corollary now follows immediately by applying Theorem 1.1. □

3. Two algorithms. There is generally no closed formula for the optimal solution F in Theorem 1.1. We next present two approximation methods in the case of Corollary 1.2.

ALGORITHM 1. The first algorithm we present is a naive one based on the coordinate descent method. We refer to [16] as a general reference on coordinate descent methods. First we need some preliminary results.

Given a matrix M we let $M^{(r,s)}$ denote the matrix M with row r and column s deleted.

LEMMA 3.1. *Let $A = (A_{ij})_{i,j=1}^n$ be a positive definite matrix and $\alpha \in \mathbf{C}$. Let $1 \leq r, s \leq n$ and let $A(z)$ be the Hermitian matrix A with the (r,s)th entry replaced by z and the (s,r)th entry replaced by $\bar{z}$. Let $z_0 \in \mathbf{C}$ be the unique complex number z such that*

$$-|z|^2 \det([[A(0)]^{(r,s)}]^{(s,r)}) + 2(-1)^{r+s}\mathrm{Re}\,(z\ \det([A(0)]^{(r,s)}) + \det A(0) > 0 \tag{4}$$

and

$$\frac{(-1)^{r+s}\det([A(0)]^{(s,r)} - z\ \det([[A(0)]^{(r,s)}]^{(s,r)})}{-|z|^2 \det([[A(0)]^{(r,s)}]^{(s,r)}) + 2(-1)^{r+s}\mathrm{Re}\,(z\ \det([A(0)]^{(r,s)}) + \det A(0)} = \alpha. \tag{5}$$

Then $A(z_0)$ is the unique positive definite matrix whose entries coincide with those in A except for the (r,s) and the (s,r) position, and whose inverse has on the (r,s) position the value α.

In case $n = 2$ the determinant of the 0×0 matrix should be interpreted as being equal to 1.

Proof. It is straightforward to check that $\det A(z)$ equals the left-hand side of (4), and using Cramer's rule one easily computes that $(A(z)^{-1})_{rs}$ is given by the left-hand side of (5). The existence and uniqueness of a solution now follows from Corollary 1.2. □

Given the conditions of Corollary 1.2, let $\{(i_k, j_k) : k = 0, 1, \ldots, s-1\}$ be an arbitrary ordering of the elements of J. For any $M \in (A + \mathcal{W}) \cap \mathbf{PD}$ define the positive definite matrices $X_k^{(M)}$, $k = 0, 1, \ldots, s$ by $X_0^{(M)} = M$ and letting $X_{k+1}^{(M)}$ be obtained by modifying the (i_k, j_k) and (j_k, i_k) entries of $X_k^{(M)}$ such that

$$(X_{k+1}^{(M)^{-1}})_{i_k, j_k} = B_{i_k, j_k}$$

as indicated in Lemma 3.1 (in the real case this comes down to finding the root of a quadratic polynomial that satisfies (4)). Define then the function

$$g : (A + \mathcal{W}) \cap \mathbf{PD} \to (A + \mathcal{W}) \cap \mathbf{PD}, g(M) = X_s^M,$$

which is continuous since $X_{k+1}^{(M)}$ depends continuously on $X_k^{(M)}$. Further note that F in Corollary 1.2 is a fixed point for g. In fact, it is the only fixed point for g, since if $P \in (A + \mathcal{W}) \cap \mathbf{PD}$ is such that $g(P) = P$, we obtain from the definition of g that $(P^{-1})_{i_k, j_k} = B_{i_k, j_k}$ for $k = 0, 1, \ldots, s-1$. Thus $P = F$ by Corollary 1.2. In addition, it should be noted that $f(g(M)) \geq f(M)$.

Define now the following sequence: $Y_0 = A$, $Y_{m+1} = g(Y_m)$ for $m \geq 0$. Since the sequence $\{Y_m\}_{m=0}^{\infty}$ lies in the compact set $(A + \mathcal{W}) \cap \mathbf{PD} \cap \{M \mid f(M) \geq f(A)\}$, the sequence has a limit point $H \in (A + \mathcal{W}) \cap \mathbf{PD}$. Consequently, since $Y_{m+1} = g(Y_m)$, and g is continuous we get that $H = F$, and thus $Y_m \to F$.

In the case that $B = 0$ and J is ordered using a perfect elimination scheme (which is only possible in the chordal case) then the algorithm can be adjusted so that the iteration stops after the first step (i.e., $Y_1 = F$). The adjustment concerns restricting the attention in Lemma 3.1 to certain submatrices that are determined by the ordering; see [11], [19], [1], and [15] for details.

ALGORITHM 2. This algorithm is based on Newton's algorithm and can be found in §1.2 of [18]. The description of the method given in §3.4 of [2] was very helpful to us.

Given is the positive definite matrix A, the matrix B, and the positions $J = \{(i_1, j_1), \ldots, (i_s, j_s)\}$.

Introduce

$$x = (A_{i_1,j_1}, \ldots, A_{i_s,j_s}), \qquad y = ((A^{-1} - B)_{i_1,j_1}, \ldots, (A^{-1} - B)_{i_s,j_s})$$

and error $= \|y\|_\infty$ (the maximal modulus among the coefficients of y). While error $>$ tol do

$$H = (H_{pq})_{p,q=1}^{s}, \qquad H_{pq} = (A^{-1})_{i_p j_q}(A^{-1})_{i_q j_p} + (A^{-1})_{i_q i_p}(A^{-1})_{j_q j_p},$$

$$v = H^{-1}y, \qquad \delta = \sqrt{v^T y},$$

$$\alpha = \begin{cases} 1 & \text{if } \delta < \frac{1}{4}, \\ \frac{1}{1+\delta} & \text{if } \delta \geq \frac{1}{4}, \end{cases}$$

$$x := x + \alpha v, \qquad A_{i_p,j_p} := x_p, p = 1, \ldots, s,$$

$$y := ((A^{-1} - B)_{i_1,j_1}, \ldots, (A^{-1} - B)_{i_s,j_s}), \qquad \text{error} = \|y\|_\infty.$$

4. Some test results. We implemented the algorithms in the previous section using Mathematica [20]. We now present some results.

4.1. Experiment 1. Here $n = 3$, $J = \{(1,2), (2,3)\}$ (with this ordering), $B_{12} = 1$, $B_{23} = 2$, and

$$A = \begin{pmatrix} 14 & 14.7 & 10 \\ 14.7 & 17.01 & 12.9 \\ 10 & 12.9 & 14 \end{pmatrix}.$$

ALGORITHM 1. After 33 iterations we get that $A_{12} = -13.2165$, $A_{23} = -14.7713$, and the inverse of the new matrix A is

$$\begin{pmatrix} .534561 & 1.00001 & .673279 \\ 1.00001 & 2.57256 & 2. \\ .673279 & 2. & 1.7007 \end{pmatrix}.$$

ALGORITHM 2. With tol $= 10^{-16}$ we obtain after 89 iterations that $A_{12} = -13.2165$, $A_{23} = -14.7713$, and the inverse of the new matrix A is

$$\begin{pmatrix} .534552 & 1. & .673272 \\ 1. & 2.57255 & 2. \\ .673272 & 2. & 1.70071 \end{pmatrix}.$$

4.2. Experiment 2. Here $n = 4$, $J = \{(1,3),(2,4)\}$ (with this ordering), $B_{13} = 1$, $B_{24} = 2$, and

$$A = \begin{pmatrix} 30 & 14.7 & 16 & 27.9 \\ 14.7 & 17.01 & 12.9 & 29.13 \\ 16 & 12.9 & 16.25 & 23.6 \\ 27.9 & 29.13 & 23.6 & 52.53 \end{pmatrix}.$$

ALGORITHM 1. After 129 iterations we get that $A_{13} = 17.3068$, $A_{24} = 8.3627$, and the inverse of the new matrix A is

$$\begin{pmatrix} .260914 & -.753857 & 1. & -.467832 \\ -.753857 & 3.31916 & -4.73663 & 2. \\ 1. & -4.73663 & 7.00058 & -2.92219 \\ -.467832 & 2. & -2.92219 & 1.26196 \end{pmatrix}.$$

ALGORITHM 2. With tol $= 10^{-16}$ we obtain after 47 iterations that $A_{13} = 17.3068$, $A_{24} = 8.3627$, and the inverse of the new matrix A is

$$\begin{pmatrix} .260913 & -.753856 & 1. & -.467832 \\ -.753856 & 3.31916 & -4.73663 & 2. \\ 1. & -4.73663 & 7.00059 & -2.92219 \\ -.467832 & 2. & -2.92219 & 1.26196 \end{pmatrix}.$$

4.3. Experiment 3. Here $n = 5$, $J = \{(1,3),(2,4),(3,5)\}$ (with this ordering), $B_{13} = 1$, $B_{24} = 2$, $B_{35} = 3$, and

$$A = \begin{pmatrix} 55 & 34.7 & 26 & 40.4 & 59.6 \\ 34.7 & 33.01 & 20.9 & 39.13 & 45.6 \\ 26 & 20.9 & 20.25 & 28.6 & 30.8 \\ 40.4 & 39.13 & 28.6 & 58.78 & 59.6 \\ 59.6 & 45.6 & 30.8 & 59.6 & 95.72 \end{pmatrix}.$$

ALGORITHM 1. After 225 iterations we get that $A_{13} = 13.6371$, $A_{24} = 42.246$, $A_{35} = 15.4163$, and the inverse of the new matrix A is

$$\begin{pmatrix} .22564 & -.969866 & .999933 & -.291023 & .341699 \\ -.969866 & 5.74027 & -6.29935 & 2.00113 & -2.36217 \\ .999933 & -6.29935 & 7.88392 & -3.03768 & 3. \\ -.291023 & 2.00113 & -3.03768 & 1.4746 & -1.20104 \\ .341699 & -2.36217 & 3. & -1.20104 & 1.18766 \end{pmatrix}.$$

ALGORITHM 2. With tol $= 10^{-16}$ we obtain after 31 iterations that $A_{13} = 13.6366$, $A_{24} = 42.2479$, and $A_{35} = 15.4148$, and the inverse of the new matrix A is

$$\begin{pmatrix} .22572 & -.970351 & 1. & -.290738 & .341707 \\ -.970351 & 5.74342 & -6.30058 & 2. & -2.36257 \\ 1. & -6.30058 & 7.88385 & -3.03661 & 3. \\ -.290738 & 2. & -3.03661 & 1.47429 & -1.2007 \\ .341707 & -2.36257 & 3. & -1.2007 & 1.18768 \end{pmatrix}.$$

4.4. Experiment 4. Here $n = 10$, $J = \{(1,8),(2,9),(3,10)\}$ (with this ordering), $B_{18} = B_{29} = 2, B_{3,10} = 0$, and

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0.6530 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 1.933 & 2 \\ 1 & 2 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 2.679 \\ 1 & 2 & 3 & 4 & 4 & 4 & 4 & 4 & 4 & 4 \\ 1 & 2 & 3 & 4 & 5 & 5 & 5 & 5 & 5 & 5 \\ 1 & 2 & 3 & 4 & 5 & 6 & 6 & 6 & 6 & 6 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 7 & 7 & 7 \\ 0.6530 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 8 & 8 \\ 1 & 1.933 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 9 \\ 1 & 2 & 2.679 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{pmatrix}.$$

The solution to this problem is given by $A = (A_{ij})$ in which $A_{ij} = \min\{i,j\}$. The inverse is given by $(A^{-1})_{ii} = 2$, $(A^{-1})_{i,i+1} = (A^{-1})_{i,i-1} = -1$, and zero elsewhere. With Algorithm 1 we obtain after only six iterations the right result with an error smaller than 10^{-5}. Algorithm 2 takes 21 iterations to get to this result.

Note that the pattern of specified entries in Experiments 2, 3, and 4 correspond to nonchordal graphs (see [11]). But even in the case of a chordal graph there is not a closed formula for the completion when the prescribed entries in the inverse are nonzero.

In the above experiments both algorithms seem acceptable. For larger problems, though, Algorithm 2 seems to outdo Algorithm 1 by far. For comparison purposes we did ten experiments with 10×10 randomly generated matrices, in which the entries $\{(1,4),(2,5),\ldots,(7,10)\}$ in the inverse were prescribed. With tolerance equal to 10^{-5} we obtained that it took Algorithm 1 on the average 1687 iterations to stop, and Algorithm 2 on the average 99 iterations. The fact that the convergence in Algorithm 1 can be quite slow is most likely due to the fact that the function f is very "flat" near the optimum. An additional 100 experiments with random 10×10 matrices as above with Algorithm 2 and tolerance equal to 10^{-7} led to an average of 144 iterations.

Acknowledgment. We wish to thank the referees for acquainting us with references [9], [10], [12], and [13].

REFERENCES

[1] M. BAKONYI, *Completion of partial operator matrices with chordal graphs*, Integral Equations and Operator Theory, 15(1992), pp. 173–185.

[2] S. BOYD AND L. EL GHAOUI, *Method of centers for minimizing generalized eigenvalues*, Linear Algebra Appl., 188/9(1993), pp. 63–111.

[3] S. BOYD AND L. VANDENBERGHE, *A primal-dual potential reduction method for problems involving matrix inequalities*, Math. Programming, to appear.

[4] R. H. BYRD AND J. NOCEDAL, *A tool for the analysis of quasi-Newton methods with application to unconstrained minimization*, SIAM J. Numer. Anal., 26(1989), pp. 727–739.

[5] A.P. DEMPSTER, *Covariance selections*, Biometrics, 28(1972), pp. 157–175.

[6] H. DYM AND I. GOHBERG, *Extensions of band matrices with band inverses*, Linear Algebra Appl., 36(1981), pp. 1–24.

[7] M. FIEDLER, *A remark on positive definite matrices* (Czech, English summary), Casopis Pro Pest. Mat., 85 (1960), pp. 75–77.

[8] ———, *Matrix inequalities*, Numer. Math., 9 (1966), pp. 109–119.

[9] J. SEDLAČEK AND A. VRBA, *Sixty years of Professor Miroslav Fiedler*, Czechoslovak Math. J., 36(1986), p. 498.

[10] R. FLETCHER, *A new variational result for quasi-Newton formulae*, SIAM J. Optim., 1 (1991), pp. 18–21.

[11] R. GRONE, C. R. JOHNSON, E. SA, AND H. WOLKOWITZ, *Positive definite completions of partial Hermitian matrices*, Linear Algebra Appl., 58 (1984), pp. 109–124.
[12] S.-P. HAN AND O. L. MANGASARIAN, *Conjugate cone characterizations of positive definite and semidefinite matrices*, Linear Algebra Appl., 56 (1984), pp. 89–103.
[13] ———, *Characterization of positive definite and semidefinite matrices via quadratic programming duality*, SIAM J. Alg. Disc., 5(1984), pp. 26–32.
[14] R. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.
[15] C. R. JOHNSON AND M. E. LUNDQUIST, *Matrices with chordal inverse zero patterns*, Linear Multilinear Algebra, 36(1993), pp. 1–17.
[16] D. G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973.
[17] M. E. LUNDQUIST AND C. R. JOHNSON, *Linearly constrained positive definite completions*, Linear Algebra Appl., 150(1991), pp. 195–207.
[18] YU. NESTEROV AND A. NEMIROVSKY, *Interior Point Polynomial Algorithms in Convex Programming*, Society for Industrial and Applied Mathematics, Philadelphia, 1994.
[19] T. P. SPEED AND H. T. KIIVERI, *Gaussian Markov distributions over finite graphs*, Ann. Statist., 14 (1986), pp. 138–150.
[20] S. WOLFRAM, *Mathematica* 2.1, Wolfram Research Inc, Champaign, IL, 1988-1992.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 377–387, April 1995

004

A FINITE PROCEDURE FOR THE TRIDIAGONALIZATION OF A GENERAL MATRIX*

A. GEORGE†, K. IKRAMOV‡, A. N. KRIVOSHAPOVA‡, AND W.-P. TANG†

Abstract. Interest in the problem of the tridiagonalization of an arbitrary square complex matrix by similarity transformation has been renewed recently through work by Geist, Parlett, Tang and others. To our knowledge, no procedure has so far been presented to compute a tridiagonal matrix similar to a general square complex matrix that requires only a finite number of operations and works for any matrix. In this paper, finite algorithms that are guaranteed to reduce an unreduced Hessenberg matrix or a general matrix to tridiagonal form via similarity transformations are presented. The algorithms are mainly of theoretical interest; that of finding a practical, cost-effective procedure for solving the problem remains an open problem.

Key word. tridiagonalization

AMS subject classifications. 65F10, 76S05

1. Introduction.

The problem of tridiagonalization of an arbitrary square complex matrix via similarity transformations is an important practical problem and has been the subject of much study. An excellent review of the theoretical background on this subject has been provided recently by Beresford Parlett [19]. Descriptions of efforts and approaches for algorithms for the problem can be found in [4], [7], [14], [16], [17], [22], [23].

It is well known that every Hermitian matrix (and every symmetric or skew-symmetric matrix in the real case) can be reduced to tridiagonal form by a similarity or unitary transformation. The Householder reduction [10] can be used to this purpose and can be regarded as a constructive proof of the following theorem.

THEOREM 1.1. *Every Hermitian (symmetric or skew-symmetric) matrix can be transformed into tridiagonal form by a finite procedure using only rational operations of the corresponding number field* **K** (**R** *or* **C**) *and extractions of square roots.*

If the matrix A is large and sparse, the Lanczos algorithm will be more appropriate for the tridiagonalization. The Lanczos procedure could be prematurely halted if the dimension of the Krylov subspace is smaller than the size of A. However, it is not difficult to show that these irregularities can still be resolved finitely [18].

When the matrix A is non-Hermitian, the Householder reduction process can only transform A to Hessenberg form. Then, the Strachey–Francis (SF) algorithm can be employed to further reduce the resulting Hessenberg form to tridiagonal form. The Lanczos algorithm can also be generalized to non-Hermitian matrices. Unfortunately, as far as is currently known, both approaches can suffer from breakdown. In particular, if a *serious breakdown* occurs, we have no choice but to repeat the Lanczos algorithm with a new set of starting vectors and with no guarantee that a new serious breakdown will not be encountered. Such a repetition could be in principle infinite. Therefore, no existing computational procedure for tridiagonalizing a general matrix can be regarded as a constructive proof of the following conjecture.

* Received by the editors July 30, 1992; accepted for publication (in revised form) by G. Cybenko, December 2, 1993. This work was supported by the Natural Sciences and Engineering Research Council of Canada, and by the Information Technology Research Centre, a Centre of Excellence funded by the Province of Ontario.

† Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1 (wptang@lady.uwaterloo.ca).

‡ Moscow State University, Faculty of Numerical Mathematics and Cybernetics, Moscow, Russia. 119899.

Every square matrix, complex or real, can be reduced to tridiagonal form by a finite procedure employing only rational operations of the corresponding number field.

In a recent paper by Parlett [19], he wrote, "No one has presented a finite algorithm that is guaranteed to compute a tridiagonal matrix similar to an arbitrary given square complex matrix while avoiding huge intermediate quantities." To our knowledge, no one has presented a finite algorithm for the goal he describes even if huge intermediate (but finite) quantities *are* allowed. Moreover, as shown in [12] where Lanczos methods are investigated in the context of solving nonsymmetric linear equations, there are matrices for which almost any choice of a starting vector leads to a serious breakdown.

The main result of this paper is stated in the following theorem.

THEOREM 1.2. *Every matrix $A \in \mathbf{R}^{n\times n}$ or $\mathbf{C}^{n\times n}$ may be reduced to tridiagonal form (over $\mathbf{R}$ or $\mathbf{C}$, respectively) by a finite procedure involving only the rational operations of the corresponding number field.*

The bulk of the proof of this theorem constitutes the examination of the case when A is an unreduced Hessenberg matrix. We have chosen, therefore, to first investigate this case in §§2 and 3, and state the result as a separate theorem.

THEOREM 1.3. *Every unreduced Hessenberg matrix $H \in \mathbf{R}^{n\times n}$ or $\mathbf{C}^{n\times n}$ may be reduced to tridiagonal form (over $\mathbf{R}$ or $\mathbf{C}$, respectively) by a finite procedure involving only the rational operations of the corresponding number field.*

For a general matrix A, we prove Theorem 1.2 when A is diagonalizable (over $\mathbf{C}$) in §4, and for the nondiagonalizable case in §5. In the Appendix, we show that the tridiagonal form is not particularly good for normal matrices that are not Hermitian.

We would like to emphasize the following three points.

1. First, the procedure described in this paper is not a practical approach for tridiagonalizing a given non-Hermitian matrix. Our motivation here is mainly theoretical. Therefore, the problem of constructing a cost-effective and stable algorithm remains open.

2. If one is prepared to give up the strict tridiagonal form, then the interesting work on look-ahead Lanczos reported in [5], [6], [11] can be very effective. For the computation of the eigenvalues of a non-Hermitian matrix, that algorithm is adequate.

3. The word "finite" is the key in this paper. In fact, our reasoning here is to some extent similar to Householder's argument, which appears on page 18 of his book [15]. However, the finiteness is crucial to the tridiagonalization problem. Otherwise, the Jordan canonical form theorem can be considered as an answer to this problem, at least for complex matrices.

2. Algorithm. Let $H \in \mathbf{R}^{n\times n}$ (or $\mathbf{C}^{n\times n}$) be an unreduced Hessenberg matrix. We recall that the unreducibility of H means that $h_{i+1,i} \neq 0$, $i = 1, 2, \ldots, n-1$. To transform H to a tridiagonal form, the SF-procedure [20] can be applied. If the procedure succeeds, namely, the pivots on all steps are nonzero, then a tridiagonal matrix T will be obtained. This transformation can be written as

$$T = V^{-1}HV,$$

where the matrix V is an upper-triangular matrix with unit diagonal and first row $e_1^T = (1, 0, \ldots, 0)$. It is clear that V^{-1} has e_1^T as its first row as well. In fact, the nonzero elements of V^{-1} are the multipliers of different steps of the SF-procedure with the sign reversed.

If we precede the SF-procedure by the similarity transformation

$$\hat{H} = U^{-1}HU,$$

where

$$U = \begin{bmatrix} 1 & -\alpha_2 & -\alpha_3 & \cdots & -\alpha_n \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & \ddots & \\ 0 & 0 & 0 & & 1 \end{bmatrix},$$

then $\hat{H}$ also has upper Hessenberg form. Moreover, the subdiagonal elements of H remain the same in $\hat{H}$. Applying the SF-procedure to $\hat{H}$ and assuming it is successful, we obtain

$$\begin{aligned} \hat{T} &= \hat{V}^{-1}\hat{H}\hat{V} \\ &= (\hat{U}\hat{V})^{-1} H (\hat{U}\hat{V}) \\ &= \widetilde{W} H \widetilde{V}, \end{aligned} \tag{1}$$

where $\hat{T}$ is the resulting tridiagonal matrix, $\widetilde{V} = \hat{U}\hat{V}$, and $\widetilde{W} = (\hat{U}\hat{V})^{-1}$. Note that the matrix $\widetilde{W}$ is still upper triangular and its first row is

$$\begin{bmatrix} 1 & \alpha_2 & \alpha_3 & \cdots & \alpha_n \end{bmatrix}.$$

It is well known (see, for example, [14]) that the SF-procedure is equivalent to applying the unsymmetric Lanczos algorithm to the matrix H with the initial vectors

$$v_1 = e_1 \qquad \text{and} \qquad w_1 = e_1.$$

If we wish to retain the subdiagonal elements $h_{i+1,i}$, $i = 1, 2, \ldots, n-1$, of the original matrix H as the subdiagonal elements of the tridiagonal form then the same matrices T, V, and W will be obtained, where V and $W = V^{-1}$ are formed by the right and left Lanczos vectors.

In the same way, the combined procedure in (1) is equivalent to applying the unsymmetric Lanczos method to the matrix H with initial vectors

$$v_1 = e_1 \qquad \text{and} \qquad w_1 = \begin{bmatrix} 1 & \alpha_2 & \alpha_3 & \cdots & \alpha_n \end{bmatrix}^T,$$

where w_1 is an arbitrary vector with the first component equal to 1.

The reduction of H to tridiagonal form will not be feasible if the pivot is zero at any of the $n-2$ steps of the SF-procedure. The exact analogue of this situation in the unsymmetric Lanczos algorithm is the so-called *breakdown* phenomenon. Suppose that the Lanczos vectors $v_1, \ldots, v_{k-1}$ and $w_1, \ldots, w_{k-1}$ are already determined. We will be able to obtain the next pair of vectors v_k and w_k only when the following Hankel determinant is nonzero [5], [21]:

$$\Delta_k = \begin{vmatrix} m_0 & m_1 & m_2 & \cdots & m_{k-1} \\ m_1 & m_2 & m_3 & & m_k \\ m_2 & m_3 & m_4 & & m_{k+1} \\ \vdots & & & & \vdots \\ m_{k-1} & m_k & m_{k+1} & \cdots & m_{2k-2} \end{vmatrix}$$

(i.e., no breakdown occurs at kth step of the Lanczos method). Here

$$m_j = w_1^T H^j v_1, \qquad j = 0, 1, \ldots,$$

and v_1, w_1 are the initial vectors of the Lanczos algorithm.

The condition

$$\Delta_1 \Delta_2 \cdots \Delta_k \neq 0$$

also assures that the first $k-1$ steps of the equivalent SF-procedure are feasible. For example, if

$$v_1^T w_1 \neq 0 \qquad \text{and} \qquad \Delta_2 = m_0 m_2 - m_1^2 \neq 0,$$

then the SF-procedure can be started.

3. Proof of Theorem 1.3. To prove our first main theorem, we need the following three lemmas.

LEMMA 3.1. *The unreduced Hessenberg matrix H can be reduced to the companion form by a finite procedure using only the rational operations of the corresponding number field.*

This reduction can actually be done by several different finite algorithms. One possibility is the Danilevski algorithm [4]. Based on this lemma, it is sufficient to prove Theorem 1.3 for the companion matrix

$$F = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0 & f_1 \\ 1 & 0 & 0 & 0 & & 0 & f_2 \\ 0 & 1 & 0 & 0 & & 0 & f_3 \\ 0 & 0 & 1 & 0 & & 0 & f_4 \\ \vdots & & & \ddots & & & \vdots \\ 0 & 0 & 0 & 0 & & 1 & f_n \end{bmatrix}, \tag{2}$$

instead of a general Hessenberg matrix H.

LEMMA 3.2. *If $v_1 = e_1$, $w_1^T = (1, \alpha_2, \alpha_3, \ldots, \alpha_n)$, and H is replaced by F, then the Hankel determinants $\Delta_1, \Delta_2, \ldots, \Delta_{n-1}$ are nontrivial polynomials (i.e., not identically zero) in the variables $\alpha_2, \alpha_3, \ldots, \alpha_n$.*

Proof. It is easy to see that

$$F^j e_1 = e_{j+1}, \qquad j = 0, 1, \ldots, n-1,$$

where $e_1, e_2, \ldots, e_n$ are the coordinate vectors in $\mathbf{K}^n$, and

$$F^n e_1 = f = (f_1, f_2, \ldots, f_n)^T.$$

Therefore,

$$\begin{aligned} m_0 &= 1, \\ m_j &= \alpha_{j+1}, \qquad j = 1, 2, \ldots, n-1, \\ m_n &= f_1 + f_2 \alpha_2 + \cdots + f_n \alpha_n. \end{aligned}$$

The moments m_j with indices greater than n are still linear functions in $\alpha_2, \alpha_3, \ldots, \alpha_n$, although they depend on coefficients $f_1, f_2, \ldots, f_n$ in a more complicated way. The Hankel determinants $\Delta_1, \Delta_2, \ldots, \Delta_{n-1}$ have the following form:

$$\Delta_1 = m_0 = 1,$$

$$\Delta_2 = \begin{vmatrix} 1 & \alpha_2 \\ \alpha_2 & \alpha_3 \end{vmatrix},$$

$$\Delta_3 = \begin{vmatrix} 1 & \alpha_2 & \alpha_3 \\ \alpha_2 & \alpha_3 & \alpha_4 \\ \alpha_3 & \alpha_4 & \alpha_5 \end{vmatrix},$$

$$\vdots \quad \vdots$$

$$\Delta_{n-1} = \begin{vmatrix} 1 & \alpha_2 & \alpha_3 & \cdots & \alpha_{n-2} & \alpha_{n-1} \\ \alpha_2 & \alpha_3 & \alpha_4 & \cdots & \alpha_{n-1} & \alpha_n \\ \alpha_3 & \alpha_4 & \alpha_5 & \cdots & \alpha_n & \alpha_{n+1} \\ \vdots & & & & & \vdots \\ \alpha_{n-2} & \alpha_{n-1} & \alpha_n & \cdots & & m_{2n-5} \\ \alpha_{n-1} & \alpha_n & \alpha_{n+1} & \cdots & m_{2n-5} & m_{2n-4} \end{vmatrix}.$$

It is obvious that Δ_1 and Δ_2 cannot be identically zero. The same is true for other polynomials, although it is less obvious. We demonstrate the proof for Δ_{n-1}. The same approach can be used for other cases.

If Δ_{n-1} is identically zero, then its value should be zero for whatever values we assign to the parameters $\alpha_2, \ldots, \alpha_n$. However, if we let

$$\alpha_2 = \alpha_3 = \cdots = \alpha_{n-1} = 0 \qquad \text{and} \qquad \alpha_n = 1,$$

then we obtain the determinant

$$\begin{vmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 1 & x \\ \vdots & & & & & \vdots \\ 0 & 0 & 1 & \cdots & x & x \\ 0 & 1 & x & \cdots & x & x \end{vmatrix} = \pm 1.$$

This is contradictory to the assumption and proves the lemma. □

COROLLARY 3.3. *If we apply the combined procedure described above to the matrix F, then all the pivots of the* SF*-procedure will be nontrivial rational functions of the parameters* $\alpha_2, \ldots, \alpha_n$. For example,

$$\Delta_2 = \alpha_3 - \alpha_2^2 \neq 0$$

is exactly the condition for the first pivot of the SF-procedure to be nonzero.

Remark 3.4. As each entry in the determinant Δ_k is a linear function in variables $\alpha_2, \ldots, \alpha_n$, the degree of Δ_k in each of α's does not exceed k, the order of Δ_k. Moreover, the total degree of Δ_k is no more than k.

Let $\Phi \in \mathbf{K}[x_1, x_2, \ldots, x_s]$, the ring of polynomials over $\mathbf{K}$, where $\mathbf{K} = \mathbf{R}$ or $\mathbf{K} = \mathbf{C}$. Suppose that Φ is nontrivial and has the degree d_j in x_j, $j = 1, \ldots, s$. We use G^s to denote a grid in $\mathbf{K}^s$ with sides parallel to the coordinate axes. On the jth axis there are $d_j + 1$ nodes in the G^s; the spacing of these nodes can be arbitrary. The grid G^s, therefore, consists of $(d_1 + 1)(d_2 + 1) \cdots (d_s + 1)$ nodes. Then, we have the following lemma.

LEMMA 3.5. *There exists a grid node* $(x_1^*, x_2^*, \ldots, x_s^*)$ *in* G^s *such that*

$$\Phi(x_1^*, x_2^*, \ldots, x_s^*) \neq 0,$$

and such a node can be found by no more than N_s evaluations of Φ at nodes of G^s, where

$$N_s = \left(\sum_{j=1}^{s} d_j \right) + s.$$

Proof. The proof is by induction. For $s = 1$, the assertion of the lemma is a slightly modified version of the well-known fact: *a nontrivial polynomial Φ in one variable x_1, having the degree d_1, cannot have more than d_1 roots.* So, evaluating this polynomial at any different $d_1 + 1$ nodes, we will certainly find amongst them one that is not a root of Φ.

Assume the lemma is true for all $s < k$. We now prove it for $s = k$. We view Φ as a polynomial in the variable x_1 with coefficients in $\mathbf{K}[x_2, \ldots, x_k]$. Let $\phi(x_2, \ldots, x_k)$ be the leading coefficient of Φ. The degree of ϕ in x_j does not exceed d_j, $j = 2, \ldots, k$, and ϕ is nontrivial by definition.

Let G^{k-1} be a subgrid of G^k formed by the nodes of any layer perpendicular to the x_1-direction in G^k. By our assumption, there is a node $(x_2^*, \ldots, x_k^*)$, such that

$$\phi(x_2^*, \ldots, x_k^*) \neq 0,$$

and this node can be found in no more than $N_{k-1} = (\sum_{j=2}^{k} d_j) + (k-1)$ evaluations of ϕ. If we let

$$x_2 = x_2^*, \ldots, x_k = x_k^*,$$

then Φ becomes a polynomial in the variable x_1 with coefficients in $\mathbf{K}$. Evaluating Φ at $d_1 + 1$ different values, we will be able to find a node for which $\Phi \neq 0$. The total number of evaluations will not exceed

$$N_{k-1} + d_1 + 1 = \left(\sum_{j=1}^{k} d_j \right) + k = N_k. \qquad \square$$

With these results, we can now complete the proof of Theorem 1.3.

Proof. By Lemma 3.1, we can assume that the Hessenberg matrix is, in fact, the companion matrix F in (2). We set

$$v_1 = e_1, \qquad w_1^T = (1, \alpha_2, \ldots, \alpha_n),$$

and denote by Φ the product of the Δ's:

$$\Phi = \Delta_2 \Delta_3 \cdots \Delta_{n-1}.$$

Obviously, Φ is a polynomial in the variables $\alpha_2, \ldots, \alpha_n$, and its degree d_j in α_j does not exceed[1]

$$d_j = 2 + 3 + \cdots + (n-1) = \frac{n^2 - n - 2}{2}, \qquad j = 2, \ldots, n.$$

According to Lemma 3.5, we can find values $\alpha_2^*, \ldots, \alpha_n^*$, for which $\Phi(\alpha_2^*, \ldots, \alpha_n^*) \neq 0$. This requires a finite number of evaluations of Φ at different nodes. Every such

[1] In fact, the closed form for the degrees of the most of α's can be obtained, but it is not important for this proof.

evaluation can be considered as an application of the unsymmetric Lanczos algorithm (or the equivalent combined SF-procedure) to the matrix F with the initial vectors e_1 and $w_1 = (1, \alpha_2, \ldots, \alpha_n)^T$ with assigned values of $\alpha_2, \ldots, \alpha_n$ at the given node. If $\Phi = 0$ at the particular node, then the Lanczos process will end prematurely by a serious breakdown. However, if $\alpha_i = \alpha_i^*$, $j = 2, \ldots, n$, the Lanczos procedure will be completed successfully and a tridiagonal form of the original matrix H will be obtained. □

4. Proof of Theorem 1.2 (diagonalizable case). We apply the Danilevski algorithm to the matrix A. If the algorithm succeeds, a companion matrix of the form (2) will be obtained. Then the assertion follows from Theorem 1.3. On the other hand, it is possible that this algorithm prematurely halts and a matrix of the form

$$\hat{A} = \begin{bmatrix} F_1 & B_1 \\ 0 & A_1 \end{bmatrix}$$

is obtained. Here F_1 is a companion matrix (2) of order $k_1 (k_1 < n)$.

Let us apply a similarity transformation with the matrix

$$S = \begin{bmatrix} I_{k_1} & X \\ 0 & I_{n-k_1} \end{bmatrix}$$

to the matrix $\hat{A}$. Then,

$$\tilde{A} = S^{-1} \hat{A} S = \begin{bmatrix} F_1 & -XA_1 + F_1 X + B_1 \\ 0 & A_1 \end{bmatrix}.$$

The diagonalizability of the matrix A implies that the matrix equation

$$XA_1 - F_1 X = B_1 \tag{3}$$

has a solution. In fact, if the matrices F_1 and A_1 have disjoint spectra the solution of (3) is unique. Using any one of the solutions for (3), the matrix A can be written in the direct sum form

$$\tilde{A} = F_1 \oplus A_1.$$

Recursively applying the same reasoning to the matrix A_1, then A_2, and so on, we can obtain a direct sum decomposition of the matrix A

$$Q^{-1} A Q = F_1 \oplus F_2 \oplus \cdots \oplus F_m,$$

where matrices F_i are in companion form. Theorem 1.3 then can be applied to each of the matrices F_i. It remains to note that a solution of the linear matrix equation (3) of Sylvester type can be obtained by a finite procedure, using only rational operations of the corresponding number field.

5. Proof of Theorem 1.2 (nondiagonalizable case). We can proceed as in proving for the diagonalizable case. However, in case of a general matrix A, there is no guarantee whatsoever that matrix equation (3) has solutions. We can express this differently and possibly more clearly in terms of Krylov subspaces. In transformation

$$A \quad \rightarrow \quad \hat{A} = R^{-1} A R$$

columns $1, 2, \ldots, k_1$ of the matrix R are

$$e_1, Ae_1, \ldots, A^{k_1-1}e_1,$$

i.e., they constitute a basis of the Krylov subspace $\mathcal{K}(A, e_1)$ of dimension k_1. But some invariant subspaces of a general matrix A may have no A-invariant complement.

Let us precede the Danilevski procedure by the similarity transformation

$$A \quad \rightarrow \quad \check{A} = Z^{-1}AZ \tag{4}$$

with z as the first column in Z. Then for a new matrix

$$\hat{A} = \check{R}^{-1}\check{A}\check{R} = (\check{R}^{-1}Z^{-1})A(Z\check{R}) = \hat{R}^{-1}A\hat{R} \tag{5}$$

columns 1, 2, and so on, of the $\hat{R}$ form the Krylov sequence

$$z, Az, A^2z, \ldots$$

i.e., a basis of the Krylov subspace $\mathcal{K}(A, z)$.

Recall that the dimension of the Krylov subspace $\mathcal{K}(A, z)$is called the *index* (or *degree* [12]) of z with respect to A and is denoted by ind$z$. The maximal possible value of ind$z$ for a given A coincides with the degree m of the minimal polynomial of A.

Now, if ind$z = n$ then in (5) $\hat{A}$ is a companion matrix, and there is nothing to prove. Therefore, assume that ind$z = m < n$. We claim that the Krylov subspace $\mathcal{K}(A, z)$ has an A-invariant complement.

Indeed, the operator induced by A on $\mathcal{K}(A, z)$ is "responsible" for that part of the Jordan structure of A which is the direct sum of the Jordan blocks of the maximal dimensions, one block for each eigenvalue. As a consequence, there should be a complementary A-invariant subspace that provides the remaining part of the Jordan structure of A.

The existence of a complementary subspace amounts to the existence of a solution X for (3), if in (4) z is a vector with index m with respect to A. To prove Theorem 1.2 it remains to show that such a vector z could be constructed finitely. We can deduce that from the following two assertions. It is convenient to formulate them as Lemmas 5.1 and 5.3.

LEMMA 5.1. *The degree m of the minimal polynomial of A can be found by means of a finite number of rational operations.*

Proof. It is sufficient to apply the orthogonalization procedure to the sequence

$$I, \ A, \ A^2, \ldots, \ A^{n-1}. \tag{6}$$

The matrices in this sequence are considered as elements of $\mathbf{C}^{n^2}$ or $\mathbf{R}^{n^2}$ with the standard scalar product. Another possibility is to find the rank of the $n^2 \times n$ matrix $\mathcal{A}$ formed by matrices (6) as columns via elementary transformations. □

Remark 5.2. The first way of finding the number m was proposed in [8].

LEMMA 5.3. *If m is the (known) degree of the minimal polynomial of A then a vector z of index m with respect to A can be found by a finite procedure involving only rational operations of the corresponding number field.*

Proof. We search for a vector z such that the vectors

$$z, Az, \ldots, A^{m-1}z$$

are independent. In other words, letting

$$B = [\, z \mid Az \mid \cdots \mid A^{m-1}z \,],$$

we must ensure that at least one of the $m \times m$ minors of this matrix is nonzero.

Working with B in field $\mathbf{C}[z_1, \ldots, z_n]$ of rational functions in variables $z_1, \ldots, z_n$, we can obtain the decomposition

$$PB = LU,$$

where P is a permutation matrix, L is a lower triangular $n \times m$ matrix, U is a unit upper triangular $m \times m$ matrix, the last two matrices with elements in $\mathbf{C}[z_1, \ldots, z_n]$. The numerator of the element l_{mm} is a nontrivial polynomial in $z_1, \cdots, z_n$. We can therefore apply Lemma 3.5, and the assertion of this lemma follows. □

Returning to Theorem 1.2, we must only add that for a real matrix A the vector z in Lemma 5.3 could be taken real. Hence we can consider $\mathcal{K}(A, z)$ and its complementary subspace as subspaces in $\mathbf{R}^n$, and the matrix $\hat{R}$ is real as well.

Appendix. Would the tridiagonal form be beneficial for a general normal matrix? The theorem provided below shows that any complex normal matrix which is at the same time unreduced tridiagonal essentially is Hermitian (in the real case, symmetric or skew-symmetric). On the other hand, most normal matrices can be transformed into the unreduced tridiagonal form. Indeed, we can apply the standard Householder procedure to a general normal matrix first. If an unreduced Hessenberg matrix is obtained, the finite procedure described in this paper then can be used for attaining the tridiagonal form. However, we have to give up the normality to acquire the tridiagonal form, with the exception of simple Hermitian-like cases mentioned above. This explains to some extent why a canonical form under unitary similarity for unitary matrices invented in [2] is not tridiagonal but rather pentadiagonal with a number of zeros inside the band.

THEOREM 5.4. *If the matrix*

$$A = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & & \cdots & 0 \\ \gamma_1 & \alpha_2 & \beta_2 & 0 & & 0 \\ 0 & \gamma_2 & \alpha_3 & \beta_3 & & 0 \\ & & \ddots & \ddots & \ddots & \\ & & & & \gamma_{n-1} & \alpha_n \end{bmatrix}$$

is unreduced, namely,

$$\beta_i \gamma_i \neq 0, \qquad i = 1, 2, \ldots, n-1,$$

and A is also normal, then

1. *If A is real, the following must be true:*

$$A = A^T \quad \text{or} \quad A = \alpha I + K,$$

where $K = -K^T$, $\alpha \in \mathbf{R}$.

2. *If A is complex, then*

$$A = \alpha I + \tau H, \tag{7}$$

where $H = H^*$, $|\tau| = 1$, $\alpha \in \mathbf{C}$.

A sketch of the proof of Theorem 5.4 can be obtained by noting first that any normal matrix A satisfies $A^* = p(A)$ for some polynomial A, followed by an argument showing that the minimal degree of p for unreduced tridiagonal matrices must be one, and further followed by showing that normal matrices satisfying this relation are exactly the matrices referred to in Theorem 5.4. A full proof of this theorem can be found in [3, Lem. 3].

Acknowledgment. The authors are grateful to M. H. Gutknecht and B. N. Parlett for drawing their attention to Rutishauser's paper [13] as this paper was going to press. Use of Rutishauser's result could have shortened the proof of Theorem 1.2 and made a separate investigation of the Hessenberg case unnecessary. Our Lemmas 3.1 and 3.2, although not as elegant as Rutishauser's proof, still have one minor advantage to the latter: they are proved by purely algebraic means. We note also that the problem of finite choice of good starting vectors (i.e., vectors secured from serious breakdown) was not raised in [13].

The second author wishes to thank A. Ja. Belyankov for the useful discussion of the preliminary version of Lemma 3.5 and Matthew Austin for literary advice.

Finally, the authors wish to thank the referees for numerous helpful comments and suggestions.

REFERENCES

[1] R. H. BARTELS AND G. W. STEWART, *Solution of the equation $ax + xb = c$*, Comm. ACM, 15(1972), pp. 820–826.

[2] A. BUNSE-GERSTER AND L. ELSNER, *Schur parameter pencils for the solution of the unitary eigenproblem*, Linear Algebra Appl., 120 (1991), pp. 741–778.

[3] V. FABER AND J. MANTEUFFEL, *Necessary and sufficient conditions for existence of a conjugate gradient method*, SIAM J. Numer. Anal., 21 (1984), pp. 352–362.

[4] V. N. FADDEEVA, *Computational Methods of Linear Algebra*, Dover, New York, 1959.

[5] R. W. FREUND, M. H. GUTKNECHT, AND N. M. NACHTIGAL, *An implementation of the look-ahead Lanczos algorithm for non-Hermitian matrices, Part* I, Tech. Report RIACS 90.45, RIACS, Moffett Field, CA, 1990.

[6] R. W. FREUND AND N. M. NACHTIGAL, *An implementation of the look-ahead Lanczos algorithm for non-Hermitian matrices, Part* II, Tech. Report RIACS 90.46, RIACS, Moffett Field, CA, 1990.

[7] G. A. GEIST, *Reduction of a general matrix to tridiagonal form*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 362–373.

[8] B. R. GELBAUM, *An algorithm for the minimal polynomial of a matrix*, Amer. Math. Monthly, 90(1983), pp. 43–44.

[9] G. H. GOLUB, S. NASH, AND C. VAN LOAN, *A Hessenberg–Schur method for the matrix problem $ax + xb = c$*, IEEE Trans. Auto. Cont., AC-24 (1979), pp. 909–913.

[10] G. H. GOLUB AND C. VAN LOAN, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

[11] M. H. GUTKNECHT *A completed theory of the unsymmetric Lanczos process and related algorithms, Part* I, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 594–639.

[12] W. JOUBERT, *Lanczos methods for the solution of nonsymmetric systems of linear equations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 926–943.

[13] H. RUTISHAUSER, *Beiträge zur Kenntnis des Biorthogonalisierungs-Algorithmus von Lanczos*, ZAMP, 4(1953), pp. 35–56.

[14] D. E. HARE AND W. P. TANG, *Toward a stable tridiagonalization algorithm for unsymmetric matrices*, Tech. Report CS-89-03, University of Waterloo, Ontario, 1989.

[15] A. S. HOUSEHOLDER, *The Theory of Matrices in Numerical Analysis*, Dover, Blaisdell, New York, 1964.

[16] M. D. KENT, *Chebychev, Krylov, Lanczos: Matrix Relationships and Computations*, Ph.D. thesis, STAN-CS-89-1271, Stanford University, Stanford, CA, 1989.

[17] C.D. LABUDDE, *The reduction of an arbitrary real sparse matrix to tridiagonal form using similarity transformations*, Math. Comp., 17 (1963), pp. 443–447.

[18] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[19] ———, *Reduction to tridiagonal form and minimal realizations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 567–593.

[20] C. STRACHEY AND J. G. F. FRANCIS, *The reduction of a matrix to codiagonal form by elimination*, Comput. J., 4 (1961), pp. 168–176.

[21] D. R. TAYLOR, *Analysis of the Look Ahead Lanczos algorithm*, Ph.D. thesis, CPAM-108, University of California, Berkeley, CA, 1982.

[22] D. WATKINS, *Use of the* LR *algorithm to tridiagonalize a general matrix*, talk at Society for Industrial and Applied Mathematics Annual Meeting, Minneapolis, MN, 1988.

[23] J.H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, UK, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 388–414, April 1995

005

A LOOK-AHEAD BLOCK SCHUR ALGORITHM FOR TOEPLITZ-LIKE MATRICES*

ALI H. SAYED† AND THOMAS KAILATH‡

Abstract. We derive a look-ahead recursive algorithm for the block triangular factorization of Toeplitz-like matrices. The derivation is based on combining the block Schur/Gauss reduction procedure with displacement structure and leads to an efficient block-Schur complementation algorithm. For an $n \times n$ Toeplitz-like matrix, the overall computational complexity of the algorithm is $O(rn^2 + \frac{n^3}{t})$ operations, where r is the matrix displacement rank and t is the number of diagonal blocks. These blocks can be of any desirable size. They may, for example, correspond to the smallest nonsingular leading submatrices or, alternatively, to numerically well-conditioned blocks.

Key words. Toeplitz-like matrices, block Schur algorithm, block triangular factorization, linear equations, singular minors, look-ahead algorithm

AMS subject classifications. 65F05, 65F30, 15A23, 15A06

1. Introduction.

The triangular factorization of a matrix is a useful tool for many problems. Such a factorization is guaranteed to exist whenever the matrices are strongly regular, i.e., all leading principal minors are nonzero [11]. The standard Gaussian elimination technique (also known as Schur reduction) may then be used to compute the triangular factors of the matrix. Also, in many applications, one is often faced with matrices that exhibit some structure, e.g., Toeplitz, Hankel, close-to-Toeplitz, close-to-Hankel, and related matrices. Such structure is nicely captured by introducing the concept of displacement structure [18], [20]. In this context, an $n \times n$ structured matrix R is characterized by an $n \times r$ matrix G (called a generator of R) with $r \ll n$ usually. The minimum column dimension of G is called the displacement rank of R. The triangular factorization of such strongly regular R can be computed efficiently and recursively in $O(rn^2)$ operations (additions and multiplications) [19], [23], [32]. This is achieved by appropriately combining Gaussian elimination with displacement structure. The resulting algorithm can then be regarded as a far-reaching generalization of an algorithm of Schur [1], [36], which was chiefly concerned with the apparently very different problem of checking whether a power series is analytic and bounded in the unit disc; hence the name generalized Schur algorithm. The reader may consult the recent survey paper [21] for detailed discussions on the topic of displacement structure.

Now most fast factorization algorithms that have been derived so far in the literature assume that the involved structured matrices are strongly regular. In several instances, however, it might be more appropriate to perform *block* Schur complementation steps. This happens, for example, when the assumption of strong regularity is dropped, which then requires the use of the smallest nonsingular leading minor, or

* Received by the editors June 3, 1992; accepted for publication (in revised form) by G. Cybenko December 16, 1993. This work was supported in part by Air Force Office of Scientific Research, Air Force Systems Command contract AFOSR91-0060, and Army Research Office contract DAAL03-89-K-0109.

† Department of Electrical and Computer Engineering, University of California, Santa Barbara, California 93106-9560 (sayed@gibran.ece.ucsb.edu). This work was performed while Dr. Sayed was a Research Associate in the Information Systems Laboratory, Stanford University, and on leave from Escola Politécnica da Universidade de São Paulo, Brazil. This work was also supported by a fellowship from Fundacao de Amparo a Pesquisa do Estado de Sao Paulo, Brazil.

‡ Information Systems Laboratory, Stanford University, Stanford, California 94305.

a numerically well-conditioned leading minor of appropriate dimensions, in order to proceed with a block Schur reduction step.

Indeed, many authors have worked on the problem of extending the fast algorithms to nonstrongly regular matrices, where the sizes of the block Schur complementation steps were determined by the sizes of the smallest nonsingular leading minors. Among these we mention the works of Heinig and Rost [15], Delsarte, Genin, and Kamp [9], and Gover and Barnett [13] who generalized the classical Levinson algorithm for solving Toeplitz systems of linear equations (or equivalently factoring the inverse of the Toeplitz coefficient matrix); a so-called split-Levinson algorithm was later considered by Ciliz and Krishna [7]. Pombra, Lev-Ari, and Kailath [28] also derived both Levinson- and Schur-type algorithms for Toeplitz matrices by generalizing the three-term recursion for polynomials orthogonal on the unit circle. The case of nonstrongly regular Hankel matrices arises in many applications as well, such as the partial realization problem and decoding of BCH codes [4], [8]. Algorithms for computing the triangular factorization and/or inversion of arbitrary Hankel matrices have been derived by Berlekamp [3], Massey [25], Kung [22], and Citron [8]. More recently, Zarowski [37] used the algorithms of Heinig and Rost [15] and Delsarte, Genin, and Kamp [9] to induce Schur-type algorithms for Hermitian Toeplitz and Hankel matrices with singular minors.

All these algorithms are applicable to Toeplitz and Hankel matrices only. Recently, Pal and Kailath [26], [27] derived recursive algorithms that are applicable to a larger class of matrices called quasi-Toeplitz and quasi-Hankel. These are congruent to Toeplitz and Hankel matrices in a certain sense. The derivation exploits this fact and, among other results, shows that the determination of the size of the smallest nonsingular minor is reduced to counting the number of repeated zeros at the origin of a certain polynomial.

But Toeplitz, Hankel, quasi-Toeplitz, and quasi-Hankel matrices are all structured matrices with displacement rank $r = 2$. In many applications, however, such as system identification, image processing, and multichannel filtering, block structured matrices arise that have displacement ranks larger than two. In these cases, the previous algorithms are not applicable. Moreover, in the varied approaches above, the sizes of the block Schur complements were set equal to the sizes of the smallest nonsingular minors, which thus requires the verification of the occurrence of exact singularities. This may pose considerable difficulties from a numerical point of view.

Alternatively, one can determine the sizes of the block steps by looking for numerically well-conditioned blocks. This has recently been studied by several authors trying to devise effective numerical algorithms for general Toeplitz systems of equations. An early paper was the one of Chan and Hansen [6]. Among many others we mention Gutknecht [14] and Freund [10], which give extensive references.

In this paper, we provide a new fast look-ahead (block-Schur) algorithm for matrices with very general displacement structure, which includes the Toeplitz case as a special instance. We study arbitrary Hermitian Toeplitz-like matrices and derive an algorithm that leads to a factorization of the form $R = LDL^*$, where L is a lower triangular matrix and D is a block diagonal matrix whose block entries are easily invertible. The overall computational complexity of the algorithm is $O(rn^2 + n^3/t)$ elementary operations (addition and multiplication), where t is the number of diagonal blocks in D. In the strongly regular case we have $t = n$ and the complexity reduces to the usual $O(rn^2)$ figure. The diagonal blocks in D can be of any desirable size. They can be chosen, for example, as the smallest nonsingular minors or as the size

of numerically well-conditioned blocks. For this reason, our development consists of two independent steps. We first derive the block Schur algorithm assuming arbitrary choices for the sizes of the blocks, thus leading to a general-purpose fast Schur complementation procedure that does not depend on the specific choices for the sizes of these blocks. We then focus in §5 on the particular choices that correspond to the smallest (exactly) nonsingular leading blocks. This is done here because, apart from numerical possibilities, the fast block-Schur complementation algorithm also has several theoretically interesting features as well. For example, the explicit formulas for the block diagonal matrix in the block triangular factorization can give simple rules for computing the inertia of general structured matrices, with important applications in root distribution of polynomials.

The paper is organized as follows. In §2 we review the class of structured matrices and describe the Schur/Gauss reduction procedure for the triangular factorization of strongly regular matrices. In §3 we combine the Schur reduction procedure with displacement structure and derive the generalized block Schur algorithm. In §4 we separately consider the special cases of strongly regular and block steps along with the corresponding computational complexities. In §5 we address the issue of determining the sizes of the smallest (exact) nonsingular minors. In §6 we show how to exploit the matrix structures in order to efficiently compute the QR factors of the blocks of D_i. In §7 we give a system (and state-space) interpretation of the derived recursions and we conclude with §8.

2. Structured matrices. The concept of displacement structure and structured matrices can be briefly motivated by considering the much-studied special case of a Hermitian Toeplitz matrix, $T = [c_{i-j}]_{i,j=0}^{n-1}, c_k = c_{-k}^*$. Since T depends only on n parameters rather than n^2, it may not be surprising that matrix problems involving T (such as triangular factorization, orthogonalization, inversion) have complexity $O(n^2)$ rather than $O(n^3)$. But what about the complexity of such problems for inverses, products, and related combinations of Toeplitz matrices such as $T^{-1}, T_1T_2, T_1 - T_2T_3^{-1}T_4, (T_1T_2)^{-1}T_3, \ldots$? Though these are not Toeplitz, they are certainly structured and the complexity of inversion and factorization is not expected to be much different from that for a pure Toeplitz matrix, T. It turns out that the appropriate common property of all these matrices is not their "Toeplitzness," but the fact that they all have low *displacement rank*. The displacement of an $n \times n$ Hermitian matrix R was originally defined by Kailath, Kung, and Morf [20] as

$$\nabla R \equiv R - ZRZ^*, \tag{1}$$

where the symbol $*$ stands for Hermitian conjugate transpose of a matrix (complex conjugation for scalars), and Z is the $n \times n$ lower shift matrix with ones on the first subdiagonal and zeros elsewhere; ZRZ^* corresponds to shifting R downward along the main diagonal by one position, explaining the name *displacement* for ∇R. If ∇R has low rank, say r, independent of n, then R is said to be *structured* with respect to the displacement defined by (1), and r is referred to as the displacement rank of R. In this case, we can (nonuniquely) factor ∇R as $\nabla R = GJG^*$, where $J = J^*$ is a signature matrix that specifies the *displacement inertia* of R: it has as many $\pm 1's$ on the diagonal as ∇R has positive and negative eigenvalues, $J = (I_p \oplus -I_q)$, $p + q = r$, and G is an $n \times r$ matrix. Here, I_p denotes the $p \times p$ identity matrix. The pair $\{G, J\}$ is called a *generator* of R. For a Hermitian Toeplitz matrix $T = [c_{i-j}]_{i,j=0}^{n-1}, c_k = c_{-k}^*$, with $c_0 = 1$, it is straightforward to verify that (1) leads to a compact description of

T. Indeed, if we subtract ZTZ^* from T we get

$$T - ZTZ^* = \begin{bmatrix} 1 & 0 \\ c_1 & c_1 \\ \vdots & \vdots \\ c_{n-1} & c_{n-1} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ c_1 & c_1 \\ \vdots & \vdots \\ c_{n-1} & c_{n-1} \end{bmatrix}^*, \tag{2}$$

which shows that $T - ZTZ^*$ has rank 2, or equivalently, T has displacement rank 2, independent of n.

To motivate more general structures, and to clarify the importance of direct factorization problems as opposed to inversion problems, we consider a simple example that shows the need for more general structures such as $R - FRF^*$, with lower triangular F.

Consider again the case of an $n \times n$ Hermitian Toeplitz matrix T for which $T - Z_n T Z_n^*$ has rank 2 (Z_n now denotes the $n \times n$ lower shift matrix), and assume we are interested in factoring T^{-1}. If we form the block matrix (see [19] for more examples and discussion)

$$M = \begin{bmatrix} -T & I \\ I & \mathbf{0} \end{bmatrix},$$

it is then straightforward to check that the displacement rank of M with respect to $M - Z_{2n} M Z_{2n}^*$ is equal to four. However, we can get a lower displacement rank by using a different definition, viz.,

$$\nabla M = M - \begin{bmatrix} Z_n & \mathbf{0} \\ \mathbf{0} & Z_n \end{bmatrix} M \begin{bmatrix} Z_n & \mathbf{0} \\ \mathbf{0} & Z_n \end{bmatrix}^*,$$

which corresponds to choosing $F = Z_n \oplus Z_n$ in the definition $R - FRF^*$ (rather than $F = Z_{2n}$, the $2n \times 2n$ lower shift matrix).

The question is then how to exploit the structure of M in order to obtain fast factorization of T^{-1}. The answer is that the (generalized) Schur algorithm operates as follows: it starts with a generator matrix G of a structured matrix (say the generator of M), and it recursively computes generator matrices of the successive Schur complements of the matrix. So the first step of the algorithm gives us G_1, which is a generator of the Schur complement of M with respect to its $(0,0)$ entry. The next step gives us G_2, which is a generator of the Schur complement of M with respect to its 2×2 leading submatrix, and so on. After n such steps, we obtain a generator of the nth Schur complement, which is T^{-1}. This procedure can be shown to provide the triangular factorization of T^{-1} (see, e.g., [19], [21]).

Hence, by performing the *direct* factorization of the extended matrix M we also obtain the factors of the inverse matrix T^{-1}; this is an alternative to the use of the Levinson algorithm for this problem. Applications with more general matrices F (such as diagonal or in Jordan form) include interpolation problems [5], [33], [34], and adaptive filtering [35].

In this paper we study $n \times n$ Hermitian matrices R with Toeplitz-like displacement structure of the form

$$R - FRF^* = GJG^*, \tag{3}$$

where F is an $n \times n$ lower triangular matrix with diagonal elements $\{f_0, f_1, \ldots, f_{n-1}\}$, G is an $n \times r$ so-called generator matrix (with $r \leq n$), and J is an $r \times r$ Hermitian

signature matrix satisfying $J^2 = I$, such as $J = (I_p \oplus -I_q)$, $p+q=r$, or some other convenient form (as will be the case in §5.5). We further assume that R is invertible but *not necessarily strongly regular*, and that $1 - f_i f_j^* \neq 0$ for every i, j. The latter condition guarantees the existence of a unique solution R of (3) (but it can be relaxed as discussed in [21]). We say that R has a Toeplitz-like structure with respect to F and $\{G, J\}$ is called a generator pair of R.

2.1. The block Schur/Gauss reduction procedure. The Gaussian elimination (or Schur reduction) procedure is a recursive algorithm that computes the triangular factors of a matrix. To clarify this, consider a Hermitian and invertible (but not necessarily strongly regular) matrix R, and let η_0 denote the desired size of the leading (invertible) block, D_0, with respect to which a Schur complementation step is to be performed. The η_0 may stand for the size of the smallest nonsingular minor of R or, alternatively, for the size of a numerically well-conditioned block (as in [6], [10], for example), or for some other convenient choice. If L_0 represents the first η_0 columns of R then

$$R - L_0 D_0^{-1} L_0^* = \begin{bmatrix} \mathbf{0}_{\eta_0 \times \eta_0} & \mathbf{0} \\ \mathbf{0} & R_1 \end{bmatrix} \equiv \tilde{R}_1,$$

where R_1 is an $(n-\eta_0) \times (n-\eta_0)$ matrix that is called the Schur complement of D_0 in R. Also, L_0 is an $n \times \eta_0$ matrix whose leading $\eta_0 \times \eta_0$ block is equal to D_0. We shall say that $\tilde{R}_1$ has one (block) zero row and one (block) zero column (the size of the block being η_0). If we further let η_1 denote the desired size of the leading (invertible) block of R_1 (denoted by D_1) and consider the corresponding first η_1 columns of R_1 (denoted by L_1), then we also have

$$R_1 - L_1 D_1^{-1} L_1^* = \begin{bmatrix} \mathbf{0}_{\eta_1 \times \eta_1} & \mathbf{0} \\ \mathbf{0} & R_2 \end{bmatrix} \equiv \tilde{R}_2,$$

where R_2 is now an $(n-\eta_0-\eta_1) \times (n-\eta_0-\eta_1)$ matrix that is the Schur complement of D_1 in R_1. Repeating this recursive procedure, viz,

$$\begin{bmatrix} \mathbf{0}_{\eta_i \times \eta_i} & \mathbf{0} \\ \mathbf{0} & R_{i+1} \end{bmatrix} = R_i - L_i D_i^{-1} L_i^*, \qquad i \geq 0, \tag{4}$$

we clearly get, say after t steps,

$$R = L D_B^{-1} L^* =$$

$$L_0 D_0^{-1} L_0^* + \begin{bmatrix} \mathbf{0}_{\eta_0 \times \eta_1} \\ L_1 \end{bmatrix} D_1^{-1} \begin{bmatrix} \mathbf{0}_{\eta_0 \times \eta_1} \\ L_1 \end{bmatrix}^* + \begin{bmatrix} \mathbf{0}_{\eta_0 \times \eta_2} \\ \mathbf{0}_{\eta_1 \times \eta_2} \\ L_2 \end{bmatrix} D_2^{-1} \begin{bmatrix} \mathbf{0}_{\eta_0 \times \eta_2} \\ \mathbf{0}_{\eta_1 \times \eta_2} \\ L_2 \end{bmatrix}^* + \ldots +,$$

where $D_B = (D_0 \oplus D_1 \oplus \cdots \oplus D_{t-1})$ is *block* diagonal, and the (nonzero parts of the) columns of the *block* lower triangular matrix L are $\{L_0, \ldots, L_{t-1}\}$. Here t is the number of reduction steps, i.e., $n = \sum_{i=0}^{t-1} \eta_i$. We also define, for later reference, $\alpha_j = \sum_{i=0}^{j-1} \eta_i, \alpha_0 = 0$. The computational complexity of the above procedure is $O(n^3)$ elementary operations and it leads to a block triangular factorization of R.

It is clear at this point that the following questions are among the major issues that arise during the block triangular factorization procedure: (i) how to efficiently

exploit any Toeplitz-like structure of R; (ii) how to efficiently compute the triangular factors L_i and D_i; (iii) how to compute (or avoid) the inversion of the diagonal blocks D_i; (iv) how to determine an alternative triangular factorization of the form $R = \hat{L}\hat{D}_B^{-1}\hat{L}^*$, with $\hat{L}$ lower (*not block*) triangular and with a block-diagonal matrix $\hat{D}_B$ whose block entries are easily invertible; (v) how to determine the sizes of the block steps, η_i.

We address the first four questions in the next two sections and postpone the discussion of the last question to §5, where we focus on a particular choice for the η_i that is determined by the sizes of the smallest nonsingular minors of R. It will be clear from the derivation that follows that, in order to increase the computational efficiency of the resulting algorithm, these questions should be answered by essentially restricting ourselves to the use of the entries of the generator matrix of R, without the need to explicitly form its successive block Schur complements, R_i.

3. Block Schur algorithm for Toeplitz-like matrices. We now exploit the fact that R is a structured (Toeplitz-like) matrix. That is, we show that the successive computation of the Schur complements of R in (4) can be carried out in a computationally efficient recursive procedure by exploiting the structure implied by (3). To begin with, we define F_i to be the submatrix obtained by ignoring the first α_i columns and rows (or the first i block columns and rows) of F (recall that $\alpha_i = \eta_0 + \cdots + \eta_{i-1}$). This means that F_{i+1} is a submatrix of F_i, viz.,

$$F_i = \begin{bmatrix} \hat{F}_i & \mathbf{0} \\ ? & F_{i+1} \end{bmatrix}, \qquad F_0 = F,$$

where ? denotes irrelevant entries and $\hat{F}_i$ is the $\eta_i \times \eta_i$ leading submatrix of F_i. In other words, F_{i+1} is obtained by deleting the first η_i rows and columns of F_i. The following theorem, first stated in general terms, shows that the successive Schur complements of a Toeplitz-like matrix inherit its structure and thus satisfy a displacement equation similar to (3).

THEOREM 3.1. *The ith Schur complement R_i of a Toeplitz-like matrix R, as in* (3) *and* (4), *is also Toeplitz-like with respect to F_i, viz., R_i satisfies a displacement equation of the form $R_i - F_i R_i F_i^* = G_i J G_i^*$, where the generator matrix G_i satisfies the following recursive construction: start with $G_0 = G, F_0 = F$, and repeat for $i = 0, 1, \ldots, t-1$:*

1. *At step i we have F_i and G_i. Let $\hat{G}_i$ denote the top η_i rows of G_i.*
2. *The ith triangular factors L_i and D_i are the solutions of the equations*

$$D_i = \hat{F}_i D_i \hat{F}_i^* + \hat{G}_i J \hat{G}_i^*, \qquad L_i = F_i L_i \hat{F}_i^* + G_i J \hat{G}_i^*. \tag{5a}$$

3. *Choose arbitrary $r \times \eta_i$ and $r \times r$ matrices $\hat{H}_i$ and $\hat{K}_i$, respectively, so as to satisfy the embedding relation*

$$\begin{bmatrix} \hat{F}_i & \hat{G}_i \\ \hat{H}_i & \hat{K}_i \end{bmatrix} \begin{bmatrix} D_i & \mathbf{0} \\ \mathbf{0} & J \end{bmatrix} \begin{bmatrix} \hat{F}_i & \hat{G}_i \\ \hat{H}_i & \hat{K}_i \end{bmatrix}^* = \begin{bmatrix} D_i & \mathbf{0} \\ \mathbf{0} & J \end{bmatrix}. \tag{5b}$$

4. *A generator for R_{i+1} is then given by*

$$\begin{bmatrix} \mathbf{0}_{\eta_i \times r} \\ G_{i+1} \end{bmatrix} = F_i L_i \hat{H}_i^* J + G_i J \hat{K}_i^* J. \tag{5c}$$

Proof. We prove the result for G_1. The same argument applies to $\{G_i,\ i>1\}$. It follows from (3) that the leading submatrix D_0 and the corresponding η_0 columns L_0 are solutions of the equations: $D_0 = \hat{F}_0 D_0 \hat{F}_0^* + \hat{G}_0 J \hat{G}_0^*$ and $L_0 = F L_0 \hat{F}_0^* + G J \hat{G}_0^*$. Substituting these expressions into the definition of $\tilde{R}_1$ in (4) and computing the difference $\tilde{R}_1 - F\tilde{R}_1 F^*$ we get

$$\begin{aligned}\tilde{R}_1 - F\tilde{R}_1 F^* &= GJ\left\{J - \hat{G}_0^* D_0^{-1}\hat{G}_0\right\} JG^* \\ &\quad + FL_0\left[D_0^{-1} - \hat{F}_0^* D_0^{-1}\hat{F}_0\right] L_0^* F^* \\ &\quad - FL_0\hat{F}_0^* D_0^{-1}\hat{G}_0 JG^* - GJ\hat{G}_0^* D_0^{-1}\hat{F}_0 L_0^* F^*.\end{aligned} \tag{6}$$

We now verify that the right-hand side of the above expression can be put into the form of a *perfect square* by introducing some auxiliary quantities. Consider an $r \times \eta_0$ matrix $\hat{H}_0$ and an $r \times r$ matrix $\hat{K}_0$ that are defined to satisfy the following relations (in terms of the quantities that appear on the right-hand side of the above expression. We shall see very soon that this is always possible).

$$\hat{H}_0^* J \hat{H}_0 = D_0^{-1} - \hat{F}_0^* D_0^{-1}\hat{F}_0,\quad \hat{K}_0^* J \hat{K}_0 = J - \hat{G}_0^* D_0^{-1}\hat{G}_0,\quad \hat{K}_0^* J \hat{H}_0 = -\hat{G}_0^* D_0^{-1}\hat{F}_0.$$

Using $(\hat{H}_0, \hat{K}_0)$ we can factor the right-hand side of (6) as $\tilde{G}_1 J \tilde{G}_1^*$, where $\tilde{G}_1 = FL_0\hat{H}_0^* J + GJ\hat{K}_0^* J$. But the first block row and block column of $\tilde{R}_1$ are zero. Hence, $\tilde{G}_1$ is of the form $\tilde{G}_1 = \begin{bmatrix} \mathbf{0}_{r\times\eta_0} & G_1^{\mathbf{T}} \end{bmatrix}^{\mathbf{T}}$. Moreover, it follows from the above expressions for $(\hat{H}_0, \hat{K}_0)$ that $\hat{F}_0$, $\hat{G}_0$, $\hat{H}_0$, and $\hat{K}_0$ satisfy the relation

$$\begin{bmatrix} \hat{F}_0 & \hat{G}_0 \\ \hat{H}_0 & \hat{K}_0 \end{bmatrix}^* \begin{bmatrix} D_0^{-1} & \mathbf{0} \\ \mathbf{0} & J \end{bmatrix} \begin{bmatrix} \hat{F}_0 & \hat{G}_0 \\ \hat{H}_0 & \hat{K}_0 \end{bmatrix} = \begin{bmatrix} D_0^{-1} & \mathbf{0} \\ \mathbf{0} & J \end{bmatrix},$$

which is equivalent to (5b) for $i = 0$. □

We still need to show how to choose matrices $(\hat{H}_i, \hat{K}_i)$ so as to satisfy the embedding relation (5b). Following an argument similar to that in [24] we get the following result.

LEMMA 3.2. *All choices of $\hat{H}_i$ and $\hat{K}_i$ that satisfy* (5b) *can be expressed in terms of $\hat{F}_i$, $\hat{G}_i$, and D_i as follows*:

$$\begin{aligned}\hat{H}_i &= \Theta_i^{-1} J \hat{G}_i^* \left[I_{\eta_i} - \tau_i \hat{F}_i^*\right]^{-1} D_i^{-1}(\tau_i I_{\eta_i} - \hat{F}_i), \\ \hat{K}_i &= \Theta_i^{-1}\left\{I_r - J\hat{G}_i^*\left[I_{\eta_i} - \tau_i\hat{F}_i^*\right]^{-1} D_i^{-1}\hat{G}_i\right\},\end{aligned} \tag{7}$$

where Θ_i is an arbitrary $J-$unitary matrix ($\Theta_i J \Theta_i^* = J$) and τ_i is an arbitrary unit-modulus scalar ($|\tau_i| = 1$).

Substituting expression (7) for $\hat{H}_i$ and $\hat{K}_i$ into the generator recursion (5c) we obtain the following algorithm, which we refer to as the *generalized block Schur algorithm.* This algorithm allows us to compute generator matrices for the successive (block) Schur complements of R, viz., $G \to G_1 \to G_2 \to \ldots$, which can then be used to solve for the triangular factors via (5a).

ALGORITHM 3.3 (BLOCK SCHUR ALGORITHM). *The generators G_i of the successive Schur complements R_i satisfy the recursion*

$$\begin{bmatrix} \mathbf{0}_{\eta_i\times r} \\ G_{i+1} \end{bmatrix} = \left\{G_i + (\tau_i^* F_i - I_{n-\alpha_i}) L_i D_i^{-1}(I_{\eta_i} - \tau_i^*\hat{F}_i)^{-1}\hat{G}_i\right\}\Theta_i, \tag{8}$$

where Θ_i is an arbitrary $J-$unitary matrix and τ_i is an arbitrary unit-modulus scalar. The ith triangular factors L_i and D_i are found by solving (5a).

4. Computational issues and simplifications. The point to stress here is that we have so far shown the following: the triangular factors L_i and D_i can be found by solving (5a), which are completely specified in terms of F_i and G_i and without the need to explicitly form R_i, since the $G_i's$ can be recursively computed via (8). To further stress this point we now take a closer look at recursion (8) and (5a).

4.1. Strongly regular steps. We first consider the special case that corresponds to $\eta_i = 1$, and which we refer to as a *strongly regular* step. In this case, it is possible to further simplify the generator recursion (8). To this effect, we note that the triangular factor L_i is now a column vector, which we denote by the lowercase letter l_i, the diagonal factor D_i is a scalar, denoted by d_i, the first η_i rows of G_i collapse to a single row vector, denoted by g_i, and the quantity $\hat{F}_i$ is also a scalar equal to f_{α_i} (we are using lower case letters to refer to quantities in a strongly regular step). A direct consequence of these facts is that we can now explicitly solve for d_i and l_i in (5a). More specifically, we get

$$d_i = \frac{g_i J g_i^*}{1 - |f_{\alpha_i}|^2}, \qquad l_i = (I_{n-\alpha_i} - f_{\alpha_i}^* F_i)^{-1} G_i J g_i^*. \tag{9a}$$

Substituting these expressions into the generator recursion (8) we readily verify that it simplifies to

$$\begin{bmatrix} \mathbf{0}_{1\times r} \\ G_{i+1} \end{bmatrix} = \left\{ G_i + (\Phi_i - I_{n-\alpha_i}) G_i \frac{J g_i^* g_i}{g_i J g_i^*} \right\} \Theta_i, \tag{9b}$$

where Φ_i is a "Blaschke matrix" or "Blaschke–Potapov" factor (see [29]) of the form

$$\Phi_i = \frac{1 - \tau_i f_{\alpha_i}^*}{\tau_i - f_{\alpha_i}} (F_i - f_{\alpha_i} I_{n-\alpha_i})(I_{n-\alpha_i} - f_{\alpha_i}^* F_i)^{-1}. \tag{9c}$$

The difference between (9b) and the general form (8) is that recursion (9b) is written in terms of F_i and G_i only, while expression (8) still involves L_i and D_i^{-1}.

We now move a step further and show that (9b) can be further simplified by conveniently choosing the free parameters Θ_i and τ_i. An obvious choice is $\Theta_i = I_r$ and

$$\tau_i = \frac{1 + f_{\alpha_i}}{1 + f_{\alpha_i}^*}$$

(this choice for τ_i leads to $\Phi_i = (F_i - f_{\alpha_i} I_{n-\alpha_i})(I_{n-\alpha_i} - f_{\alpha_i}^* F_i)^{-1}$). There are other convenient choices for Θ_i as well, such as the one we describe next: a strongly regular step clearly implies that $d_i \neq 0$ and consequently $g_i J g_i^* \neq 0$. That is, g_i has nonzero $J-$norm. Hence, we can always find a $J-$unitary rotation Θ_i that reduces g_i to the form

$$g_i \Theta_i = \begin{bmatrix} 0 & \dots & 0 & x_i^{(j)} & 0 & \dots & 0 \end{bmatrix}, \tag{10}$$

with a nonzero entry in a single (convenient) column, say the jth column. So assume we use this choice for Θ_i, which can be implemented in a variety of ways: we may use elementary rotations such as Givens or hyperbolic [12] or Householder transformations [12], [30]. Using the above choice leads to the following algorithm in the strongly regular case.

ALGORITHM 4.1 (Strongly regular step). *The generator recursion for a strongly regular step is given by*

$$\begin{bmatrix} \mathbf{0} \\ G_{i+1} \end{bmatrix} = G_i\Theta_i \begin{bmatrix} I_j & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_{r-j-1} \end{bmatrix} + \Phi_i G_i \Theta_i \begin{bmatrix} \mathbf{0}_j & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0}_{r-j-1} \end{bmatrix}, \tag{11}$$

where $\Phi_i = (F_i - f_{\alpha_i} I_{n-\alpha_i})(I_{n-\alpha_i} - f_{\alpha_i}^* F_i)^{-1}$. *That is,* G_{i+1} *is obtained as follows: choose a convenient* $J-$*unitary rotation that reduces the first row of* G_i *to the form* (10); *multiply the* j*th column of* $G_i\Theta_i$ *by* Φ_i *and keep all other columns unchanged; these steps result in a generator* G_{i+1}. *The triangular factors are given by*

$$d_i = g_i J g_i^*/(1 - |f_{\alpha_i}|^2) \quad \text{and} \quad l_i = (I_{n-\alpha_i} - f_{\alpha_i}^* F_i)^{-1} G_i \Theta_i J \begin{bmatrix} \mathbf{0} & x_i^{*(j)} & \mathbf{0} \end{bmatrix}^*.$$

An alternative form for the generator recursion that corresponds to using $\Theta_i = I_r$, *instead of* (10), *is given by*

$$\begin{bmatrix} \mathbf{0}_{1\times r} \\ G_{i+1} \end{bmatrix} = G_i + (\Phi_i - I_{n-\alpha_i}) G_i \frac{J g_i^* g_i}{g_i J g_i^*}.$$

In this case, we compute l_i *via* $l_i = (I_{n-\alpha_i} - f_{\alpha_i}^* F_i)^{-1} G_i J g_i^*$, *and* d_i *is the leading entry of* l_i.

We assume throughout that F is a sparse matrix in the sense that computing Fx, for any $n \times 1$ column vector x, requires $O(n)$ operations. It can then be checked that each step of recursion (11) requires $O(r(n-\alpha_i))$ operations. Furthermore, we may not need to explicitly compute the inverse matrix $(I_{n-\alpha_i} - f_{\alpha_i}^* F_i)^{-1}$ that appears in the expressions for Φ_i and l_i. We can instead, in the case of l_i for example, solve a *triangular* system of linear equations of the form $(I_{n-\alpha_i} - f_{\alpha_i}^* F_i)x = G_i\Theta_i J \begin{bmatrix} \mathbf{0} & 1 & \mathbf{0} \end{bmatrix}^{\mathbf{T}}$. Moreover, in many applications the matrix F has zero diagonal entries (i.e., $f_{\alpha_i} = 0$), in which case computing l_i and Φ_i is trivialized since the inverse term disappears.

As remarked above, a strongly regular step corresponds to $d_i \neq 0$, or equivalently, $g_i J g_i^* \neq 0$. There is however a trivial special case with $d_i = 0$, which can still be incorporated into a strongly regular step. This happens when g_i is itself a zero row vector. That is, G_i is of the form

$$G_i = \begin{bmatrix} \mathbf{0} \\ \bar{G}_i \end{bmatrix}.$$

This implies that the first row and column of R_i are zero. Going back to the description of the Schur reduction procedure in §2.1, we see that we can proceed in this special case by choosing $l_i = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}^{\mathbf{T}}$ and by setting $G_{i+1} = \bar{G}_i$ and F_{i+1} equal to the submatrix obtained by deleting the first row and column of F_i.

4.2. Block steps. We now consider the case $\eta_i > 1$ that we refer to as a *block* step. In this case, the triangular factors L_i and D_i are block matrices and it is not possible, in general, to solve for L_i and D_i and write down simple explicit expressions in terms of F_i and G_i only, as in the strongly regular case (see (9a)).

We can however proceed with (8) and use the simple choices $\Theta_i = I_r$ and $\tau_i = 1$. Under these conditions, we can rewrite the generator recursion (8) in the following form.

ALGORITHM 4.2 (Block step). *The generator recursion for a block step can be expressed as*

$$\begin{bmatrix} \mathbf{0}_{\eta_i \times r} \\ G_{i+1} \end{bmatrix} = G_i + X_i, \tag{12a}$$

where $X_i = (F_i - I_{n-\alpha_i}) L_i D_i^{-1} (I_{\eta_i} - \hat{F}_i)^{-1} \hat{G}_i$. *The triangular factor* L_i *is obtained by solving the equation (the leading* $\eta_i \times \eta_i$ *submatrix of* L_i *provides* D_i*)*

$$L_i = F_i L_i \hat{F}_i^* + G_i J \hat{G}_i^*. \tag{12b}$$

Expression (12a) shows that G_{i+1} is obtained by adding the last $(n - \alpha_{i+1})$ rows of X_i and G_i, while the top rows of X_i should cancel the top rows of G_i.

4.2.1. Computing L_i. Solving for L_i in (12b) is not a major problem in most applications such as linear prediction, inverse scattering, solution of (structured) linear systems, least-squares problems, interpolation problems, etc., because the matrix F arises in sparse forms, e.g., $F = Z$, $F = Z + \lambda I$, $F =$ diagonal $\{f_0, f_1, \ldots, f_{n-1}\}$, $F = (Z + \lambda_0 I) \oplus (Z + \lambda_1 I) \oplus \ldots$, $F = Z +$ diagonal $\{f_0, \ldots, f_{n-1}\}$. Consider, for instance, this last bidiagonal form. Denote the η_i columns of L_i by $L_i = \begin{bmatrix} l_{i0} & l_{i1} & \ldots & l_{i,\eta_i - 1} \end{bmatrix}$, and the η_i rows of $\hat{G}_i$ by $\{g_{i0}, g_{i1}, \ldots, g_{i,\eta_i - 1}\}$ $(g_i = g_{i0})$. It is then straightforward to check, using (12b), that the columns of L_i can be recursively computed as follows:

$$\begin{aligned} l_{i0} &= (I_{n-\alpha_i} - f_{\alpha_i}^* F_i)^{-1} G_i J g_{i0}^*, \\ l_{ij} &= (I_{n-\alpha_i} - f_{\alpha_i + j}^* F_i)^{-1} \left[G_i J g_{ij}^* + F_i l_{i,j-1} \right] \quad \text{for } j = 1, \ldots, \eta_i - 1. \end{aligned}$$

Once again, the inversion $(I_{n-\alpha_i} - f_{\alpha_i+j}^* F_i)^{-1}$ can be avoided by solving a sparse triangular system of linear equations. The computational complexity needed in computing L_i is $O(r\eta_i(n - \alpha_i))$.

4.2.2. Computing X_i. We now consider the operation count for one possibility for computing X_i (other possibilities clearly exist). Recall that L_i has D_i as its leading block. To show this explicitly we partition L_i as follows: $L_i = \begin{bmatrix} D_i^{\mathbf{T}} & W_i^{\mathbf{T}} \end{bmatrix}^{\mathbf{T}}$. Then $L_i D_i^{-1} = \begin{bmatrix} I_{\eta_i} & (W_i D_i^{-1})^{\mathbf{T}} \end{bmatrix}^{\mathbf{T}}$. At this stage we introduce the QR decomposition of D_i, viz., $D_i = Q_i P_i$, where Q_i is an $\eta_i \times \eta_i$ unitary matrix $(Q_i Q_i^* = I_{\eta_i})$ and P_i is an $\eta_i \times \eta_i$ nonsingular upper triangular matrix. Invoking the fact that D_i is Hermitian (i.e., $Q_i P_i = P_i^* Q_i^*$) we conclude that $D_i^{-1} = Q_i P_i^{-*}$. The point is that we show later in §6 that Q_i and P_i can be efficiently computed with $O(\eta_i^2)$ operations by using only *strongly regular steps* (this is despite the fact that the leading minors of D_i may be singular). Assume, for the moment, that this is indeed the case. We can then rewrite X_i in the form

$$X_i = (F_i - I_{n-\alpha_i}) \begin{bmatrix} P_i^* \\ W_i Q_i \end{bmatrix} P_i^{-*} (I_{\eta_i} - \hat{F}_i)^{-1} \hat{G}_i. \tag{13}$$

We now evaluate the operation count needed in computing X_i. The term $Y_1 = (I_{\eta_i} - \hat{F}_i)^{-1} \hat{G}_i$ can be evaluated in $O(r\eta_i)$ operations (by solving r lower triangular linear systems, for instance). The product $Y_2 = P_i^{-*} Y_1$ can also be reduced to the solution of r triangular linear systems, viz., $P_i^* Y_2 = Y_1$, and thus requires $O(r\eta_i^2)$ operations. The term $Y_3 = W_i Q_i Y_2$ requires $O((\eta_i^2 + r\eta_i)(n - \alpha_{i+1}))$ operations.

Finally computing the last $(n-\alpha_{i+1})$ rows of $(F_i - I_{n-\alpha_i})[\begin{smallmatrix} Y_1 \\ Y_3 \end{smallmatrix}]$ requires $O(\eta_i(n-\alpha_{i+1}))$ operations.

It is not necessary to perform these computations in the above specified order. Other orders are possible and may be more suitable depending on the problem at hand. We may even ignore the QR factorization of D_i altogether and simply invert D_i. But we opted here for introducing the QR representation of D_i simply because, as we shall show in a later section, this factorization can be computed rather efficiently due to the Toeplitz-like structure of R and, moreover, it will also lead to an alternative convenient factorization for R itself, as shown in the next section.

But for now we note that the computational cost involved in computing G_{i+1} and L_i in the block case is $O((n-\alpha_{i+1})(\eta_i^2+\eta_i+2r\eta_i)+r\eta_i^2+r\eta_i+r\eta_i\eta_{i-1})$ operations. To get an idea of the overall computational complexity, i.e., for $i=0,1,\ldots,t-1$, we assume that the $\eta_i's$ are equal, viz., $\eta_0=\eta_1=\cdots=\eta_{t-1}=n/t$. It is then straightforward to verify that the above operation count reduces to $O(rn^2+\frac{n^3}{t})$. (In the strongly regular case we have $t=n$ and $\eta_i=1$, in which case the complexity reduces to the usual $O(rn^2)$ figure.)

4.3. An alternative triangular factorization. The factors L_i and D_i lead to a triangular factorization of the form $R=LD_B^{-1}L^*$, as discussed in §2.1, where D_B is block diagonal with entries equal to D_i and L_i is block lower triangular. We can instead use the QR factors of D_i to write an alternative factorization for R, where L is replaced by a *lower* triangular matrix $\hat{L}$, and D_B is replaced by a block diagonal matrix $\hat{D}_B$ with *unitary* and *triangular* blocks. To clarify this, we introduce the block-diagonal unitary matrix $Q=Q_0\oplus Q_1\oplus\cdots\oplus Q_{t-1}$ and the block diagonal matrix $P=P_0^{-*}\oplus P_1^{-*}\oplus\cdots\oplus P_{t-1}^{-*}$, where the diagonal blocks P_i^{-*} are lower triangular. Then $LD_B^{-1}L^*=LQQ^*D_B^{-1}QQ^*L^*$. If we define $\hat{L}=LQ$ then we obtain the alternative factorization

$$R=\hat{L}\underbrace{PQ}_{\hat{D}_B^{-1}}\hat{L}^*,$$

where $\hat{L}$ is lower triangular. In fact, the (nonzero part of the) ith block column of $\hat{L}$ has the form

$$\begin{bmatrix} P_i^* \\ W_iQ_i \end{bmatrix},$$

where P_i^* is lower triangular and the term W_iQ_i has already been computed in the generator recursion. We further remark that the inverses P_i^{-*} in P may not be needed explicitly since using the factorization $R=\hat{L}\hat{D}_B^{-1}\hat{L}^*$ to solve a linear system of equations, for example, requires knowledge of the $P_i's$ only. In summary, we get the following algorithm.

ALGORITHM 4.3 (Fast block triangular factorization). *Consider a Hermitian invertible and Toeplitz-like matrix R, viz., R satisfies $R-FRF^*=GJG^*$. A triangular factorization for R can be recursively computed in $O(rn^2+\frac{n^3}{t})$ operations as follows: start with $G_0=G, F_0=F$, and repeat for $i\geq 0$.*

1. *At step i we have F_i and G_i.*
2. *Choose the size η_i of block Schur complementation step.*
3. *If $\eta_i=1$ then update G_i to G_{i+1} using Algorithm 4.1 and compute the corresponding $l_i=\begin{bmatrix} d_i & w_i^{\mathbf{T}} \end{bmatrix}^{\mathbf{T}}$. A QR factorization for d_i can be trivially chosen as $q_i=1$ and $p_i=d_i$.*

4. *If* $\eta_i > 1$ *then compute* $L_i = \begin{bmatrix} D_i^{\mathbf{T}} & W_i^{\mathbf{T}} \end{bmatrix}^{\mathbf{T}}$ *as described in* §4.2.1 *and determine the QR factors of* D_i, *viz.,* $D_i = Q_i P_i$ *as described in* §6. *Also update* G_i *to* G_{i+1} *using Algorithm* 4.2.

5. *Construct the (nonzero parts of the) columns of* $\hat{L}$ *via*

$$\begin{bmatrix} p_i^* \\ w_i q_i \end{bmatrix}$$

or

$$\begin{bmatrix} P_i^* \\ W_i Q_i \end{bmatrix}.$$

This leads to a triangular factorization of the form $R = \hat{L}PQ\hat{L}^*$ *where* $Q = Q_0 \oplus Q_1 \oplus \cdots \oplus Q_{t-1}$ *and* $P = P_0^{-*} \oplus P_1^{-*} \oplus \cdots \oplus P_{t-1}^{-*}$.

The standard block triangular factorization, $R = LD_B^{-1}L^*$, *can also be obtained by simply ignoring the QR factorizations specified above and directly using the* L_i *and* D_i.

5. One possibility for choosing the block sizes η_i: The exact case. As mentioned earlier, the sizes of the block steps (η_i) can be determined in different ways. They may denote the smallest (exact) nonsingular minors, or the sizes of numerically well-conditioned blocks, or some other convenient choices. In this section we focus, however, on the first choice in order to highlight some theoretically interesting features that arise in the *exactly* singular case. But we hasten to add that the block factorization algorithm of the previous section is equally applicable to other choices for the η_i.

5.1. Checking for $\eta_i = 1, 2, 3$. We first remark that for a Toeplitz-like matrix R as in (3), determining whether $\eta_i = 1, 2$, or 3 in the exactly singular case is a simple task. To clarify this, recall from Theorem 3.1 that the successive Schur complements of R are also Toeplitz-like, viz., they satisfy displacement equations of the form

$$R_i - F_i R_i F_i^* = G_i J G_i^*, \tag{14}$$

where F_i is lower triangular with diagonal entries equal to $\{f_{\alpha_i}, f_{\alpha_i+1}, \ldots, f_{n-1}\}$. It thus follows that the top-left corner element of R_i is given by (where g_{i0} denotes the first row of G_i) $d_i = g_{i0} J g_{i0}^* / (1 - f_{\alpha_i} f_{\alpha_i}^*)$. If $d_i \neq 0$, or equivalently, $g_{i0} J g_{i0}^* \neq 0$, then $\eta_i = 1$. If this is not the case, then we must check for the nonsingularity of the 2×2 leading submatrix of R_i, which must be of the form

$$\begin{bmatrix} 0 & r_{01}^{(i)} \\ r_{01}^{*(i)} & r_{11}^{(i)} \end{bmatrix}.$$

Using (14) it is easy to verify that $r_{01}^{(i)} = g_{i0} J g_{i1}^* / (1 - f_{\alpha_i} f_{\alpha_i+1}^*)$, which implies that $\eta_i = 2$ if, and only if, $g_{i0} J g_{i0}^* = 0$ and $g_{i0} J g_{i1}^* \neq 0$. If this test fails then we proceed to check for the leading 3×3 submatrix of R_i, viz.,

$$\begin{bmatrix} 0 & 0 & r_{02}^{(i)} \\ 0 & r_{11}^{(i)} & r_{12}^{(i)} \\ r_{02}^{*(i)} & r_{12}^{*(i)} & r_{22}^{(i)} \end{bmatrix}, \tag{15}$$

where, using (14) again, $r_{02}^{(i)} = g_{i0}Jg_{i2}^*/(1-f_{\alpha_i}f_{\alpha_i+2}^*)$, $r_{11}^{(i)} = g_{i1}Jg_{i1}^*/(1-f_{\alpha_i+1}f_{\alpha_i+1}^*)$, $r_{12}^{(i)} = g_{i1}Jg_{i2}^*/(1-f_{\alpha_i+1}f_{\alpha_i+2}^*)$, and $r_{22}^{(i)} = g_{i2}Jg_{i2}^*/(1-f_{\alpha_i+2}f_{\alpha_i+2}^*)$. Hence, for $\eta_i = 3$ we need $g_{i0}Jg_{i2}^* \neq 0$ and $g_{i1}Jg_{i1}^* \neq 0$ In summary we have the following lemma.

LEMMA 5.1. *The following are simple tests for $\eta_i = 1, 2,$ or 3 in the exactly singular case:*

If $g_{i0}Jg_{i0}^* \neq 0$ then $\eta_i = 1$
else if $g_{i0}Jg_{i1}^* \neq 0$ then $\eta_i = 2$
else if $g_{i0}Jg_{i2}^* \neq 0$ and $g_{i1}Jg_{i1}^* \neq 0$ then $\eta_i = 3$
else $\eta_i \geq 4$.

Observe that for $\eta_i \leq 3$ the leading nonsingular submatrix of R_i has a reversed lower triangular form. The inversion or QR factorization of these submatrices can be easily evaluated. For example, the QR decomposition of (15) is

$$D_i = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} r_{02}^{*(i)} & r_{12}^{*(i)} & r_{22}^{(i)} \\ 0 & r_{11}^{(i)} & r_{12}^{(i)} \\ 0 & 0 & r_{02}^{(i)} \end{bmatrix}.$$

Moreover, $\hat{G}_i J \hat{G}_i^*$ also has the same reversed lower triangular form for $\eta_i = 1, 2, 3$ ($\hat{G}_i$ being the first η_i rows of G_i). For example, the conditions for $\eta_i = 3$ mean that $\hat{G}_i J \hat{G}_i^*$ has to be of the form

$$\hat{G}_i J \hat{G}_i^* = \begin{bmatrix} 0 & 0 & x \\ 0 & x & x \\ x & x & x \end{bmatrix}.$$

The above discussion suggests the following result.

LEMMA 5.2. *For some k, the leading $k \times k$ submatrix of $G_i J G_i^*$ has a nonsingular reversed lower triangular form with antidiagonal entries $\{m_{0,k-1}, m_{1,k-2}, \ldots, m_{k-1,0}\}$, if and only if $\eta_i = k$ and the leading nonsingular submatrix D_i has the same reversed lower triangular form.*

Proof. The claim is certainly sufficient and necessary for $k = 1, 2, 3$, as discussed prior to the statement of the lemma. To verify the claim for larger values of k we consider a general $k \times k$ matrix E in reversed lower triangular form with antidiagonal entries $\{e_{0,k-1}, e_{1,k-2}, \ldots, e_{k-1,0}\}$, and let $\hat{F}_i$ denote the leading $k \times k$ submatrix of F_i. It is then easy to check that we can find a matrix E of this form that solves the equation

$$E - \hat{F}_i E \hat{F}_i^* = \begin{bmatrix} \mathrm{O} & & m_{0,k-1} \\ & \cdot & \\ m_{k-1,0} & & \mathrm{X} \end{bmatrix}.$$

In fact, we can write down explicit formulas for the desired entries of E in terms of the known entries on the right-hand side of the above equality. For example, the diagonal entries of E are given by

$$e_{0,k-1} = \frac{m_{0,k-1}}{1 - f_{\alpha_i} f_{\alpha_i+k-1}^*}, \qquad e_{1,k-2} = \frac{m_{1,k-2}}{1 - f_{\alpha_i+1} f_{\alpha_i+k-2}^*}, \ldots,$$

which shows that we can always find an invertible solution E. But the leading $k \times k$ minor of R_i satisfies the same equation as E. It follows from the uniqueness condition

$(1 - f_i f_j^* \neq 0$, for all $i, j)$ that we must have $D_i = E$. Conversely, assume that D_i has the suggested reversed lower triangular form, then it readily follows that $D_i - \hat{F}_i D_i \hat{F}_i^*$ is nonsingular and has the same reversed lower triangular form. □

We should stress that the lemma does *not* state that the nonsingular submatrices D_i always have a reversed lower triangular form. It only states that if D_i happens to have this form then $\hat{G}_i J \hat{G}_i^*$ also has the same form (and vice versa). In fact, the triangular structure of D_i is not necessarily valid for higher sizes η_i as can be easily checked. For example, a nonsingular 4×4 leading submatrix of R_i may have one of the following forms:

$$\begin{bmatrix} 0 & 0 & 0 & x \\ 0 & x & x & x \\ 0 & x & x & x \\ x & x & x & x \end{bmatrix}, \quad \begin{bmatrix} 0 & 0 & 0 & x \\ 0 & 0 & x & x \\ 0 & x & x & x \\ x & x & x & x \end{bmatrix}, \quad \text{or} \quad \begin{bmatrix} 0 & 0 & x & x \\ 0 & 0 & x & x \\ x & x & x & x \\ x & x & x & x \end{bmatrix}.$$

We can, however, give a stronger statement in the important special case of displacement rank $r = 2$.

5.2. Displacement rank r = 2. We now consider the special case of structured matrices R as in (3) but with displacement rank $r = 2$, i.e., G has two columns. We further assume that $J = (1 \oplus -1)$ and that F is a stable matrix, or equivalently, that its diagonal entries have less than unit-modulus magnitude,

$$R - FRF^* = \begin{bmatrix} \mathbf{u}_0 & \mathbf{v}_0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \mathbf{u}_0 & \mathbf{v}_0 \end{bmatrix}^*. \tag{16}$$

Our purpose is to show that for this class of structured matrices we can derive an explicit test for all $\eta_i's$ in the exactly singular case. Special cases of (16) were studied earlier in [16], [27]. Iohvidov [16] studied the special case of Toeplitz matrices, which corresponds to the special choice $F = Z$, and a special generator matrix of the form (recall expression (2))

$$G = \begin{bmatrix} 1 & c_1 & \cdots & c_{n-1} \\ 0 & c_1 & \cdots & c_{n-1} \end{bmatrix}^{\mathbf{T}}.$$

Pal and Kailath [26], [27] considered the wider class of so-called quasi-Toeplitz matrices, which still corresponds to $F = Z$, but one where the columns $\mathbf{u}_0$ and $\mathbf{v}_0$ of G are arbitrary and not as restricted as in the Toeplitz case above. Such matrices can be shown to be congruent to Toeplitz matrices in a certain sense, hence the name quasi-Toeplitz. The derivation in [26], [27] exploits this fact and, among other results, shows that the determination of the size of the smallest nonsingular minor is reduced to counting the number of repeated zeros at the origin of a certain polynomial.

We provide here a general statement that goes beyond the $F = Z$ case. We follow a matrix-based argument that also reveals under what conditions on F the derived test is not applicable. (See also [2] for generalizations of the Iohvidov laws using the theory of reproducing kernel Hilbert spaces.)

We start again with the displacement equation of the ith Schur complement, viz.,

$$R_i - F_i R_i F_i^* = G_i J G_i^*, \tag{17}$$

and denote the entries of the now two-column generator G_i by

$$G_i = \begin{bmatrix} u_{ii} & u_{i+1,i} & u_{i+2,i} & \cdots \\ v_{ii} & v_{i+1,i} & v_{i+2,i} & \cdots \end{bmatrix}^{\mathbf{T}} = \begin{bmatrix} \mathbf{u}_i & \mathbf{v}_i \end{bmatrix}.$$

Assume we encounter a singularity $d_i = 0$, or equivalently, $g_i J g_i^* = |u_{ii}|^2 - |v_{ii}|^2 = 0$. Then either of the following two cases could have happened: g_i is a zero row, which corresponds to the trivial case discussed at the end of §4.1, or g_i is a nonzero row, which corresponds to a *block step* that we now discuss in more detail.

5.3. A preliminary result and definitions. Before proving the main theorem we first state an easily verifiable result that follows from the following type of argument: an equality such as $g_i J g_i^* = 0$ clearly implies that v_{ii} and u_{ii} are related via $v_{ii} = u_{ii} e^{j\xi}$ for some phase angle $\xi \in [0, 2\pi]$. More generally, we have the following lemma.

LEMMA 5.3. *The entries of the first k rows of G_i satisfy*

$$v_{l+i,i} = u_{l+i,i} e^{j\xi}, \qquad l = 0, 1, \ldots, k-1,$$

for some phase angle $\xi \in [0, 2\pi]$ if and only if the leading $2k \times 2k$ submatrix of $G_i J G_i^*$ has the form

$$\begin{bmatrix} \mathbf{0}_{k\times k} & M_{k\times k} \\ M^*_{k\times k} & X \end{bmatrix}, \tag{18}$$

where M is a rank 1 matrix and X denotes irrelevant entries. That is, $G_i J G_i^*$ has a $k \times k$ leading zero block.

For a column vector $\mathbf{x}$ and a square matrix A, we let $K^m(A, \mathbf{x})$ denote the Krylov matrix $K^m(A, \mathbf{x}) = \begin{bmatrix} \mathbf{x} & A\mathbf{x} & \ldots & A^{m-1}\mathbf{x} \end{bmatrix}$. We further define some auxiliary quantities that will be used in the statement and proof of the next theorem: for a positive number k, we define the column vectors $\{\mathbf{a}, \mathbf{b}, \mathbf{x}, \mathbf{y}\}$ as follows:

(19a)

$$\begin{bmatrix} \mathbf{a} & \mathbf{b} \end{bmatrix} = \begin{bmatrix} u_{ii} & v_{ii} \\ u_{i+1,i} & v_{i+1,i} \\ \vdots & \vdots \\ u_{i+k-1,i} & v_{i+k-1,i} \end{bmatrix}, \quad \begin{bmatrix} \mathbf{x} & \mathbf{y} \end{bmatrix} = \begin{bmatrix} u_{i+k,i} & v_{i+k,i} \\ u_{i+k+1,i} & v_{i+k+1,i} \\ \vdots & \vdots \\ u_{i+2k-1,i} & v_{i+2k-1,i} \end{bmatrix}.$$

That is, $\{\mathbf{a}, \mathbf{b}\}$ contain the entries of the first k rows of G_i, while $\{\mathbf{x}, \mathbf{y}\}$ contain the entries of the next k rows of G_i. Recall that g_i is a nonzero row vector with zero $J-$norm. Consequently, both u_{ii} and v_{ii} must be nonzero since if one of them is zero then the other one must be zero, due to the relation $v_{ii} = u_{ii} e^{j\xi}$. We also define the column vectors

$$\rho = \frac{\mathbf{a} + e^{-j\xi}\mathbf{b}}{\sqrt{2}}, \qquad \nu = \frac{\mathbf{x} - e^{-j\xi}\mathbf{y}}{\sqrt{2}} \quad \text{for a given } \xi, \tag{19b}$$

and partition F_i as follows

$$F_i = \begin{bmatrix} \hat{F}_i & & \mathrm{O} \\ ? & \hat{A}_i & \\ ? & ? & ? \end{bmatrix}, \tag{19c}$$

where $\hat{F}_i$ and $\hat{A}_i$ are $k \times k$ lower triangular matrices.

5.4. Main result for displacement rank r = 2. The next result gives an explicit test for the determination of the sizes of the nonsingular minors for the class of structured matrices as in (16), with extra conditions on the entries of F. This extends earlier results in [16], [27].

THEOREM 5.4. *The size of the smallest nonsingular leading submatrix of R_i is $2k$ and has the block form*

$$\begin{bmatrix} \mathbf{0}_{k\times k} & N_{k\times k} \\ N^*_{k\times k} & C_{k\times k} \end{bmatrix}, \tag{20a}$$

where N is invertible if and only if the $k\times k$ matrix $K^{\infty}(\hat{F}_i,\rho)K^{*\infty}(\hat{A}_i,\nu)$ is invertible and the entries of the first k rows of G_i satisfy

$$v_{l+i,i} = u_{l+i,i}e^{j\xi}, \qquad l = 0,1,\ldots,k-1, \tag{20b}$$

for some phase angle $\xi \in [0,2\pi]$.

Proof. If $u_{l+i,i}$ and $v_{l+i,i}$ satisfy (20b) then it is straightforward to verify that the leading $2k \times 2k$ submatrix of R_i has a leading zero block as in (20a) (similar to the argument in Lemma 5.3). The converse is also true. If the leading $2k \times 2k$ submatrix of R_i has a leading zero block as in (20a) then $u_{l+i,i}$ and $v_{l+i,i}$ satisfy (20b). We still need to prove that (20a) is the smallest nonsingular minor. For this purpose, it is enough to verify that N is invertible.

It follows from (17) that N satisfies the (non-Hermitian) displacement equation

$$N - \hat{F}_i N \hat{A}_i^* = \begin{bmatrix} \mathbf{a} & \mathbf{b} \end{bmatrix} J \begin{bmatrix} \mathbf{x} & \mathbf{y} \end{bmatrix}^*,$$

where $\{\mathbf{a},\mathbf{b},\mathbf{x},\mathbf{y}\}$ were defined in (19a). But conditions (20b) imply that $\mathbf{b} = e^{j\xi}\mathbf{a}$. Also, the eigenvalues of the lower triangular matrices $\hat{F}_i$ and $\hat{A}_i$ are strictly less than unit magnitude. Hence, we can write

$$\begin{aligned} N &= K^{\infty}(\hat{F}_i,\mathbf{a})K^{*\infty}(\hat{A}_i,\mathbf{x}) - K^{\infty}(\hat{F}_i,\mathbf{b})K^{*\infty}(\hat{A}_i,\mathbf{y}) \\ &= \frac{1}{2}\Big\{K^{\infty}(\hat{F}_i,\mathbf{a}+e^{-j\xi}\mathbf{b})K^{*\infty}(\hat{A}_i,\mathbf{x}-e^{-j\xi}\mathbf{y}) \\ &\quad + K^{\infty}(\hat{F}_i,\mathbf{a}-e^{-j\xi}\mathbf{b})K^{*\infty}(\hat{A}_i,\mathbf{x}+e^{-j\xi}\mathbf{y})\Big\} \\ &= K^{\infty}(\hat{F}_i,\rho)K^{*\infty}(\hat{A}_i,\nu), \end{aligned}$$

where ρ and ν were defined in (19b). We thus conclude that N is full rank. □

5.4.1. Remarks. The last theorem states that, provided the following condition is satisfied,

$$K^{\infty}(\hat{F}_i,\rho)K^{*\infty}(\hat{A}_i,\nu) \text{ is invertible,} \tag{21}$$

the determination of η_i reduces to checking the proportionality condition (20b), viz., whether the first k elements of $\mathbf{v}_i$ are unit-modulus multiples of the first k elements of $\mathbf{u}_i$. It is clear that *necessary* conditions for (21) to hold are

$$K^{\infty}(\hat{F}_i,\rho) \text{ and } K^{*\infty}(\hat{A}_i,\nu) \text{ must have full rank.}$$

For those familiar with system theory [17], the above necessary conditions are equivalent to saying that the pair $(\hat{F}_i,\rho)$ must be controllable and the pair $(\hat{A}_i^*,\nu^*)$

must be observable. For example, if $\hat{F}_i$ is similar to a Jordan structure with repeated Jordan blocks for the same eigenvalue then the pair $(\hat{F}_i, \rho)$ will not be controllable. A similar remark holds for $\hat{A}_i$.

Furthermore, condition (21) is automatically satisfied in the special case $F = Z$ studied in [16], [27]. Indeed, $F = Z$ implies that $K^{\infty}(\hat{F}_i, \rho) = K^{\infty}(Z, \rho) = \begin{bmatrix} \mathbf{L}(\rho) & \mathbf{0} \end{bmatrix}$ and $K^{\infty}(\hat{A}_i, \nu) = K^{\infty}(Z, \nu) = \begin{bmatrix} \mathbf{L}(\nu) & \mathbf{0} \end{bmatrix}$, where the notation $\mathbf{L}(x)$ denotes a lower triangular Toeplitz matrix whose first column is x. But $\mathbf{L}(\rho)$ and $\mathbf{L}(\nu)$ are full rank matrices since the top entries of ρ and ν are nonzero. Hence, $\mathbf{L}(\rho)\mathbf{L}^*(\nu)$ is invertible and (21) is satisfied. It also follows that N is strongly regular.

Moreover, using (20a) we get

$$D_i^{-1} = \begin{bmatrix} -N^{-*}CN^{-1} & N^{-*} \\ N^{-1} & \mathbf{0} \end{bmatrix},$$

which shows that inverting D_i essentially reduces to inverting a strongly regular matrix N, which has a non-Hermitian Toeplitz-like structure. This can be done in strongly regular (i.e., scalar) steps. Following this reasoning we can show that in this case ($F = Z$), the inversion of D_i (or N) and the generator recursion (12a) reduce to the algorithm derived in [27], which involves only scalar operations. We do not go into the details here mainly because the derivation (and simplifications thereof) relies heavily on the special structure in question ($r = 2$ and $F = Z$). We focus instead on the case of higher displacement ranks ($r > 2$).

5.5. A recursive test for displacement ranks r >2. A conventional rank test for determining whether an arbitrary $n \times n$ matrix is invertible or not requires $O(n^3)$ operations. This figure can be reduced to $O(rn^2)$ in the case of structured matrices as discussed in §5.5.1. The following lemma states that if we are given a structured matrix R (*not necessarily strongly regular*), then checking whether R is invertible or not can be achieved by using only *strongly regular* Schur steps that are applied to an appropriately defined extended generator matrix.

LEMMA 5.5. *Let T be any $n \times n$ positive-definite matrix. Then an $n \times n$ Hermitian matrix R (not necessarily strongly regular) is invertible if and only if the extended $2n \times 2n$ matrix $\hat{R}$,*

$$\hat{R} = \begin{bmatrix} -T & R \\ R & \mathbf{0} \end{bmatrix},$$

is strongly regular.

Proof. The leading $n \times n$ submatrix of $\hat{R}$ is strongly regular since T is positive definite ($T > 0$). The Schur complement with respect to the leading $n \times n$ block is $RT^{-1}R$. The claim now follows by observing that $RT^{-1}R$ is positive-definite if and only if R is invertible. □

In other words, if we apply the generalized Schur algorithm to a generator of $\hat{R}$ and a singularity is (not) encountered then we conclude that the original R is (not) singular. But we first need to check whether the extended matrix $\hat{R}$ is structured. For this purpose, recall that R is Toeplitz-like, viz., $R - FRF^* = GJG^*$. It then follows that

$$\hat{R} - \begin{bmatrix} F & \mathbf{0} \\ \mathbf{0} & F \end{bmatrix} \hat{R} \begin{bmatrix} F & \mathbf{0} \\ \mathbf{0} & F \end{bmatrix}^* = \begin{bmatrix} FTF^* - T & GJG^* \\ GJG^* & \mathbf{0} \end{bmatrix}, \tag{22}$$

which shows that $\hat{R}$ has a Toeplitz-like structure if $(FTF^* - T)$ has low rank, say β. So assume that this is the case. Then we can (nonuniquely) factor $(FTF^* - T)$ as

follows: $FTF^* - T = VJ_\beta V^*$, where V is an $n \times \beta$ generator matrix and J_β is a $\beta \times \beta$ signature matrix with $\beta \ll n$. This means that we need to choose a positive-definite matrix T that has low displacement rank with respect to F. We show later in this section how such choices (of T and, consequently, of V and J_β) can be made. Then we can factor the right-hand side of (22) as follows:

$$\begin{bmatrix} FTF^* - T & GJG^* \\ GJG^* & \mathbf{0} \end{bmatrix} = \begin{bmatrix} V & \mathbf{0} & G \\ \mathbf{0} & G & \mathbf{0} \end{bmatrix} \mathcal{J} \begin{bmatrix} V & \mathbf{0} & G \\ \mathbf{0} & G & \mathbf{0} \end{bmatrix}^*, \quad \mathcal{J} = \begin{bmatrix} J_\beta & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & J \\ \mathbf{0} & J & \mathbf{0} \end{bmatrix},$$

where $\mathcal{J}$ satisfies $\mathcal{J}^2 = I$. We thus conclude that a possible (not necessarily minimal) $2n \times (2r + \beta)$ generator for $\hat{R}$ is

$$\hat{H} = \begin{bmatrix} V & \mathbf{0} & G \\ \mathbf{0} & G & \mathbf{0} \end{bmatrix}.$$

We can now proceed by applying Algorithm 4.1 to $\hat{R}$ with the initial conditions $G_0 = \hat{H}$, $F_0 = (F \oplus F)$, and $J = \mathcal{J}$. The first n steps of the algorithm will clearly yield negative diagonal entries $\{d_i,\ i = 0, 1, \ldots, n-1, d_i < 0\}$ since $-T$ is negative definite. The nth generator, G_n, will be a generator of the Schur complement $RT^{-1}R$ of $\hat{R}$ with respect to its leading $n \times n$ submatrix $(-T)$. If in the subsequent generator steps $(i = n, n+1, \ldots, 2n-1)$ we obtain a zero d_i, (i.e., a row vector g_i with a zero $\mathcal{J}$–norm), then we stop and conclude that the original matrix R is singular. Otherwise, R is nonsingular. This test requires at most $O((2r + \beta)n^2)$ operations (which is the computational effort due to applying the strongly regular Schur algorithm to $\hat{R}$). This should be compared with a conventional $O(n^3)$ rank test applied to R. A computational advantage results when $(2r + \beta) \ll n$.

5.5.1. Specializing to the η_i's. We now show how to recursively use the above procedure to compute the $\eta_i's$. Recall that the successive Schur complements R_i of the Toeplitz-like matrix R satisfy displacement equations of the form (14), and our objective is to determine the size η_i of the smallest nonsingular submatrix of R_i. We already know how to check whether $\eta_i \leq 3$ (as discussed in §5.1). For higher values of η_i we can proceed as suggested by the result of Lemma 5.5.

For this purpose, assume we have already chosen a positive-definite matrix T_i that has low displacement rank with respect to F_i (as described ahead), and introduce the factorization

$$F_i T_i F_i^* - T_i = V_i J_\beta V_i^*.$$

We further define E_k, T_k, $\hat{F}_k$, $\hat{G}_k$, and V_k to denote the leading $k \times k$, $k \times k$, $k \times k$, $k \times r$, and $k \times \beta$ submatrices of R_i, T_i, F_i, G_i, and V_i, respectively. It follows from (14) that E_k is also a Toeplitz-like matrix since $E_k - \hat{F}_k E_k \hat{F}_k^* = \hat{G}_k J \hat{G}_k^*$. We can now use the result of Lemma 5.5 to check whether E_k is nonsingular by forming the corresponding extended matrix $\hat{E}_k$,

$$\hat{E}_k = \begin{bmatrix} -T_k & E_k \\ E_k & \mathbf{0} \end{bmatrix},$$

and checking for its strong regularity. If E_k turns out to be invertible, then we set $\eta_i = k$, otherwise we check for the next submatrix E_{k+1}, and so on. A generator for

$\hat{E}_k$ is given by

$$\hat{H}_k = \begin{bmatrix} V_k & 0 & \hat{G}_k \\ 0 & \hat{G}_k & 0 \end{bmatrix}, \qquad \mathcal{J} = \begin{bmatrix} J_\beta & 0 & 0 \\ 0 & 0 & J \\ 0 & J & 0 \end{bmatrix}, \tag{23a}$$

and we thus apply the generator recursion of Algorithm 4.1 with the initial conditions $G_0 = \hat{H}_k, F_0 = (\hat{F}_k \oplus \hat{F}_k)$, $J = \mathcal{J}$. More precisely, we can rewrite recursion (11) for the present case as follows: start with $\hat{H}_{k,0} = \hat{H}_k$ and repeat for $i = 0, 1, \ldots, 2k-1$,

$$\begin{bmatrix} 0_{1\times(2r+\beta)} \\ \\ \hat{H}_{k,i+1} \end{bmatrix} = \hat{H}_{k,i}\Theta_i \begin{bmatrix} I_j & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I \end{bmatrix} + \Phi_i \hat{H}_{k,i}\Theta_i \begin{bmatrix} 0_j & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{23b}$$

where Φ_i is as defined in Algorithm 4.1 with $F_0 = (\hat{F}_k \oplus \hat{F}_k)$, and Θ_i is a $\mathcal{J}-$unitary rotation that reduces the first row of $\hat{H}_{k,i}$ (denoted by $h_{k,i}$) to the form

$$h_{k,i}\Theta_i = \begin{bmatrix} 0 & \kappa_{k,j}^{(i)} & 0 \end{bmatrix},$$

where $\kappa_{k,j}^{(i)}$ is a scalar at a convenient jth column position.

The test starts by applying the above recursion to $\hat{H}_k$. Schematically, we form (Θ_0, Φ_0) and apply the recursion to obtain $\hat{H}_{k,1}$. We then form (Θ_1, Φ_1) and apply the recursion again to obtain $\hat{H}_{k,2}$, and so on. Each such step corresponds to a transformation determined by the pair (Θ_i, Φ_i). After the first k transformations ($i = 0, 1, \ldots, k-1$), we obtain $\hat{H}_{k,k}$, which is a generator for $E_k T_k^{-1} E_k$. We then proceed by applying (at most) k more steps of the recursion. E_k will then be declared singular if, at any of the steps $i = k, \ldots, 2k-1$ we encounter a row $h_{k,i}$ with a zero $\mathcal{J}-$norm, viz., $h_{k,i}\mathcal{J}h_{k,i}^* = |\kappa_{k,j}^{(i)}|^2 = 0$, for some $i \geq k$.

If the procedure ends without encountering a singularity then $\eta_i = k$, otherwise we must check for the next leading submatrix E_{k+1}. Now, the generators of $\hat{E}_{k+1}$ and $\hat{E}_k$ are closely related since V_k and $\hat{G}_k$ are submatrices of V_{k+1} and $\hat{G}_{k+1}$, respectively. That is,

$$V_{k+1} = \begin{bmatrix} V_k \\ b_k \end{bmatrix}, \qquad \hat{G}_{k+1} = \begin{bmatrix} \hat{G}_k \\ a_k \end{bmatrix},$$

for some row vectors a_k and b_k. Hence, $\hat{H}_{k+1}$ and $\hat{H}_k$ differ only at rows $(k+1)$ and $2(k+1)$, viz.,

$$\hat{H}_{k+1} = \begin{bmatrix} V_k & 0 & \hat{G}_k \\ b_k & 0 & a_k \\ 0 & \hat{G}_k & 0 \\ 0 & a_k & 0 \end{bmatrix} = \begin{bmatrix} V_{k+1} & 0 & \hat{G}_{k+1} \\ 0 & \hat{G}_{k+1} & 0 \end{bmatrix}.$$

Therefore, $\hat{H}_k$ and $\hat{H}_{k+1}$ share the same first k Schur reduction steps. This means that in order to obtain a generator for $E_{k+1}T_{k+1}^{-1}E_{k+1}$, we first apply the *last* $(k+2)$ rows of $\hat{H}_{k+1}$, viz.,

$$\begin{bmatrix} b_k & 0 & a_k \\ 0 & \hat{G}_k & 0 \\ 0 & a_k & 0 \end{bmatrix},$$

through the first k transformations $\{(\Theta_i, \Phi_i),\ i = 0, \ldots, k-1\}$ that were applied to $\hat{H}_k$. This leads to $\hat{H}_{k+1,k}$. We now apply one more transformation (Θ_k, Φ_k) to $\hat{H}_{k+1,k}$ in order to get $\hat{H}_{k+1,k+1}$, which is a generator for $E_{k+1}T_{k+1}^{-1}E_{k+1}$. We then proceed by applying at most $(k+1)$ steps in order to check for the positive-definiteness of $E_{k+1}T_{k+1}^{-1}E_{k+1}$, and so on. The size η_i is determined when, for some k, we are able to complete the whole recursive procedure without encountering a singularity. In this case, we get $k = \eta_i$ and hence $E_k = E_{\eta_i} = D_i$. The η_i transformations $\{\Theta_i, \Phi_i,\ i = 0, \ldots, \eta_i - 1\}$ used in this last test will be relevant in §6 while computing the QR factors of D_i.

It can be verified that $O(k^2(r+\beta))$ operations are needed for each k. This should be compared with the following alternative procedure: For each k, compute the leading $k \times k$ submatrix and check whether it is singular using a conventional rank test. This requires $O(k^3)$ operations and does not exploit the underlying (displacement) structure. A computational advantage results when $(r+\beta)$ is smaller than k.

ALGORITHM 5.6. *To check whether the $k \times k$ leading submatrix of R_i is nonsingular we proceed as follows.*

1. *Form a generator pair $(\hat{H}_k, \mathcal{J})$ as in* (23a).

2. *Apply k steps of recursion* (23b) *starting with $\hat{H}_{k,0} = \hat{H}_k, F_0 = (\hat{F}_k \oplus \hat{F}_k)$, and $J = \mathcal{J}$. This leads to $\hat{H}_{k,k}$.*

3. *Apply more steps of recursion* (23b) *to $\hat{H}_{k,k}$. If $h_{k,j}$ is found to have zero $\mathcal{J}-$norm, for some $k \le j \le 2k-1$, then E_k is declared singular $(\eta_i > k)$. Otherwise $\eta_i = k$.*

4. *To check for the higher order $(k+1) \times (k+1)$ submatrix we essentially repeat the same procedure, except that we exploit the fact that $\hat{H}_k$ and $\hat{H}_{k+1}$ differ only in two rows as follows:*

   a. *Apply the last $(k+2)$ rows of $\hat{H}_{k+1}$ through the k transformations $\{(\Theta_i, \Phi_i),\ i = 0, \ldots, k-1\}$ that were applied to $\hat{H}_k$. This leads to $\hat{H}_{k+1,k}$.*

   b. *Apply one more step to get $\hat{H}_{k+1,k+1}$.*

   c. *Go back to step* 3 *and repeat.*

5.5.2. Choosing T. We now show how to choose a positive-definite matrix T that has low displacement rank with respect to an F. This choice is rather trivial in some special (though frequent) cases such as $F = Z$ or $F = Z \oplus Z \oplus \cdots \oplus Z$. For these cases, a simple choice is $T = I$. For example, choosing $T = I$ in the $F = Z$ case leads to $\beta = 1$, $J_\beta = -1$, and $V = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}^{\mathbf{T}}$, viz.,

$$ZZ^* - I = -\begin{bmatrix} 1 \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} 1 \\ \mathbf{0} \end{bmatrix}^*.$$

On the other hand, for a diagonal or bidiagonal matrix F with distinct diagonal entries, the choice $T = I$ would usually lead to a full displacement rank $\beta = n$, i.e., $FF^* - I$ would generally have rank n, which substantially increases the computational cost of the recursive tests. However, for such cases, it is still possible to choose a positive-definite matrix T that leads to a low displacement rank β. For this purpose, we exploit connections with analytic interpolation theory.

Assume, for instance, that we have an $n \times n$ diagonal matrix F with distinct and stable entries f_i ($|f_i| < 1$), and choose *any* scalar function $s(z)$ that is analytic and strictly bounded by unity inside the open unit disc $|z| < 1$, viz., $\sup_{|z|<1} |s(z)| < 1$. We say that $s(z)$ is a Schur-type function [1], [36]. We further introduce the matrices

V and J_β given by

$$V = \begin{bmatrix} 1 & s(f_0) \\ 1 & s(f_1) \\ \vdots & \vdots \\ 1 & s(f_{n-1}) \end{bmatrix}, \qquad J_\beta = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

and define T to be the solution of the displacement equation

$$T - FTF^* = V \begin{bmatrix} 1 & \\ & -1 \end{bmatrix} V^*.$$

It is then a standard result in analytic interpolation theory (see, e.g., [1], [32], [34]) that T is a positive-definite matrix since $s(z)$ is of Schur-type. So all we need to do is to choose a Schur function $s(z)$ and define V and J_β as above. We do not even need to explicitly determine the corresponding T since the recursive algorithm described in the previous section uses (V, J_β) and not T.

For a bidiagonal matrix $F = Z + \text{diag.}\{f_0, \ldots, f_{n-1}\}$ with distinct stable entries f_i ($|f_i| < 1$), we again choose a Schur function $s(z)$ and define

$$V = \begin{bmatrix} 1 & \phi_0 \\ 0 & \phi_1 \\ \vdots & \vdots \\ 0 & \phi_{n-1} \end{bmatrix}, \qquad J = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

where the $\phi_i's$ denote the first n Newton-series coefficients associated with $s(z)$. These coefficients can be recursively determined via the so-called *divided difference recursion* as follows: start with $s_0(z) = s(z)$ and then use

$$s_i(z) = \frac{s_{i-1}(z) - \phi_{i-1}}{z - f_{i-1}}, \qquad \phi_i = s_i(f_i).$$

It also follows that the associated matrix T is positive-definite [31], [34]. For a more general matrix F with $r_i \times r_i$ Jordan blocks, viz., $F = (Z + f_0 I) \oplus (Z + f_1 I) \oplus (Z + f_2 I) \oplus \ldots$, with f_i distinct and $|f_i| < 1$, we define [31], [34]

$$V = \begin{bmatrix} 1 & s(f_0) \\ 0 & s^{(1)}(f_0) \\ \vdots & \vdots \\ 0 & \frac{1}{(r_0-1)!} s^{(r_0-1)}(f_0) \\ 1 & s(f_1) \\ \vdots & \vdots \\ 0 & \frac{1}{(r_1-1)!} s^{(r_1-1)}(f_1) \\ \vdots & \vdots \end{bmatrix}, \qquad J = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

where $s^{(j)}(f_i)$ denotes the jth derivative of $s(z)$ at f_i.

6. QR factorization of the D_i's. Once the sizes η_i have been chosen, say as described in the above sections for the exactly singular case or as numerically well-conditioned blocks, we still need to show how to compute the QR factors of D_i, viz., $D_i = Q_i P_i$, where Q_i is an $\eta_i \times \eta_i$ unitary matrix and P_i is an $\eta_i \times \eta_i$ nonsingular upper triangular matrix. This is useful if the alternative triangular factorization of §4.3 is desired. The discussion that follows assumes, for brevity of argument and notation, that the η_i have been chosen as described in the above section. But it is rather immediate to see that the result is equally applicable for other choices of the η_i. The main point is simply the following: to compute the QR factors of D_i we form a $3\eta_i \times 3\eta_i$ extended block matrix and apply $2\eta_i$ steps of the (strongly regular) Schur algorithm to it. Once this is done, the QR factors can be read out from the resulting triangular factors.

So we first assume that F is such that the matrix $T = I$ has low displacement rank with respect to it. We then consider the $3\eta_i \times 3\eta_i$ extended matrix

$$\hat{D}_i = \begin{bmatrix} -I & D_i & \mathbf{0} \\ D_i & \mathbf{0} & D_i \\ \mathbf{0} & D_i & \mathbf{0} \end{bmatrix},$$

which also turns out to be Toeplitz-like with respect to $(\hat{F}_i \oplus \hat{F}_i \oplus \hat{F}_i)$ and with a generator matrix of the form

$$\text{(24a)} \qquad \begin{bmatrix} V_{\eta_i} & \mathbf{0} & \hat{G}_i \\ \mathbf{0} & \hat{G}_i & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \hat{G}_i \end{bmatrix}, \qquad \mathcal{J} = \begin{bmatrix} J_\beta & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & J \\ \mathbf{0} & J & \mathbf{0} \end{bmatrix},$$

where $\hat{F}_i \hat{F}_i^* - I = V_{\eta_i} J_\beta V_{\eta_i}^*$. The first two block rows of the above generator are the same block rows of the generator $\hat{H}_{\eta_i}$ of $\hat{E}_{\eta_i}$ (refer to (23a)), viz.,

$$\hat{H}_{\eta_i} = \begin{bmatrix} V_{\eta_i} & \mathbf{0} & \hat{G}_i \\ \mathbf{0} & \hat{G}_i & \mathbf{0} \end{bmatrix}.$$

Therefore, if we apply to the generator (24a) of $\hat{D}_i$ the same η_i transformations $\{(\Theta_i, \Phi_i),\ i = 0, 1, \ldots, \eta_i - 1\}$ that were applied to $\hat{H}_{\eta_i}$, we then obtain a generator matrix for the Schur complement of the leading block matrix in $\hat{D}_i$, which is equal to $\bar{D}_i$ below:

$$\text{(24b)} \qquad \bar{D}_i = \begin{bmatrix} D_i D_i & D_i \\ D_i & \mathbf{0} \end{bmatrix}.$$

If we denote this generator of $\bar{D}_i$ by $\bar{S}_i$ then $\bar{S}_i$ is clearly of the form

$$\text{(24c)} \qquad \bar{S}_i = \begin{bmatrix} \hat{H}_{\eta_i, \eta_i} \\ \tilde{S}_i \end{bmatrix},$$

where $\tilde{S}_i$ results from the application of the above η_i transformations $\{\Theta_i, \Phi_i\}$ to the last block row in the generator of $\hat{D}_i$, viz., $\begin{bmatrix} \mathbf{0} & \mathbf{0} & \hat{G}_i \end{bmatrix}$. In summary, we already know how to obtain a generator for (the $2\eta_i \times 2\eta_i$ matrix) $\bar{D}_i$ in (24b): just update the block row $\begin{bmatrix} \mathbf{0} & \mathbf{0} & \hat{G}_i \end{bmatrix}$ via the transformations (Θ_i, Φ_i) and construct $\bar{S}_i$.

Once a generator for $\bar{D}_i$ is available, we can then use it to determine the first η_i triangular factors of $\bar{D}_i$. For this purpose, we need only apply η_i steps of the

strongly regular Algorithm 4.1 starting with $G_0 = \bar{S}_i, F_0 = (\hat{F}_i \oplus \hat{F}_i)$, and $J = \mathcal{J}$. These steps however, are completely specified in terms of the same transformations $\{(\Theta_i, \Phi_i),\ i = \eta_i, \ldots, 2\eta_i - 1\}$ that were applied to $\hat{H}_{\eta_i,\eta_i}$ while checking the positive definiteness of $E_{\eta_i} E_{\eta_i}$. So we just need to update the last block row $\tilde{S}_i$ via the same transformations.

The point is that we can read out the desired QR factors Q_i and P_i from these first η_i triangular factors of $\bar{D}_i$. To see this, we denote the first η_i triangular factors by

$$L_d = \begin{bmatrix} l_{d0} & l_{d1} & \ldots & l_{d,\eta_i-1} \end{bmatrix}, \qquad D_d = \text{diagonal}\{d_{d0}, \ldots, d_{d,\eta_i-1}\}.$$

Then we can write, using the Schur reduction procedure (4),

$$\bar{D}_i = \bar{L}_d \bar{L}_d^* + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -I_{\eta_i} \end{bmatrix}, \tag{24d}$$

where $\bar{L}_d = L_d D_d^{-1/2}$. Comparing (24b) and (24d) we can easily conclude that $\bar{L}_d$ can be partitioned into a top lower triangular block equal to P_i^* and a lower block equal to Q_i, viz.,

$$\bar{L}_d = \begin{bmatrix} P_i^* \\ Q_i \end{bmatrix}.$$

ALGORITHM 6.1. *The QR factors Q_i and P_i^* can be computed in strongly regular steps as follows.*

1. *Apply the transformations $\{(\Theta_i, \Phi_i),\ i = 0, \ldots, \eta_i - 1\}$ that were applied to $\hat{H}_{\eta_i}$ to the block row $\begin{bmatrix} \mathbf{0} & \mathbf{0} & \hat{G}_i \end{bmatrix}$, and construct $\bar{S}_i$ as in (24c).*
2. *Apply the last block row $\tilde{S}_i$ through the next η_i transformations $\{(\Theta_i, \Phi_i),\ i = \eta_i, \ldots, 2\eta_i - 1\}$ that were applied to $\hat{H}_{\eta_i,\eta_i}$ while checking the positive definiteness of $E_{\eta_i} E_{\eta_i}$. This determines the first η_i triangular factors of $\bar{D}_i$.*
3. *Partition $\bar{L}_d$ as shown above and read out Q_i and P_i^*.*

What about the choice $T_{\eta_i} \neq I$? In this case we need to consider the extended (also Toeplitz-like) matrix

$$\hat{D}_i = \begin{bmatrix} -T_{\eta_i} & D_i & \mathbf{0} \\ D_i & \mathbf{0} & D_i \\ \mathbf{0} & D_i & \mathbf{0} \end{bmatrix},$$

which still leads, after the first η_i recursive steps, to a generator for the matrix

$$\bar{D}_i = \begin{bmatrix} D_i T_{\eta_i}^{-1} D_i & D_i \\ D_i & \mathbf{0} \end{bmatrix},$$

and which is of the form

$$\bar{S}_i = \begin{bmatrix} \hat{H}_{\eta_i,\eta_i} \\ \tilde{S}_i \end{bmatrix}.$$

The point, however, is that the first η_i triangular factors of $\bar{D}_i$ now lead to a factorization of the form $D_i = Q_i P_i$, where P_i is still upper triangular but Q_i now satisfies $Q_i Q_i^* = T_{\eta_i}$. That is, Q_i is no longer a unitary matrix. But T_{η_i} is a positive-definite

and structured matrix. Hence, its Cholesky factorization $T_{\eta_i} = \bar{L}_T \bar{L}_T^*$, can be efficiently evaluated in $O(\beta\eta_i^2)$ operations by using the strongly regular Algorithm 4.1. In this case, and following the argument in §4.3, we are instead led to a triangular factorization for R of the form $R = \hat{L}PQ\hat{L}^*$, where we now define $Q = Q_0 \oplus \cdots \oplus Q_{t-1}$, $P = P_0^{-*} \oplus \cdots \oplus P_{t-1}^{-*}$, $T = T_{\eta_0}^{-1} \oplus \cdots \oplus T_{\eta_{t-1}}^{-1}$, and $\hat{L} = LTQ$. The matrix $\hat{L}$ is still lower triangular with block columns of the form

$$\begin{bmatrix} P_i^* \\ W_i T_{\eta_i}^{-1} Q_i \end{bmatrix}.$$

The inverses $T_{\eta_i}^{-1}$ are not needed explicitly because, once we have the Cholesky factor of T_{η_i}, the products $T_{\eta_i}^{-1} Q_i$ can be computed by solving linear triangular systems. Also, the generator recursion has the same form as before (12a), viz.,

$$\begin{bmatrix} \mathbf{0}_{\eta_i \times r} \\ G_{i+1} \end{bmatrix} = G_i + X_i, \qquad X_i = (F_i - I_{n-\alpha_i}) L_i D_i^{-1} (I_{\eta_i} - \hat{F}_i)^{-1} \hat{G}_i,$$

and where X_i can now be rewritten as (compare with (13))

$$X_i = (F_i - I_{n-\alpha_i}) \begin{bmatrix} P_i^* \\ S_i T_{\eta_i}^{-1} Q_i \end{bmatrix} P_i^{-*} (I_{\eta_i} - \hat{F}_i)^{-1} \hat{G}_i.$$

7. System interpretation. The generator recursions of Algorithms 4.1 and 4.2 have an interpretation as a cascade of linear state-space systems of orders $\{\eta_0, \eta_1, \ldots\}$. To clarify this, observe that the expressions for L_i and G_{i+1} in Theorem 3.1 can be combined together as follows:

$$\begin{bmatrix} L_i & \mathbf{0} \\ & G_{i+1} \end{bmatrix} = \begin{bmatrix} F_i L_i & G_i \end{bmatrix} \begin{bmatrix} \hat{F}_i^* & \hat{H}_i^* J \\ J\hat{G}_i^* & J\hat{K}_i^* J \end{bmatrix}.$$

Hence, each recursive step involves an η_i–order discrete-time system that arises in state-space form on the right-hand side of the above expression, viz.,

$$\begin{bmatrix} \mathbf{x}_{j+1} & \mathbf{y}_j \end{bmatrix} = \begin{bmatrix} \mathbf{x}_j & \mathbf{w}_j \end{bmatrix} \begin{bmatrix} \hat{F}_i^* & \hat{H}_i^* J \\ J\hat{G}_i^* & J\hat{K}_i^* J \end{bmatrix},$$

where $\mathbf{x}_j$ is a $1 \times \eta_i$ state-vector and $\mathbf{w}_j$ and $\mathbf{y}_j$ are $1 \times r$ (row) input and output vectors, respectively, at time j. The above system matrix can also be regarded as a state-space realization of the inverse system

$$\begin{bmatrix} \hat{F}_i & \hat{G}_i \\ \hat{H}_i & \hat{K}_i \end{bmatrix}^{-1},$$

since it follows from the embedding relation (5b) that

$$\begin{bmatrix} \hat{F}_i & \hat{G}_i \\ \hat{H}_i & \hat{K}_i \end{bmatrix}^{-1} = \begin{bmatrix} D_i \hat{F}_i^* D_i^{-1} & D_i \hat{H}_i^* J \\ J\hat{G}_i^* D_i^{-1} & J\hat{K}_i^* J \end{bmatrix}.$$

The corresponding $r \times r$ transfer matrix $\Theta_i(z)$ is given by

$$\Theta_i(z) = J\hat{K}_i^* J + J\hat{G}_i^* \left[z^{-1} I_{\eta_i} - \hat{F}_i^* \right]^{-1} \hat{H}_i^* J.$$

It also follows from the embedding relation (5b) that $\Theta_i(z)$ satisfies the normalization condition $\Theta_i(z)J\Theta_i^*(z) = J$ on $|z| = 1$ and that, using (7), we can rewrite $\Theta_i(z)$ in the form

$$\Theta_i(z) = \left\{ I - (1 - z\tau_i^*)J\hat{G}_i^*(I_{\eta_i} - z\hat{F}_i^*)^{-1}D_i^{-1}(I - \tau_i^*\hat{F}_i)^{-1}\hat{G}_i \right\} \Theta_i.$$

Therefore, t recursive steps lead to a cascade $\boldsymbol{\Theta}(z) = \Theta_0(z)\Theta_1(z)\dots\Theta_{t-1}(z)$, which also satisfies $\boldsymbol{\Theta}(z)J\boldsymbol{\Theta}^*(z) = J$ on $|z| = 1$. In fact, we can further show that the cascade admits a state-space realization in terms of the original matrices F and G [24] [31].

THEOREM 7.1. *The cascade $\boldsymbol{\Theta}(z)$ admits an $n-$dimensional state-space description of the form*

$$\begin{bmatrix} \mathbf{x}_{j+1} & \mathbf{y}_j \end{bmatrix} = \begin{bmatrix} \mathbf{x}_j & \mathbf{w}_j \end{bmatrix} \begin{bmatrix} F^* & H^*J \\ JG^* & JK^*J \end{bmatrix},$$

where H and K are $r \times n$ and $r \times r$ matrices that satisfy the embedding relation

$$\begin{bmatrix} F & G \\ H & K \end{bmatrix} \begin{bmatrix} R & \mathbf{0} \\ \mathbf{0} & J \end{bmatrix} \begin{bmatrix} F & G \\ H & K \end{bmatrix}^* = \begin{bmatrix} R & \mathbf{0} \\ \mathbf{0} & J \end{bmatrix}.$$

It also follows that the matrices H and K can be expressed in terms of R, F, and G as follows:

$$H = \boldsymbol{\Theta}^{-1}JG^*\left[I - \tau F^*\right]^{-1}R^{-1}(\tau I - F),$$
$$K = \boldsymbol{\Theta}^{-1}\left\{ I - JG^*\left[I - \tau F^*\right]^{-1}R^{-1}G \right\},$$

and that $\boldsymbol{\Theta}(z) = \left\{ I - (1 - z\tau^*)JG^*(I - zF^*)^{-1}R^{-1}(I - \tau^*F)^{-1}G \right\} \boldsymbol{\Theta}$, where τ is a unit-modulus scalar and $\boldsymbol{\Theta}$ is a $J-$unitary matrix.

8. Concluding remarks. We derived a block Schur algorithm for the block triangular factorization of Hermitian Toeplitz-like matrices. We also provided tests for the determination of the sizes of the nonsingular minors in the exactly singular case. We also presented a system interpretation of the algorithm in terms of a cascade of elementary sections. We further remark that the results can be extended to non-Hermitian Toeplitz-like matrices, as well as Hankel-like matrices, and may be discussed elsewhere; see [31].

Some issues deserve further consideration and may simplify the development of the algorithm. We have limited ourselves in the block case, for example, to the obvious choice $\Theta_i = I$. Other choices may be considered and could lead to an array form of the generator recursion (12a) in the same spirit as (11). Also, explicit tests for determining the sizes of the nonsingular minors in the general case of displacement ranks larger than two, along the lines of the special cases discussed in §5.2, deserve further investigation. These issues will be addressed elsewhere.

REFERENCES

[1] N. I. AKHIEZER, *The Classical Moment Problem and Some Related Questions in Analysis*, Hafner Publishing Company, New York, 1965.

[2] D. ALPAY AND H. DYM, *Structured invariant spaces of vector valued rational functions, Hermitian matrices, and a generalization of the Iohvidov laws*, Linear Algebra Appl., 137–138 (1990), pp. 137–181.

[3] E. R. BERLEKAMP, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.
[4] R. E. BLAHUT, *Theory and Practice of Error Control Codes*, Addison-Wesley, Reading, MA, 1983.
[5] T. BOROS, A. H. SAYED, AND T. KAILATH, *Structured matrices and unconstrained rational interpolation*, Linear Algebra Appl., 203–204 (1994), pp. 155–188.
[6] T. F. CHAN AND P. C. HANSEN, *A look-ahead Levinson algorithm for general Toeplitz systems*, IEEE Trans. Signal Processing, 40 (1992), pp. 1079–1090.
[7] M. K. CILIZ AND H. KRISHNA, *Split Levinson algorithm for Toeplitz matrices with singular submatrices*, IEEE Transactions on Circuits and Systems, 36 (1989), pp. 922–924.
[8] T. CITRON, *Algorithms and Architectures for Error Correcting Codes*, Ph.D. thesis, Stanford University, Stanford, CA, 1986.
[9] P. DELSARTE, Y. GENIN, AND Y. KAMP, *A generalization of the Levinson algorithm for Hermitian Toeplitz matrices with any rank profile*, IEEE Transactions on Acoustics, Speech and Signal Processing, 33 (1985), pp. 964–971.
[10] R. W. FREUND, *A look-ahead Bareiss algorithm for general Toeplitz matrices*, Numer. Math., 60 (1994), pp. 35–69
[11] F. R. GANTMACHER, *The Theory of Matrices*, Chelsea Publishing Company, New York, 1959.
[12] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, second ed., The Johns Hopkins University Press, Baltimore, 1989.
[13] M. J. C. GOVER AND S. BARNETT, *Inversion of Toeplitz matrices which are not strongly nonsingular*, IMA J. Numer. Anal., 5 (1985), pp. 101–110.
[14] M. GUTKNECHT, *Stable row recurrences for the Padé table and generically superfast lookahead solvers for non-Hermitian Toeplitz systems*, Linear Algebra Appl., 188/189 (1993), pp. 351–422.
[15] G. HEINIG AND K. ROST, *Algebraic Methods for Toeplitz-like Matrices and Operators*, Operator Theory: Advances and Applications, Vol. 13, Birkhäuser-Verlag, Basil, Boston, 1984.
[16] I. S. IOHVIDOV, *Hankel and Toeplitz Matrices and Forms*, Birkhäuser, Boston, MA, 1982.
[17] T. KAILATH, *Linear Systems*, Prentice Hall, Englewood Cliffs, NJ, 1980.
[18] ———, *Signal processing applications of some moment problems*, in Moments in Mathematics, H. Landau, ed., Amer. Math. Soc., Providence, 1987, pp. 71–109.
[19] T. KAILATH AND J. CHUN, *Generalized displacement structure for block-Toeplitz, Toeplitz-block, and Toeplitz-derived matrices*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 114–128.
[20] T. KAILATH, S. Y. KUNG, AND M. MORF, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979), pp. 395–407.
[21] T. KAILATH AND A. H. SAYED, *Displacement structure: Theory and applications*, SIAM Rev., submitted.
[22] S. Y. KUNG, *Multivariable and Multidimensional Systems*, Ph.D. thesis, Stanford University, Stanford, CA, June 1977.
[23] H. LEV-ARI AND T. KAILATH, *Triangular factorization of structured Hermitian matrices*, Operator Theory: Advances and Applications, 18 (1986), pp. 301–324.
[24] ———, *State-space approach to factorization of lossless transfer functions and structured matrices*, Linear Algebra Appl., 162–164 (1992), pp. 273–295.
[25] J. L. MASSEY, *Shift-register synthesis and* BCH *decoding*, IEEE Trans. Information Theory, 15 (1969), pp. 122–127.
[26] D. PAL, *Fast Algorithms for Structured Matrices with Arbitrary Rank Profile*, Ph.D. thesis, Stanford University, Stanford, CA, May 1990.
[27] D. PAL AND T. KAILATH, *Fast triangular factorization of Hermitian Toeplitz and related matrices with arbitrary rank profile*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 451–478.
[28] S. POMBRA, H. LEV-ARI, AND T. KAILATH, *Levinson and Schur algorithms for Toeplitz matrices with singular minors*, in Proc. IEEE Internat. Conf. Acoustics, Speech and Signal Processing, NY, April 1988, pp. 1643–1646.
[29] Y. POTAPOV, *The multiplicative structure of J-contractive matrix functions*, Amer. Math. Soc. Transl., 15 (1960), pp. 131–244.
[30] C. M. RADER AND A. O. STEINHARDT, *Hyperbolic Householder transformations*, IEEE Trans. Acoustics, Speech and Signal Processing, 34 (1986), pp. 1589–1602.
[31] A. H. SAYED, *Displacement Structure in Signal Processing and Mathematics*, Ph.D. thesis, Stanford University, Stanford, CA, August 1992.
[32] A. H. SAYED, T. CONSTANTINESCU, AND T. KAILATH, *Square-Root Algorithms for Structured Matrices, Interpolation, and Completion Problems*, Vol. 69, IMA Volumes in Mathematics and Its Applications, Springer-Verlag. To appear.
[33] ———, *Time-variant displacement structure and interpolation problems*, IEEE Trans. Automatic Control, 39 (1994), pp. 960–976.

[34] A. H. SAYED, T. KAILATH, H. LEV-ARI, AND T. CONSTANTINESCU, *Recursive solutions to rational interpolation problems via fast matrix factorization*, Integral Equations Operator Theory, 20 (1994), pp. 84–118.

[35] A. H. SAYED, H. LEV-ARI, AND T. KAILATH, *Time-variant displacement structure and triangular arrays*, IEEE Trans. Signal Processing, 42 (1994), pp. 1052–1062.

[36] I. SCHUR, *Über potenzreihen die im Inneren des Einheitskreises beschränkt sind*, Journal für die Reine und Angewandte Mathematik, 147 (1917), pp. 205–232. English translation in Operator Theory: Advances and Applications, I. Gohberg, ed., Vol. 18, pp. 31–88, Birkhäuser, Boston, 1986.

[37] C. J. ZAROWSKI, *Schur algorithms for Hermitian Toeplitz, and Hankel matrices with singular leading principal submatrices*, IEEE Trans. Signal Processing, 39 (1991), pp. 2464–2480.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 415–420, April 1995

006

COMPUTING EXACT COMPONENTWISE BOUNDS ON SOLUTIONS OF LINEAR SYSTEMS WITH INTERVAL DATA IS NP-HARD*

JIRI ROHN† AND VLADIK KREINOVICH‡

Abstract. We prove that it is NP-hard to compute the exact componentwise bounds on solutions of all the linear systems that can be obtained from a given linear system with a nonsingular matrix by perturbing all the data independently of each other within prescribed tolerances.

Key words. linear equations, perturbation, componentwise bounds

AMS subject classifications. 15A06, 65G10, 68Q25

1. Introduction.

Given a system of linear equations

$$Ax = b, \tag{1}$$

where $A \in R^{n\times n}$ is nonsingular and $b \in R^n$, consider the perturbed system

$$A'x' = b' \tag{2}$$

with data A', b' satisfying

$$|A' - A| \leq \Delta \tag{3}$$

and

$$|b' - b| \leq \delta, \tag{4}$$

where $\Delta \in R^{n\times n}_+$ and $\delta \in R^n_+$ are correspondingly the matrix and vector of perturbation bounds (the absolute value of a matrix $B = (b_{ij})$ is defined by $|B| = (|b_{ij}|)$, and the inequality (3) is understood componentwise; similarly for vectors). Let X denote the set of solutions of all the perturbed systems, i.e.,

$$X = \{x'; A'x' = b' \text{ for some } A', b' \text{ satisfying } (3), (4)\}.$$

Naturally, we are interested in knowing the exact range of the components of the solution under the allowed perturbations, i.e., in computing the numbers

$$\underline{x}_i = \min_{x'\in X} x'_i, \tag{5}$$

$$\overline{x}_i = \max_{x'\in X} x'_i \tag{6}$$

$(i = 1, \ldots, n)$; we call them the *exact componentwise bounds* on solutions of the perturbed systems.

* Received by the editors June 28, 1993; accepted for publication (in revised form) by R. Brualdi, December 20, 1993. This work was partially supported by National Science Foundation grant CDA-9015006, Czech Republic Grant Agency grant GACR 201/93/0429 and a grant from the Institute for Materials and Manufacturing Management.

† Faculty of Mathematics and Physics, Charles University, Malostranske nam. 25, 11800 Prague, Czech Republic (rohn@kam.ms.mff.cuni.cz).

‡ Department of Computer Science, University of Texas at El Paso, El Paso, Texas 79968 (vladik@cs.utep.edu).

During the last 30 years, the problem of computing the exact componentwise bounds (formulated often in the framework of systems of linear interval equations) has received much attention. General methods (assuming only nonsingularity of each matrix A' satisfying (3)) were given by Oettli [7], Rohn [9], and Shary [12]; however, all of them require in the worst case an amount of operations that is exponential in n. As a result, these methods are not applicable to problems of large dimension n. Therefore, a number of articles deal with special cases (such as M-matrices [2], H-matrices [6], inverse stable matrices [9], matrices satisfying a spectral condition [11], or diagonally dominant [4]) for which there exist polynomial algorithms for computing the exact componentwise bounds (or their enclosures). For surveys of such methods, see the monographs by Alefeld and Herzberger [1] or Neumaier [6].

In this paper we show that computing the exact componentwise bounds is NP-hard (see Garey and Johnson [3] for basic concepts of the complexity theory). Thus, unless $P = NP$ (which is currently widely believed to be not true), we cannot expect an existence of polynomial-time algorithms for solving our problem. The NP-hardness of the computation of (5), (6) for overdetermined systems (A of size $m \times n$, $m > n$) was recently established by Kreinovich, Lakeyev, and Noskov [5], but the idea of the proof, which reduces 3-satisfiability to computation of the exact componentwise bounds for linear systems with matrices of size about $3n \times n$, cannot be used for the square case.

We carry out the proof of our result by studying a special instance of *constant* componentwise perturbations. We show that in this case the optimal value of a specially chosen linear function over X can be expressed in terms of the reciprocal value of the so-called radius of nonsingularity, which has been recently shown to be NP-hard to compute (Poljak and Rohn [8]). Then adding one more row and column to the original system to make the linear function depend on a single variable only, we obtain the desired result.

2. Auxiliary results. For a given nonsingular matrix $A \in R^{n \times n}$ and the linear system

$$Ax = 0$$

(which has a unique solution $x = 0$), consider the perturbed systems

$$A'x' = b'$$

with

$$|A' - A| \leq \beta e e^T \tag{7}$$

and

$$|b'| \leq \beta e, \tag{8}$$

where $e = (1, 1, \ldots, 1)^T \in R^n$ and β is a real parameter. To underline the dependence on the parameter, let us denote the solution set by X_β:

$$X_\beta = \{x'; A'x' = b' \text{ for some } A', b' \text{ satisfying } (7), (8)\}.$$

We first give a description of the set X_β; throughout the following text, we use the norm $\|x\| = \|x\|_1 = e^T|x| = \sum_i |x_i|$.

PROPOSITION 2.1. *Let A be nonsingular, and let β satisfy*

$$0 < \beta < \frac{1}{e^T|A^{-1}|e}. \tag{9}$$

Then each A' satisfying (7) is nonsingular and we have

$$X_\beta = \left\{ x'; x' = \frac{\beta}{1 - \beta\|A^{-1}c\|} A^{-1}c, -e \le c \le e \right\}. \tag{10}$$

Proof. (i) Let $x' \in X_\beta$, i.e., $A'x' = b'$ for some A', b' satisfying (7), (8). Then we have $|Ax'| = |(A - A')x' + b'| \le \beta ee^T|x'| + \beta e = \beta(\|x'\| + 1)e$; hence if we take

$$c = \frac{1}{\beta(\|x'\| + 1)} Ax',$$

then we have $-e \le c \le e$ and $Ax' = \beta(\|x'\| + 1)c$, which implies

$$x' = \beta(\|x'\| + 1)A^{-1}c, \tag{11}$$

hence

$$\|x'\| = \beta(\|x'\| + 1)\|A^{-1}c\|. \tag{12}$$

Since

$$\beta\|A^{-1}c\| = \beta e^T|A^{-1}c| \le \beta e^T|A^{-1}|e < 1 \tag{13}$$

due to (9), from (12) we obtain

$$\|x'\| = \frac{\beta\|A^{-1}c\|}{1 - \beta\|A^{-1}c\|}.$$

Substituting this equality into (11) leads to

$$x' = \frac{\beta}{1 - \beta\|A^{-1}c\|} A^{-1}c, \tag{14}$$

hence x' is of the form described in (10).

(ii) Conversely, let x' be of the form (14) for some c satisfying $-e \le c \le e$. Define a vector $z \in R^n$ as follows: $z_j = 1$ if $x'_j \ge 0$ and $z_j = -1$ otherwise $(j = 1, \ldots, n)$. Then $z^Tx' = e^T|x'| = \|x'\|$, hence

$$(A - \beta cz^T)x' = \frac{1}{1 - \beta\|A^{-1}c\|}(\beta c - \beta^2\|A^{-1}c\|c) = \beta c,$$

which means that x' is a solution of the system

$$(A - \beta cz^T)x' = \beta c,$$

where $|(A - \beta cz^T) - A| = \beta|c| \cdot |z|^T \le \beta ee^T$ and $|\beta c| \le \beta e$. Hence, $x' \in X_\beta$.

(iii) From (13) and (14), we conclude that

$$\|x'\| \le \frac{\beta e^T|A^{-1}|e}{1 - \beta e^T|A^{-1}|e}$$

for each $x' \in X_\beta$, hence X_β is bounded. If some A' satisfying (7) was singular, then we would have $A'x' = 0$ for some $x' \neq 0$, hence $\lambda x' \in X_\beta$ for each $\lambda \in R^1$, which would contradict the boundedness of X_β. Hence, each A' satisfying (7) is nonsingular. □

Before proceeding further, let us introduce, for a matrix $B \in R^{n\times n}$, the number

$$r(B) = \max\{\|By\|; y \in \{-1,1\}^n\}.$$

A simple reasoning shows that it can be also written as

$$r(B) = \max\{z^T By; z, y \in \{-1,1\}^n\},$$

which is the form in which it was originally introduced in [8]. Then, we have the following result.

PROPOSITION 2.2. *Let A be nonsingular and let β satisfy* (9). *Then for each $i \in \{1,\dots,n\}$ we have*

$$\max_{x'\in X_\beta} (Ax')_i = \frac{\beta}{1-\beta r(A^{-1})}. \tag{15}$$

Proof. (i) First, we prove that

$$\|A^{-1}c\| \le r(A^{-1}) \tag{16}$$

holds for each c, $|c| \le e$. For every c that satisfies this inequality $|c| \le e$, we define vectors $z, y \in \{-1,1\}^n$ as follows: $z_j = 1$ if $(A^{-1}c)_j \ge 0$ and $z_j = -1$ otherwise $(j = 1,\dots,n)$, and $y_j = 1$ if $(z^TA^{-1})_j \ge 0$ and $y_j = -1$ otherwise $(j = 1,\dots,n)$. Then, we have $\|A^{-1}c\| = e^T|A^{-1}c| = z^TA^{-1}c \le z^TA^{-1}y \le \max\{z^TA^{-1}y; z, y \in \{-1,1\}^n\} = r(A^{-1})$, i.e., (16).

(ii) Let us fix an $i \in \{1,\dots,n\}$ and let $x' \in X_\beta$. According to Proposition 2.1, we have

$$x' = \frac{\beta}{1-\beta\|A^{-1}c\|} A^{-1}c$$

for some c such that $|c| \le e$. Since the denominator is positive (due to (9) and (13)), we have

$$(Ax')_i \le |(Ax')_i| \le \frac{\beta}{1-\beta\|A^{-1}c\|} \le \frac{\beta}{1-\beta r(A^{-1})}$$

(due to (16)). Hence,

$$\max_{x'\in X_\beta} (Ax')_i \le \frac{\beta}{1-\beta r(A^{-1})}. \tag{17}$$

(iii) Take $\overline{y} \in \{-1,1\}^n$ such that

$$\|A^{-1}\overline{y}\| = \max\{\|A^{-1}y\|; y \in \{-1,1\}^n\} = r(A^{-1}).$$

Since $\|A^{-1}(-y)\| = \|A^{-1}y\|$, $\overline{y}$ can be chosen in such a way that $\overline{y}_i = 1$. According to Proposition 2.1, the vector

$$x' = \frac{\beta}{1-\beta\|A^{-1}\overline{y}\|} A^{-1}\overline{y}$$

belongs to X_β and satisfies the equality

$$(Ax')_i = \frac{\beta}{1 - \beta r(A^{-1})}.$$

Hence the upper bound in (17) is achieved, which proves (15). □

3. NP-hardness. Now we are able to prove the main result.

THEOREM 3.1. *For an instance n, A, b, Δ, δ, and $i \in \{1, \ldots, n\}$ such that each matrix A' satisfying* (3) *is nonsingular, computing both $\underline{x}_i$ and $\overline{x}_i$ given by* (5) *and* (6) *is* NP-*hard.*

Comment. Since checking nonsingularity of all matrices A' satisfying (3) is already NP-hard [8], we must include nonsingularity into the assumptions to separate the two problems.

Proof. In [8, Thm. 2.6] it is proved that computing $r(B)$ is NP-hard for $B \in R^{n \times n}$. The result was stated there for general matrices, but it remains valid if we confine ourselves to nonsingular matrices only (since the proof employs a diagonally dominant matrix, which is nonsingular). We will show that computing $r(B)$ can be polynomially reduced to the computation of an exact componentwise bound.

For a given nonsingular $B \in R^{n \times n}$, choose a β satisfying

$$0 < \beta < \frac{1}{e^T |B| e} \tag{18}$$

and compute $A = B^{-1}$ (this can be done in polynomial time). Now, construct the $(n+1) \times (n+1)$ matrices

$$\tilde{A} = \begin{pmatrix} A & 0 \\ A_{n\cdot} & -1 \end{pmatrix},$$

where $A_{n\cdot}$ denotes the nth row of A, and

$$\Delta = \begin{pmatrix} \beta e e^T & 0 \\ 0 & 0 \end{pmatrix},$$

and let

$$b = 0$$

and

$$\delta = \begin{pmatrix} \beta e \\ 0 \end{pmatrix}$$

($e \in R^n$). Then each $A' \in R^{(n+1)\times(n+1)}$ with $|A' - \tilde{A}| \leq \Delta$ is nonsingular by (18) and by Proposition 2.1, and for the solution set of the perturbed systems we have

$$X = \{(x, x_{n+1})^T; x \in R^n, x \in X_\beta, x_{n+1} = A_{n\cdot} x\}.$$

Hence, for the exact componentwise bound on x_{n+1}, we conclude from Proposition 2.2 that

$$\overline{x}_{n+1} = \max_{x' \in X_\beta} (Ax')_n = \frac{\beta}{1 - \beta r(B)}.$$

So, the computation of $r(B)$ has been polynomially reduced to the computation of $\overline{x}_{n+1}$. Thus, since computing $r(B)$ is NP-hard, the same must be true for $\overline{x}_{n+1}$ as well. In this way we have proved the NP-hardness of computing the exact upper bound on the highest index variable; now by permutation of variables we easily extend this result to an arbitrary variable. The statement for lower bounds follows immediately from the result just proved if we observe that the lower bounds differ only in their signs from the upper bounds for the system $Ax = -b$ under the same Δ and δ. □

Final note. The result can be made more understandable if we point out that (15) is, in general, a nonconvex optimization problem. Indeed, a lengthy argument (which we omit here) based on Theorems 1 and 2 in [10] proves that *if* $n \geq 3$, A *is nonsingular and* β *satisfies* (9), *then* X_β *is a nonconvex set whose convex hull has* 2^n *vertices which are exactly those points* x' *in* (10) *that correspond to parameter values* $t \in \{-1, 1\}^n$.

Acknowledgments. The authors would like to thank the anonymous referees for their valuable editorial suggestions.

REFERENCES

[1] G. ALEFELD AND J. HERZBERGER, *Introduction to Interval Computations*, Academic Press, New York, 1983.

[2] W. BARTH AND E. NUDING, *Optimale Lösung von Intervallgleichungssystemen*, Computing, 12 (1974), pp. 117–125.

[3] M. E. GAREY AND D. S. JOHNSON, *Computers and Intractability: A Guide to the Theory of* NP-*Completeness*, Freeman, San Francisco, 1979.

[4] E. R. HANSEN, *Bounding the solution of interval linear equations*, SIAM J. Numer. Anal., 29 (1992), pp. 1493–1503.

[5] V. KREINOVICH, A. V. LAKEYEV, AND S. I. NOSKOV, *Optimal solution of interval linear systems is intractable* (NP-*hard*), Interval Computations, 1 (1993), pp. 6–14.

[6] A. NEUMAIER, *Interval Methods for Systems of Equations*, Cambridge University Press, Cambridge, 1990.

[7] W. OETTLI, *On the solution set of a linear system with inaccurate coefficients*, SIAM J. Numer. Anal., 2 (1965), pp. 115–118.

[8] S. POLJAK AND J. ROHN, *Checking robust nonsingularity is* NP-*hard*, Math. Control, Signals Systems, 6 (1993), pp. 1–9.

[9] J. ROHN, *Systems of linear interval equations*, Linear Algebra Appl., 126 (1989), pp. 39–78.

[10] ———, *On nonconvexity of the solution set of a system of linear interval equations*, BIT, 30 (1989), pp. 161–165.

[11] S. M. RUMP, *Solving algebraic problems with high accuracy*, in A New Approach to Scientific Computation, U. Kulisch and W. Miranker, eds., Academic Press, NY, 1983, pp. 51–120.

[12] S. P. SHARY, *A new class of algorithms for optimal solution of interval linear systems*, Interval Computations, 2 (1992), pp. 18–29.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 421–435, April 1995

007

HAMILTON AND JACOBI MEET AGAIN: QUATERNIONS AND THE EIGENVALUE PROBLEM*

NILOUFER MACKEY†

Abstract. The algebra isomorphism between $\mathcal{M}_4(\mathcal{R})$ and $\mathcal{H}\otimes\mathcal{H}$, where $\mathcal{H}$ is the algebra of quaternions, has unexpected computational payoff: it helps construct an orthogonal similarity that 2×2 block-diagonalizes a 4×4 symmetric matrix. Replacing plane rotations with these more powerful 4×4 rotations leads to a quaternion-Jacobi method in which the "weight" of four elements (in a 2×2 block) is transferred all at once onto the diagonal. Quadratic convergence sets in sooner, and the new method requires at least one fewer sweep than plane-Jacobi methods. An analogue of the sorting angle for plane rotations is developed for these 4×4 rotations.

Key words. eigenvalues, symmetric matrix, Jacobi method, quaternion, tensor product

AMS subject classifications. 65F15, 15A18, 15A21, 15A69

1. Introduction.

One hundred and fifty years ago, on 16 October 1843, W. R. Hamilton carved the equations defining the algebra of quaternions on the stones of Brougham Bridge, Dublin [15], [9], [30]. Two years later, in an unrelated piece of work, C. G. J. Jacobi described an iterative method for solving the eigenproblem of an $n \times n$ symmetric matrix. This method appeared the following year in Crelle's journal [22].[1]

To the student of mechanics, the names Hamilton and Jacobi are already closely linked: in 1837, Jacobi extended Hamilton's work in dynamics, giving rise to what is today known as the Hamilton–Jacobi theory.[2] Now, a century and a half later, we bring the work of these two men together in a new way: we show how Hamilton's quaternions enhance Jacobi's algorithm for solving the symmetric eigenproblem. Jacobi diagonalizes a symmetric matrix by performing a sequence of orthogonal similarity transformations. Each transformation is a plane rotation, chosen so that the induced similarity diagonalizes some 2×2 principal submatrix, moving the weight of the annihilated elements onto the diagonal. Can one explicitly specify an orthogonal transformation that diagonalizes a larger submatrix?

We show that the algebra isomorphism between $\mathcal{M}_4(\mathcal{R})$ and $\mathcal{H}\otimes\mathcal{H}$, where $\mathcal{H}$ is the algebra of quaternions, has direct and unexpected computational payoff: it leads to the construction of an orthogonal similarity to (2×2)-block diagonalize a 4×4 symmetric matrix. The quaternion-Jacobi method thus obtained produces four times as many zeros at each step, and hence converges in fewer iterations. The price we pay for this abundance of zeros is the cost of computing one left-right singular vector pair of a 3×3 matrix, whose entries are simple linear combinations of the entries of the 4×4 symmetric matrix being diagonalized. The quaternion-Jacobi method is at least quadratically convergent, and experimental evidence strongly suggests that it requires at least one fewer sweep than the traditional Jacobi method.

* Received by the editors October 8, 1993; accepted for publication (in revised form) by N. J. Higham, January 10, 1994. An earlier version of this paper won the 1993 SIAM Student Paper Competition.

† Department of Mathematics and Statistics, Western Michigan University, Kalamazoo, Michigan 49008 (nil.mackey@wmich.edu).

[1] The key ideas for the method were introduced by Jacobi in an earlier paper [21]. Thanks to Sven Hammarling for pointing this out.

[2] The two mathematicians first met in 1842 at the meeting of the British Association for the Advancement of Science held in Manchester, England. According to Hankins [17], Jacobi refers to Hamilton as "the illustrious Astronomer Royal of Dublin" and later, as "the Lagrange" of Ireland.

There is renewed interest in Jacobi-type methods today, because they are easily parallelizable [3], [11], [12], [28], and compute small eigenvalues with greater accuracy than the QR method [8], [25].

2. The quaternions. Recently, Hacon [14] showed that one can use quaternions to construct an orthogonal similarity transformation that will directly reduce any given 4×4 skew-symmetric matrix to its real Schur (2×2 block diagonal) form, and thereby obtain a Jacobi-type algorithm for $n \times n$ skew-symmetric matrices. The method is intriguing, and it is natural to ask whether a 4×4 symmetric matrix can be (block) diagonalized in a similar manner. To uncover the symmetric algorithm, we start with some algebraic preliminaries.

The quaternions, $\mathcal{H}$, are a four-dimensional, associative, but noncommutative, division algebra over $\mathcal{R}$, with the standard basis $\{1, i, j, k\}$. Multiplication is determined by the rules

$$i^2 = j^2 = k^2 = ijk = -1,$$

which imply $jk = -kj = i, ki = -ik = j, ij = -ji = k$. The typical quaternion is

$$q = q_0 + q_1 i + q_2 j + q_3 k, \qquad q_0, q_1, q_2, q_3 \in \mathcal{R}.$$

The real part of q is q_0 and the pure quaternion part is $q_1 i + q_2 j + q_3 k$. The *conjugate* of q is given by $\overline{q} = q_0 - q_1 i - q_2 j - q_3 k$ and the *norm* $|q|$, is defined as

$$|q|^2 = q_0^2 + q_1^2 + q_2^2 + q_3^2 = q\overline{q} = \overline{q}q.$$

Thus one can compute the multiplicative inverse of any nonzero quaternion,

$$q^{-1} = \frac{\overline{q}}{(|q|)^2}.$$

As a vector space, $\mathcal{H}$ is identified with $\mathcal{R}^4$ via the customary isomorphism,

$$q_0 + q_1 i + q_2 j + q_3 k \longleftrightarrow (q_0, q_1, q_2, q_3)^t,$$

which in turn induces an isomorphism between the subspace $\mathcal{P}$ of pure quaternions and $\mathcal{R}^3$,

$$q_1 i + q_2 j + q_3 k \longleftrightarrow (q_1, q_2, q_3)^t.$$

Motivated by these isomorphisms we will, when convenient, denote the elements $1, i, j, k$ of $\mathcal{H}$ by e_0, e_1, e_2, e_3, respectively. We will also make use of the standard decomposition,

$$\mathcal{H} = \text{span}\{1\} \oplus \text{span}\{i, j, k\} = \mathcal{R} \oplus \mathcal{P}. \tag{1}$$

3. Tensor products. For the convenience of the reader we include a concrete definition of tensor products of finite-dimensional algebras [1], [27].

DEFINITION 1. *Let V and W be vector spaces over a field $\mathcal{F}$, of dimension m and n, respectively. Let $\{e_r : 1 \le r \le m\}$ be a basis for V and $\{f_s : 1 \le s \le n\}$ be a basis for W. Then the tensor product $V \otimes W$ is an mn dimensional vector space over $\mathcal{F}$ with basis $\{e_r \otimes f_s : 1 \le r \le m, 1 \le s \le n\}$, and is equipped with a bilinear map*

$$\otimes : V \times W \longrightarrow V \otimes W,$$

$$\left(\sum_{r=1}^{m} v_r e_r \, , \sum_{s=1}^{n} w_s f_s \right) \overset{\otimes}{\longmapsto} \sum_{r,s} v_r w_s (e_r \otimes f_s). \tag{2}$$

It is customary to denote $\otimes(v, w)$ by $v \otimes w$, where $v \in V, w \in W$.

Now let $\mathcal{A}$ and $\mathcal{B}$ be *algebras* over a field $\mathcal{F}$. The tensor product $\mathcal{A} \otimes \mathcal{B}$ over $\mathcal{F}$ can be made into an algebra that is essentially the tensor product of the underlying vector spaces, enhanced by an additional multiplicative structure:

$$(3) \qquad * (a \otimes b, a' \otimes b') = aa' \otimes bb', \quad \forall\, a, a' \in A, \quad \forall\, b, b' \in B$$

By extending the above $*$ bilinearly to all the other elements of $\mathcal{A} \otimes \mathcal{B}$, we make $\mathcal{A} \otimes \mathcal{B}$ into an algebra. We henceforth denote $*(a \otimes b, a' \otimes b')$ by simply $(a \otimes b)(a' \otimes b')$.

4. The isomorphisms.

4.1. $\mathcal{P} \otimes \mathcal{P}$ and $\mathcal{M}_3(\mathcal{R})$. A useful isomorphism between the 9-dimensional vector spaces $\mathcal{P} \otimes \mathcal{P}$ and $\mathcal{M}_3(\mathcal{R})$ is obtained by first defining a bilinear map f on the Cartesian product $\mathcal{P} \times \mathcal{P}$:

$$f : \mathcal{P} \times \mathcal{P} \longrightarrow \mathcal{M}_3(\mathcal{R}), \quad f(p, q) = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} \begin{pmatrix} q_1 & q_2 & q_3 \end{pmatrix} = pq^t.$$

We remark that $f(p, q)$ is often called the Kronecker product of p and q [20, §4.2]. From the basic properties of tensor product [1], f induces a unique *linear* map

$$\psi : \mathcal{P} \otimes \mathcal{P} \longrightarrow \mathcal{M}_3(\mathcal{R})$$

with the property that

$$(4) \qquad \psi(p \otimes q) = pq^t.$$

Showing that ψ is a vector space isomorphism is now a simple exercise.

4.2. $\mathcal{H} \otimes \mathcal{H}$ and $\mathcal{M}_4(\mathcal{R})$. An analogous Kronecker product clearly gives a bilinear map from $\mathcal{H} \otimes \mathcal{H}$ to $\mathcal{M}_4(\mathcal{R})$, which then induces a linear isomorphism between $\mathcal{H} \otimes \mathcal{H}$ and $\mathcal{M}_4(\mathcal{R})$. However, this linear isomorphism does *not* preserve the multiplicative structure and hence fails to be an algebra isomorphism. (Since $\mathcal{P} \otimes \mathcal{P}$ is not an algebra, this question does not arise in §4.1.)

Construct an algebra isomorphism between the 16-dimensional algebras $\mathcal{H} \otimes \mathcal{H}$ and $\mathcal{M}_4(\mathcal{R})$ as follows. First, to every ordered pair (p, q) in $\mathcal{H} \times \mathcal{H}$, associate the real 4×4 matrix that represents the linear transformation

$$\begin{aligned} &\mathcal{H} \to \mathcal{H}, \\ &v \mapsto pv\overline{q} \end{aligned}$$

with respect to the standard basis. Denote this matrix by $\mu(p, q)$. Clearly this defines a bilinear map $\mu : \mathcal{H} \times \mathcal{H} \longrightarrow \mathcal{M}_4(\mathcal{R})$, which, from the basic properties of tensor product [1] induces a unique *linear* map

$$\phi : \mathcal{H} \otimes \mathcal{H} \to \mathcal{M}_4(\mathcal{R})$$

with the property that $\phi(p \otimes q) = \mu(p, q)$. It can be shown that ϕ is a bijection that preserves not only the vector space structure, but also the multiplicative structure.[3]

[3] This isomorphism between $\mathcal{H} \otimes \mathcal{H}$ and $\mathcal{M}_4(\mathcal{R})$ has also been used in an entirely different context: the characterization of linear maps which preserve the Ky Fan norms [23].

The proof is omitted, as it is largely a straightforward exercise in formal algebra [4]. Another way to prove that these algebras are isomorphic is to use the fact that the Brauer group of $\mathcal{R}$ is $\mathcal{Z}/2\mathcal{Z}$ [27].

From (3) defining multiplication in a tensor product, we get

$$p \otimes q = (p \otimes 1)(1 \otimes q). \tag{5}$$

Applying ϕ gives $\phi(p \otimes q) = \phi(p \otimes 1)\phi(1 \otimes q)$, or in terms of matrices

$$\phi(p \otimes q) = \begin{pmatrix} p_0 & -p_1 & -p_2 & -p_3 \\ p_1 & p_0 & -p_3 & p_2 \\ p_2 & p_3 & p_0 & -p_1 \\ p_3 & -p_2 & p_1 & p_0 \end{pmatrix} \begin{pmatrix} q_0 & q_1 & q_2 & q_3 \\ -q_1 & q_0 & -q_3 & q_2 \\ -q_2 & q_3 & q_0 & -q_1 \\ -q_3 & -q_2 & q_1 & q_0 \end{pmatrix}. \tag{6}$$

Since $(p \otimes 1)(1 \otimes q) = (1 \otimes q)(p \otimes 1)$, a nonobvious fact that now follows immediately is that the matrices on the right-hand side of (6) commute!

Define conjugation in $\mathcal{H} \otimes \mathcal{H}$ by

$$\overline{p \otimes q} = \overline{p} \otimes \overline{q} \qquad \forall p, q \in \mathcal{H} \tag{7}$$

and extend linearly to all of $\mathcal{H} \otimes \mathcal{H}$. Examining the matrix $\phi(p \otimes 1)$, displayed as the first factor in (6), one sees that

$$\phi(\overline{p} \otimes 1) = (\phi(p \otimes 1))^t.$$

Similarly, $\phi(1 \otimes \overline{q}) = (\phi(1 \otimes q))^t$. Thus we have

PROPOSITION 1. *Conjugation in $\mathcal{H} \otimes \mathcal{H}$ corresponds, via ϕ, to transpose in $\mathcal{M}_4(\mathcal{R})$.*

By the usual abuse of notation, we will sometimes use $p \otimes q$ to stand for $\phi(p \otimes q)$. The reason for this indulgence is twofold: to simplify notation and to emphasize that we will be freely moving between $\mathcal{H} \otimes \mathcal{H}$ and $\mathcal{M}_4(\mathcal{R})$, sometimes even appearing to be in both places at once!

5. Strategy. We want to translate the problem of orthogonally diagonalizing a 4×4 symmetric matrix into a corresponding problem in $\mathcal{H} \otimes \mathcal{H}$. To this end, we investigate each of the questions listed below.

- What does a symmetric matrix look like in $\mathcal{H} \otimes \mathcal{H}$?
- What does a diagonal matrix look like in $\mathcal{H} \otimes \mathcal{H}$?
- What does an orthogonal matrix correspond to in $\mathcal{H} \otimes \mathcal{H}$?
- How does one represent an orthogonal similarity in $\mathcal{H} \otimes \mathcal{H}$?

6. Quaternion representations. The decomposition of $\mathcal{H}$ as the direct sum $\mathcal{R} \oplus \mathcal{P}$ induces a decomposition of $\mathcal{H} \otimes \mathcal{H}$:

$$\begin{aligned} \mathcal{H} \otimes \mathcal{H} &\cong (\mathcal{R} \oplus \mathcal{P}) \otimes (\mathcal{R} \oplus \mathcal{P}) \\ &\cong \Big\{(\mathcal{R} \otimes \mathcal{R}) \oplus (\mathcal{P} \otimes \mathcal{P})\Big\} \ \oplus \ \Big\{(\mathcal{P} \otimes \mathcal{R}) \oplus (\mathcal{R} \otimes \mathcal{P})\Big\}. \end{aligned} \tag{8}$$

Let $\mathcal{S} = (\mathcal{R} \otimes \mathcal{R}) \oplus (\mathcal{P} \otimes \mathcal{P})$, and $\mathcal{K} = (\mathcal{P} \otimes \mathcal{R}) \oplus (\mathcal{R} \otimes \mathcal{P})$. Then $\mathcal{S}$ and $\mathcal{K}$ are eigenspaces for the conjugation map on $\mathcal{H} \otimes \mathcal{H}$ corresponding to the eigenvalues 1 and -1, respectively. Since conjugation translates to transpose in $\mathcal{M}_4(\mathcal{R})$, every element of $\mathcal{S}$ represents a 4×4 symmetric matrix, and every element of $\mathcal{K}$ represents a 4×4

skew-symmetric matrix. Observing that $\mathcal{S}$ is a 10-dimensional subspace of $\mathcal{H}\otimes\mathcal{H}$ and $\mathcal{K}$ is a 6-dimensional subspace, it follows that

$$\phi(\mathcal{S}) = \{S \in \mathcal{M}_4(\mathcal{R}) : S \text{ is symmetric}\}$$
$$\text{and } \phi(\mathcal{K}) = \{K \in \mathcal{M}_4(\mathcal{R}) : K \text{ is skew-symmetric}\}.$$

For notational convenience $\mathcal{S}$ will be used freely to denote either $\mathcal{S}$ or $\phi(\mathcal{S})$; context will make it clear which is intended. (Similarly for $\mathcal{K}$, S, K.) (*Note.* See Remark at end of §4.2.) We have thus established the following propositions.

PROPOSITION 2. *The following are equivalent.*

1. S *is* 4×4 *symmetric.*
2. $S \in (\mathcal{R}\otimes\mathcal{R}) \oplus (\mathcal{P}\otimes\mathcal{P})$.
3. *There exist pure quaternions* $p, q,$ *and* $r,$ *and* $c \in \mathcal{R}$ *such that*
$$S = c1\otimes 1 + p\otimes i + q\otimes j + r\otimes k.$$

PROPOSITION 3 (HACON). *The following are equivalent.*

1. K *is* 4×4 *skew-symmetric.*
2. $K \in (\mathcal{P}\otimes\mathcal{R}) \oplus (\mathcal{R}\otimes\mathcal{P})$.
3. *There exist pure quaternions* p, q *such that* $K = p\otimes 1 + 1\otimes q$.

Indeed, the standard basis $\{e_r \otimes e_s : 1 \le r, s \le 4\}$ for $\mathcal{H}\otimes\mathcal{H}$ gives, via ϕ, a beautiful basis for $\mathcal{M}_4(\mathcal{R})$ comprised entirely of *orthogonal* matrices, ten of them symmetric and the remaining six skew-symmetric.[4] One can use this basis, which is listed in the Appendix, to calculate the quaternion representation of any 4×4 matrix. For a symmetric matrix $S = [s_{\ell m}] = c1\otimes 1 + p\otimes i + q\otimes j + r\otimes k$, the pure quaternions p, q, r, and the scalar c are given by

$$c = \frac{1}{4}\operatorname{trace}(S), \tag{9}$$

$$p_1 = \frac{1}{4}(s_{11} + s_{22} - s_{33} - s_{44}), \quad p_2 = \frac{1}{2}(s_{23} + s_{14}), \quad p_3 = \frac{1}{2}(s_{24} - s_{13}), \tag{10}$$

$$q_1 = \frac{1}{2}(s_{23} - s_{14}), \quad q_2 = \frac{1}{4}(s_{11} - s_{22} + s_{33} - s_{44}), \quad q_3 = \frac{1}{2}(s_{34} + s_{12}), \tag{11}$$

$$r_1 = \frac{1}{2}(s_{24} + s_{13}), \quad r_2 = \frac{1}{2}(s_{34} - s_{12}), \quad r_3 = \frac{1}{4}(s_{11} - s_{22} - s_{33} + s_{44}). \tag{12}$$

The corresponding calculation for a skew-symmetric matrix $K = [k_{\ell m}] = p\otimes 1 + 1\otimes q$ is even simpler:

$$p_1 = -\frac{1}{2}(k_{12} + k_{34}), \quad p_2 = \frac{1}{2}(k_{24} - k_{13}), \quad p_3 = -\frac{1}{2}(k_{14} + k_{23}), \tag{13}$$

$$q_1 = \frac{1}{2}(k_{12} - k_{34}), \quad q_2 = \frac{1}{2}(k_{24} + k_{13}), \quad q_3 = \frac{1}{2}(k_{14} - k_{23}). \tag{14}$$

Further examination of the quaternion basis reveals the following simple characterizations of the canonical forms of interest to us.

PROPOSITION 4. (a) *A* 4×4 *matrix is diagonal if and only if it can be expressed as*

$$c_0 1\otimes 1 + c_1 i\otimes i + c_2 j\otimes j + c_3 k\otimes k, \quad \textit{for some } c_0, c_1, c_2, c_3 \in \mathcal{R}.$$

[4] Any pair of elements from this basis either commute or anticommute. While this property may at first appear striking, it is a natural consequence of the way multiplication is defined in a tensor product, and the fact that the basis elements $\{1, i, j, k\}$ of $\mathcal{H}$ either commute or anticommute.

(b) (Hacon) *A 4×4 skew-symmetric matrix is in real Schur (2×2 block diagonal) form if and only if it can be expressed as $si\otimes 1 + 1\otimes ti$, for some $s,t\in\mathcal{R}$.*

7. Orthogonal similarities in $\mathcal{H}\otimes\mathcal{H}$.

7.1. Rotations. Within a year of the discovery of the quaternions, Hamilton and Cayley had found a connection between orthogonal maps of $\mathcal{R}^3$ and quaternions [16], [5], [10]. About a decade later, in 1855, Cayley observed that every orthogonal map of $\mathcal{R}^4$ can be represented by pairs of quaternions [6], [10]. We now outline these results in modern terminology.

Let $\mathcal{GL}_n(\mathcal{R})$ denote the group of $n\times n$ nonsingular matrices. The orthogonal and special orthogonal subgroups are, respectively,

$$\mathcal{O}(n) = \{R \in \mathcal{M}_n(\mathcal{R}) : R^tR = RR^t = I_n\},$$
$$\mathcal{SO}(n) = \{R \in \mathcal{O}(n) : \det R = +1\}.$$

We refer to elements of $\mathcal{SO}(n)$ as n-dimensional rotations. Let $\mathcal{U}$ denote the set of unit quaternions,

$$\mathcal{U} = \{u \in \mathcal{H} : u\overline{u} = 1\},$$

which is a subgroup of $\mathcal{H}$ under multiplication.

PROPOSITION 5. *The map*

$$\mathcal{U}\times\mathcal{U} \longrightarrow \mathcal{GL}_4(\mathcal{R}),$$
$$(x,y) \mapsto \phi(x\otimes y),$$

is a group homomorphism with image $\mathcal{SO}(4)$ and kernel $\{(1,1),(-1,-1)\}$.

COROLLARY 1. *The map*

$$\mathcal{U} \longrightarrow \mathcal{GL}_4(\mathcal{R}),$$
$$x \mapsto \phi(x\otimes x)$$

is a group homomorphism with kernel $\{1,-1\}$ and image the set of all matrices in $\mathcal{SO}(4)$ of the form

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot \end{pmatrix},$$

which can be interpreted as rotations of $\mathcal{P}\cong\mathcal{R}^3$.

We resist the temptation to provide a detailed proof of these results here as they are "well known" [7], [10]. Instead we confine ourselves to pointing out why $\phi(x\otimes y)$ is in $\mathcal{SO}(4)$. Observe that $(x\otimes y)(\overline{x}\otimes\overline{y}) = (x\overline{x}\otimes y\overline{y}) = 1\otimes 1$. But conjugation in $\mathcal{H}\otimes\mathcal{H}$ corresponds to transpose in $\mathcal{M}_4(\mathcal{R})$, so $\phi(x\otimes y)$ must be an orthogonal matrix. The continuity of the determinant, together with the connectedness of $\mathcal{U}\times\mathcal{U}$ imply that $\det(\phi(x\otimes y))$ is positive.

To see why $\phi(x\otimes x)$ has the form stated in the corollary, recall that $\phi(x\otimes x)$ is the matrix encoding the map $q\mapsto xq\overline{x}$ in the standard basis (see §4.2). The first column of this matrix is the vector representing the quaternion $x1\overline{x} = x\overline{x} = 1$, i.e., e_0. Orthogonality of the matrix now forces the rest.

7.2. Similarities. Let $R \in \mathcal{SO}(4)$. Then by Proposition 5, there exist $x, y \in \mathcal{U}$ such that $\phi(x \otimes y) = R$. Consider an element of $\mathcal{H}\otimes\mathcal{H}$ of the form $p \otimes q$. By Proposition 1, the similarity $R\,\phi(p \otimes q)R^t$ in $\mathcal{M}_4(\mathcal{R})$ translates to

$$(x \otimes y)(p \otimes q)(\overline{x} \otimes \overline{y})$$

in $\mathcal{H}\otimes\mathcal{H}$. Using the definition of multiplication in a tensor product, this becomes

$$(xp\overline{x}) \otimes (yq\overline{y}).$$

Now comes the crucial observation: $xp\overline{x}$ is just the image of p under the rotation $\phi(x \otimes x)$, while $yq\overline{y}$ is the image of q under the rotation $\phi(y \otimes y)$. By Corollary 1, both of these rotations act in a nontrivial way only on the pure quaternion parts of p and q. Since a general element of $\mathcal{H}\otimes\mathcal{H}$ is a linear combination of elements of the form $p \otimes q$, we have:

The effect of a 4×4 *special orthogonal similarity on a* 4×4 *matrix can be reduced to the* independent *action of two* 3*-dimensional rotations.*

8. The skew-symmetric algorithm. We have everything we need to outline how Hacon's method [14] for transforming a 4×4 skew-symmetric matrix K directly into its real Schur form works. By Proposition 3, there exist pure quaternions p, q such that $K = p \otimes 1 + 1 \otimes q$. Let $R = x \otimes y \in \mathcal{SO}(4)$. Then

$$RKR^t = xp\overline{x} \otimes 1 + 1 \otimes yq\overline{y}.$$

By Proposition 4(b), all that remains is to find two 3-dimensional rotations: $\phi(x \otimes x)$ to rotate the 3-vector p into the i-direction and $\phi(y \otimes y)$ to rotate the 3-vector q into the i-direction. Hacon [14] shows that taking

$$x = \frac{|p| - ip}{|\,|p| - ip\,|}, \qquad y = \frac{|q| - iq}{|\,|q| - iq\,|} \tag{15}$$

achieves this goal. That is,

$$RKR^t = si \otimes 1 + 1 \otimes ti = \begin{pmatrix} 0 & t-s & 0 & 0 \\ s-t & 0 & 0 & 0 \\ 0 & 0 & 0 & -s-t \\ 0 & 0 & s+t & 0 \end{pmatrix},$$

where $s = |p|$, $t = |q|$.

The matrix $R = x \otimes y$ can be computed from (15) and (6) as the product of the matrices

$$\frac{1}{|\,|p| - ip\,|} \begin{pmatrix} |p| + p_1 & 0 & -p_3 & p_2 \\ 0 & |p| + p_1 & p_2 & p_3 \\ p_3 & -p_2 & |p| + p_1 & 0 \\ -p_2 & -p_3 & 0 & |p| + p_1 \end{pmatrix} \tag{16}$$

and

$$\frac{1}{|\,|q| - iq\,|} \begin{pmatrix} |q| + q_1 & 0 & q_3 & -q_2 \\ 0 & |q| + q_1 & q_2 & q_3 \\ -q_3 & -q_2 & |q| + q_1 & 0 \\ q_2 & -q_3 & 0 & |q| + q_1 \end{pmatrix}, \tag{17}$$

where the pure quaternions p and q are computed from K via (13) and (14).

9. The symmetric algorithm. Let $S \in \mathcal{S}$. By Proposition 2, there exist $p, q, r \in \mathcal{P}$, and $c \in \mathcal{R}$ such that

$$S - c1 \otimes 1 \;=\; S - cI \;=\; p \otimes i + q \otimes j + r \otimes k \quad \in \quad \mathcal{P} \otimes \mathcal{P}.$$

Certainly it suffices to diagonalize $\widehat{S} = S - cI$, since RSR^t is diagonal $\Leftrightarrow$ $R(S - cI)R^t$ is diagonal. Now for $R = x \otimes y \in \mathcal{SO}(4)$ we have

$$R\widehat{S}R^t \;=\; xp\overline{x} \otimes yi\overline{y} \;+\; xq\overline{x} \otimes yj\overline{y} \;+\; xr\overline{x} \otimes yk\overline{y}. \tag{18}$$

If R is to diagonalize $\widehat{S}$, then by Proposition 4, the three-dimensional rotation $\phi(x \otimes x)$ must align the *triple of* 3*-vectors* $\{p, q, r\}$ along the *orthogonal* triple $\{i, j, \pm k\}$. Since $\{p, q, r\}$ is in general an oblique triple, this is clearly impossible. In the skew-symmetric case, by contrast, the problem reduces to rotating just *one* given 3-vector into a specified direction (granted, this must be done twice, but the rotations are completely independent of one another). This in a nutshell is why the symmetric case is more complicated and requires a new idea.

What we need is a different tensor decomposition of $\widehat{S}$,

$$a_1 \otimes b_1 + a_2 \otimes b_2 + a_3 \otimes b_3 \quad \in \quad \mathcal{P} \otimes \mathcal{P},$$

with the property that *both* triples $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$ are *orthogonal.* Then clearly it will be possible to independently rotate each triple into alignment with $\{i, j, \pm k\}$, thus diagonalizing $\widehat{S}$. This is where the vector space isomorphism $\psi : \mathcal{P} \otimes \mathcal{P} \to \mathcal{M}_3(\mathcal{R})$, introduced in §4.1, and the singular value decomposition (SVD) [13, pp. 70–72] work together beautifully to produce the "right" tensor decomposition of $\widehat{S}$.

Since the canonical inclusion map embeds $\mathcal{P} \otimes \mathcal{P}$ into $\mathcal{H} \otimes \mathcal{H}$, every element of $\mathcal{P} \otimes \mathcal{P}$ can be associated with a 3×3 matrix as well as a quite different 4×4 matrix:[5]

$$\begin{array}{ccc} \mathcal{P} \otimes \mathcal{P} & \xrightarrow{\;\psi\;} & \mathcal{M}_3(\mathcal{R}) \\ \downarrow & & \\ \mathcal{H} \otimes \mathcal{H} & \xrightarrow{\;\phi\;} & \mathcal{M}_4(\mathcal{R}). \end{array}$$

Successively exploiting the properties of ψ, we get

$$\begin{aligned} \psi(\widehat{S}) &= \psi(p \otimes i + q \otimes j + r \otimes k) \\ &= pe_1^t + qe_2^t + re_3^t && \text{(linearity of } \psi \text{ and (5))} \\ &= \begin{pmatrix} p_1 & q_1 & r_1 \\ p_2 & q_2 & r_2 \\ p_3 & q_3 & r_3 \end{pmatrix} \\ &= \sigma_1 u_1 v_1^t + \sigma_2 u_2 v_2^t + \sigma_3 u_3 v_3^t && \text{(SVD)} \\ &= \psi(\sigma_1 u_1 \otimes v_1 + \sigma_2 u_2 \otimes v_2 + \sigma_3 u_3 \otimes v_3) && \text{(linearity of } \psi \text{ and (5))} \end{aligned}$$

But ψ is bijective, so we must have

$$\widehat{S} = p \otimes i \;+\; q \otimes j \;+\; r \otimes k \;=\; \sigma_1 u_1 \otimes v_1 \;+\; \sigma_2 u_2 \otimes v_2 \;+\; \sigma_3 u_3 \otimes v_3.$$

---

[5] For example, $\psi(i \otimes i) = e_1 e_1^t = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, but $\phi(i \otimes i) = \begin{pmatrix} I_2 & 0 \\ 0 & -I_2 \end{pmatrix}$.

Since $\{u_1, u_2, u_3\}$ and $\{v_1, v_2, v_3\}$ are orthonormal triples, the "right" tensor decomposition of a 4×4 symmetric matrix has thus been obtained from the SVD of its associated 3×3 matrix.

Although it is geometrically clear that a rotation of $\mathcal{P}$ that aligns a given orthonormal triple $\{u_1, u_2, u_3\}$ with $\{i, j, \pm k\}$ exists, trying to compute it all at once is somewhat complicated. Instead we can align these triples in stages. Begin by choosing $R = x \otimes y$ so that $\phi(x \otimes x)$ rotates any one of the left singular vectors into i, and $\phi(y \otimes y)$ rotates the *corresponding* right singular vector into i. Then the remaining singular vectors will perforce be moved into the jk-plane. Does it matter which left-right singular vector pair is rotated into i? The next section provides the surprising answer.

For now, let σ_1, σ_2, and σ_3 denote the singular values of $\psi(\widehat{S})$, not necessarily in decreasing order. And let R align u_1 and v_1 with i. Then

$$R\widehat{S}R^t = \sigma_1 i \otimes i + \sigma_2 x u_2 \overline{x} \otimes y v_2 \overline{y} + \sigma_3 x u_3 \overline{x} \otimes y v_3 \overline{y}, \tag{19}$$

where $xu_2\overline{x}, yv_2\overline{y}, xu_3\overline{x}$, and $yv_3\overline{y}$ are each linear combinations of j and k only. Thus $R\widehat{S}R^t$ is a linear combination of $i \otimes i, j \otimes j, k \otimes k, j \otimes k$, and $k \otimes j$. A quick look at the quaternion basis in the appendix reveals that $R\widehat{S}R^t$ and hence, RSR^t is already 2×2 block diagonal!

One may continue the alignment process by constructing a second orthogonal matrix $Q = z \otimes w$, where $\phi(z \otimes z)$ is a three-dimensional rotation with axis i that aligns $xu_2\overline{x}$ with j; and $\phi(w \otimes w)$ is a three-dimensional rotation with axis i that rotates $yv_2\overline{y}$ into j. Then $xu_3\overline{x}$ and $yv_3\overline{y}$ will necessarily be aligned with $\pm k$, and thus $QR\widehat{S}R^tQ^t$ (and hence $QRSR^tQ^t$) will be diagonal. (Alternatively, two ordinary Jacobi rotations in the $(1, 2)$ and $(3, 4)$ planes could be used to achieve diagonalization.)

We remark that the matrix $R = x \otimes y$ is given by the product of the matrices in (16) and (17), with the left singular vector that R aligns with i playing the role of p and the corresponding right singular vector playing the role of q. By exploiting the special structure of these matrices and noting that $|p| = 1 = |q|$ in this case, their product can be computed for the small price of 14 additions, 14 multiplications, and one square root.

The major cost of computing R lies in finding a left-right singular vector pair of a 3×3 matrix. Note that the orthogonality of R does not depend on the accuracy of the singular vector pair used to construct it—as long as the vectors are unit, R will be orthogonal. During the early iterations of the quaternion-Jacobi method, it may perhaps not be necessary to compute the vectors with hair-splitting accuracy, since the annihilated elements will shortly be resurrected. Bear in mind also that the complete SVD is *not* needed—only one left-right singular vector pair is called for. We see in the next section that the pair corresponding to the largest singular value is the one to compute. What then is the best way to carry out this task? Among the many schemes to be evaluated are several iterative techniques and one direct method, due to Bojanczyk and Lutoborski [2], that gives closed form formulae for the eigenvectors of a 3×3 symmetric matrix.

A matrix representation of Q can be obtained analogously. However, we see in §12 that in practice block diagonalization suffices, so Q need never be computed.

10. Eigenvalues from singular values. Assume that $R, Q \in \mathcal{SO}(4)$ are chosen as described at the end of §9, that is,

$$R \text{ rotates } u_1, v_1 \text{ into } i, \tag{20}$$

(21) QR rotates u_1, v_1 into i, and u_2, v_2 into j.

Equation (19) suggests there is a connection between the singular values of $T = \psi(\widehat{S}) = [p\ q\ r]$ and the eigenvalues of S.

PROPOSITION 6. *If $\sigma_1 \geq \sigma_2 \geq \sigma_3$ are the singular values of $T = [p\ q\ r]$, then the eigenvalues of $S - cI = p \otimes i + q \otimes j + r \otimes k$ are*

$$\sigma_1 + \sigma_2 + \tau\sigma_3, \quad \sigma_1 - \sigma_2 - \tau\sigma_3, \quad -\sigma_1 + \sigma_2 - \tau\sigma_3, \quad -\sigma_1 - \sigma_2 + \tau\sigma_3,$$

where $\tau = \operatorname{sign}(\det(T))$. (If $\det(T) = 0$, set $\tau = 0$).

Proof. Let $\sigma_1 u_1 v_1^t + \sigma_2 u_2 v_2^t + \sigma_3 u_3 v_3^t$ be the SVD of T. The sign of $\det(T)$ determines whether T is orientation preserving or reversing. Since $T(v_\ell) = \sigma_\ell u_\ell$ for $1 \leq \ell \leq 3$, the triples $\{u_1, u_2, u_3\}$ and $\{v_1, v_2, v_3\}$ will have the *same* "handedness" when $\tau > 0$. That is, either both can be rotated into $\{i, j, k\}$ or both can be rotated into $\{i, j, -k\}$. On the other hand, $\tau < 0$ means that $\{u_1, u_2, u_3\}$ and $\{v_1, v_2, v_3\}$ have *opposite* handedness, i.e., one triple can be rotated into $\{i, j, k\}$ and the other into $\{i, j, -k\}$. We have seen that

$$\begin{aligned} S - cI &= p \otimes i + q \otimes j + r \otimes k \\ &= \sigma_1 u_1 \otimes v_1 + \sigma_2 u_2 \otimes v_2 + \sigma_3 u_3 \otimes v_3, \end{aligned}$$

so choosing rotations $R, Q \in \mathcal{SO}(4)$ as described in (20)–(21) gives

$$\begin{aligned} QR(S - cI)R^tQ^t &= \begin{cases} \sigma_1 i \otimes i + \sigma_2 j \otimes j + \sigma_3 k \otimes k & \text{if } \tau > 0, \\ \sigma_1 i \otimes i + \sigma_2 j \otimes j - \sigma_3 k \otimes k & \text{if } \tau < 0, \end{cases} \\ &= \operatorname{diag}(\sigma_1 + \sigma_2 + \tau\sigma_3, \sigma_1 - \sigma_2 - \tau\sigma_3, -\sigma_1 + \sigma_2 - \tau\sigma_3, \\ &\qquad -\sigma_1 - \sigma_2 + \tau\sigma_3). \end{aligned}$$

The case $\tau = 0$ is even simpler. In this situation $\sigma_3 = 0$, so

$$\begin{aligned} QR(S - cI)R^tQ^t &= \sigma_1 i \otimes i + \sigma_2 j \otimes j \\ &= \operatorname{diag}(\sigma_1 + \sigma_2, \sigma_1 - \sigma_2, -\sigma_1 + \sigma_2, -\sigma_1 - \sigma_2), \end{aligned}$$

which establishes the result. □

The alert reader may notice that, in the above proof, the hypothesis $\sigma_1 \geq \sigma_2 \geq \sigma_3$ is used only in the case when $\det(T) = 0$, and then only to conclude that σ_3 must be zero. When T is nonsingular, the expressions for the eigenvalues of $S - cI$ in terms of the singular values of T remain valid under any permutation of the σ_i's. That is, the ordering of the σ_i's has no effect on the *set* of eigenvalues of $S - cI$ (as expected), but it does affect the order in which these eigenvalues appear on the diagonal of $QR(S - cI)R^tQ^t$. It is to this issue that we turn next.

11. Sorting similarities. For notational convenience, denote the eigenvalues of $S - cI$ by λ_ℓ, $1 \leq \ell \leq 4$, where

(22) $$\lambda_1 = \sigma_1 + \sigma_2 + \tau\sigma_3,$$

(23) $$\lambda_2 = \sigma_1 - \sigma_2 - \tau\sigma_3,$$

(24) $$\lambda_3 = -\sigma_1 + \sigma_2 - \tau\sigma_3,$$

(25) $$\lambda_4 = -\sigma_1 - \sigma_2 + \tau\sigma_3.$$

Note that in the course of proving the previous proposition, we showed that

$$QR(S - cI)R^tQ^t = \operatorname{diag}(\lambda_1, \lambda_2, \lambda_3, \lambda_4).$$

We now ask, what conditions on R and Q will ensure that the eigenvalues of $S - cI$ appear in decreasing order on the diagonal? This question is particularly relevant because Mascarenhas [24] has recently shown that choosing the sorting angle for the plane rotations used in the Jacobi method can lead to significant improvement in performance when orderings with higher than quadratic rates of convergence are used. (The sorting angle eventually becomes the usual (small) angle in the plane-Jacobi methods.) If one could construct a 4×4 analogue of the 2×2 sorting similarity, one would expect a similar improvement in performance. Happily, the construction is unexpectedly simple. But first, a definition.

DEFINITION 2. *Let $A \in \mathcal{M}_n(\mathcal{R})$ have real eigenvalues. Let $B \in \mathcal{M}_n(\mathcal{R})$ block diagonalize A, i.e., $BAB^{-1} = \operatorname{diag}(\widetilde{A}_1, \widetilde{A}_2, \ldots, \widetilde{A}_m)$ where each $\widetilde{A}_\ell$ is some $n_\ell \times n_\ell$ matrix. If every eigenvalue of $\widetilde{A}_\ell$ is greater than or equal to every eigenvalue of $\widetilde{A}_{\ell+1}$, for $1 \le \ell \le m-1 \le n-1$, then BAB^{-1} is called a sorting similarity.*

PROPOSITION 7. *Suppose $\sigma_1 \ge \sigma_2 \ge \sigma_3$. If R and Q are chosen according to* (20)–(21), *then QR is a sorting similarity for S.*

Proof. $\lambda_1 \ge \lambda_2 \Leftrightarrow \sigma_1 + \sigma_2 + \tau\sigma_3 \ge \sigma_1 - \sigma_2 - \tau\sigma_3 \Leftrightarrow \sigma_2 \ge -\tau\sigma_3$, which follows from $\sigma_2 \ge \sigma_3$. Next, $\lambda_2 \ge \lambda_3 \Leftrightarrow \sigma_1 - \sigma_2 - \tau\sigma_3 \ge -\sigma_1 + \sigma_2 - \tau\sigma_3 \Leftrightarrow \sigma_1 \ge \sigma_2$. Finally, $\lambda_3 \ge \lambda_4 \Leftrightarrow -\sigma_1 + \sigma_2 - \tau\sigma_3 \ge -\sigma_1 - \sigma_2 + \tau\sigma_3 \Leftrightarrow \sigma_2 \ge \tau\sigma_3$, which again follows from $\sigma_2 \ge \sigma_3$. □

PROPOSITION 8. *Suppose $\sigma_1 \ge \sigma_2 \ge \sigma_3$. If R is chosen according to* (20), *then R is a sorting similarity for S. That is, the eigenvalues of the upper 2×2 block of RSR^t are larger than those of the lower 2×2 block.*

Proof. We know from §9 that

$$R(S - cI)R^t = \sigma_1 i \otimes i + c_{jj} j \otimes j + c_{kk} k \otimes k + c_{jk} j \otimes k + c_{kj} k \otimes j,$$

where the $c_{\ell m}$'s are real constants. Hence by the discussion following (19), $R(S-cI)R^t$ is 2×2 block diagonal. Expressing this in matrix form we have

$$R(S - cI)R^t = \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix},$$

where

$$S_1 = \begin{pmatrix} \sigma_1 + c_{jj} + c_{kk} & -c_{jk} + c_{kj} \\ -c_{jk} + c_{kj} & \sigma_1 - c_{jj} - c_{kk} \end{pmatrix}$$

and

$$S_2 = \begin{pmatrix} -\sigma_1 + c_{jj} - c_{kk} & c_{jk} + c_{kj} \\ c_{jk} + c_{kj} & -\sigma_1 - c_{jj} + c_{kk} \end{pmatrix}.$$

The 2×2 block diagonal form implies that the eigenvalues of S_1 must be two of the eigenvalues $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ of $S - cI$. From Propositions 6 and 7, the sum of the two largest eigenvalues of $S - cI$ is $\lambda_1 + \lambda_2 = 2\sigma_1$; of course the sum of any other pair $\lambda_i + \lambda_j$ can be no bigger than this. But we also have $\operatorname{trace}(S_1) = \Sigma(\text{eigenvalues of } S_1) = 2\sigma_1$. Therefore λ_1 and λ_2 must be eigenvalues of S_1. Consequently λ_3, λ_4 must be eigenvalues of S_2, so by Proposition 7 we are done. □

12. Preliminary experimental results. In practice, one can partition a $2n \times 2n$ symmetric matrix A into contiguous 2×2 blocks, denoted by $A_{\ell m}$, with $1 \le \ell, m \le n$. Then annihilate each off-diagonal block $A_{\ell m}$ with $\ell \ne m$ according to some order,

thereby obtaining a Jacobi-type method. Complete diagonalization of the target 4×4 submatrix is unnecessary (hence wasteful): block diagonalization incurs no penalty in the form of additional sweeps or additional iterations per sweep. This is because the weight of the annihilated (pivot) block moves onto the two corresponding 2×2 diagonal blocks, and stays there! In other words, if we define

$$\text{BlockOff}^2(A) = \|A\|_F^2 - \sum_{i=1}^{n} \|A_{ii}\|_F^2,$$

then just as in traditional Jacobi methods,

$$\text{BlockOff}^2(A^{(\ell+1)}) = \text{BlockOff}^2(A^{(\ell)}) - 2\|A_{ij}^{(\ell)}\|_F^2.$$

Here $A^{(\ell)}$ denotes the matrix A after ℓ block annihilations, and ij denotes the position of the ℓth pivot block. Thus $\{\text{BlockOff}^2(A^{(\ell)})\}$ is a decreasing sequence.

It is observed experimentally that the sequence $\{A^{(\ell)}\}$ converges to a (2×2)- block-diagonal matrix, which can then be diagonalized with negligible cost. We present below the results of a MATLAB implementation of Jacobi and quaternion-Jacobi methods on a random 64×64 symmetric matrix with entries uniformly distributed in $[-1, 1]$. This behavior seems to be typical, but more experimentation is needed.

TABLE 1

$n = 64$	Odd-even ordering; sorting similarity			
	Jacobi		quaternion-Jacobi	
Sweep	Number of 2×2 Rotations	BlockOff	Number of 4×4 Rotations	BlockOff
0		$1.800e+01$		$1.800e+01$
1	2016	$1.067e+01$	496	$1.032e+01$
2	2016	$2.883e+00$	496	$2.609e+00$
3	2016	$4.633e-01$	496	$3.301e-01$
4	2016	$1.840e-02$	496	$5.025e-03$
5	2016	$2.877e-05$	496	$1.261e-06$
6	2016	$8.111e-11$	490	$6.702e-14$
7	541	$1.218e-13$		

TABLE 2

$n = 64$	Row-cyclic ordering; sorting similarity			
	Jacobi		quaternion-Jacobi	
Sweep	Number of 2×2 Rotations	BlockOff	Number of 4×4 Rotations	BlockOff
0		$1.800e+01$		$1.800e+01$
1	2016	$1.120e+01$	496	$1.066e+01$
2	2016	$3.664e+00$	496	$3.171e+00$
3	2016	$8.714e-01$	496	$6.249e-01$
4	2016	$1.208e-01$	496	$7.069e-02$
5	2016	$3.321e-03$	496	$1.332e-03$
6	1976	$4.324e-06$	468	$5.580e-07$
7	1279	$8.445e-12$	158	$1.529e-13$
8	1270	$1.230e-13$		

The same matrix is used in both tables; Table 1 compares the two methods when the odd-even ordering is used, Table 2 when the row-cyclic order is used.

13. Convergence. Sorting similarities lead to a quaternion-Jacobi method that converges under rather general quasicyclic orderings. Quasicyclic orderings were described by Henrici [18], who credits their invention to Hestenes. They can be characterized as orderings that have no arbitrarily long gaps between successive annihilations of any individual element.

THEOREM 1. *Let A_0 be an $n \times n$ symmetric matrix and let $A_{\ell+1} = R_\ell A_\ell R_\ell^t$ be the sequence of iterates of a quaternion-Jacobi method. If each R_ℓ effects a sorting, diagonalizing similarity on its target 4×4 submatrix, then the following are true.*

1. *The vector $L^\infty = \lim_{\ell\to\infty} \operatorname{diag}(A_\ell)$ always exists. That is, the sequence of diagonal vectors converges for any ordering, including those that fail to be quasicyclic.*
2. *If L^∞ has distinct entries, i.e., $L_r^\infty \neq L_s^\infty$ for all $r \neq s$, $1 \leq r, s \leq n$, and if the pivot ordering is quasi-cyclic, then*

$$\lim_{\ell\to\infty} A_\ell = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

where $\lambda_1 > \lambda_2 > \cdots > \lambda_n$ are the eigenvalues of A. Furthermore, the rate of convergence is asymptotically quadratic.

The proof of this theorem is quite long and will be published in a later paper. It contains results that encompass extremely general Jacobi-type methods and that grew out of an attempt to extend the ideas found in Mascarenhas' thesis [24]. Among the tools used is a majorization result due to Schur [19, pp. 192–193] and an SVD perturbation theorem due to Wedin [29], [31].

14. Extension to normal matrices. Combining the two algorithms leads to a quaternion-Jacobi method for real normal matrices. First, write the matrix A as the sum of a symmetric matrix S and a skew-symmetric matrix K. Since A is normal, S and K commute. Diagonalize S using the symmetric quaternion-Jacobi algorithm with sorting transformations at every iteration. This gives an orthogonal matrix U such that $UAU^t = D_s + UKU^t$, with the entries of the diagonal matrix D_s appearing in decreasing order. The skew-symmetric matrix $B_k = UKU^t$ is block-diagonal: since B_k commutes with D_s, the only nonzero elements of B_k occur in diagonal blocks corresponding to equal eigenvalues of D_s. Now each diagonal block of B_k can be independently transformed into its real Schur form using Hacon's skew-symmetric algorithm, without affecting D_s.

As is well known, the plane rotations that are so effective in diagonalizing 2×2 symmetric matrices leave all 2×2 skew-symmetric matrices unscathed.[6] The remarkable correspondence between $\mathcal{H} \otimes \mathcal{H}$ and $\mathcal{M}_4(\mathcal{R})$ gives us, for the first time, a Jacobi method that works in a similar manner for *both* symmetric and skew-symmetric matrices, exploiting and preserving their special structure while lending a unity to the eigenproblem for these two classes that has hitherto been lacking.

[6] Paardekooper [26] deserves mention for first addressing the skew-symmetric case.

A. The quaternion basis for $\mathcal{M}_4(\mathcal{R})$.

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad 1 \otimes 1 \qquad \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad 1 \otimes i \qquad \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} \quad 1 \otimes j \qquad \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix} \quad 1 \otimes k$$

$$\begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad i \otimes 1 \qquad \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \quad i \otimes i \qquad \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix} \quad i \otimes j \qquad \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad i \otimes k$$

$$\begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} \quad j \otimes 1 \qquad \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad j \otimes i \qquad \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \quad j \otimes j \qquad \begin{pmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad j \otimes k$$

$$\begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad k \otimes 1 \qquad \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad k \otimes i \qquad \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad k \otimes j \qquad \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad k \otimes k$$

Acknowledgments. I owe special thanks to my advisor, Professor P. J. Eberlein, for introducing me to Hacon's work and for suggesting the problem of extending it to the symmetric case. It is also a pleasure to thank Professor Eberlein, Nick Higham, Tom Laffey, Steve Mackey, and Don Schack for stimulating discussions, and Cleve Moler and the MATLAB team for providing such a convenient environment in which to test matrix algorithms.

REFERENCES

[1] G. Birkhoff and S. MacLane, *A Survey of Modern Algebra*, Macmillan, New York, 1977.
[2] A. W. Bojanczyk and A. Lutoborski, *Computation of the Euler angles of a symmetric* 3×3 *matrix*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 41–48.
[3] R. P. Brent and F. T. Luk, *The solution of singular-value and symmetric eigenvalue problems on multiprocessor arrays*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 69–84.
[4] T. Bröcker and T. tom Dieck, *Representations of Compact Lie Groups*, Springer-Verlag, New York, 1985.
[5] A. Cayley, *On certain results relating to quaternions*, in The Collected Mathematical Papers, Vol. 1, Johnson Reprint Company, New York, 1963, pp. 123–126. Originally published in The Philos. Mag., 26 (1845), pp. 141–145.
[6] ———, *Recherches ultérieures sur les déterminants gauches*, in The Collected Mathematical Papers, Vol. 2, Johnson Reprint Company, New York, 1963, pp. 202–215. Originally published in J. Reine Angew. Math., 50 (1855), pp. 299–313.
[7] H. S. M. Coxeter, *Quaternions and reflections*, Amer. Math. Monthly, 53 (1946), pp. 136–146.
[8] J. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.
[9] R. Dimitric and B. Goldsmith, *The mathematical tourist*, Math. Intelligencer, 11 (1989), pp. 29–30.
[10] H. D. Ebbinghaus, H. Hermes, F. Hirzebruch, M. Keocher, K. Mainzer, J. Neukirch, A. Prestel, and R. Remmert, *Numbers*, Springer-Verlag, New York, 1991.

[11] P. J. EBERLEIN, *On one-sided Jacobi methods for parallel computation*, SIAM J. Alg. Disc. Meth., 8 (1987), pp. 790–796.

[12] ———, *On using the Jacobi method on the hypercube*, in Proc. 2nd Conference on Hypercube Multiprocessors, M. T. Heath, ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1987, pp. 605–611.

[13] G. H. GOLUB AND C. VAN LOAN, *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, 1989.

[14] D. HACON, *Jacobi's method for skew-symmetric matrices*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 619–628.

[15] W. R. HAMILTON, in The Collected Mathematical Papers, H. Halberstam and R. E. Ingram, eds., Vol. 3, Cambridge University Press, 1967, pp. xx–xvi. Letter to his son Archibald Hamilton.

[16] ———, *On quaternions, or a new system of imaginaries in algebra; with some geometrical illustrations*, in The Collected Mathematical Papers, H. Halberstam and R. E. Ingram, eds., Vol. 3, Cambridge University Press, Cambridge, 1967, pp. 355–362. Originally published in Proc. Royal Irish Academy, 3 (1847), pp. 1–16; communicated November 1844.

[17] T. L. HANKINS, *Sir William Rowan Hamilton*, Johns Hopkins University Press, Baltimore, MD, 1980.

[18] P. HENRICI, *On the speed of convergence of cyclic and quasicyclic Jacobi methods for computing eigenvalues of Hermitian matrices*, J. Soc. Indust. Appl. Math, 6 (1958), pp. 144–162.

[19] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, New York, 1990.

[20] ———, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.

[21] C. G. J. JACOBI, *Über eine neue Auflösungart der bei Methode der kleinsten Quadrate vorkommenden lineären Gleichungen*, Astronomische Nachrichten, 22 (1845). Translated by G. W. Stewart, University of Maryland, Tech. Report TR-92-42, April 1992.

[22] ———, *Über ein leichtes Verfahren die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen*, J. Reine Angew. Math., 30 (1846), pp. 51–94.

[23] C. R. JOHNSON, T. LAFFEY, AND C.-K. LI, *Linear transformations on $M_n(R)$ that preserve the Ky Fan k-norm and a remarkable special case when $(n,k) = (4,2)$*, Linear Multilinear Algebra, 23 (1988), pp. 285–298.

[24] W. F. MASCARENHAS, *On the convergence of the Jacobi method for arbitrary orderings*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, 1991.

[25] R. MATHIAS, *Accurate eigensystem computations by Jacobi methods*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 565–578.

[26] M. PAARDEKOOPER, *An eigenvalue algorithm for skew-symmetric matrices*, Numer. Math., 17 (1971), pp. 189–202.

[27] R. S. PIERCE, *Associative Algebras*, Springer-Verlag, New York, 1982.

[28] A. H. SAMEH, *On Jacobi and Jacobi-like algorithms for a parallel computer*, Math. Comp., 25 (1971), pp. 579–590.

[29] G. W. STEWART, *Perturbation theory for the singular value decomposition*, Tech. Report TR-90-124, University of Maryland, College Park, 1990.

[30] B. L. VAN DER WAERDEN, *Hamilton's discovery of quaternions*, Math. Mag., 49 (1976), pp. 227–234.

[31] P. -Å. WEDIN, *Perturbation bounds in connection with singular value decomposition*, BIT, 12 (1972), pp. 99–111.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 436–447, April 1995

008

ON THE INDEX OF BLOCK UPPER TRIANGULAR MATRICES*

RAFAEL BRU†, JOAN JOSEP CLIMENT‡, AND MICHAEL NEUMANN§

Abstract. Let M be an upper block triangular matrix with A and B singular diagonal blocks. It is known that $\max\{\text{index}(A), \text{index}(B)\} \leq \text{index}(M) \leq \text{index}(A) + \text{index}(B)$. Recently, a necessary and sufficient condition has been given so that $\text{index}(M) = \text{index}(A) + \text{index}(B)$. In this paper we find various characterizations for $\text{index}(M)$ to take any specific values between $\max\{\text{index}(A), \text{index}(B)\}$ and $\text{index}(A) + \text{index}(B)$ which generalize previous results.

Key words. index, block triangular matrix, generalized eigenspace, generalized eigenvector, Drazin inverse, height, depth

AMS subject classifications. 15A21, 15A18

1. Introduction.

It is well known that the *index* of a singular matrix P is the size of the largest Jordan block corresponding to the eigenvalue zero. An equivalent definition is the smallest nonnegative integer k such that $\text{Ker}(P^k) = \text{Ker}(P^{k+1})$, where $\text{Ker}(\cdot)$ denotes the kernel of a matrix. In this paper we are concerned with the index of a block triangular matrix.

Let A and B be two singular matrices of dimensions $m \times m$ and $n \times n$, respectively. Given an $m \times n$ matrix X, without loss of generality, we consider the upper block triangular matrix

$$M = \begin{bmatrix} A & X \\ O & B \end{bmatrix}. \tag{1}$$

The determination of the index of the matrix M given in (1) in terms of the index of A and the index of B has been studied by different authors who used different approaches to the problem. For example, Meyer and Rose [8, Thm. 2.1] established that

$$\max\{\text{index}(A), \text{index}(B)\} \leq \text{index}(M) \leq \text{index}(A) + \text{index}(B). \tag{2}$$

For each nonnegative integer p, let X_p denote the matrix

$$X_p = \begin{cases} \displaystyle\sum_{i=1}^{p} A^{p-i} X B^{i-1} & \text{if } p \geq 1, \\ O & \text{if } p = 0. \end{cases} \tag{3}$$

Meyer and Rose studied the index of M by using the Drazin inverses of A and B and the matrix X_p defined by (3). They also obtained additional results which will permit

* Received by the editors August 5, 1992; accepted for publication (in revised form) by Carl Meyer, January 12, 1994.

† Departament de Matemàtica Aplicada, Universitat Politècnica de València, 46071 València, Spain (rbru@mat.upv.es). This research was supported by Spanish Dirección General de Investigación Cientifica y Técnica grant PB91-0535.

‡ Departament de Tecnologia Informàtica i Computació, Universitat d'Alacant, 03071 Alacant, Spain (jcliment@dtic.ua.es). This research was supported by Spanish Dirección General de Investigación Cientifica y Técnica grant PB91-0535.

§ Department of Mathematics, University of Connecticut, Storrs, Connecticut 06269–3009 (neumann@uconnvm.bitnet). This research was supported by National Science Foundation grants DMS-8901860 and DMS-9306357.

us here to sharpen the upper bound in (2). Another set of authors who considered the index of M as a function of X are Hershkowitz, Rothblum, and Schneider in [4]. They determined a necessary and sufficient condition on X so that $\operatorname{index}(M) = \operatorname{index}(A) + \operatorname{index}(B)$. Next, Johnson, Schreiner, and Elsner [5] considered the index of M when A and B have exactly one Jordan block for the eigenvalue zero in terms of the eigenvectors corresponding to Jordan chains of A^T and B. Finally, Johnson and Schreiner related the index of M and the Jordan structure of M when A has various Jordan blocks corresponding to the eigenvalue zero, but B has only one or two Jordan blocks in [6]. They further permit the Jordan structure of B corresponding to zero to also be arbitrary in [7].

Let k be an arbitrary but fixed integer in the permissible range of (2). In this paper we determine necessary and sufficient conditions under which the index of M is k. We develop our main results, viz. Theorems 2.6 and 2.7, in §2. These results permit us to generalize the result of Hershkowitz, Rothblum, and Schneider in [4, Thm. 6.8]. In §3 we consider the relationship between the results of §2 and the Drazin inverses of A and B. In so doing we generalize some of the results of Meyer and Rose [8, Thm. 2.2]. In §4 the results of Johnson, Schreiner, and Elsner [5, Thm. 8] are generalized in terms of the height and depth of the generalized eigenvectors of A^T and B without a need to limit the number of Jordan blocks of A^T and B. In §5 we summarize all the equivalences that have been established in the paper.

To make the paper more self contained we introduce now the concepts of height and depth. Let P be an $n \times n$ singular matrix. The *height* of a vector x with respect to P, $\operatorname{ht}_P(x)$, is the minimum nonnegative integer k such that $P^k x = 0$. The *depth* of a vector $x \neq 0$ with respect to P, $\operatorname{dp}_P(x)$, is the maximum nonnegative integer k such that $x = P^k y$ for some vector y. Finally, we denote by $E(P)$ the *generalized eigenspace* of P associated with the eigenvalue zero. We comment that the height and the depth of a vector in relation to a matrix and its properties have been studied by Bru, Rodman, and Schneider in [2].

2. Main results. We begin by observing that for each nonegative integer p,

$$M^p = \begin{bmatrix} A^p & X_p \\ O & B^p \end{bmatrix}, \qquad p \geq 0, \tag{4}$$

where X_p is given in (3). We further recall the following result concerning the index of the upper block triangular matrix M given in (1).

THEOREM 2.1 (Theorem 6.8 of [4]). *Let M be given by* (1). *Let $a = \operatorname{index}(A)$ and $b = \operatorname{index}(B)$. Then $\operatorname{index}(M) = a + b$ if and only if*

$$X[\operatorname{Im}(B^{b-1}) \cap \operatorname{Ker}(B)] \not\subseteq \operatorname{Im}(A) + \operatorname{Ker}(A^{a-1}).$$

Thus Theorem 2.1 gives a necessary and sufficient condition for the index of M to take the maximum possible value given in (2). Next we develop some results that allow us to prove Theorems 2.6 and 2.7 that establish the necessary and sufficient conditions so that the index of M is allowed to take any specific value between the upper and the lower bounds of (2), in similar terms to Theorem 2.1. For this purpose we require another result of Hershkowitz, Rothblum, and Schneider [4] which we state here.

LEMMA 2.2 (Lemma 3.2 of [4]). *Let M be given by* (1) *and let $v \in E(B)$. Then there exists a vector $x \in E(M)$ such that $x = \begin{bmatrix} u \\ v \end{bmatrix}$ for some m-vector u.*

The above lemma allows us to determine a certain lower bound on the index of M.

THEOREM 2.3. *Let M be given by* (1). *Let p and q be two positive integers and suppose that $1 \leq k \leq \min\{p, q\}$. Let X_k be given by* (3). *If*

$$X_k[\mathrm{Im}(B^{q-k}) \cap \mathrm{Ker}(B^k)] \not\subseteq \mathrm{Im}(A^k) + \mathrm{Ker}(A^{p-k}),$$

then $\mathrm{index}(M) \geq p + q - (k-1)$.

Proof. Let $y \in \mathrm{Im}(B^{q-k}) \cap \mathrm{Ker}(B^k)$ be a vector such that $X_k y \notin \mathrm{Im}(A^k) + \mathrm{Ker}(A^{p-k})$. Obviously $y \neq 0$, $y = B^{q-k}v$ for some n–vector v, and $B^k y = 0$. As

$$B^q v = B^k y = 0,$$

we have that $v \in E(B)$. Thus, by Lemma 2.2, there is a vector $x = \left[\begin{smallmatrix} u \\ v \end{smallmatrix}\right] \in E(M)$ for some m–vector u. But then, from (4),

$$\begin{aligned} M^q x &= \begin{bmatrix} A^k & X_k \\ O & B^k \end{bmatrix} \begin{bmatrix} A^{q-k} & X_{q-k} \\ O & B^{q-k} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \\ &= \begin{bmatrix} A^k(A^{q-k}u + X_{q-k}v) + X_k B^{q-k} v \\ B^k B^{q-k} v \end{bmatrix} \\ &= \begin{bmatrix} A^k z + X_k y \\ 0 \end{bmatrix}, \end{aligned}$$

where $z = A^{q-k}u + X_{q-k}v$. If $A^k z + X_k y = w \in \mathrm{Ker}(A^{p-k})$, then

$$X_k y = -A^k z + w \in \mathrm{Im}(A^k) + \mathrm{Ker}(A^{p-k}),$$

which is a contradiction. Thus $A^k z + X_k y = w \notin \mathrm{Ker}(A^{p-k})$. This means that

$$M^{p-k} M^q x = \begin{bmatrix} A^{p-k} & X_{p-k} \\ O & B^{p-k} \end{bmatrix} \begin{bmatrix} A^k z + X_k y \\ 0 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

so that $\mathrm{index}(M) \geq p + q - (k-1)$. □

A further simple, but necessary, lemma at this point is the following.

LEMMA 2.4. *Let P be a singular matrix. Let $p = \mathrm{index}(P)$ and suppose that $1 \leq k \leq p$. If $w \in \mathrm{Ker}(P^{p-k+1}) \setminus \mathrm{Ker}(P^{p-k})$, then $w \notin \mathrm{Im}(P^k) + \mathrm{Ker}(P^{p-k})$.*

Proof. Assume that $w \in \mathrm{Im}(P^k) + \mathrm{Ker}(P^{p-k})$. Then $w = P^k u + v$ with $P^{p-k} v = 0$. In this case,

$$P^{p-k} w = P^p u. \tag{5}$$

On the other hand,

$$0 = P^{p-k+1} w = P^{p+1} u.$$

Thus $u \in \mathrm{Ker}(P^{p+1}) = \mathrm{Ker}(P^p)$ and, by (5), $w \in \mathrm{Ker}(P^{p-k})$, which contradicts the hypothesis. Consequently $w \notin \mathrm{Im}(P^k) + \mathrm{Ker}(P^{p-k})$. □

When $k = 1$, Theorem 2.3 and Lemma 2.4 become Proposition 6.2 and Lemma 6.1 of [4], respectively. In fact the ideas we use in proving Theorem 2.3 and Lemma 2.4 follow closely the ideas used in [4].

Lemma 2.4 allows us to prove a partial converse to Theorem 2.3 as follows.

THEOREM 2.5. *Let M be given by* (1). *Let $a = \mathrm{index}(A)$ and $b = \mathrm{index}(B)$. Suppose $1 \leq k \leq \min\{a, b\}$ and let X_k be given by* (3). *If* $\mathrm{index}(M) = a + b - (k-1)$, *then*

$$X_k[\mathrm{Im}(B^{b-k}) \cap \mathrm{Ker}(B^k)] \not\subseteq \mathrm{Im}(A^k) + \mathrm{Ker}(A^{a-k}).$$

Proof. Since $\operatorname{index}(M) = a + b - (k-1)$, there is a vector $x \in E(M)$ such that

$$M^{a+b-(k-1)}x = 0 \quad \text{and} \quad M^{a+b-k}x \neq 0. \tag{6}$$

Suppose that $x = \begin{bmatrix} u \\ v \end{bmatrix}$, where u is an m–vector and v is an n–vector. From (4) and (6) we have that $B^{a+b-(k-1)}v = 0$ and hence $v \in E(B)$. Thus $B^b v = 0$. But then from (4) we have that

$$\begin{aligned} \begin{bmatrix} A^b u + X_b v \\ 0 \end{bmatrix} &= M^k M^{b-k} x \\ &= \begin{bmatrix} A^k & X_k \\ O & B^k \end{bmatrix} \begin{bmatrix} A^{b-k}u + X_{b-k}v \\ B^{b-k}v \end{bmatrix} \\ &= \begin{bmatrix} A^k z + X_k y \\ 0 \end{bmatrix}. \end{aligned}$$

This shows that

$$A^b u + X_b v = A^k z + X_k y, \tag{7}$$

where $y = B^{b-k}v$ and $z = A^{b-k}u + X_{b-k}v$. Clearly $y \in \operatorname{Im}(B^{b-k}) \cap \operatorname{Ker}(B^k)$. Now from (6) we have that

$$A^b u + X_b v \in \operatorname{Ker}(A^{a-k+1}) \setminus \operatorname{Ker}(A^{a-k}). \tag{8}$$

But then by (8), (7), and Lemma 2.4, we have that $A^k z + X_k y \notin \operatorname{Im}(A^k) + \operatorname{Ker}(A^{a-k})$, from which it follows that $X_k y \notin \operatorname{Im}(A^k) + \operatorname{Ker}(A^{a-k})$. This means that $X_k[\operatorname{Im}(B^{b-k}) \cap \operatorname{Ker}(B^k)] \not\subseteq \operatorname{Im}(A^k) + \operatorname{Ker}(A^{a-k})$. □

Remark 1. Observe that the case $k = 0$ is also possible in the above theorem since we have the trivial inclusion

$$X_0[\operatorname{Im}(B^{b-0}) \cap \operatorname{Ker}(B^0)] \subseteq \operatorname{Im}(A^0) + \operatorname{Ker}(A^{a-0}).$$

THEOREM 2.6. *Let M be given by* (1). *Suppose that $a = \operatorname{index}(A)$ and $b = \operatorname{index}(B)$ and assume that $1 \leq k \leq \min\{a, b\}$. Then $\operatorname{index}(M) = a + b - (k-1)$ if and only if the following conditions hold:*

1. $X_i[\operatorname{Im}(B^{b-i}) \cap \operatorname{Ker}(B^i)] \subseteq \operatorname{Im}(A^i) + \operatorname{Ker}(A^{a-i})$, $i = 0, 1, 2, \ldots, k-1$;
2. $X_k[\operatorname{Im}(B^{b-k}) \cap \operatorname{Ker}(B^k)] \not\subseteq \operatorname{Im}(A^k) + \operatorname{Ker}(A^{a-k})$, *where* X_j, $j = 0, 1, 2, \ldots, k$, *is given by* (3).

Proof. We prove the theorem by induction. For $k = 1$ the equivalence holds by Theorem 2.1 and Remark 1. Suppose then that the theorem holds for $1 \leq t \leq k-1$ and that Conditions 1 and 2 hold for such values of t. Then from Condition 1 we have that $\operatorname{index}(M) \leq a + b - (k-1)$. But from Condition 2 and Theorem 2.3 we see that $\operatorname{index}(M) \geq a + b - (k-1)$. Hence $\operatorname{index}(M) = a + b - (k-1)$.

Conversely, if $\operatorname{index}(M) = a + b - (k-1)$, then from Theorem 2.5 we obtain that

$$X_k[\operatorname{Im}(B^{b-k}) \cap \operatorname{Ker}(B^k)] \not\subseteq \operatorname{Im}(A^k) + \operatorname{Ker}(A^{a-k}),$$

and so Condition 2 holds. Next, if some of the inclusions in Condition 1 do not hold, then, by the induction hypothesis, we must have that $\operatorname{index}(M) = a + b - (t-1)$ for some t with $1 \leq t \leq k-1$. Hence,

$$\operatorname{index}(M) = a + b - (t-1) > a + b - (k-1) = \operatorname{index}(M),$$

a contradiction. Whence, Condition 1 holds. □

We observe that when $k = 1$, the results of Theorem 2.6 yield Theorem 2.1, whereas, if $k \leq \min\{a, b\}$, then Theorem 2.6 yields necessary and sufficient conditions so that $\text{index}(M) \geq \max\{a, b\} + 1$. In other words, Theorem 2.6 provides necessary and sufficient conditions for $\text{index}(M)$ to take an arbitrary, but fixed, value between $\max\{a, b\}+1$ and $a+b$. In the next theorem we determine the necessary and sufficient condition for $\text{index}(M)$ to attain the lower bound in (2).

THEOREM 2.7. *Let M be given by* (1). *Let $a = \text{index}(A)$ and $b = \text{index}(B)$. Then $\text{index}(M) = \max\{a, b\}$ if and only if*

$$X_i[\text{Im}(B^{b-i}) \cap \text{Ker}(B^i)] \subseteq \text{Im}(A^i) + \text{Ker}(A^{a-i}), \qquad i = 0, 1, 2, \ldots, \min\{a, b\},$$

where the X_i's, $i = 0, 1, 2, \ldots, \min\{a, b\}$, are given in (3).

Proof. Suppose that $\text{index}(M) = \max\{a, b\}$, but that at least one of the above inclusions does not hold. Let $1 \leq k \leq \min\{a, b\}$ be an integer such that

$$X_i[\text{Im}(B^{b-i}) \cap \text{Ker}(B^i)] \subseteq \text{Im}(A^i) + \text{Ker}(A^{a-i}), \qquad i = 0, 1, 2, \ldots, k-1,$$
$$X_k[\text{Im}(B^{b-k}) \cap \text{Ker}(B^k)] \not\subseteq \text{Im}(A^k) + \text{Ker}(A^{a-k}).$$

Then, by Theorem 2.6,

$$\text{index}(M) = a + b - (k-1) \geq \max\{a, b\} + 1 > \text{index}(M),$$

which is a contradiction.

Conversely, if $\text{index}(M) \neq \max\{a, b\}$, then by (2) we have that $\text{index}(M) > \max\{a, b\}$ and so, by Theorem 2.5,

$$X_k[\text{Im}(B^{b-k}) \cap \text{Ker}(B^k)] \not\subseteq \text{Im}(A^k) + \text{Ker}(A^{a-k}),$$

for some k with $1 \leq k \leq \min\{a, b\}$. This contradicts our hypothesis. Hence $\text{index}(M) = \max\{a, b\}$. □

3. The index and the Drazin inverse. In this section we obtain characterizations for values of $\text{index}(M)$, where M is given in (1), in terms of the Drazin inverses of A and B (see [1] and [3] for properties of such inverses) and of the matrix X_p defined by (3). These conditions then provide additional equivalences to the results already developed in Theorems 2.6 and 2.7. For clarity we denote the identity matrices of sizes $m \times m$ and $n \times n$ by I_m and I_n, respectively. We begin by quoting the following result of Meyer and Rose that we shall use subsequently.

THEOREM 3.1 (Theorem 2.2 of [8]). *Let M be given by* (1). *For each positive integer q, $\text{index}(M) \leq q$ if and only if the following conditions hold:*

1. $\text{index}(A) \leq q$;
2. $\text{index}(B) \leq q$;
3. $(I_m - AA^D)X_q(I_n - BB^D) = O$, *where A^D and B^D are the Drazin inverses of A and B, respectively, and X_q is given by* (3).

Note that if $q = a + b$, then by (2) and Theorem 3.1, we have that

$$(I_m - AA^D)X_{a+b}(I_n - BB^D) = O. \tag{9}$$

Theorem 3.1 is a characterization for a positive integer to be an upper bound on $\text{index}(M)$. We now attempt to determine the exact value of $\text{index}(M)$. For this purpose we also require the following lemma.

LEMMA 3.2. *Let P be a singular matrix with $p = \text{index}(P)$. Suppose that $1 \leq k \leq p$. Then* $\text{Im}([P^{p-k}(I - PP^D)]) = \text{Im}(P^{p-k}) \cap \text{Ker}(P^k)$.

Proof. Let $y \in \text{Im}([P^{p-k}(I-PP^D)])$. Then $y = P^{p-k}(I-PP^D)x$ for some vector x. Clearly $y \in \text{Im}P^{p-k}$. Since $P^p(I - PP^D) = O$, we have that

$$P^k y = P^p(I - PP^D)x = 0.$$

This means that $y \in \text{Im}(P^{p-k}) \cap \text{Ker}(P^k)$ and so

$$\text{Im}([P^{p-k}(I - PP^D)]) \subseteq \text{Im}(P^{p-k}) \cap \text{Ker}(P^k). \tag{10}$$

To show the reverse containment let $y \in \text{Im}(P^{p-k}) \cap \text{Ker}(P^k)$. Then $P^k y = 0$ and $y = P^{p-k}x$ for some vector x which now must lie in $\text{Ker}(P^p)$. Since $I - PP^D$ is a projection on $\text{Ker}(P^p)$ along $\text{Im}(P^p)$ we see that

$$(I - PP^D)x = x.$$

Hence

$$P^{p-k}(I - PP^D)x = P^{p-k}x = y$$

and so

$$\text{Im}(P^{p-k}) \cap \text{Ker}(P^k) \subseteq \text{Im}([P^{p-k}(I - PP^D)]). \tag{11}$$ □

Note that the above lemma also holds for the case when $k = 0$ because then

$$\text{Im}([P^{p-0}(I - PP^D)]) = \{0\}$$

and

$$\text{Im}(P^{p-0}) \cap \text{Ker}(P^0) = \text{Im}(P^p) \cap \{0\} = \{0\}.$$

Consider again the matrix M given by (1). On comparing the top right-hand blocks on both sides of the equality, $M^{s+t} = M^s M^t$ shows that

$$X_{s+t} = A^s X_t + X_s B^t. \tag{12}$$

We next prove an additional auxiliary lemma.

LEMMA 3.3. *Let M be given by* (1). *Set $a = \text{index}(A)$ and $b = \text{index}(B)$. Suppose further that $1 \leq k \leq \min\{a, b\}$. Then*

$$(I_m - AA^D)X_{a+b-k}(I_n - BB^D) = (I_m - AA^D)A^{a-k}X_k B^{b-k}(I_n - BB^D), \tag{13}$$

where X_{a+b-k} and X_k are given as in (3).

Proof. Applying (12) twice in succession we obtain that

$$X_{a+b-k} = A^a X_{b-k} + A^{a-k}X_k B^{b-k} + X_{a-k}B^b.$$

Substituting this expression for X_{a+b-k} in the left-hand side of (13) and using the facts that $(I_m - AA^D)A^a = O$ and $(I_n - BB^D)B^b = O$ yields inmediately (13). □

We comment that from (9) and (3) we have that

$$(I_m - AA^D)X_{a+b}(I_n - BB^D) = O = (I_m - AA^D)A^a X_0 B^b(I_n - BB^D). \tag{14}$$

Thus Lemma 3.3 is also valid for the case for $k = 0$.

We are now ready for our first main characterizations of this section.

THEOREM 3.4. *Let M be given by* (1). *Suppose that $a = \text{index}(A)$ and that $b = \text{index}(B)$. Assume that $1 \leq k \leq \min\{a, b\}$ and that X_k is given by* (3). *Then the following conditions are equivalent:*

1. $X_k[\mathrm{Im}(B^{b-k}) \cap \mathrm{Ker}(B^k)] \subseteq \mathrm{Im}(A^k) + \mathrm{Ker}(A^{a-k})$;
2. $(I_m - AA^D)A^{a-k}X_kB^{b-k}(I_n - BB^D) = O$.

Proof. Suppose that Condition 1 holds and let v be an n–vector. By Lemma 3.2,

$$B^{b-k}(I_n - BB^D)v \in \mathrm{Im}(B^{b-k}) \cap \mathrm{Ker}(B^k).$$

Thus by Condition 1,

$$X_kB^{b-k}(I_n - BB^D)v = A^kx + y,$$

for some vectors x and y with $A^{a-k}y = 0$. Now $(I_m - AA^D)A^a = O$ and so we have that

$$\begin{aligned}(I_m - AA^D)A^{a-k}X_kB^{b-k}(I_n - BB^D)v &= (I_m - AA^D)A^{a-k}(A^kx + y)\\ &= (I_m - AA^D)(A^ax + A^{a-k}y) = 0.\end{aligned}$$

Conversely, suppose that Condition 2 holds and let v be an n–vector. Then

$$(I_m - AA^D)A^{a-k}X_kB^{b-k}(I_n - BB^D)v = 0.$$

Now since $I_m - AA^D$ is the projection on $\mathrm{Ker}(A^a)$ along $\mathrm{Im}(A^a)$, we have that

$$A^{a-k}X_kB^{b-k}(I_n - BB^D)v \in \mathrm{Im}A^a.$$

This means that

$$A^{a-k}X_kB^{b-k}(I_n - BB^D)v = A^ax = A^{a-k}A^kx$$

for some m–vector x. Therefore

$$A^{a-k}(X_kB^{b-k}(I_n - BB^D)v - A^kx) = 0,$$

and so the vector

$$y := X_kB^{b-k}(I_n - BB^D)v - A^kx$$

is in $\mathrm{Ker}(A^{a-k})$. But then,

$$X_kB^{b-k}(I_n - BB^D)v = A^kx + y \in \mathrm{Im}(A^k) + \mathrm{Ker}(A^{a-k})$$

so that

$$X_k[\mathrm{Im}(B^{b-k}(I_n - BB^D))] \subseteq \mathrm{Im}(A^k) + \mathrm{Ker}(A^{a-k}).$$

Thus Condition 1 obtains on applying Lemma 3.2 to the left-hand side in the above inclusion. □

From Remark 1 and (14), Theorem 3.4 is also applicable in the case that $k = 0$. Therefore, on using Theorems 2.6 and 3.4 we can establish the following result.

THEOREM 3.5. *Let M be given by* (1) *and suppose that* $a = \mathrm{index}(A)$ *and* $b = \mathrm{index}(B)$. *Assume that* $1 \leq k \leq \min\{a, b\}$. *Then* $\mathrm{index}(M) = a + b - (k - 1)$ *if and only if the following conditions hold:*

1. $(I_m - AA^D)A^{a-i}X_iB^{b-i}(I_n - BB^D) = O$, $i = 0, 1, 2, \ldots, k-1$;
2. $(I_m - AA^D)A^{a-k}X_kB^{b-k}(I_n - BB^D) \neq O$.

Here, as before, X_j, $j = 0, 1, 2, \ldots, k$, *is given by* (3).

Theorems 2.7 and 3.4 together yield the theorem that follows.

THEOREM 3.6. *Let* M *be given by* (1) *and suppose that* $a = \text{index}(A)$ *and* $b = \text{index}(B)$. *Then* $\text{index}(M) = \max\{a, b\}$ *if and only if*

$$(I_m - AA^D)A^{a-i}X_iB^{b-i}(I_n - BB^D) = O, \qquad i = 0, 1, 2, \ldots, \min\{a, b\},$$

where X_i, $i = 0, 1, 2, \ldots, \min\{a, b\}$, *is given by* (3).

We comment that on considering the results in Lemma 3.3, we see that Theorem 3.1 is a particular case of Theorems 3.5 and 3.6.

4. The index of a matrix and the height and depth of generalized eigenvectors. Let M be given by (1). Let $a = \text{index}(A)$ and $b = \text{index}(B)$. In Johnson, Schreiner, and Elsner [5], the index of M, where A and B have exactly one Jordan block for the eigenvalue zero, is studied using the following notations.

Let $u_1, u_2, \ldots, u_a$ *be a left Jordan chain of* A, *that is* $u_i^T = u_1^T A^{i-1}$, $i = 1, 2, \ldots, a-1, a$, *and* $u_a^T A = 0^T$. *Similarly, let* $v_1, v_2, \ldots, v_b$ *be a right Jordan chain of* B, *that is* $v_i = B^{b-i}v_b$, $i = 1, 2, \ldots, b-1, b$, *and* $Bv_1 = 0$.

THEOREM 4.1 (Theorem 8 of [5]). *Under the above-mentioned conditions, if* $u_a^T X v_1 \neq 0$, *then* $\text{index}(M) = a + b$.

We observe that for the vectors u_a and v_1 we have that $\text{ht}_{A^T}(u_a) = 1$, $\text{dp}_{A^T}(u_a) = a - 1$, $\text{ht}_B(v_1) = 1$, and $\text{dp}_B(v_1) = b - 1$.

The results that we prove below allow us to characterize the index of M in terms of the height and depth of certain vectors in $E(A^T)$ and $E(B)$. This, in combination with the results of §2, yield a strengthening and an extension of the results of Johnson, Schreiner, and Elsner [5] and Johnson and Schreiner [6], [7].

We begin by proving a sequence of lemmas.

LEMMA 4.2. *Let* P *be a square matrix. Let* r *and* s *be two positive integers. Then*

$$\text{Im}(P^r) + \text{Ker}(P^s) = [\text{Ker}((P^T)^r) \cap \text{Im}((P^T)^s)]^{\perp}.$$

Proof. The lemma follows from the facts that $\text{Im}(P) = (\text{Ker}P^T)^{\perp}$ and that if U and V are two subspaces, then $(U \cap V)^{\perp} = U^{\perp} + V^{\perp}$ and $(V^{\perp})^{\perp} = V$. □

LEMMA 4.3. *Let* M *be given by* (1). *Suppose that* $a = \text{index}(A)$ *and* $b = \text{index}(B)$ *and let* $0 \leq k \leq \min\{a, b\}$. *Then for* X_k *be given by* (3), *the following conditions are equivalent.*

1. $X_k[\text{Im}(B^{b-k}) \cap \text{Ker}(B^k)] \subseteq \text{Im}(A^k) + \text{Ker}(A^{a-k})$.
2. $y^T X_k z = 0$ *for all* $y \in \text{Im}((A^T)^{a-k}) \cap \text{Ker}((A^T)^k)$ *and for all* $z \in \text{Im}(B^{b-k}) \cap \text{Ker}(B^k)$.

Proof. Suppose that Condition 1 holds and let $z \in \text{Im}(B^{b-k}) \cap \text{Ker}(B^k)$ so that

$$X_k z = A^k u + v$$

for some vectors u and v with $A^{a-k}v = 0$. Let $y \in \text{Im}((A^T)^{a-k}) \cap \text{Ker}((A^T)^k)$. Then $y^T = w^T A^{a-k}$ for some vector w and $y^T A^k = 0^T$. Therefore

$$y^T X_k z = y^T(A^T u + v) = y^T A^k u + y^T v = 0^T u + w^T A^{a-k} v = w^T 0 = 0$$

and so Condition 2 holds.

Conversely, if Condition 2 holds, but Condition 1 is not valid, then there exists a vector $v \in \text{Im}(B^{b-k}) \cap \text{Ker}(B^k)$ such that

$$X_k v \notin \text{Im}(A^k) + \text{Ker}(A^{a-k}).$$

Then, by Lemma 4.2, there exists a vector $u \in \mathrm{Im}((A^T)^{a-k}) \cap \mathrm{Ker}((A^T)^k)$ such that

$$u^T X_k v \neq 0$$

and Condition 2 does not hold; a contradiction. □

LEMMA 4.4. *Let M be given by* (1). *Suppose that $a = \mathrm{index}(A)$ and $b = \mathrm{index}(B)$ and let $1 \leq k \leq \min\{a, b\}$. Let X_k be given by* (3). *Assume that $y \in E(A^T)$ and $z \in E(B)$. If $\mathrm{ht}_{A^T}(y) + \mathrm{ht}_B(z) \leq k$, then $y^T X_k z = 0$.*

Proof. Suppose that $\mathrm{ht}_{A^T}(y) = p$ and $\mathrm{ht}_B(z) = q$. Then $y^T A^r = 0^T$ for all integers $r \geq p$ and $B^s z = 0$ for all integers $s \geq q$. But then

$$y^T X_k z = y^T (A^p X_{k-p} + X_p B^{k-p}) z = 0$$

because $q \leq k - p$. □

LEMMA 4.5. *Let M be given by* (1). *Assume that $a = \mathrm{index}(A)$ and $b = \mathrm{index}(B)$ and suppose that $1 \leq k \leq \min\{a, b\}$. If $y \in E(A^T)$ and $z \in E(B)$ are vectors such that $\mathrm{ht}_{A^T}(y) \leq r$ and $\mathrm{ht}_B(z) \leq s$ and if $k < r + s \leq 2k$, then*

$$y^T X_k z = y^T A^{k-s} X_{r+s-k} B^{k-r} z,$$

where X_k and X_{r+s-k} are given as in (3).

Proof. Clearly $0 < r + s - k \leq k$. Then

$$\begin{aligned} y^T X_k z &= y^T X_r B^{k-r} z = y^T (A^{k-s} X_{r-(k-s)} + X_{k-s} B^{r-(k-s)}) B^{k-r} z \\ &= y^T A^{k-s} X_{r+s-k} B^{k-r} z + y^T X_{k-s} B^s z = y^T A^{k-s} X_{r+s-k} B^{k-r} z \end{aligned}$$

because $B^s z = 0$. □

In the following theorem we consider some implications of the two conditions which appear in the equivalence of Theorem 2.6.

THEOREM 4.6. *Let M be given by* (1) *and assume that $a = \mathrm{index}(A)$ and $b = \mathrm{index}(B)$. Let $1 \leq k \leq \min\{a, b\}$. Suppose that the following conditions hold:*

1. $X_i[\mathrm{Im}(B^{b-i}) \cap \mathrm{Ker}(B^i)] \subseteq \mathrm{Im}(A^i) + \mathrm{Ker}(A^{a-k})$, $i = 0, 1, 2, \ldots, k-1$;
2. $X_k[\mathrm{Im}(B^{b-k}) \cap \mathrm{Ker}(B^k)] \not\subseteq \mathrm{Im}(A^k) + \mathrm{Ker}(A^{a-k})$,

where X_j, $j = 0, 1, 2, \ldots, k$, is given by (3). *Then there exist vectors $u \in E(A^T)$ and $v \in E(B)$ such that*

(a) $u^T X_k v \neq 0$,
(b) $k < \mathrm{ht}_{A^T}(u) + \mathrm{ht}_B(v) \leq 2k$,
(c) $\mathrm{ht}_{A^T}(u) \leq k$ *and* $\mathrm{dp}_{A^T}(u) = a - k$,
(d) $\mathrm{ht}_B(v) \leq k$ *and* $\mathrm{dp}_B(v) = b - k$.

Proof. From Condition 2 and Lemma 4.2 there exist a vector $u \in \mathrm{Im}((A^T)^{a-k}) \cap \mathrm{Ker}((A^T)^k)$ and a vector $v \in \mathrm{Im}(B^{b-k}) \cap \mathrm{Ker}(B^k)$ such that

$$u^T X_k v \neq 0.$$

Clearly $u \in E(A^T)$ and $v \in E(B)$ and so Condition (a) holds. Moreover,

$$\mathrm{ht}_{A^T}(u) \leq k, \qquad \mathrm{dp}_{A^T}(u) \geq a - k$$

and

$$\mathrm{ht}_B(v) \leq k, \qquad \mathrm{dp}_B(v) \geq b - k.$$

Obviously $\mathrm{ht}_{A^T}(u) + \mathrm{ht}_B(v) \leq 2k$ and so by Lemma 4.4, $\mathrm{ht}_{A^T}(u) + \mathrm{ht}_B(v) > k$ and Condition (b) holds.

Next, as $a = \mathrm{index}(A)$, we have that $\mathrm{ht}_{A^T}(u) + \mathrm{dp}_{A^T}(u) \leq a$. Hence the vector u satisfies one of the following conditions:

- $\mathrm{ht}_{A^T}(u) = k$, $\mathrm{dp}_{A^T}(u) = a - k$;
- $\mathrm{ht}_{A^T}(u) = p < k$, $\mathrm{dp}_{A^T}(u) = a - k$;
- $\mathrm{ht}_{A^T}(u) = p < k$, $\mathrm{dp}_{A^T}(u) = a - p > a - k$;
- $\mathrm{ht}_{A^T}(u) = p < r < k$, $a - p > \mathrm{dp}_{A^T}(u) = a - r > a - k$.

Similarly, the vector v satisfies one of the following conditions:

- $\mathrm{ht}_B(v) = k$, $\mathrm{dp}_B(v) = b - k$;
- $\mathrm{ht}_B(v) = q < k$, $\mathrm{dp}_B(v) = b - k$;
- $\mathrm{ht}_B(v) = q < k$, $\mathrm{dp}_B(v) = b - q > b - k$;
- $\mathrm{ht}_B(v) = q < s < k$, $b - q > \mathrm{dp}_B(v) = b - s > b - k$.

To make our analysis easier to follow, we have summarized the possible cases in Table 1. To consider which possibilities in the table are feasible, we first deal with the last

TABLE 1

Case	$\mathrm{ht}_{A^T}(u)$	$\mathrm{dp}_{A^T}(u)$	$\mathrm{ht}_B(v)$	$\mathrm{dp}_B(v)$	Observations
1	k	$a-k$	k	$b-k$	
2	k	$a-k$	q	$b-k$	$q<k$
3	k	$a-k$	q	$b-q$	$q<k$
4	k	$a-k$	q	$b-s$	$q<s<k$
5	p	$a-k$	k	$b-k$	$p<k$
6	p	$a-k$	q	$b-k$	$p<k, q<k$
7	p	$a-k$	q	$b-q$	$p<k, q<k$
8	p	$a-k$	q	$b-s$	$q<s<k$
9	p	$a-p$	k	$b-k$	$p<k$
10	p	$a-p$	q	$b-k$	$p<k, q<k$
11	p	$a-p$	q	$b-q$	$p<k, q<k$
12	p	$a-p$	q	$b-s$	$p<k, q<s<k$
13	p	$a-r$	k	$b-k$	$p<r<k$
14	p	$a-r$	q	$b-k$	$p<r<k, q<k$
15	p	$a-r$	q	$b-q$	$p<r<k, q<k$
16	p	$a-r$	q	$b-s$	$p<r<k, q<s<k$

entry in the table. Suppose that entry were possible. Then by Lemma 4.5,

$$u^T X_k v = u^T A^{k-s} X_{r+s-k} B^{k-r} v. \tag{15}$$

Clearly $\mathrm{ht}_{A^T}((A^T)^{k-s}u) = p+s-k$ and $\mathrm{dp}_{A^T}((A^T)^{k-s}u) \geq a-(r+s-k)$. Therefore,

$$\begin{aligned}(A^T)^{k-s}u &\in \mathrm{Im}((A^T)^{a-(r+s-k)}) \cap \mathrm{Ker}((A^T)^{p+s-k}) \\ &\subseteq \mathrm{Im}((A^T)^{a-(r+s-k)}) \cap \mathrm{Ker}((A^T)^{r+s-k})\end{aligned}$$

because $p + s - k < r + s - k$. Similarly, it holds that

$$B^{k-r}v \in \mathrm{Im}B^{b-(r+s-k)} \cap \mathrm{Ker}B^{r+s-k}.$$

Now, as $r + s - k < k$, by Condition 1 we have that

$$X_{r+s-k}[\mathrm{Im}(B^{b-(r+s-k)}) \cap \mathrm{Ker}(B^{r+s-k})] \subseteq \mathrm{Im}(A^{r+s-k}) + \mathrm{Ker}(A^{a-(r+s-k)}).$$

Thus, by Lemma 4.3 and (15) we conculde that

$$u^T X_k v = 0. \tag{16}$$

This contradicts the validity of Condition (a) above.

Note that cases 1–15 in Table 1 can be obtained from case 16 by replacing some of the $<$ signs by $=$ signs in the Observations column. Then with similar reasoning, we can obtain that the equality (16) for most of the cases holds, making these cases incompatible with Condition (a). Because of Condition (a), we need only consider those situations in which

$$u^T A^{k-s} X_{r+s-k} B^{k-r} v \neq 0.$$

These situations, in turn, can only occur when $r + s - k = k$. Notice then that Condition (b) is satisfied and $r = s = k$. Considering the table again, we see that $r = s = k$ in cases 1, 2, 5, and 6. These are precisely the cases in which

$$\mathrm{ht}_{A^T}(u) \leq k, \quad \mathrm{dp}_{A^T}(u) = a - k, \quad \mathrm{ht}_B(v) \leq k, \quad \mathrm{dp}_B(v) = b - k,$$

and Conditions (c) and (d) hold. □

We can strengthen Theorem 4.6 and prove the following equivalence.

THEOREM 4.7. *Let M be given by* (1). *Let $a = \mathrm{index}(A)$ and and let $b = \mathrm{index}(B)$. Suppose that $1 \leq k \leq \min\{a, b\}$. For $j = 0, 1, 2, \ldots, k$, let X_j be given by* (3). *Consider the following statements:*

1. $X_i[\mathrm{Im}(B^{b-i}) \cap \mathrm{Ker}(B^i)] \subseteq \mathrm{Im}(A^i) + \mathrm{Ker}(A^{a-i})$, $0 \leq i \leq k-1$;
2. $X_k[\mathrm{Im}(B^{b-k}) \cap \mathrm{Ker}(B^k)] \not\subseteq \mathrm{Im}(A^k) + \mathrm{Ker}(A^{a-k})$;
3. *Let $0 \leq i \leq k-1$. Then $y_i^T X_i z_i = 0$ for all $y_i \in \mathrm{Im}((A^T)^{a-i}) \cap \mathrm{Ker}((A^T)^i)$ and for all $z_i \in \mathrm{Im}(B^{b-i}) \cap \mathrm{Ker}(B^i)$;*
4. *There exists a vector $u \in E(A^T)$ and a vector $v \in E(B)$ such that*
   (a) $u^T X_k v \neq 0$,
   (b) $k < \mathrm{ht}_{A^T}(u) + \mathrm{ht}_B(v) \leq 2k$,
   (c) $\mathrm{ht}_{A^T}(u) \leq k$ *and* $\mathrm{dp}_{A^T}(u) = a - k$,
   (d) $\mathrm{ht}_B(v) \leq k$ *and* $\mathrm{dp}_B(v) = b - k$.

Then Conditions 4.7 *and* 4.7 *are equivalent to Conditions* 4.7 *and* 4.7.

Proof. If Conditions 4.7 and 4.7 hold, then by Theorem 4.6, Condition 4.7 holds, while Condition 4.7 holds by Lemma 4.3.

Conversely, if Condition 4.7 holds, then by Lemma 4.3, Condition 4.7 holds. Furthermore, if Condition 4.7 holds, then from Conditions 4.7 and 4.7 we have that

$$u \in \mathrm{Im}((A^T)^{a-k}) \cap \mathrm{Ker}((A^T)^k) \quad \text{and} \quad v \in \mathrm{Im}(B^{b-k}) \cap \mathrm{Ker}(B^k).$$

Thus, from the Condition 4.7 and Lemma 4.3 we see that Condition 4.7 holds. □

Taking into consideration the notation introduced at the beginning of this section, where A and B where assumed to have exactly one Jordan block for the eigenvalue zero, we see that Theorem 4.7 has the implication that the conditions

1. $u_a^T X v_1 \neq 0$,
2. $X[\mathrm{Im}(B^{b-1}) \cap \mathrm{Ker}(B)] \not\subseteq \mathrm{Im}(A) + \mathrm{Ker}(A^{a-1})$

are equivalent. This together with Theorem 2.1 allows us to strengthen Theorem 4.1 as follows.

THEOREM 4.8. *Under the notations and conditions of Theorem* 4.1, *then $u_a^T X v_1 \neq 0$ if and only if* $\mathrm{index}(M) = a + b$.

5. Conclusions. We can summarize the main results of this paper by the following theorem, which consists of a string of equivalences that come from by Theorems 2.6, 3.5, and 4.7.

THEOREM 5.1. *Let M be given by* (1). *Let $a = \mathrm{index}(A)$ and $b = \mathrm{index}(B)$. Suppose that $1 \leq k \leq \min\{a, b\}$ and that X_j, $j = 0, 1, 2, \ldots, k$, are given as in* (3). *Then the following four conditions are equivalent*

A. $\operatorname{index}(M) = a + b - (k-1)$.

B. 1. $X_i[\operatorname{Im}(B^{b-i}) \cap \operatorname{Ker}(B^i)] \subseteq \operatorname{Im}(A^i) + \operatorname{Ker}(A^{a-i})$, $0 \le i \le k-1$;
   2. $X_k[\operatorname{Im}(B^{b-k}) \cap \operatorname{Ker}(B^k)] \not\subseteq \operatorname{Im}(A^k) + \operatorname{Ker}(A^{a-k})$

C. 1. $(I_m - AA^D)A^{a-i}X_iB^{b-i}(I_n - BB^D) = O$, $0 \le i \le k-1$.;
   2. $(I_m - AA^D)A^{a-k}X_kB^{b-k}(I_n - BB^D) \neq O$.

D. 1. *Let* $0 \le i \le k-1$. *Then* $y_i^T X_i z_i = 0$ *for all* $y_i \in \operatorname{Im}((A^T)^{a-i}) \cap \operatorname{Ker}((A^T)^i)$ *and for all* $z_i \in \operatorname{Im}(B^{b-i}) \cap \operatorname{Ker}(B^i)$;
   2. *There exists a vector* $u \in E(A^T)$ *and a vector* $v \in E(B)$ *such that*
      (a) $u^T X_k v \neq 0$,
      (b) $k < \operatorname{ht}_{A^T}(u) + \operatorname{ht}_B(v) \le 2k$,
      (c) $\operatorname{ht}_{A^T}(u) \le k$ *and* $\operatorname{dp}_{A^T}(u) = -k$,
      (d) $\operatorname{ht}_B(v) \le k$ *and* $\operatorname{dp}_B(v) = b - k$.

We see that for $k = 1$, the theorem gives necessary and sufficient conditions so that $\operatorname{index}(M) = a + b$, whereas if $k \le \min\{a, b\}$ it stipulates necessary and sufficient conditions so that $\operatorname{index}(M) \ge \max\{a, b\} + 1$. Finally, by Theorems 2.7 and 3.6 and Lemma 4.3, the following four conditions are also equivalent:

E. $\operatorname{index}(M) = \max\{a, b\}$.

F. $X_i[\operatorname{Im}(B^{b-i}) \cap \operatorname{Ker}(B^i)] \subseteq \operatorname{Im}(A^i) + \operatorname{Ker}(A^{a-i})$, $0 \le i \le \min\{a, b\}$.

G. $(I_m - AA^D)A^{a-i}X_iB^{b-i}(I_n - BB^D) = O$, $0 \le i \le \min\{a, b\}$.

H. Let $0 \le i \le \min\{a, b\}$. Then $y_i^T X_i z_i = 0$ for all $y_i \in \operatorname{Im}((A^T)^{a-i}) \cap \operatorname{Ker}((A^T)^i)$ and for all $z_i \in \operatorname{Im}(B^{b-i}) \cap \operatorname{Ker}(B^i)$.

Acknowledgment. We would like to thank Professor H. Schneider for some helpful discussions on parts of this work.

REFERENCES

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

[2] R. Bru, L. Rodman, and H. Schneider, *Extensions of Jordan bases for invariant subspaces of a matrix*, Linear Algebra Appl., 150 (1991), pp. 209–225.

[3] S. L. Campbell and C. D. Meyer, Jr., *Generalized Inverses of Linear Transformations*, Dover Publications, New York, 1991.

[4] D. Hershkowitz, U. G. Rothblum, and H. Schneider, *The combinatorial structure of the generalized nullspace of a block triangular matrix*, Linear Algebra Appl., 116 (1989), pp. 9–26.

[5] C. R. Johnson, E. A. Schreiner, and L. Elsner, *Eigenvalue neutrality in block triangular matrices*, Linear Multilinear Algebra, 27 (1990), pp. 289–297.

[6] C. R. Johnson and E. A. Schreiner, *Explicit Jordan form for certain block triangular matrices*, Linear Algebra Appl., 150 (1991), pp. 297–314.

[7] ———, *Explicit Jordan form for certain block triangular matrices* II, Linear Algebra Appl., 162–164 (1992), pp. 601–613.

[8] C. D. Meyer, Jr. and N. J. Rose, *The index and the Drazin inverse of block triangular matrices*, SIAM J. Appl. Math., 33 (1977), pp. 1–7.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 448–461, April 1995

009

SOME REMARKS CONCERNING ITERATIVE METHODS FOR LINEAR SYSTEMS *

FRED B. WEISSLER†

Abstract. Let A be a Hermitian, positive definite matrix. It is well known that if $A = M - N$, where M is invertible and $M^* + N$ is also positive definite, then the iterative method for solving $Au = b$, based on this decomposition, is convergent. A new proof of this convergence is given, providing an explicit estimate for the spectral radius of $M^{-1}N$. Also, a new method to estimate the optimal relaxation parameter for certain large matrices is suggested.

Key words. spectral radius, positive definite matrix, iterative methods, successive overrelaxation, optimal relaxation parameter

AMS subject classification. 65F10

1. Introduction and historical remarks.

In this note we consider a class of iterative methods for a linear system $Au = b$, where A is an invertible square matrix, b a given vector, and u the unknown vector. If $A = M - N$, where M is also invertible, one can define the sequence $\{u_k\colon\ k = 0, 1, 2, 3, \ldots\}$ by $Mu_{k+1} = Nu_k + b$, where u_0 is an arbitrary initial vector. It is well known that this method converges, i.e., the sequence u_k converges to the solution u (for all possible u_0), if and only if $\rho(M^{-1}N) < 1$, where ρ denotes the spectral radius. In this case, $-\log \rho(M^{-1}N)$ is the asymptotic average rate of convergence of u_k toward the solution (Young [21, p. 88]). The following theorem gives a classical result concerning the convergence of this method.

THEOREM A. *Let A be a Hermitian, positive definite matrix, and let $A = M - N$, where M is an invertible matrix. If, in addition, the (necessarily Hermitian) matrix $M^* + N$ is positive definite, then $\rho(M^{-1}N) < 1$.*

This result is used to prove the Ostrowski–Reich Theorem, i.e., if A is a positive definite matrix, then the successive overrelaxation method converges for all values of the relaxation parameter in the interval $(0, 2)$.

The purpose of this note is to present a new proof of Theorem A that provides an explicit estimate for $\rho(M^{-1}N)$ and to examine the consequences of this estimate. First, however, some historical remarks are in order.

Theorem A is sometimes stated and proved in the slightly stronger form as an if and only if statement, and indeed that is done in two different ways. More precisely, in Lascaux and Théodor [9, Thm. 22, §7.4, p. 426], Householder [4, Thm. 4.17, p. 227], and Householder [5, p. 111] one finds the following result.

THEOREM A1. *Let A be an invertible Hermitian matrix, and let $A = M - N$, where M is an invertible matrix. Suppose in addition that the (necessarily Hermitian) matrix $M^* + N$ is positive definite. Then $\rho(M^{-1}N) < 1$ if and only if A is positive definite.*

Also, Theorem A1 is stated in Ortega and Plemmons [11, p. 179] and is called the Householder–John Theorem. In the special case of the Gauss–Seidel decomposition,

* Received by the editors May 4, 1992; accepted for publication (in revised form) by M. H. Gutknecht, January 17, 1994.

† Laboratoire Analyse Géométrie et Applications, Institut Galilée–Université Paris XIII, Av. J.-B. Clément, 93430 Villetaneuse, France (weissler@math.univ-paris13.fr).

Theorem A1 is stated and proved in Reich [13] and for the successive overrelaxation method in Ostrowski [12]. On the other hand, in Young [21, Thm. 5.3, §3.5, p. 79] one finds the following slightly different result.

THEOREM A2. *Let A be a Hermitian positive definite matrix, and let $A = M - N$, where M is an invertible matrix. Then $\rho(M^{-1}N) < 1$ if and only if the (necessarily Hermitian) matrix $M^* + N$ is positive definite.*

Theorem A alone can be found in Weissinger [20, p. 160], Householder [3, Cor. to Thm. 18, pp. 31–32], Wachspress [19, Thm. 1-11, §1.4, p. 17], John [6, p. 21], Ciarlet [1, Thm. 5.3-1, §5.3, p. 102], Schatzman [16, Thm. 10.9, p. 101], and in a slightly different form (the P-regular splitting theorem) in Ortega [10, Thm. 7.1.9, §7.1, p. 123].

In examining the history of these results, one must keep in mind that they developed more or less as a generalization of the Ostrowski–Reich Theorem. One finds today at least four identifiably different proofs of Theorem A. My approach to the history is to give a relatively modern reference for each proof and then to describe previous versions of that or similar arguments.

The proof of Theorem A1 in Lascaux and Théodor (1987) [9, Thm. 22, §7.4, p. 426] is based on the formula

$$\langle u, Au\rangle - \langle Bu, ABu\rangle = \langle (I - B)u, (M^* + N)(I - B)u\rangle, \tag{1.1}$$

where A is assumed to be Hermitian and $B = M^{-1}N$. If now u is an eigenvector of B with eigenvalue λ, then (1.1) implies that

$$(1 - |\lambda|^2)\langle u, Au\rangle = |1 - \lambda|^2\langle u, (M^* + N)u\rangle. \tag{1.2}$$

Furthermore, $\lambda \neq 1$, since if $\lambda = 1$ then $M^{-1}Au = (I - B)u = 0$, and so $u = 0$. If both A and $M^* + N$ are positive definite, then $|\lambda|^2 < 1$. The reverse implication in Theorem A1 is proved using formula (1.1) directly. Formula (1.2) can also be found in the proof of Theorem A by John (1967/1956) [6, pp.21–22]. Formulas (1.1) and (1.2) (for the special case of successive overrelaxation) appear in Varga's proof of the Ostrowski–Reich Theorem (Varga (1962) [18, Thm. 3.6, §3.4, p. 77]), which is based on Ostrowski's (1954) original argument [12]. See the bibliographic remarks in Varga [18, p. 96]. Finally, formula (1.1) (for the special case of the Gauss–Seidel decomposition) appears in Reich's proof of Theorem A1 for Gauss–Seidel (1949) [13]. (Reich attributes Theorem A for Gauss–Seidel to Seidel (1874) [15].) Reich proves the necessity part of Theorem A1 (for the Gauss–Seidel decomposition) using yet another formula of which (1.2) is a special case. Also, formula (1.1) is implicit in the proof of Theorem 2 in Stein (1952) [17].

The proof of Theorem A in Ortega (1972) [10, Thm. 7.1.9, §7.1, p. 123] is based on the formula

$$A - B^*AB = (I - B)^*(M^* + N)(I - B), \tag{1.3}$$

where A again is assumed to be Hermitian. (Actually, Ortega works with real symmetric matrices instead of complex Hermitian matrices. To simplify the exposition, since there is no substantial mathematical difference, I will discuss all of the results as if they applied to complex Hermitian matrices. Also, Ortega's P-regular splitting theorem is not exactly the same as Theorem A in that the hypotheses concern $M + N$ rather than $M^* + N$, but it is close enough to be referred to as Theorem A, retrospectively.

See Ortega and Plemmons [11, p. 187] for a discussion of this point.) Formula (1.3) implies formula (1.1), and therefore formula (1.2), but Ortega does not go that route. Instead, he proves Theorem A by applying Stein's Theorem to formula (1.3). Stein's Theorem, i.e., part of Theorem 1 in Stein (1952) [17], says that if A is Hermitian, positive definite and if $A - B^*AB$ is also positive definite (it is clearly Hermitian), then $\rho(B) < 1$, and Ortega proves Stein's Theorem by considering $\langle (A - B^*AB)u, u\rangle$, where u is an eigenvector of B. This argument is close to Stein's proof, which Stein attributes to L. J. Paige (personal communication, apparently; see Stein (1952) [17, footnote 3, p. 82]). Formula (1.3) can also be found earlier in Householder (1958) [4, Thm. 4.17, p. 227] and (1964) [5, p. 111], where proof of Theorem A is essentially the same as in Ortega (1972) [10]. Both Ortega and Householder credit Weissinger (1953) [20] for Theorem A as well as for formula (1.3). See the footnote in Ortega (1972) [10, p. 123] and the historical comments in Householder (1964) [5, p. 115]. However, in Ortega and Plemmons (1979) [11, p. 181] an earlier work of Householder (1955) [3], as well as that of Weissinger, is credited with formula (1.3). Indeed, one finds formula (1.3), generalized to the case where A is not necessarily Hermitian, in Householder (1955) [3, pp. 27–28], where it is called "Weissinger's Lemma." This same formula appears as formula (2.5) in Weissinger (1953) [20, p. 156] and Theorem A is given in Weissinger (1953) [20, p. 160].

Concerning the proof of the necessity part of Theorem A1, Householder (1958) [4, Thm. 4.17, p. 227] and (1964) [5, p. 111] gives an iteration argument based on formula (1.3) and credits Reich (1949) [13] (as I have just done two paragraphs above) for the same result in the special case of Gauss–Seidel. (Again, see the historical comments in Householder [5, p. 115].) I have not found any author who noticed that the necessity part of Theorem A1 follows immediately from formula (1.3) and Theorem 2 in Stein (1952) [17].

Young's proof (1971) [21, Thm. 5.3, §3.5, p. 79] of Theorem A2 is based on yet a different formula,

$$(A^{1/2}BA^{-1/2})(A^{1/2}BA^{-1/2})^* = I - A^{1/2}M^{-1}(M^* + N)(M^{-1})^*A^{1/2}, \tag{1.4}$$

which becomes, if $M^* + N$ is positive definite,

$$(A^{1/2}BA^{-1/2})(A^{1/2}BA^{-1/2})^* = I - [A^{1/2}M^{-1}(M^* + N)^{1/2}][A^{1/2}M^{-1}(M^* + N)^{1/2}]^*. \tag{1.5}$$

where $A^{1/2}$ and $(M^* + N)^{1/2}$ are the Hermitian, positive definite square root matrices of the Hermitian, positive definite matrices A and $M^* + N$, respectively. Formula (1.4) is actually not so different from formula (1.3): if formula (1.3) is multiplied on the left and right by $A^{1/2}$, the result is formula (1.4), except that the Hermitian adjoints are on the left instead of the right. (Notice that the hypotheses of Theorem A2, as opposed to Theorem A1, allow the use of $A^{-1/2}$ in the proof.) Young [21, p. 94] credits the calculations and a similar result to Wachspress (1966) [19], and indeed Wachspress (1966) [19, Thm. 1-11, §1.4, p. 17] states and proves Theorem A using formulas (1.4) and (1.5). Wachspress in turn credits the calculation to Habetler (1959) [2], but I have not obtained a copy of this article. Interestingly, Young states Stein's Theorem just after proving Theorem A2 and gives a proof analogous to Wachspress' proof of Theorem A. However, Young does not prove Theorem A by reducing it to Stein's Theorem, though he does use Stein's Theorem to establish other criteria for convergence, thereby providing another proof of the Ostrowski–Reich Theorem (Young

[21, p. 84]). Also, Young [21, p. 80] notes that the $A^{1/2}$-operator norm (defined below) of $B = M^{-1}N$ is less than 1.

The fourth proof of Theorem A can be found, for example, in Ciarlet (1985) [1, Thm. 5.3-1, §5.3, p. 102] and, more recently, in Schatzman (1991) [16, Thm. 10.9, p. 101]. It is more functional analytic in flavor and seems to be drawn from ideas in the Habetler–Wachspress–Young proof. Some of the technicalities are avoided by appealing to compactness of the unit ball in a finite dimensional vector space. Since the proof presented here is an extension of the proof in Ciarlet [1], the details are given below.

Finally, I would like to emphasize that this historical survey is not necessarily complete. Theorem A, along with the Ostrowski–Reich Theorem, has an extraordinarily rich history, and I invite the reader to consult the bibliographies and historical comments in the works cited in the present article. Theorem A seems to be due to a number of different researchers independently.

The plan of this paper is as follows. In §2, the new estimate for $\rho(M^{-1}N)$ is proved. In §3, this estimate is examined for some examples, in particular for the method of overrelaxation applied to a certain tridiagonal matrix. In §4, the infinite dimensional version of this matrix is studied, as well as the convergence as the dimension goes to infinity of results for the finite dimensional matrices. This analysis suggests a new method for estimating the optimal relaxation parameter for certain large matrices (§5).

2. A new estimate for $\rho(M^{-1}N)$. I present here a proof of Theorem A that provides an explicit estimate for $\rho(M^{-1}N)$ in terms of the spectral properties of the matrices A, M, and N (in various combinations). To emphasize the fact that the result does not rely on finite dimensionality, the theorem is formulated in the context of selfadjoint operators on a Hilbert space $\mathcal{H}$. It should be noted, however, that the proof of Theorem A given by Wachspress [19] and, later, Young [21] is essentially valid as stated in the infinite dimensional case.

To fix the terminology, a *positive* selfadjoint operator is one whose spectrum is contained in $[0, \infty)$. An invertible, positive bounded selfadjoint operator is one whose spectrum is contained in some interval $[\delta, \gamma]$, where $0 < \delta < \gamma < \infty$; this is the appropriate infinite dimensional analogue of a Hermitian, positive definite matrix. The spectral radius of an operator H is denoted $\rho(H)$. Finally, the norm on $\mathcal{H}$, as well as the induced operator norm, is denoted by $\|\ \ \|$ and the inner product by $\langle , \rangle$. Recall that $\rho(H) \leq \|H\|$ for any bounded operator H, this inequality being true if $\|\ \ \|$ is replaced by an operator norm induced by any equivalent norm on $\mathcal{H}$ (see, Kreyszig [8, Thm. 7.3-4]).

THEOREM B. *Let A be an invertible, positive bounded selfadjoint operator on the Hilbert Space $\mathcal{H}$. Suppose that $A = M - N$, where M and N are bounded operators, with M invertible. If, in addition, the (necessarily selfadjoint) operator $M^* + N$ is positive and invertible, then $\rho(M^{-1}N) < 1$. More precisely,*

$$\rho(M^{-1}N)^2 \leq I - \frac{\inf\{\text{spectrum}(M^* + N)\}}{\rho(M^* A^{-1} M)}$$

This inequality becomes an equality when A, M, and N are all scalar multiples of the identity operator.

Proof. The proof begins as in Ciarlet [1, Thm. 5.3-1, §5.3, p. 102], and I will indicate at what point it departs from this proof. Consider a new norm on $\mathcal{H}$ (referred

to above as the $A^{1/2}$-norm) defined by

$$|||v|||^2 = \langle Av, v\rangle = \|A^{1/2}v\|^2,$$

where $A^{1/2}$ is the positive, selfadjoint square root of A. Since A is an invertible, positive bounded selfadjoint operator, it follows that $|||\ \ |||$ is a (Hilbert) norm on $\mathcal{H}$, equivalent to the original norm. We use $|||\ \ |||$ also to denote the induced operator norm. Thus,

$$\rho(M^{-1}N) \leq |||M^{-1}N|||;$$

and so to estimate $\rho(M^{-1}N)$, it suffices to estimate $|||M^{-1}N|||$.

By definition of the induced operator norm,

$$|||M^{-1}N|||^2 = |||I - M^{-1}A|||^2 = \sup\{|||v - M^{-1}Av|||^2 : |||v||| = 1\}.$$

For simplicity of notation, we let $w = M^{-1}Av$, so $Av = Mw$. If $|||v||| = 1$, we see that

$$\begin{aligned} |||v - w|||^2 &= \langle A(v - w), v - w\rangle \\ &= 1 - \langle Aw, v\rangle - \langle Av, w\rangle + \langle Aw, w\rangle \\ &= 1 - \langle w, Mw\rangle - \langle Mw, w\rangle + \langle Aw, w\rangle \\ &= 1 - \langle M^*w, w\rangle - \langle Mw, w\rangle + \langle Aw, w\rangle \\ &= 1 - \langle (M^* + N)w, w\rangle. \end{aligned}$$

(This last identity is equivalent to formula (1.1).) At this point we need to estimate $\langle (M^* + N)w, w\rangle$ from below independently of v with $|||v||| = 1$, and it is here that we no longer follow the proof in [1]. (The proof in [1] concludes—in the finite dimensional case—by observing that the continuous function $v \to |||v - M^{-1}Av|||^2$ attains its maximum on the unit sphere with respect to the norm $|||\ \ |||$, and that this maximum is less that 1 since $M^* + N$ is positive definite.)

By hypothesis, the infimum of the spectrum of $M^* + N$ is a positive number, which we denote by $\delta > 0$. It follows (see, Kreyszig [8, Thm. 9.2-1]) that

$$\langle (M^* + N)w, w\rangle \geq \delta\|w\|^2.$$

Furthermore

$$1 = |||v||| = |||A^{-1}Mw||| = \|A^{-1/2}Mw\| \leq \|A^{-1/2}M\|\,\|w\|,$$

and so

$$\|w\| \geq \frac{1}{\|A^{-1/2}M\|}.$$

We see therefore that

$$\langle (M^* + N)w, w\rangle \geq \frac{\delta}{\|A^{-1/2}M\|^2},$$

from which we conclude that

$$|||M^{-1}N|||^2 \leq 1 - \frac{\delta}{\|A^{-1/2}M\|^2}.$$

To complete the proof, it suffices to note that

$$\|A^{-1/2}M\|^2 = \sup_{x\neq 0} \frac{\langle A^{-1/2}Mx, A^{-1/2}Mx\rangle}{\langle x,x\rangle} = \sup_{x\neq 0} \frac{\langle M^*A^{-1}Mx, x\rangle}{\langle x,x\rangle} = \rho(M^*A^{-1}M),$$

the last equality being justified since $M^*A^{-1}M = (A^{-1/2}M)^*A^{-1/2}M$ is selfadjoint and positive (see [8, Thms. 9.2-1, 9.2-2, and 9.2-3]). (This calculation shows that $\|T\|^2 = \rho(T^*T)$ for any bounded operator T.) □

Remarks. In the above proof, if we make the additional assumption that $M^* + N = rI$ for some (positive) real number r, then the chain of inequalities in the proof almost becomes an exact equality. To see this, note first that $\langle (M^* + N)w, w\rangle = r\langle w, w\rangle$. Next, for any small $\varepsilon > 0$, let v_ε and w_ε be such that $v_\varepsilon = A^{-1}Mw_\varepsilon$ and

$$1 = |||v_\varepsilon||| = |||A^{-1}Mw_\varepsilon||| = \|A^{-1/2}Mw_\varepsilon\| > (\|A^{-1/2}M\| - \varepsilon)\|w_\varepsilon\|.$$

It follows that

$$|||v_\varepsilon - w_\varepsilon|||^2 = 1 - r\|w_\varepsilon\|^2 > 1 - \frac{r}{(\|A^{-1/2}M\| - \varepsilon)^2},$$

this for all small $\varepsilon > 0$. One concludes easily that

$$|||M^{-1}N|||^2 = 1 - \frac{r}{\|A^{-1/2}M\|^2} = 1 - \frac{r}{\rho(M^*A^{-1}M)} = 1 - \frac{\inf\{\text{spectrum}(M^* + N)\}}{\rho(M^*A^{-1}M)}.$$

It remains then to determine how accurate an estimate $|||M^{-1}N|||^2$ is for $\rho(M^{-1}N)^2$. This point will be discussed below for a specific example. Note that if the diagonal elements of A are all equal, then the decomposition for successive overrelaxation gives $M^* + N = rI$ for some real r.

Finally, although a calculation similar to the above proof but based on formula (1.5) can also be used to obtain Theorem B, it was the functional analytic formulation in [1] that led naturally to the calculation in the proof of Theorem B.

3. Finite dimensional examples. It is, of course, interesting to see what the estimate of Theorem B gives for specific examples. In this section we consider the finite dimensional case, i.e., where A is a matrix.

The first example is the case where $M = \gamma I$, where $\gamma > 0$. Thus, $A = \gamma I - N$ and $M^* + N = \gamma I + N = 2\gamma I - A$. In this case, the hypotheses of Theorem A are simply that A is a Hermitian matrix whose eigenvalues belong to the open interval $(0, 2\gamma)$. We denote by λ_m and λ_M, respectively, the smallest and largest eigenvalues of A. Theorem B then implies that

$$\rho(M^{-1}N)^2 \leq 1 - \frac{\inf\{\text{spectrum}(2\gamma I - A)\}}{\gamma^2\rho(A^{-1})} = 1 - \frac{(2\gamma I - \lambda_M)\lambda_m}{\gamma^2} = 1 - \frac{2\lambda_m}{\gamma} + \frac{\lambda_m\lambda_M}{\gamma^2}$$

On the other hand, it is easy to see that

$$\rho(M^{-1}N)^2 = \rho(I - \gamma^{-1}A)^2 = \max\left\{1 - \frac{2\lambda_m}{\gamma} + \frac{(\lambda_m)^2}{\gamma^2}, 1 - \frac{2\lambda_M}{\gamma} + \frac{(\lambda_M)^2}{\gamma^2}\right\},$$

where the maximum is attained with λ_m if $\lambda_m \leq 2\gamma - \lambda_M$, and with λ_M if $\lambda_m \geq 2\gamma - \lambda_M$.

Next, we examine the consequences of Theorem B for the method of successive overrelaxation. Thus, we write

$$A = D - E - F,$$

where D is a diagonal matrix and E and F are strictly lower and upper triangular matrices, respectively. (For simplicity we do not treat block decompositions.) If $\omega \neq 0$, the matrices $M = M_\omega$ and $N = N_\omega$ are defined by

$$M = M_\omega = \frac{1}{\omega}(D - \omega E),$$

$$N = N_\omega = \frac{1}{\omega}(\omega F + (1 - \omega)D).$$

It is clear that $A = M_\omega - N_\omega$, this being the decomposition for successive overrelaxation. (The case $\omega = 1$ gives the Gauss–Seidel method.) The matrix $(M_\omega)^{-1}N_\omega$ is then given by

$$\mathcal{L}_\omega = (M_\omega)^{-1}N_\omega = (D - \omega E)^{-1}(\omega F + (1 - \omega)D).$$

Suppose now that A is a Hermitian, positive definite matrix. It follows that $E^* = F$ and therefore that

$$(M_\omega)^* + N_\omega = \frac{1}{\omega}(D - \omega E^*) + \frac{1}{\omega}(\omega F + (1 - \omega)D) = \frac{(2 - \omega)}{\omega}D.$$

Since A being positive definite implies that all its diagonal elements are positive, it follows immediately from Theorem A that if $0 < \omega < 2$, then the successive overrelaxation method converges, i.e., $\rho(\mathcal{L}_\omega) < 1$. Theorem B gives the following estimate:

$$\rho(\mathcal{L}_\omega)^2 \leq 1 - \frac{\omega(2 - \omega)\inf(a_{ii})}{\rho((D - \omega F)A^{-1}(D - \omega E))}, \tag{3.1}$$

where we denote the elements of A by (a_{ij}).

In evaluating the usefulness of this estimate, one may ask two different questions. First, how good is the estimate for a given value of ω? Second, what is the predicted optimal value of ω obtained from these estimates, i.e., the value of ω which minimizes the right side of the estimate, and how close is it to the real optimal value?

As a first test case, it is natural to apply this estimate to the $n \times n$ matrix $A = (a_{ij})$ given by

$$a_{ij} = \begin{cases} \gamma & \text{if} \quad i = j, \\ -1 & \text{if} \quad |i - j| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

where, for example, $\gamma \geq 2$. To apply the estimate (3.1), one needs to calculate (or estimate) $\rho((D - \omega F)A^{-1}(D - \omega E))$. For this particular matrix A, one can easily verify that

$$\lambda \in sp((D - \omega F)A^{-1}(D - \omega E))$$

if and only if

$$(D - \omega E)(D - \omega F) - \lambda A = (\omega^2 + \gamma^2 - \lambda\gamma)I + (\lambda - \omega\gamma)(E + F) + \omega^2 G$$

is not bijective, where $G = (g_{ij})$ is the matrix whose only nonzero entry is $g_{11} = -1$. It follows that

$$\lambda \in sp((D - \omega F)A^{-1}(D - \omega E)) \Leftrightarrow -(\omega^2 + \gamma^2 - \lambda\gamma) \in sp((\lambda - \omega\gamma)(E + F) + \omega^2 G).$$

Rather than calculate $sp((\lambda - \omega\gamma)(E + F) + \omega^2 G)$ exactly, we use an estimate that has the advantage of being simple and also independent of the dimension n (perhaps that is a disadvantage). Indeed, for any matrix $H = (h_{ij}), \rho(H) \leq \|H\|_\infty = \max_{1 \leq i \leq n} \sum_{1 \leq j \leq n} |h_{ij}|$, where $\| \quad \|_\infty$ denotes the matrix norm induced by the l^∞ vector norm. It follows that

$$\rho((\lambda - \omega\gamma)(E + F) + \omega^2 G) \leq \max[2|\lambda - \omega\gamma|, \omega^2 + |\lambda - \omega\gamma|]. \tag{3.2}$$

Thus, we finally see that if $\lambda \in sp((D - \omega F)A^{-1}(D - \omega E))$, then

$$|\omega^2 + \gamma^2 - \lambda\gamma| \leq \max[2|\lambda - \omega\gamma|, \omega^2 + |\lambda - \omega\gamma|]. \tag{3.3}$$

To simplify the calculations, we will assume from now on that γ is sufficiently large, e.g., $\gamma > 8$, so that special cases do not need to be considered. Also, since $(D - \omega F)A^{-1}(D - \omega E)$ is a positive definite matrix, we only need to consider $\lambda > 0$. A straightforward, albeit tedious, calculation shows that (3.3) implies

$$\lambda \leq \max\left[\frac{(\gamma - \omega)^2}{\gamma - 2}, \frac{2\omega^2 - \omega\gamma + \gamma^2}{\gamma - 1}, \frac{2\omega^2 + \omega\gamma + \gamma^2}{\gamma + 1}, \frac{(\gamma + \omega)^2}{\gamma + 2}\right]. \tag{3.4}$$

For simplicity of notation, let us denote by $f(\omega)$ the right-hand side of (3.4), and so $\rho((D - \omega F)A^{-1}(D - \omega E)) \leq f(\omega)$. Another straightforward, yet even more tedious, calculation shows that

$$f(\omega) = \begin{cases} \dfrac{(\gamma - \omega)^2}{\gamma - 2} & 0 \leq \omega \leq \omega_1, \\ \dfrac{2\omega^2 - \omega\gamma + \gamma^2}{\gamma - 1} & \omega_1 \leq \omega \leq \omega_2, \\ \dfrac{2\omega^2 + \omega\gamma + \gamma^2}{\gamma + 1} & \omega_2 \leq \omega \leq \omega_3, \\ \dfrac{(\gamma + \omega)^2}{\gamma + 2} & \omega_3 \leq \omega \leq 2, \end{cases}$$

where

$$\omega_1 = \frac{2}{1 + \sqrt{1 + \frac{4(\gamma - 3)}{\gamma^2}}}, \quad \omega_2 = \frac{2}{1 + \sqrt{1 - \frac{8}{\gamma^2}}}, \quad \omega_3 = \frac{2}{1 + \sqrt{1 - \frac{4(\gamma + 3)}{\gamma^2}}}.$$

Theorem B in the form of (3.1) now implies that

$$\rho(\mathcal{L}_\omega)^2 \leq 1 - \frac{\gamma\omega(2 - \omega)}{f(\omega)}. \tag{3.5}$$

It is relatively easy to see that $\gamma\omega(2 - \omega)/f(\omega)$ is increasing on the interval $[0, \omega_2]$ and decreasing on the interval $[\omega_2, 2]$, and so the value of ω for which (3.5) provides the lowest estimate of $\rho(\mathcal{L}_\omega)$ is

$$\omega_2 = \frac{2}{1 + \sqrt{1 - \frac{8}{\gamma^2}}}. \tag{3.6}$$

This value is slightly larger than the real optimal value of ω (see [18, §4.3], [21, Chap. 6], or [1, Thm. 5.3-6, §5.3]):

$$\omega_{\text{opt}} = \frac{2}{1+\sqrt{1-\frac{\rho(E+F)^2}{\gamma^2}}} = \frac{2}{1+\sqrt{1-\frac{4\cos^2(\pi/(n+1))}{\gamma^2}}}. \tag{3.7}$$

Note that $\omega_2 - \omega_{\text{opt}} = O(\gamma^{-2})$ for large γ.

It is also interesting to compare the actual spectral radius with the upper bound given by Theorem B. In the Gauss–Seidel case ($\omega = 1$), it is known (see, for example, [1, Thm. 5.3-4]) that

$$\rho(\mathcal{L}_1) = \rho(\gamma^{-1}(E+F))^2 = \frac{4\cos^2(\pi/(n+1))}{\gamma^2}. \tag{3.8}$$

On the other hand, Theorem B gives the estimate

$$\rho(\mathcal{L}_1)^2 \leq 1 - \frac{\gamma}{f(1)} = 1 - \frac{\gamma(\gamma-1)}{2-\gamma+\gamma^2} = \frac{2}{\gamma^2-\gamma+2}, \tag{3.9}$$

which is not a very good estimate of the true value, especially for large γ.

In other words, while the optimal value of the relaxation parameter predicted by Theorem B is quite accurate, the individual estimates for $\rho(\mathcal{L}_\omega)$ are not good at all.

There is, however, an explanation for this discrepancy. Theorem B is valid in Hilbert spaces and can therefore be applied to the infinite matrix having the same form as above, considered as a bounded selfadjoint operator on l^2. All the estimates obtained above from Theorem B apply equally well for the infinite matrix. However, as we shall presently see, while $\rho(\gamma^{-1}(E+F)) = 2/\gamma$ for the infinite matrix, *it is no longer true that* $\rho(\mathcal{L}_1) = \rho(\gamma^{-1}(E+F))^2$.

4. An infinite dimensional example. Let $A = A_\gamma\,(\gamma \in R)$ be the linear operator on $l^2 = l^2(\{0,1,2,3,\ldots\})$ whose infinite matrix $(a_{ij})_{ij=0,1,2,3,\ldots}$ is given by

$$a_{ij} = \begin{cases} \gamma & \text{if} \quad i = j, \\ -1 & \text{if} \quad |i-j| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

If we set $A = D - E - F$, where D is a diagonal operator ($D = \gamma I$), E is strictly lower triangular, and F strictly upper triangular, then E is the well-known shift operator

$$E(\xi_0, \xi_1, \xi_2, \xi_3, \ldots) = (0, \xi_0, \xi_1, \xi_2, \xi_3, \ldots).$$

We observe immediately that A is a bounded selfadjoint operator, and that

$$\|E+F\| \leq \|E\| + \|F\| = 2,$$

where $\|\ \|$ still denotes the operator norm, as well as the vector norm in l^2. The spectrum of $E+F$ (which is real since $E+F$ is selfadjoint) is thus contained in the closed interval $[-2,2]$. It follows that if $\gamma \geq 2$, then $A = A_\gamma$ is a *positive* selfadjoint operator, invertible at least if $\gamma > 2$. We will see shortly that A_2 is, in fact, not invertible.

As remarked above, the Theorem B estimates (3.5) and (3.9) are still valid for this example (again with $\gamma > 8$), and the "predicted optimal value" of ω is still given

by (3.6). To justify this assertion, recall first that the spectrum of a bounded operator H on a Hilbert space $\mathcal{H}$ is given by

$$sp(H) = \{\lambda \in \mathbf{C} : H - \lambda I \text{ is not bijective on } \mathcal{H}\}.$$

Indeed, if $H-\lambda I$ is bijective, then $(H-\lambda I)^{-1}$ is bounded by the open mapping theorem for bounded operators [8, Thm. 4.12-2] or [14, Thms. 5.9 and 5.10]. Next, to justify formula (3.2), we need to verify $\rho(H) \leq \|H\|_\infty$ for the infinite real symmetric matrix $H = (\lambda - \omega\gamma)(E + F) + \omega^2 G$. For any finite real symmetric matrix, we have $\|H\| = \rho(H) \leq \|H\|_\infty$ and so, passing to the limit, we see that $\|H\| \leq \|H\|_\infty$ for any infinite real symmetric matrix. This implies that $\rho(H) \leq \|H\|_\infty$. We see below, however, that the exact values of $\rho(\mathcal{L}_\omega)$ are not the limits as $n \to \infty$ of the corresponding finite dimensional values.

The deeper properties of the operator A are best studied under the unitary transformation between l^2 and the Hardy space $H^2 = H^2(U)$ of analytic functions on the unit disc $U = \{z \in \mathbf{C} : |z| \leq 1\}$ given by

$$(\xi_0, \xi_1, \xi_2, \xi_3, \ldots) \to f(z) = \sum_{k=0}^{\infty} \xi_k z^k.$$

In particular, H^2 contains precisely those holomorphic functions $\sum_{k=0}^{\infty} \xi_k z^k$ for which $\sum_{k=0}^{\infty} |\xi_k|^2$ is finite. For details on the space H^2, as well as this isomorphism, the reader may consult [14, Chap. 17]. By abuse of notation, we denote by A, D, E, F, and $\mathcal{L}_\omega$ the operators on H^2 induced by the corresponding matrix operators on l^2 under this Hilbert space isomorphism. It follows that for any $f \in H^2$,

$$\begin{aligned}
(Ef)(z) &= zf(z).\\
(Ff)(z) &= \frac{f(z) - f(0)}{z},\\
[(A - \lambda I)f](z) &= \frac{[-z^2 + (\gamma - \lambda)z - 1]f(z) + f(0)}{z},\\
[(\mathcal{L}_\omega - \lambda I)f](z) &= \frac{[\omega\lambda z^2 + (1 - \omega - \lambda)\gamma z + \omega]f(z) - \omega f(0)}{z(\gamma - \omega z)}.
\end{aligned}$$

It is now in principle a straightforward matter to determine the spectra of the operators A and $\mathcal{L}_\omega$.

PROPOSITION 4.1. *$sp(A_\gamma) = [\gamma - 2, \gamma + 2]$ for all $\gamma \in \mathrm{R}$. Furthermore, A_γ has no eigenvalues.*

Proof. It is clear that $(A - \lambda I)f = g$ if and only if for all $z \in U$,

$$f(z) = \frac{zg(z) - f(0)}{-z^2 + (\gamma - \lambda)z - 1} = \frac{zg(z) - f(0)}{-(z - \zeta_1)(z - \zeta_2)},$$

where ζ_1 and ζ_2 are the two roots of $z^2 - (\gamma - \lambda)z + 1$. In particular $\zeta_1\zeta_2 = 1$. I claim that $A - \lambda I$ is bijective on H^2 if and only if one of the two roots belongs to U. Indeed, if $|\zeta_1| < 1$, then given $g \in H^2$, there exists a unique $f \in H^2$ such that $(A - \lambda I)f = g$, and f is given by

$$f(z) = \frac{zg(z) - \zeta_1 g(\zeta_1)}{-(z - \zeta_1)(z - \zeta_2)}.$$

Since $|\zeta_2| > 1$, the function f is holomorphic in U and by Theorem 17.9 in [14], it is in H^2. Also, any choice of $f(0)$ other than $\zeta_1 g(\zeta_1)$ would result in a function f having a singularity at $z = \zeta_1$. In this case, therefore, $A - \lambda I$ is bijective. On the other hand, if $|\zeta_1| = |\zeta_2| = 1$, then $A - \lambda I$ is not bijective. Indeed, if $g(z) \equiv 1$ and $(A - \lambda I)f = g$, then $f(z)$ must have a singularity either at ζ_1 or at ζ_2 : the choice of $f(0)$ can only eliminate one of them, and if $\zeta_1 = \zeta_2$, then setting $f(0) = \zeta_1$ still cannot eliminate the singularity. By Theorem 17.10 in [14], this singularity on the unit circle prevents f from belonging to H^2 (since f restricted to the circle is not square integrable). Therefore, if $|\zeta_1| = |\zeta_2| = 1$, then $A - \lambda I$ is not surjective. A straightforward calculation shows that $|\zeta_1| = |\zeta_2| = 1$ if and only if $\gamma - \lambda \in [-2, 2]$, proving the first statement of the proposition. In particular, A_2 is not an invertible positive operator, i.e., it is semidefinite. Also, setting $\gamma = 0$, we see that $sp(E + F) = [-2, 2]$.

Next we show that A has no eigenvalues. Indeed, λ is an eigenvalue of A if and only if there exists $f \in H^2, f$ not identically zero, such that $(A - \lambda I)f \equiv 0$, which implies that

$$f(z) = \frac{-f(0)}{-z^2 + (\gamma - \lambda)z - 1} = \frac{-f(0)}{-(z - \zeta_1)(z - \zeta_2)}.$$

Such a function is in H^2 only if $|\zeta_1| > 1$ and $|\zeta_2| > 1$, which is impossible since $\zeta_1 \zeta_2 = 1$. □

PROPOSITION 4.2. *Let* $0 < \omega < 2 < \gamma$. *Then*

$$\rho(\mathcal{L}_\omega)^2 = 1 - \gamma\omega(2 - \omega) \min\left[\frac{\gamma - 2}{(\gamma - \omega)^2}, \frac{\gamma + 2}{(\gamma + \omega)^2}\right].$$

Furthermore, the interior of $sp(\mathcal{L}_\omega)$ *is nonempty, and every interior spectral value is an eigenvalue.*

Proof. Since $0 < \omega < 2$ and $\gamma > 2$, we know from Theorem B and the previous proposition that $\rho(\mathcal{L}_\omega) < 1$; and so we need only consider complex λ such that $|\lambda| < 1$. Now $(\mathcal{L}_\omega - \lambda I)f = g$ if and only if for all $z \in U$,

$$f(z) = \frac{z(\gamma - \omega z)g(z) + \omega f(0)}{\omega\lambda z^2 + (1 - \omega - \lambda)\gamma z + \omega} = \frac{z(\gamma - \omega z)g(z) + \omega f(0)}{\omega\lambda(z - \zeta_1)(z - \zeta_2)},$$

where ζ_1 and ζ_2 are the two roots of

$$P(z) = P_{\omega,\lambda}(z) = z^2 + \frac{(1 - \omega - \lambda)\gamma}{\omega\lambda} z + \frac{1}{\lambda}.$$

(We suppose that $\lambda \neq 0$.) As in the previous proof, we use Theorems 17.9 and 17.10 in [14] to determine when $\mathcal{L}_\omega - \lambda I$ is bijective on $H^2(U)$. First, if $|\zeta_1| > 1$ and $|\zeta_2| > 1$, then any choice of $f(0)$ is acceptable and $\mathcal{L}_\omega - \lambda I$ is not injective. Next, if $|\zeta_1| < 1$ and $|\zeta_2| > 1$, then precisely one choice of $f(0)$ will work, i.e., $f(0) = -\zeta_1(\gamma - \omega\zeta_1)g(\zeta_1)/\omega$, and so $\mathcal{L}_\omega - \lambda I$ is bijective. Finally, if $|\zeta_1| = 1$ and $|\zeta_2| > 1$, then for some $g \in H^2$ (one which is singular at ζ_1) no appropriate f can be found, and so $\mathcal{L}_\omega - \lambda I$ is not surjective. (Here we use the fact that $\gamma/\omega > 1$.) Since $|\lambda| < 1$, we have $|\zeta_1\zeta_2| > 1$, and thus all cases have been considered. In particular, if λ is a spectral value, it is necessarily an eigenvalue except if one of ζ_1 and ζ_2 has modulus 1. (If $\lambda = 0$, the same conclusions hold if the two roots are considered to be $-\omega/[(1 - \omega)\gamma]$ and ∞, which are the limits of the two roots as $\lambda \to 0$. We keep this convention throughout the proof.)

If λ is such that both $|\zeta_1| > 1$ and $|\zeta_2| > 1$, then the same is true for nearby values of λ, which implies that λ is in the interior of $sp(\mathcal{L}_\omega)$. For example, this is the case

if $\lambda = 1 - \omega$; and so it follows that the spectrum of $\mathcal{L}_\omega$ always contains a nonempty open set of complex numbers. (In contrast to the finite dimensional case, this implies that $\omega - 1$ can never be the spectral radius of $\mathcal{L}_\omega$.) It also follows that the boundary of $sp(\mathcal{L}_\omega)$ is contained in the set of λ such that one of the two roots ζ_1 and ζ_2 has modulus equal to 1. Therefore, the spectral radius $\rho(\mathcal{L}_\omega)$ is achieved by a spectral value λ such that one of the two roots ζ_1 and ζ_2 has modulus equal to 1, i.e., is equal to $e^{i\theta}$ for some real θ.

From the quadratic formula, it is a straightforward calculation to see that one of the roots of $P_{\omega,\lambda}$ is equal to $e^{i\theta}$, for some real θ, if and only if

$$\lambda e^{i\theta} = \frac{(1-\omega)\gamma e^{i\theta} + \omega}{\gamma - \omega e^{i\theta}}. \tag{4.1}$$

Thus,

$$\begin{aligned}
\rho(\mathcal{L}_\omega)^2 &= \sup\{|\lambda|^2 : \lambda \text{ verifies (4.1) for some } \theta \in \mathbf{R}\} \\
&= \sup_{\theta \in \mathbf{R}} \frac{|(1-\omega)\gamma e^{i\theta} + \omega|^2}{|\gamma - \omega e^{i\theta}|^2} \\
&= \sup_{\theta \in [0,\pi]} \frac{(1-\omega)^2\gamma^2 + 2\omega(1-\omega)\gamma(\cos\theta) + \omega^2}{\gamma^2 - 2\omega\gamma(\cos\theta) + \omega^2} \\
&= \sup_{t \in [-1,1]} \left[1 - \gamma\omega(2-\omega)\frac{\gamma - 2t}{\omega^2 - 2\omega\gamma t + \gamma^2}\right],
\end{aligned}$$

where we have substituted $t = \cos\theta$. It is straightforward to check that the fraction in this last expression is a monotone function of t, and so the above supremum is realized either with $t = 1$ or with $t = -1$. The desired formula for $\rho(\mathcal{L}_\omega)^2$ follows immediately.

To complete the proof, we need to show that if λ is a spectral value such that one of the two roots of $P_{\omega,\lambda}$ is of modulus 1, then λ is on the boundary of $sp(\mathcal{L}_\omega)$. For such a λ, the two roots are distinct and therefore locally holomorphic and non-constant functions of λ. The open mapping theorem for analytic functions [14, Thm. 10.32] implies that for some nearby λ, one of the roots is inside U. Such λ are not in $sp(\mathcal{L}_\omega)$. □

The particular case $\omega = 1$ in Proposition 4.2 gives

$$\rho(\mathcal{L}_1)^2 = 1 - \gamma\frac{\gamma - 2}{(\gamma - 1)^2} = \frac{1}{(\gamma - 1)^2}. \tag{4.2}$$

One notes immediately that the value of $\rho(\mathcal{L}_1)$ given in (4.2) is *not* the limit of the finite dimensional values given by (3.8). On the other hand, the estimate (3.9) is a much more reasonable approximation to (4.2) than it is to (3.8). In addition, the estimate (3.5) *coincides* with the computed value of $\rho(\mathcal{L}_\omega)^2$ in Proposition 4.2 if $0 < \omega < \omega_1$ and if $\omega_3 < \omega < 2$. Thus, at least for these values of ω, the estimate provided by Theorem B in this example gives the exact value of $\rho(\mathcal{L}_\omega)$. I suspect this is true for all $\omega \in (0, 2)$. To verify this assertion, one would have to improve the estimate (3.2), which apparently is already sharp for the above values of ω in the infinite dimensional case.

It is straightforward to show using Proposition 4.2 that the optimal value of ω, i.e., the value for which $\rho(\mathcal{L}_\omega)$ is minimized, is given by

$$\omega_{\text{opt}} = \frac{2}{1 + \sqrt{1 - \frac{4}{\gamma^2}}}, \tag{4.3}$$

which *is* the limit as $n \to \infty$ of the finite dimensional optimal values given by (3.7).

So what goes wrong as $n \to \infty$? The spectral radii $\rho(\mathcal{L}_\omega)$ do not converge to the corresponding values for the infinite matrix; but miraculously, the optimal relaxation parameter does converge to the optimal parameter for the infinite matrix. In finite dimensions, one knows that $\rho(\mathcal{L}_1) = \rho(\gamma^{-1}(E+F))^2$, and that the nonzero eigenvalues of $\mathcal{L}_1$ are the squares of the eigenvalues of $\gamma^{-1}(E+F)$. (See, for example, [1, Thm. 5.3-4] and its proof.) As $n \to \infty$, the spectrum of $\gamma^{-1}(E+F)$, a discrete subset of $[-2/\gamma, 2/\gamma]$, becomes this entire interval; and the infinite matrix $\gamma^{-1}(E+F)$ has no eigenvalues. On the other hand, not only is it no longer true that $\rho(\mathcal{L}_1) = \rho(\gamma^{-1}(E+F))^2$, but also the interior of the spectrum of $\mathcal{L}_1$ consists entirely of eigenvalues and includes some nonreal complex eigenvalues.

The finite dimensional proof that $\rho(\mathcal{L}_1) = \rho(\gamma^{-1}(E+F))^2$ is based on the formula

$$\lambda Q_\lambda (E + F - \lambda\gamma I) Q_{1/\lambda} = \lambda^2 E + F - \lambda^2 \gamma I,$$

where Q_λ is the diagonal matrix whose diagonal elements are successively $1, \lambda, \lambda^2, \lambda^3$, etc. While this formula is still correct in infinite dimensions, the matrix $Q_{1/\lambda}$ is no longer a bounded operator on l^2, and so conclusions about the spectrum are suspect. The matter becomes clear if we interpret the matrices Q_λ and $Q_{1/\lambda}$ as operators on analytic functions. Indeed $(Q_\lambda f)(z) = f(\lambda z)$ and $(Q_{1/\lambda} f)(z) = f(z/\lambda)$. In particular, Q_λ is a bijection $H^2(U_{|\lambda|}) \to H^2(U)$ and $Q_{1/\lambda}$ is a bijection $H^2(U) \to H^2(U_{|\lambda|})$, where U_r is the set of complex numbers with $|z| < r$. It follows (for $\lambda \neq 0$) that $\lambda^2 E + F - \lambda^2\gamma I$ is *not* a bijection on $H^2(U)$, i.e., that λ^2 is a spectral value of $\mathcal{L}_1$, if and only if $(E + F - \lambda\gamma I)$ is *not* a bijection on $H^2(U_{|\lambda|})$. Since $|\lambda| < 1, U_{|\lambda|} \subset U$ (strict inclusion), and so $H^2(U_{|\lambda|})$ is a larger space than $H^2(U)$: it includes, for example, *all* functions analytic in U. There exist values of λ for which $(E + F - \lambda\gamma I)$ is a bijection on $H^2(U)$, but not on $H^2(U_{|\lambda|})$. To verify this last assertion, note that $(E + F - \lambda\gamma I)$ is just $-A_{\lambda\gamma}$ and is a bijection on $H^2(U_r)$ if and only if exactly one of the two roots of $z^2 - \lambda\gamma z + 1$ is contained in U_r. There certainly exist values of λ (such as $\lambda = (2+\varepsilon)/\gamma$ for $\varepsilon > 0$ small) for which this is true in $U = U_1$, but not in $U_{|\lambda|}$. (If it is true in $U_{|\lambda|}$, it is true in U, since the product of the roots is 1.) Thus, the spectrum of $\mathcal{L}_1$ contains squares of much more than the spectral values of $\gamma^{-1}(E+F)$: it contains all the squares of spectral values of $\gamma^{-1}(E+F)$ acting on the larger space $H^2(U_{|\lambda|})$. Also, $(E+F)$ is not selfadjoint on this larger space, and so there is no reason to have only real spectral values.

5. Concluding remarks. It seems that the optimal relaxation parameter is more stable than the individual spectral radii. This suggests the following approach to estimating the optimal parameter for certain large matrices. Suppose that A is an infinite matrix whose optimal parameter ω_∞ can be explicitly computed using techniques similar to those used above. Suppose furthermore that the optimal relaxation parameters ω_n for the principal $n \times n$ submatrices A_n of A tend to ω_∞ as $n \to \infty$. Then ω_∞ would provide a good approximation to the optimal parameter of A_n if n is large. The example studied above suggests that this might hold true under some general hypotheses.

Acknowledgments. I thank the referees for several helpful comments and suggestions that led, in particular, to the above historical survey, as well as to the application of Theorem B to successive overrelaxation. Also, I thank P. Ciarlet and G. Tronel for their interest in this work.

REFERENCES

[1] P. G. CIARLET, *Introduction à l'Analyse Numérique Matricielle et à l'Optimisation,* Masson, Paris, 1985, English translation, Introduction to Numerical Linear Algebra and Optimisation, Cambridge University Press, Cambridge, 1989.

[2] G. J. HABETLER, *Concerning the Implicit Alternating-Direction Method,* Report KAPL-2040, Knolls Atomic Power Laboratory, Schenectady, NY, 1959.

[3] A. S. HOUSEHOLDER, *On the Convergence of Matrix Iterations,* Tech. Report No. 1883, Oak Ridge National Laboratory, Oak Ridge, TN, 1955.

[4] ———, *The approximate solution of matrix problems,* J. Assoc. Comput. Mach., 5 (1958), pp. 204–243.

[5] ———, *The Theory of Matrices in Numerical Analysis,* Blaisdell, New York, 1964.

[6] F. JOHN, *Lectures on Advanced Numerical Analysis,* Gordon and Breach, New York, 1967; based on lecture notes, Institute of Mathematical Sciences, New York University, 1956–1957.

[7] H. B. KELLER, *On the solution of singular and semidefinite linear systems by iteration,* SIAM J. Numer. Anal., 2 (1965), pp. 281–290.

[8] E. KREYSZIG, *Introductory Functional Analysis with Applications,* John Wiley & Sons, New York, 1978.

[9] P. LASCAUX AND R. THÉODOR, *Analyse Numérique Matricielle Appliquée à l'Art de l'Ingénieur,* Vol. 2, Masson, Paris, 1987.

[10] J. M. ORTEGA, *Numerical Analysis, A Second Course,* Academic Press, New York, 1972.

[11] J. M. ORTEGA AND R. J. PLEMMONS, *Extensions of the Ostrowski–Reich theorem for* SOR *iterations,* Linear Algebra Appl., 28 (1979), pp. 177–191.

[12] A. M. OSTROWSKI, *On the linear iteration procedures for symmetric matrices,* Rend. Mat. Appl., 14 (1954), pp. 140–163.

[13] E. REICH, *On the convergence of the classical iterative method of solving linear simultaneous equations,* Ann. Math. Statist., 20 (1949), pp. 448–451.

[14] W. A. RUDIN, *Real and Complex Analysis,* McGraw-Hill, New York, 1966.

[15] L. SEIDEL, *Über ein Verfahren die Gleichungen, auf welche die Methode der kleinsten Quadrate führt, sowie lineare Gleichungen überhaupt, durch successive Annäherung aufzulösen,* Abhandlungen der Mathematisch-Physikalischen Classe der Königlich Bayerischen Akademie der Wissenschaften, II (1874), pp. 81–108.

[16] M. SCHATZMAN, *Analyse Numérique: Cours et exercices pour la licence,* InterEditions, Paris, 1991.

[17] P. STEIN, *Some general theorems on iterants,* J. Res. Nat. Bur. Standards, 48 (1952), pp. 82–83.

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

[19] E. L. WACHSPRESS, *Iterative Solution of Elliptic Systems, and Applications to the Neutron Diffusion Equations of Reactor Physics,* Prentice-Hall, Englewood Cliffs, NJ, 1966.

[20] J. WEISSINGER, *Verallgemeinerungen des Seidelschen Iterationsverfahrens,* Z. Angew. Math. Mech., 33 (1953), pp. 155–163.

[21] D. L. YOUNG, *Iterative Solution of Large Linear Systems,* Academic Press, New York, 1971.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 462–468, April 1995

010

A PRACTICAL UPPER BOUND FOR DEPARTURE FROM NORMALITY *

STEVEN L. LEE†

Abstract. The departure from normality of a matrix is a real scalar that is impractical to compute if the matrix is large and its eigenvalues are unknown. A simple formula is presented for computing an upper bound for departure from normality in the Frobenius norm. This new upper bound is cheaper to compute than the one derived by Henrici [*Numer. Math.*, 4 (1962), pp. 24 – 40]. Moreover, the new bound is sharp for Hermitian matrices, skew-Hermitian matrices and, in general, any matrix with eigenvalues that are horizontally or vertically aligned in the complex plane. In terms of applications, the new bound can be used in computing bounds for the spectral norm of matrix functions or bounds for the sensitivity of eigenvalues to matrix perturbations.

Key words. nonnormal matrix, departure from normality, condition numbers

AMS subject classifications. 65F35, 15A60, 15A12

1. Introduction.

The departure from normality of a matrix, like the condition number of a matrix, is a real scalar that can be used to compute bounds for various matrix computations. For example, departure from normality can be used to bound the powers, inverses, spectral variation, and fields of values of nonnormal matrices [8] or the spectral norm of matrix functions [1]. Unfortunately, the departure from normality of a matrix is impractical to compute if the matrix is large and its eigenvalues are unknown. The main result of this paper is a simple formula for computing an upper bound for departure from normality in the Frobenius norm. This new upper bound is cheaper to compute than the one derived by Henrici [8], and it is sharp for any matrix with eigenvalues that are horizontally or vertically aligned in the complex plane. The practical significance is that the new upper bound can be used in computing bounds for many of the matrix computations described in [1], [8].

The outline of this paper is as follows. In §2, we establish notation, motivate the definition of departure from normality, and give Henrici's upper bound [8]. In §3, we derive a new upper bound and prove that it is sharp for certain classes of matrices. In §4, we conclude with some numerical results that compare the tightness of Henrici's bound and the new one.

2. Preliminaries.

Let $A = (a_{ij})$ denote an $n \times n$ complex matrix and let $A^H = (\bar{a}_{ji})$ denote the conjugate transpose of A. (Herein, all matrices are square matrices of order n with real or complex entries.) Several important classes of matrices are defined in terms of their conjugate transpose: for example, A is Hermitian if and only if (iff) $A^H = A$, A is skew-Hermitian iff $A^H = -A$, and A is unitary iff $A^HA = AA^H = I$. Let M and N denote the Hermitian and skew-Hermitian parts of A, respectively. Indeed, let the functions $\mathcal{H}(\cdot)$ and $\mathcal{S}(\cdot)$ extract the Hermitian and

* Received by the editors September 9, 1993; accepted for publication (in revised form) by N. J. Higham, January 13, 1994. This research was supported by Applied Mathematical Sciences Research Program, Office of Energy Research, U.S. Department of Energy contract DE-AC05-84OR21400 with Martin Marietta Energy Systems Inc., National Science Foundation grant NSF DMS 90-15533, National Center for Supercomputing Applications and HPCC program under contract NASA NAG 5-2201.

† Mathematical Sciences Section, Oak Ridge National Laboratory, P.O. Box 2008, Building 6012, Oak Ridge, Tennessee 37831-6367 (na.slee@na-net.ornl.gov).

skew-Hermitian part of any square matrix. Then, with

$$\mathcal{H}(A) := \frac{1}{2}(A + A^H) \equiv M \tag{1}$$

and

$$\mathcal{S}(A) := \frac{1}{2}(A - A^H) \equiv N, \tag{2}$$

A has the splitting

$$A = M + N. \tag{3}$$

Let R denote an upper triangular matrix, T a strictly upper triangular matrix, U a unitary matrix, and Λ a diagonal matrix whose entries are the eigenvalues, λ_i, of A. Let $\mathrm{Re}(\Lambda)$ and $\mathrm{Im}(\Lambda)$ denote the real and imaginary parts of Λ so that

$$\Lambda = \mathrm{Re}(\Lambda) + i\,\mathrm{Im}(\Lambda). \tag{4}$$

Finally, recall that A is normal iff, for example, [7] the following are true.

(5a) A has a complete, orthogonal set of eigenvectors,

$$\|A\|_F = \|\Lambda\|_F = \sum |\lambda_i|^2, \text{ or} \tag{5b}$$

$$A^H A = A A^H \tag{5c}$$

The set of normal matrices includes the Hermitian, skew-Hermitian, and unitary matrices and, in general, any matrix that is unitarily similar to a diagonal matrix. Thus, any Schur decomposition of a normal matrix gives

$$U^H A U = R = \Lambda + T, \tag{6}$$

where $T = 0$. For a matrix that is not normal, it is convenient to quantify its departure from normality in terms of a norm of T.

DEFINITION 2.1 (Departure from Normality [8]). For any $n \times n$ matrix A,

$$\mathrm{dep}_F(A) := \|T\|_F = \left(\|A\|_F^2 - \|\Lambda\|_F^2\right)^{1/2}. \tag{7}$$

It is easily seen that $\mathrm{dep}_F(A)$ is independent of the choice of U and invariant with respect to complex shifts and rotations. That is,

$$\mathrm{dep}_F(A) = \mathrm{dep}_F(e^{-i\theta}(A - \alpha I)) \tag{8}$$

for any complex scalar α and $0 \le \theta < 2\pi$. Later, we show the significance of this observation.

More than a dozen measures of nonnormality have been proposed [3]. The choice $\mathrm{dep}_F(A)$ is especially useful since the most natural measure of nonnormality,

$$\nu_F(A) = \min\{\|E\|_F : A + E \text{ is normal}\}, \tag{9}$$

can be bounded from below [11] and above [3], [13] via

$$\mathrm{dep}_F(A)/\sqrt{n} \le \nu_F(A) \le \mathrm{dep}_F(A) \tag{10}$$

The difficult problem of finding the closest normal matrix to A in the Frobenius norm has been completely solved by Gabriel [4], [5] and, independently, by Ruhe [13]. A recent treatment of this and other matrix nearness problems is given in [9].

For normal matrices, $\mathrm{dep}_F(A) = 0$ via (5b). For nonnormal matrices, $\mathrm{dep}_F(A)$ is the nonzero quantity defined by (7). To be clear, a small example helps to summarize the main ideas up to this point.

Example 1. Two different Schur decompositions of

$$A = \begin{pmatrix} 4 & \sqrt{1/2} & -3 \\ \sqrt{2} & 2 & -\sqrt{2} \\ 1 & \sqrt{1/2} & 0 \end{pmatrix} \tag{11}$$

can be written

$$U_1^H A U_1 = R_1 = \begin{pmatrix} 1 & 1 & 4 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{pmatrix} \quad \text{and} \quad U_2^H A U_2 = R_2 = \begin{pmatrix} 2 & -1 & 3\sqrt{2} \\ 0 & 1 & \sqrt{2} \\ 0 & 0 & 3 \end{pmatrix}. \tag{12}$$

Evidently A has eigenvalues $\lambda(A) = \{1, 2, 3\}$, and it is nonnormal since R_1 and R_2 are not diagonal. The strictly upper triangular parts of R_1 and R_2 have equal norms, namely, $\mathrm{dep}_F(A) := \|T\|_F = \sqrt{21}$.

The value of $\mathrm{dep}_F(A)$ is impractical to compute if A is large and its eigenvalues are unknown. Lower bounds for $\mathrm{dep}_F(A)$ have been derived by Eberlein [2]

$$\mathrm{dep}_F(A) \geq \frac{\|A^H A - A A^H\|_F}{\sqrt{6}\,\|A\|_F} \tag{13}$$

and Loizou [12, Thm. 2]

$$\mathrm{dep}_F(A) \geq \frac{\|A^H A - A A^H\|_F}{\beta^{1/2} + \left(\beta + \sqrt{2}\,\|A^H A - A A^H\|_F\right)^{1/2}}, \tag{14}$$

where

$$\beta = \|A\|_F^2 - \frac{1}{n}\,|\mathrm{tr}(A)|^2, \tag{15}$$

$\mathrm{tr}(A)$ is the trace of A, and $A \neq 0$. The following upper bound is due to Henrici.

THEOREM 2.2. [8, Thm. 1]. *For any $n \times n$ matrix A,*

$$\mathrm{dep}_F(A) \leq \left(\frac{n^3 - n}{12}\right)^{1/4} \left(\|A^H A - A A^H\|_F\right)^{1/2}. \tag{16}$$

The bounds (13), (14), and (16) reduce to zero when A is normal. Unfortunately, all of them involve the matrix-matrix computation $A^H A - A A^H$, which generally requires $O(n^3)$ multiplications.

3. A practical upper bound. A trivial upper bound for $\mathrm{dep}_F(A)$ comes from its definition:

$$\mathrm{dep}_F(A) = \left(\|A\|_F^2 - \|\Lambda\|_F^2\right)^{1/2} \leq \|A\|_F. \tag{17}$$

We can obtain better upper bounds by manipulating expressions that arise after splitting A into its Hermitian and skew-Hermitian parts.

LEMMA 3.1. *If $A = M + N$ has the Schur decomposition $U^H A U = R = \Lambda + T$, then*

$$\|M\|_F^2 = \|\mathrm{Re}(\Lambda)\|_F^2 + \frac{1}{2}\|T\|_F^2, \tag{18}$$

$$\|N\|_F^2 = \|\mathrm{Im}(\Lambda)\|_F^2 + \frac{1}{2}\|T\|_F^2. \tag{19}$$

Proof. We compute

$$\|M\|_F^2 = \|\mathcal{H}(A)\|_F^2 = \|\mathcal{H}(R)\|_F^2 = \|\mathcal{H}(\Lambda)\|_F^2 + \|\mathcal{H}(T)\|_F^2 = \|\mathrm{Re}(\Lambda)\|_F^2 + \frac{1}{2}\|T\|_F^2, \tag{20}$$

$$\|N\|_F^2 = \|\mathcal{S}(A)\|_F^2 = \|\mathcal{S}(R)\|_F^2 = \|\mathcal{S}(\Lambda)\|_F^2 + \|\mathcal{S}(T)\|_F^2 = \|\mathrm{Im}(\Lambda)\|_F^2 + \frac{1}{2}\|T\|_F^2. \tag{21}$$

These equalities can also be found in [14, p. 495]. □

Equations (18) and (19) show that $\mathrm{dep}_F(A)$ can be defined in two different, but equivalent, ways:

$$\mathrm{dep}_F(A) = \sqrt{2}\ \left(\|M\|_F^2 - \|\mathrm{Re}(\Lambda)\|_F^2\right)^{1/2} \tag{22}$$

or

$$\mathrm{dep}_F(A) = \sqrt{2}\ \left(\|N\|_F^2 - \|\mathrm{Im}(\Lambda)\|_F^2\right)^{1/2}. \tag{23}$$

A simple upper bound for $\mathrm{dep}_F(A)$ follows directly from these equalities.

LEMMA 3.2. *If $A = M + N$, where M is the Hermitian part of A and N is the skew-Hermitian part of A, then*

$$\mathrm{dep}_F(A) \le \sqrt{2}\ \min\left\{\|M\|_F, \|N\|_F\right\}. \tag{24}$$

Proof. The upper bound is obtained from (22) and (23) by dropping the terms $\|\mathrm{Re}(\Lambda)\|_F^2$ and $\|\mathrm{Im}(\Lambda)\|_F^2$, respectively. □

We now consider the use of a complex shift αI for improving the bound (24). As in Lemma 3.1, we can split $A - \alpha I$ into its Hermitian part $M - \mathrm{Re}(\alpha)I$ and skew-Hermitian part $N - i\,\mathrm{Im}(\alpha)I$ and then rearrange terms to obtain

$$\mathrm{dep}_F(A) = \mathrm{dep}_F(A - \alpha I) = \sqrt{2}\ \left(\|M - \mathrm{Re}(\alpha)I\|_F^2 - \|\mathrm{Re}(\Lambda) - \mathrm{Re}(\alpha)I\|_F^2\right)^{1/2} \tag{25}$$

and

$$\mathrm{dep}_F(A) = \mathrm{dep}_F(A - \alpha I) = \sqrt{2}\ \left(\|N - i\,\mathrm{Im}(\alpha)I\|_F^2 - \|\mathrm{Im}(\Lambda) - \mathrm{Im}(\alpha)I\|_F^2\right)^{1/2}. \tag{26}$$

A tighter bound can be obtained by minimizing the terms

$$\|\mathrm{Re}(\Lambda) - \mathrm{Re}(\alpha)I\|_F^2 \quad \text{and} \quad \|\mathrm{Im}(\Lambda) - \mathrm{Im}(\alpha)I\|_F^2 \tag{27}$$

before dropping them from (25) and (26), respectively. In particular,

$$f_1(\mathrm{Re}(\alpha)) = \|\mathrm{Re}(\Lambda) - \mathrm{Re}(\alpha)I\|_F^2 = \sum |\mathrm{Re}(\lambda_i) - \mathrm{Re}(\alpha)|^2 \tag{28}$$

and

$$f_2(\mathrm{Im}(\alpha)) = \|\mathrm{Im}(\Lambda) - \mathrm{Im}(\alpha)I\|_F^2 = \sum |\mathrm{Im}(\lambda_i) - \mathrm{Im}(\alpha)|^2 \tag{29}$$

are quadratic functions that can be minimized using standard calculus techniques. By solving $f_1' = 0$ and $f_2' = 0$, we find that

$$\mathrm{Re}(\alpha) = \frac{\sum \mathrm{Re}(\lambda_i)}{n} \quad \text{and} \quad \mathrm{Im}(\alpha) = \frac{\sum \mathrm{Im}(\lambda_i)}{n}. \tag{30}$$

These values minimize (28) and (29), respectively, since f_1'' and f_2'' are positive. Thus, both terms of (27) are minimized by choosing

$$\alpha = \mathrm{Re}(\alpha) + i\,\mathrm{Im}(\alpha) = \frac{\sum \mathrm{Re}(\lambda_i)}{n} + \frac{i\sum \mathrm{Im}(\lambda_i)}{n} = \frac{\sum \lambda_i}{n} = \frac{\mathrm{tr}(A)}{n}. \tag{31}$$

THEOREM 3.3. *If $A = M + N$, where M is the Hermitian part of A and N is the skew-Hermitian part of A, then*

$$\mathrm{dep}_F(A) \le \sqrt{2}\, \min\{\|M - \mathrm{Re}(\alpha)I\|_F, \|N - i\,\mathrm{Im}(\alpha)I\|_F\}, \tag{32}$$

where the upper bound is minimized for

$$\alpha = \frac{\mathrm{tr}(A)}{n}. \tag{33}$$

Moreover, the bound is sharp (i.e., equality holds) iff the eigenvalues of A are horizontally or vertically aligned in the complex plane.

Proof. The upper bound is obtained from (25) and (26) by dropping the terms $\|\mathrm{Re}(\Lambda) - \mathrm{Re}(\alpha)I\|_F^2$ and $\|\mathrm{Im}(\Lambda) - \mathrm{Im}(\alpha)I\|_F^2$, respectively. The bound is sharp iff

$$\|\mathrm{Re}(\Lambda) - \mathrm{Re}(\alpha)I\|_F^2 = 0 \tag{34}$$

or

$$\|\mathrm{Im}(\Lambda) - \mathrm{Im}(\alpha)I\|_F^2 = 0. \tag{35}$$

The first condition (34) says that the real parts of the eigenvalues of A are constant (i.e., the eigenvalues are vertically aligned). The second condition (35) says that the imaginary parts of the eigenvalues of A are constant (i.e., the eigenvalues are horizontally aligned). □

Equations (25) and (26) also show that

$$\mathrm{dep}_F(A) \approx \sqrt{2}\, \min\{\|M - \mathrm{Re}(\alpha)I\|_F, \|N - i\,\mathrm{Im}(\alpha)I\|_F\} \tag{36}$$

iff

$$\|\mathrm{Re}(\Lambda) - \mathrm{Re}(\alpha)I\|_F^2 \ll \|M - \mathrm{Re}(\alpha)I\|_F^2 \tag{37}$$

or

$$\|\mathrm{Im}(\Lambda) - \mathrm{Im}(\alpha)I\|_F^2 \ll \|N - i\,\mathrm{Im}(\alpha)I\|_F^2. \tag{38}$$

Thus, the new bound (32) is a good approximation when the eigenvalues of A are relatively close to being horizontally or vertically aligned; otherwise, the bound is weak.

Example 2. Let us compare the Henrici bound (16) and the new bound (32) for the matrix A in Example 1, in which $\mathrm{dep}_F(A) = \sqrt{21}$. Using the intermediate quantities

$$\|A^HA - AA^H\|_F = \sqrt{996},\ \alpha = 2,\ \|M - 2I\|_F = \sqrt{25/2} \text{ and } \|N\|_F = \sqrt{21/2}, \tag{39}$$

the Henrici bound gives

$$\mathrm{dep}_F(A) \le (2)^{1/4}\,(\sqrt{996})^{1/2} = (\sqrt{1992})^{1/2} \approx \sqrt{44.63} \tag{40}$$

and the new bound gives

$$\mathrm{dep}_F(A) \le \sqrt{2}\,\min\left\{\sqrt{25/2}, \sqrt{21/2}\right\} = \sqrt{21}, \tag{41}$$

which is sharp since the eigenvalues of A are real. For completeness, the Eberlein (13) and Loizou (14) lower bounds are approximately $\sqrt{4.74}$ and $\sqrt{5.88}$, respectively.

Numerous examples can be contrived for which the new bound is tighter than the Henrici bound or vice versa. In general, the new bound is preferable since the Henrici bound is an $O(n^3)$ computation and the new bound is an $O(n^2)$ computation. It is sometimes possible to further improve the new bound by rotating $A - \alpha I$. For complex matrices, the eigenvalues of $A - \alpha I$ can be arbitrarily distributed and the best rotation θ cannot be determined a priori. For real matrices, the eigenvalues of $A - \alpha I$ occur in complex-conjugate pairs and the new bound is minimized for $\theta = 0$. Note that the new bound reduces to zero for Hermitian and skew-Hermitian matrices. Unfortunately, for normal matrices whose eigenvalues are not horizontally or vertically aligned, the new bound does not reduce to zero.

TABLE 1
Departure from normality results for Trefethen [15] *nonnormal test matrices.*

Test matrix	Henrici bound (16)	New bound (32)	$\mathrm{dep}_F(A)$ (7)	Ratio (16)/(7)	Ratio (32)/(7)
Jordan block	8.594	5.568	5.568	1.54	1.
Limaçon	13.980	7.810	7.810	1.79	1.
Grcar	18.398	7.681	6.007	3.06	1.28
Wilkinson	8.659	5.568	5.568	1.56	1.
Frank	1.821e+3	2.772e+2	2.772e+2	6.57	1.
Kahan	38.092	4.982	4.982	7.65	1.
Demmel	1.236e+9	1.438e+8	1.438e+8	8.59	1.
Lenferink–Spijker	2.384e+2	1.067e+2	1.067e+2	2.23	1.
Companion	5.281e+5	6.145e+4	6.145e+4	8.59	1.
Gauss–Seidel	4.555	2.149	2.149	2.12	1.
Chebyshev spectral	4.392	0.572	0.570	7.71	1.00

4. Numerical results. Table 1 compares the upper bounds (16) and (32) for some of the 32×32 nonnormal test matrices studied by Trefethen [15]. In each case, the new upper bound is tighter than the Henrici upper bound. Moreover, the new bound is sharp for the matrices entitled: Jordan block, Limaçon, Wilkinson, Frank, Kahan, Demmel, Lenferink–Spijker, Companion, and Gauss–Seidel. Such good results are predicted by Theorem 3.3 since the aforementioned matrices have strictly real eigenvalues and the other matrices (Grcar, Chebyshev spectral) have eigenvalues that are almost vertically aligned.

Acknowledgments. I wish to thank Steve Ashby, Linda Petzold, and Paul Saylor for their many suggestions and encouragements. I also thank Ed D'Azevedo and Faisal Saied for many helpful discussions and for helping to strengthen the new bound. Finally, I am grateful to June Donato, Gene Golub, Paul Saylor, Nick Trefethen, and the referees for valuable comments that have improved the presentation of this paper. Many of the matrices in Table 1 were obtained or adapted from the Test Matrix Toolbox for MATLAB [10].

Note added in proof. After this paper was accepted for publication, a referee noticed another $O(n^2)$ upper bound for departure from normality in the Frobenius norm [6, p. 66]

$$\operatorname{dep}_F(A) \leq \left(\|A\|_F^2 - |\operatorname{tr}(A^2)|\right)^{1/2}. \tag{42}$$

The bound is sharp iff 0 and the eigenvalues of A are collinear in the complex plane. The bound is a good approximation when zero and the eigenvalues of A are nearly collinear. Unfortunately, the bound does not reduce to zero for all normal matrices. Finally, we note that examples can be contrived for which the new bound (32) is tighter than (42) or vice versa.

REFERENCES

[1] J. DESCLOUX, *Bounds for the spectral norm of functions of matrices*, Numer. Math., (1963), pp. 185–190.
[2] P. J. EBERLEIN, *On measures of non-normality for matrices*, Amer. Math. Mon., 72 (1965), pp. 995–996.
[3] L. ELSNER AND M. H. C. PAARDEKOOPER, *On measures of nonnormality of matrices*, Linear Algebra Appl., 92 (1987), pp. 107–124.
[4] R. GABRIEL, *Matrizen mit maximaler Diagonale bei unitärer Similarität*, J. Reine. Angew. Math., 307/308 (1979), pp. 31–52.
[5] ———, *The normal $\triangle H$-matrices with connection to some Jacobi-like methods*, Linear Algebra Appl., 91 (1987), pp. 181–194.
[6] M. GIL', *Estimate for the norm of matrix-valued functions*, Linear Multilinear Algebra, 35 (1993), pp. 65–73.
[7] R. GRONE, C. R. JOHNSON, E. M. SÁ, AND H. WOLKOWICZ, *Normal matrices*, Linear Algebra Appl., 87 (1987), pp. 213–225.
[8] P. HENRICI, *Bounds for iterates, inverses, spectral variation and fields of values of non-normal matrices*, Numer. Math., 4 (1962), pp. 24–40.
[9] N. J. HIGHAM, *Matrix nearness problems and applications*, in Applications of Matrix Theory, M. J. C. Gover and S. Barnett, eds., Oxford University Press, Oxford, 1989, pp. 1–27.
[10] ———, *The test matrix toolbox for MATLAB*, Numerical Analysis Report 237, University of Manchester, Manchester, England, Dec. 1993.
[11] L. LÁSZLÓ, *An attainable lower bound for the best normal approximation*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1035–1043.
[12] G. LOIZOU, *Nonnormality and Jordan condition numbers of matrices*, J. Assoc. Comput. Mach., 16 (1969), pp. 580–584.
[13] A. RUHE, *Closest normal matrix finally found!*, BIT, 27 (1987), pp. 585–598.
[14] I. SCHUR, *Über die charakteristischen Wurzeln einer linearen Substitution mit einer Anwendung auf die Theorie der Integralgleichungen*, Math. Ann., 66 (1909), pp. 488–510.
[15] L. N. TREFETHEN, *Pseudospectra of matrices*, in Numerical Analysis 1991, D. F. Griffiths and G. A. Watson, eds., Longman Sci. Tech., Harlow, 1992, pp. 234–266.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 469–487, April 1995

011

FORWARD STABILITY AND TRANSMISSION OF SHIFTS IN THE QR ALGORITHM*

DAVID S. WATKINS†

Abstract. The QR algorithm is one of the most popular methods for calculating the eigenvalues of a matrix. In the course of iterations of the implicitly shifted QR algorithm on an upper Hessenberg matrix, it is crucial to check for zeros on the subdiagonal of the matrix. A zero on the subdiagonal allows the problem to be split into two independent subproblems. Moreover, if the splitting is not carried out, the subdiagonal zero will cause the subsequent QR iterations to break down. In practice exact zeros are rare; instead one normally sees very tiny numbers like 10^{-19}. It is reasonable to set such numbers to zero and split the problem. Indeed it is widely believed that it is crucial to carry out a splitting in such cases. Although such small entries, if left in place, will not cause the QR iterations to break down outright, they will (or so it is thought) trigger a breakdown of forward stability; small roundoff errors will be magnified dangerously, and the QR step will degenerate to a random similarity transformation. The first objective of this paper is to show that this widespread belief is mistaken; tiny subdiagonal entries do not normally cause forward instability or interfere in any way with the convergence of the algorithm. The second objective is to show that even in situations where forward instability does occur, the QR step is not normally rendered ineffective. On the contrary, the shift is transmitted accurately through the region of instability in such a way that a QR step with the chosen shift is performed on the trailing submatrix.

Key words. eigenvalues, QR algorithm, rounding, deflation, forward stability, shifts

AMS subject classifications. 65F15, 15A18

1. Introduction.

The most popular algorithm for finding all eigenvalues of a real or complex square matrix A is the QR algorithm. This is an iterative process that produces a sequence of unitarily similar matrices (A_k) that (nearly always) converges to quasitriangular form, thereby revealing the eigenvalues.

The QR algorithm has other applications as well, for example, solving the inverse eigenvalue problem [3]. However, we do not discuss any of these other applications; we study the QR algorithm strictly as a method for calculating eigenvalues.

In this paper we are concerned with what happens during a single step of the QR algorithm, so let us drop the subscript k and consider a single QR step from A to $\hat{A}$. The step can be described as follows. First a shift μ is chosen, then the shifted matrix $A - \mu I$ is factored into a product

$$A - \mu I = QR, \tag{1}$$

where Q is unitary and R is upper triangular. Then $\hat{A}$ is formed by performing a similarity transformation on A by Q:

$$\hat{A} = Q^{-1}AQ. \tag{2}$$

Let us assume that A is in *upper Hessenberg* form, which means that $a_{ij} = 0$ when $i > j + 1$. Thus A is almost upper triangular; below the main diagonal only the *subdiagonal* elements $a_{i,i-1}$ can be nonzero. If any of these subdiagonal entries is

* Received by the editors March 23, 1993; accepted for publication (in revised form) by P. Van Dooren January 17, 1994.

† Department of Pure and Applied Mathematics, Washington State University, Pullman, Washington 99164-3113 (na.watkins@na-net.ornl.gov). Current mailing address: 6835 24th Ave. NE, Seattle, Washington 98115-7037.

zero, then A is block upper triangular:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \tag{3}$$

and the eigenvalues of A_{11} and A_{22} can be calculated separately. Thus there is no loss of generality in assuming that $a_{i,i-1} \neq 0$ for $i = 2, \ldots, n$. We call a matrix with this property a *proper* upper Hessenberg matrix. If A is a proper upper Hessenberg matrix, then $\hat{A}$ is also upper Hessenberg. For details see [9] or [2], for example.

There are explicit and implicit implementations of the QR algorithm. An *explicit* implementation will carry out the decomposition (1) explicitly. Then $\hat{A}$ is formed by the equation

$$\hat{A} = RQ + \mu I, \tag{4}$$

which is equivalent to (2) in the absence of round-off errors. An *implicit* implementation manages to carry out the similarity transformation (2) without actually performing the QR decomposition (1). A small "bulge" in the Hessenberg form is created at the top of the matrix and "chased" to the bottom, where it falls off the end of the matrix. The details are spelled out in §2. Usually the Implicit-Q Theorem [9], [2] is invoked to verify that the implicit and explicit versions are, in the absence of round-off errors, equivalent. A more general and more revealing (but also lengthier) way to demonstrate equivalence was outlined by Miminis and Paige [4]. See also the generalization in [11]. These developments are also implied by the earlier paper [8].

The equivalence of the implicit and explicit implementations relies on the assumption that A is properly upper Hessenberg. If one of the entries $a_{i,i-1}$ is zero, the equivalence breaks down. To see what happens, think about performing a QR step on an upper Hessenberg matrix of the form (3). One easily checks that the explicit formulation performs QR steps independently on the submatrices A_{11} and A_{22}, yielding new submatrices $\hat{A}_{11}$ and $\hat{A}_{22}$. In contrast, (as we see later) the implicit formulation chases a bulge through A_{11}, thereby transforming it to $\hat{A}_{11}$; but the bulge disappears when it gets to the zero on the subdiagonal, and A_{22} is left unchanged.

Most implementations of the QR algorithm are implicit. For such implementations it is evidently crucial to check before each step that there are no zeros on the subdiagonal of A. If there are zeros, the problem must be broken into subproblems that are handled independently.

In numerical practice exact zeros seldom arise. One sees instead very tiny numbers like 10^{-19}. Suppose a number of this magnitude appears on the subdiagonal. (We assume here that most of the entries of A, including the largest ones, are of order 1.) What effect will this small entry have on an implicit QR step? The conventional wisdom is that the tiny number will act much like a zero, causing the QR step to be "washed out." Although the bulge will not vanish when it gets to the small number, serious rounding errors will (somehow) take place, causing all subsequent transformations to be inaccurate. Therefore the trailing submatrix A_{22} will be transformed in a random way, or so it is widely believed. Consequently it is generally considered essential to check for tiny entries on the subdiagonal, set any such entries to zero, and split the problem.

The conventional wisdom can be traced back at least to Wilkinson, who wrote [12, p. 536], "The total savings resulting from taking advantage of small subdiagonal elements in the two ways we have described is often very substantial. However, it is worth pointing out that we should not think of these devices merely in terms of the

saving of computation in those iterations in which they are used. The reduction to Hessenberg form is unique only if subdiagonal elements are nonzero and therefore it is important to use only the lower submatrix when a subdiagonal element is negligible." This notion was subsequently amplified by Parlett [5, p. 162], and most of the experts believe it (including, until recently, the author of this paper). Indeed, it is so obviously right that most of us have accepted it without question.

Although we see no harm in taking advantage of opportunities to split the problem when they arise, we would like to point out at least that the conventional wisdom is wrong. Let us consider for example the symmetric matrix

$$\begin{bmatrix} 5 & 1 & & & \\ 1 & 5 & \epsilon & & \\ & \epsilon & 2 & 1 & \\ & & 1 & 2 & 1 \\ & & & 1 & 2 \end{bmatrix},$$

where we take ϵ to be the ridiculously small number 10^{-60}. Any reasonable implementation of the QR algorithm will set this number to zero and deal with the 2×2 and 3×3 submatrices separately. We attacked this matrix with an implicit QR code that was ordinary in every way, except that it did not check for the possibility of splitting the problem. The code uses the Wilkinson shift strategy, which is designed to cause an eigenvalue to emerge rapidly at the lower right-hand corner of the matrix. The Wilkinson strategy is guaranteed to converge (in the absence of round-off errors). The entry in the $(n, n-1)$ position tends to zero at least quadratically and usually cubically or better [5].

According to the conventional wisdom, the entry 10^{-60} should cause round-off errors that "wash out" the step and prevent convergence. Table 1 shows what really happens.

TABLE 1

Iteration	Subdiagonal entries of A			
	$a_{2,1}$	$a_{3,2}$	$a_{4,3}$	$a_{5,4}$
0	1.0×10^{-0}	1.0×10^{-60}	1.0×10^{-0}	1.0×10^{-0}
1	6.0×10^{-1}	1.1×10^{-60}	7.1×10^{-1}	7.1×10^{-1}
2	1.6×10^{-1}	3.9×10^{-60}	3.8×10^{-1}	3.0×10^{-2}
3	3.6×10^{-2}	1.8×10^{-59}	2.0×10^{-1}	4.5×10^{-7}
4	8.2×10^{-3}	8.7×10^{-59}	1.0×10^{-1}	2.1×10^{-22}

We see that the small entry $a_{3,2}$ does not prevent $a_{5,4}$ from converging to zero cubically. After four iterations the entry $a_{5,5}$ is an eigenvalue to full precision (almost 16 decimal places in double-precision IEEE floating-point arithmetic), and it is time to deflate. Later on we look at other examples that show that the same thing happens in the nonsymmetric case even if there are several consecutive tiny subdiagonal entries.

This phenomenon is related to the forward stability (or lack thereof) of the QR algorithm. The standard implementations of the QR algorithm are (normwise) backward stable but not forward stable. These are statements about the behavior of the algorithm in the presence of the errors that are inevitably associated with floating-point arithmetic. Let $\hat{A}$ denote the matrix that is actually produced by the inexact floating-point computation, and let $\tilde{A}$ denote the theoretical result that would be obtained if all of the computations were performed exactly. Backward stability means that $\hat{A}$ is exactly unitarily similar to a matrix $A + E$ that is a small perturbation of

A in the sense that $\|E\|/\|A\|$ is small.[1] This means that if we take the eigenvalues of $\hat{A}$ as approximations to the eigenvalues of A, we will in fact get the eigenvalues of a matrix that is close to A.

Lack of forward stability means simply that $\hat{A}$ need not be anywhere near $\tilde{A}$. This is an inevitable consequence of the fact that the QR step, which is a continuous map $A \rightarrow \tilde{A}$, can be extremely sensitive (ill conditioned). That is, small perturbations in the entries of A or in the shift μ can make a big change in $\tilde{A}$. Whenever this happens, we can expect the computed matrix $\hat{A}$ to differ substantially from $\tilde{A}$.

We wish to emphasize that forward instability does not happen on every QR step; it is something that can happen but usually does not. Obviously forward instability is a necessary condition for a washout of the QR step, for if the computed matrix $\hat{A}$ is very close to the theoretical $\tilde{A}$, the step must have been a success, at least from a numerical standpoint. The conventional wisdom, stated in the language of forward instability, would imply that small entries on the subdiagonal are reliable indicators of the onset of forward instability. In this paper we will see that they are not.

Although forward instability is a necessary condition for a washout of a QR step, it turns out (amazingly) not to be sufficient. We see in §7 that even when a loss of forward stability occurs, the QR step will not normally be rendered ineffective. On the contrary, it will transmit the shift through the region of instability in such a way that the transformation performed on the lower submatrix is exactly a QR step with the chosen shift.

Our approach is informal. Without making any precise claims, we try to show how round-off errors typically affect a QR step. Our analyses address only the case of a single step, but our numerical experience indicates that double steps behave similarly.

The only study of forward instability of which we are aware is the work of Parlett and Le [6], who analyze the sensitivity of a QR step for a symmetric tridiagonal matrix as a function of the shift. They conclude that ill conditioning (forward instability) with respect to perturbations in the shift can occur only in association with a phenomenon they call "premature deflation," and they give a simple criterion for detecting the onset of forward instability. Our work has some points of contact with that of Parlett and Le, and we point them out as they arise. However, our context and approach are different from theirs. We consider the nonsymmetric case and attempt to study the propagation of rounding errors.

Our findings may have implications for parallel QR codes that perform many steps at once in pipeline fashion. The need to check diligently for possible splittings can constitute a bottleneck that seriously degrades the efficiency of such codes. This is especially true of the Francis test for two consecutive small subdiagonal entries [1], [12, pp. 526–528, 535–537], which is used in the standard double QR codes. Our findings show that the need to perform splittings is not nearly so urgent as had been believed previously. It should be possible to write successful parallel QR codes that check for splittings relatively infrequently.

2. The explicit and implicit QR algorithms. In order to establish our notation, let us recall the details of a step of the QR algorithm. First consider an explicit QR step. The QR decomposition (1) can be performed by reducing $A - \mu I$ to upper triangular form by a sequence of $n-1$ plane rotators:

$$R = Q_{n-1}^* \cdots Q_2^* Q_1^* (A - \mu I).$$

[1] The fact that all careful implementations of the QR algorithm are normwise backward stable follows from the error analysis of unitary transformations carried out in Chapter 3 of [12].

For each i, Q_i^* is a rotator[2] operating in the $(i+1, i)$ plane that annihilates $a_{i+1,i}$. The form of Q_i is

$$\begin{bmatrix} I_{i-1} & & & \\ & c_i & -\bar{s}_i & \\ & s_i & \bar{c}_i & \\ & & & I_{n-i-1} \end{bmatrix}, \tag{5}$$

where $|c_i|^2 + |s_i|^2 = 1$, and I_j is the identity matrix of order j.

We have $A - \mu I = QR$, where $Q = Q_1 \cdots Q_{n-1}$. Thus (4) takes the form

$$\hat{A} = RQ_1Q_2 \cdots Q_{n-1} + \mu I.$$

One easily checks that $\hat{A}$ is upper Hessenberg. The assumption that A is a proper upper Hessenberg matrix implies that the QR decomposition of A and the resulting matrix $\hat{A}$ are almost uniquely determined [9]. For example, Q is determined up to right multiplication by a unitary diagonal matrix, the class of which is quite trivial. The rotators Q_i are similarly almost uniquely specified. We routinely ignore this trivial nonuniqueness and speak as if all of these entities are uniquely determined.

Now we consider an implicit QR step. Suppose A is a proper upper Hessenberg matrix. By (2) we can equally well calculate $\hat{A}$ from

$$\hat{A} = Q_{n-1}^* \cdots Q_1^* A Q_1 \cdots Q_{n-1},$$

provided we have the Q_i at hand. First of all, one easily checks that Q_1 is the rotator in the (2,1) plane such that Q_1^* transforms the vector $\begin{bmatrix} a_{11}-\mu \\ a_{21} \end{bmatrix}$ to the form $\begin{bmatrix} * \\ 0 \end{bmatrix}$. This gets the transformation started. Since $a_{21} \neq 0$, Q_1 is a nontrivial rotator. Let $A_1 = Q_1^* A Q_1$, and generally

$$A_i = Q_i^* A_{i-1} Q_i,$$

so that $\hat{A} = A_{n-1}$. We will let $a_{j,k}^{(i)}$ denote the (j,k) entry of A_i. The matrix A_1 is not quite upper Hessenberg; the entry $a_{3,1}^{(1)}$ is nonzero. This is the *bulge*. We know that it is nonzero because Q_2 is a nontrivial rotator and $a_{3,2} \neq 0$. Now consider the transformation $A_2 = Q_2^* A_1 Q_2$. The left multiplication by Q_2^* operates on rows 2 and 3, and the right multiplication by Q_2 operates on columns 2 and 3. Following Stewart [8], we note that the premultiplication by Q_2^* must annihilate the bulge. The reason for this is that by the time the step is finished, we must have returned the matrix to upper Hessenberg form. Of all the transformations that remain to be done in the step, the only one that can transform the $(3,1)$ entry to zero without creating nonzero entries further down in the first column is Q_2^*. Thus Q_2 must be the rotator that transforms

$$\begin{bmatrix} a_{2,1}^{(1)} \\ a_{3,1}^{(1)} \end{bmatrix}$$

to the form $\begin{bmatrix} * \\ 0 \end{bmatrix}$. This determines Q_2 (essentially) uniquely; using the notation of (5), we have

$$c_2 = \frac{a_{2,1}^{(1)}}{m_2} \quad \text{and} \quad s_2 = \frac{a_{3,1}^{(1)}}{m_2},$$

[2] We could equally well take the Q_i to be reflectors. This would make no significant difference.

where $m_2 = \sqrt{|a_{2,1}^{(1)}|^2 + |a_{3,1}^{(1)}|^2}$. Q_2 is a nontrivial rotator because $a_{3,1}^{(1)} \neq 0$. Once the bulge has been annihilated by Q_2^*, the right multiplication by Q_2 creates a new bulge at $a_{4,2}^{(2)}$. This entry is certain to be nonzero because Q_2 is a nontrivial rotator and $a_{4,3}^{(1)} = a_{4,3} \neq 0$. Reasoning as before, we find that the left multiplication by Q_3^* must annihilate this bulge. Then the right multiplication by Q_3 creates a new bulge at $a_{5,3}^{(3)}$. In general, A_i has a bulge at position $(i+2, i)$, $i = 1, \ldots, n-2$. Each of these bulge entries is nonzero, and each rotator is nontrivial. By the time we get to $i = n-1$, the bulge entry has been pushed off of the end of the matrix, and the QR step is complete. Generalizations of this process are discussed in [4] and [11].

2.1. The effect of an exact zero on the subdiagonal. So far we have been assuming that A is a proper upper Hessenberg matrix. Now let us see how things change when one of the subdiagonal entries, say $a_{i,i-1}$, is zero. In the step $A_{i-2} = Q_{i-2}^* A_{i-3} Q_{i-2}$, the right multiplication by Q_{i-2} is supposed to create a bulge in position $(i, i-2)$. However, since $a_{i,i-1}^{(i-3)} = a_{i,i-1} = 0$, we get $a_{i,i-2}^{(i-2)} = 0$ instead. At this point we can say that the step is complete, for the matrix A_{i-2} is in upper Hessenberg form. But let us see what happens if we continue the step. Since the bulge $a_{i,i-2}^{(i-2)}$ is zero, Q_{i-1} is a trivial (zero degree) rotation, so $A_{i-1} = A_{i-2}$. This transformation would normally create a new bulge at $a_{i+1,i-1}^{(i-1)}$, but in this case the new bulge is also zero. The next rotator Q_i is normally chosen so that it transforms

$$\begin{bmatrix} a_{i,i-1}^{(i-1)} \\ a_{i+1,i-1}^{(i-1)} \end{bmatrix}$$

to the form $\begin{bmatrix} * \\ 0 \end{bmatrix}$. In this case both $a_{i,i-1}^{(i-1)}$ and $a_{i+1,i-1}^{(i-1)}$ are zero, so Q_i can be chosen arbitrarily. The simplest thing to do is take Q_i to be a zero-degree rotation. This ends the step with the lower-right corner of the matrix (called A_{22} in (3)) untouched. If, on the other hand, Q_i is taken to be some nontrivial rotator, the rest of the QR step effects a somewhat arbitrary similarity transformation on A_{22}. The arbitrariness of Q_i is worrisome. If this is what happens when $a_{i,i-1}$ is zero, might not something similar happen if $a_{i,i-1}$ is close to zero?

3. Passing through a tiny subdiagonal entry. Let us suppose that A that has a tiny subdiagonal entry $a_{i,i-1} = \epsilon$. We have just seen that if $a_{i,i-1}$ is exactly zero, then both $a_{i,i-1}^{(i-1)}$ and $a_{i+1,i-1}^{(i-1)}$ are exactly zero. Now we will not have exact zeros, but we will have $a_{i,i-1}^{(i-1)} = O(\epsilon)$ and $a_{i+1,i-1}^{(i-1)} = O(\epsilon)$. Since these entries are small, it is natural to expect that they be poorly determined and that Q_i will be poorly determined as a consequence. We wish to demonstrate that this is not the case.

We do not propose to prove a theorem here; we just want to show what typically happens. Thus we do not hesitate to make simplifying assumptions along the way. The transformations $A_{j-1} \rightarrow A_j$ for $j = 1, \ldots, i-2$ are independent of the value of $a_{i,i-1}$, so let us assume that all of these transformations have been accurate. Thus the computed A_{i-2} has a tiny (normwise) error. By this we mean that the errors in the elements of A_{i-2} are of the order $u\|A\|$, where u is the unit round-off of the computer.

We assume that our computer satisfies the following simple model: Each floating-point operation is performed with a tiny relative error. That is, if we compute $x * y$ on the computer, where $*$ is any one of the four binary arithmetic operations, the computed result will be $x * y(1+\delta)$ for some δ satisfying $|\delta| \leq u$. Computers with

IEEE floating-point arithmetic satisfy this model. We also assume that the computed square root of x satisfies $\sqrt{x}(1+\delta)$, where $|\delta| = O(u)$.

The transformations $A_{i-2} \rightarrow A_{i-1}$ and $A_{i-1} \rightarrow A_i$ carry the bulge through the region of the tiny entry. We wish to show that the rotators Q_{i-1} and Q_i are normally accurate, for this will imply that the transformations $A_{i-2} \rightarrow A_{i-1}$ and $A_{i-1} \rightarrow A_i$ are accurate. The bulge in A_{i-2} is located at position $(i, i-2)$, so Q_{i-1} is determined by the values of $a_{i-1,i-2}^{(i-2)}$ and $a_{i,i-2}^{(i-2)}$. By assumption these both have tiny errors relative to $\|A\|$. Let us take a closer look at $a_{i,i-2}^{(i-2)}$, which was formed by the operation $a_{i,i-2}^{(i-2)} = s_{i-2}a_{i,i-1} = s_{i-2}\epsilon$. (The notation s_{i-2} is from (5).) Thus $a_{i,i-2}^{(i-2)} = O(\epsilon)$. Since s_{i-2} has a tiny absolute error, the error in $a_{i,i-2}^{(i-2)}$ must be tiny relative to ϵ. The rotator Q_{i-1} is determined by the proportions $s_{i-1}a_{i-1,i-2}^{(i-2)} = c_{i-1}a_{i,i-2}^{(i-2)}$. Indeed we can take

$$c_{i-1} = \frac{a_{i-1,i-2}^{(i-2)}}{m_{i-1}} \quad \text{and} \quad s_{i-1} = \frac{a_{i,i-2}^{(i-2)}}{m_{i-1}}, \tag{6}$$

where

$$m_{i-1} = \sqrt{|a_{i-1,i-2}^{(i-2)}|^2 + |a_{i,i-2}^{(i-2)}|^2}.$$

Making the simplifying assumption that $a_{i-1,i-2}^{(i-2)} = O(\|A\|)$, we conclude that $s_{i-1} = O(\epsilon)$, and its relative error must be tiny. Furthermore, the absolute error in c_{i-1} is small. Since $|c_{i-1}| \approx 1$, the absolute error is the same as the relative error. Thus Q_{i-1} is determined accurately.

Now let us consider Q_i, which is determined by the entries $a_{i,i-1}^{(i-1)}$ and $a_{i+1,i-1}^{(i-1)}$. First of all, $a_{i+1,i-1}^{(i-1)} = s_{i-1}a_{i+1,i}$. Since $a_{i+1,i}$ has zero error, and $s_{i-1} = O(\epsilon)$ and has a tiny relative error, $a_{i+1,i-1}^{(i-1)}$ must also be $O(\epsilon)$ and have tiny relative error.

Now consider the operations that are performed in transforming $a_{i,i-1}$ to $a_{i,i-1}^{(i-1)}$. The first transformation to affect $a_{i,i-1}$ is the multiplication on the right by Q_{i-2}. This gives

$$a_{i,i-1}^{(i-2)} = \bar{c}_{i-2}a_{i,i-1}.$$

Thus $a_{i,i-1}^{(i-2)} = O(\epsilon)$. Since $\bar{c}_{i-2}$ has a tiny absolute error, the error in $a_{i,i-1}^{(i-2)}$ is tiny relative to ϵ. The left multiplication by Q_{i-1}^* gives the intermediate results

$$a'_{i,i-1} = -s_{i-1}a_{i-i,i-1}^{(i-2)} + c_{i-1}a_{i,i-1}^{(i-2)} \tag{7}$$

and

$$a'_{i,i} = -s_{i-1}a_{i-i,i}^{(i-2)} + c_{i-1}a_{i,i}^{(i-2)}. \tag{8}$$

Finally, the right multiplication by Q_{i-1} gives

$$a_{i,i-1}^{(i-1)} = c_{i-1}a'_{i,i-1} + s_{i-1}a'_{i,i}. \tag{9}$$

Each of the terms on the right-hand side of (7) has a factor of order ϵ with tiny error relative to ϵ. Consequently, $a'_{i,i-1}$ is also $O(\epsilon)$ and has tiny error relative to ϵ. We

can now apply the same argument to (9) to conclude that $a_{i,i-1}^{(i-1)} = O(\epsilon)$ and has tiny error relative to ϵ. The point of this is that although $a_{i,i-1}^{(i-1)}$ is tiny, it did not become so through cancellation; it is tiny because both terms on the right-hand side of (9) are tiny to begin with. Cancellation could occur in (9), in which case $a_{i,i-1}^{(i-1)}$ would be ultra tiny. It would then likely have a large relative error, but its error relative to ϵ would still be tiny. Let us make the simplifying assumption that cancellation does not occur in (9), so that $a_{i,i-1}^{(i-1)}$ is truly of order ϵ (not smaller) and has tiny relative error.

Since $a_{i,i-1}^{(i-1)}$ and $a_{i+1,i-1}^{(i-1)}$ both have tiny relative errors, the rotator Q_i is perfectly well determined, and the step proceeds with no loss of accuracy. This is just what happened in the example in the introduction. The tiny entry $a_{3,2}$ caused the rotator Q_2 to have a rotation angle of order 10^{-60}, which caused both $a_{4,2}^{(2)}$ and $a_{3,2}^{(2)}$ to be of order 10^{-60}. In spite of their small size, these entries were accurate to some sixteen decimal places. Therefore Q_3 was accurate and the step continued with no ill effects.

3.1. Consecutive tiny subdiagonal entries. The presence of several consecutive tiny subdiagonal entries need not cause any problems. To see this, suppose $a_{i,i-1} = \epsilon$ and $a_{i+1,i} = \delta$. Then $a_{i,i-1}^{(i-1)}$ will normally be $O(\epsilon)$ and have tiny relative error as well. The bulge $a_{i+1,i-1}^{(i-1)}$ will have tiny relative error, as always, but now it will be $O(\epsilon\delta)$, since it is the product of the tiny numbers s_{i-1} and $a_{i+1,i}$. Consequently the rotator Q_i will also have a tiny rotation angle, i.e., $|s_i| = O(\delta)$. The argument we used to show that $a_{i,i-1}^{(i-1)}$ normally has small relative error can now be applied with i replaced by $i+1$ to show that $a_{i+1,i}^{(i)}$ normally has a tiny relative error. Thus Q_{i+1} is determined accurately. If $a_{i+2,i+1}$ is also tiny, the argument can be repeated, and so on.

Consider the 6×6 symmetric tridiagonal matrix whose main diagonal entries are all 2.0 and whose off-diagonal entries are as in the first row of Table 2.

TABLE 2

Iteration	Subdiagonal entries of A				
	$a_{2,1}$	$a_{3,2}$	$a_{4,3}$	$a_{5,4}$	$a_{6,5}$
0	1.0×10^{-80}	1.0×10^{-40}	1.0×10^{-60}	1.0×10^{-0}	1.0×10^{-0}
1	1.0×10^{-80}	1.0×10^{-40}	1.4×10^{-60}	7.1×10^{-1}	7.1×10^{-1}
2	1.0×10^{-80}	1.0×10^{-40}	2.6×10^{-60}	3.8×10^{-1}	3.0×10^{-2}
3	1.0×10^{-80}	1.0×10^{-40}	5.0×10^{-60}	2.0×10^{-1}	4.5×10^{-7}
4	1.0×10^{-80}	1.0×10^{-40}	9.9×10^{-60}	1.0×10^{-1}	5.3×10^{-22}

Table 2 shows that the three consecutive tiny subdiagonal entries do not prevent cubic convergence of $a_{6,5}$ to zero.

We have noted that when there are two consecutive tiny subdiagonal entries, the bulge has magnitude $O(\epsilon\delta)$ as it passes between them. This means that in this example the bulge got as small as 10^{-120}. This is well above the underflow point for IEEE double-precision arithmetic. If the offdiagonal entries are made small enough to cause the bulge to underflow to zero, the QR step will, of course, die out.

4. More examples. One might feel that the two examples given so far are artificial. What happens in "real" problems when some of the subdiagonal entries become small? In an attempt to shed light on this question, we calculated the eigenvalues of the symmetric, tridiagonal matrix of order 300 given by $a_{i,i} = 2$ and $a_{i,i-1} = a_{i-1,i} = 1$. We used a symmetric QR code that makes no attempt to split the

problem, except that it performs a deflation whenever the bottommost subdiagonal entry becomes sufficiently small. After 213 iterations the matrix had been deflated to 198×198. At this point it was noted that $a_{191,190} \approx 10^{-13}$. In the course of the next twelve iterations, this entry shrank to approximately 10^{-46}. During these iterations, the six subdiagonal entries that lay below $a_{191,190}$ converged swiftly to zero, one after the other. Thus the small entry had absolutely no adverse effect on convergence. At iteration 226, $a_{191,190}$ was itself deflated from the matrix.

We now consider a few nonsymmetric examples. We modified a standard double-shift QR code so that it does not check for possible splittings except in the two bottommost subdiagonal entries. We tried it on the upper Hessenberg matrix

$$\begin{bmatrix} 8 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 7 & 1 & 1 & 1 & 1 & 1 & 1 \\ & -1 & 6 & 1 & 1 & 1 & 1 & 1 \\ & & \epsilon_1 & 5 & 1 & 1 & 1 & 1 \\ & & & \epsilon_2 & 4 & 1 & 1 & 1 \\ & & & & -1 & 3 & 1 & 1 \\ & & & & & -1 & 2 & 1 \\ & & & & & & -1 & 1 \end{bmatrix},$$

where $\epsilon_1 = 10^{-60}$ and $\epsilon_2 = 10^{-70}$, with the results shown in Table 3. For clarity we do not list the mantissas.

TABLE 3

Iteration	Subdiagonal entries of A						
	$a_{2,1}$	$a_{3,2}$	$a_{4,3}$	$a_{5,4}$	$a_{6,5}$	$a_{7,6}$	$a_{8,7}$
0	10^{-0}	10^{-0}	10^{-60}	10^{-70}	10^{-0}	10^{-0}	10^{-0}
1	10^{-1}	10^{-1}	10^{-61}	10^{-71}	10^{-1}	10^{-1}	10^{-0}
2	10^{-1}	10^{-1}	10^{-61}	10^{-71}	10^{-1}	10^{-2}	10^{-0}
3	10^{-1}	10^{-1}	10^{-61}	10^{-72}	10^{-1}	10^{-4}	10^{-1}
4	10^{-1}	10^{-1}	10^{-62}	10^{-72}	10^{-1}	10^{-8}	10^{-0}
5	10^{-1}	10^{-1}	10^{-62}	10^{-73}	10^{-1}	10^{-16}	10^{-0}

We see that the $(7,6)$ entry converges quadratically to zero in spite of the tiny entries in positions $(4,3)$ and $(5,4)$. This is the rate of convergence we normally expect in the nonsymmetric case [10]. After five iterations a 2×2 block containing a complex conjugate pair of eigenvalues can be deflated from the matrix. After four more iterations, all eigenvalues have been found. This matrix has two real eigenvalues and three complex conjugate pairs.

We also tried our double QR code on some random matrices. We generated full matrices with normally distributed entries, and then reduced them to upper Hessenberg form using the EISPACK [7] code ELMHES. Matrices of this type have evenly distributed eigenvalues. Consequently, as the QR iterations proceed, those subdiagonal entries that tend to zero do so very slowly, except for the two at the bottom. The matrices tend not to split apart. We did, however, observe one 200×200 matrix that developed a very small entry in the $(7,6)$ position. After some 323 iterations, the matrix had been deflated to 44×44. At that point the $(7,6)$ entry was smaller than 10^{-12}. It remained small from then on and became as small as 1.5×10^{-15} at one point. This is only about ten times the unit round-off error. The small entry did not in any way interfere with the convergence of the algorithm. After 391 (double) steps, all eigenvalues had been found.

In order to obtain a matrix that has some subdiagonal entries (other than those near the bottom) that become extremely small, we built a random 60×60 matrix as follows: All entries were normally distributed random numbers with mean 0. However, those in the upper left-hand 30×30 submatrix were from a distribution with standard deviation 1, and the rest were from a distribution with standard deviation 10^{-2}. The resulting matrix has 30 "large" eigenvalues and 30 "small" eigenvalues. We reduced the matrix to upper Hessenberg form by ELMHES and applied the double QR algorithm, not testing for possible deflations except in the two bottommost subdiagonal entries. The entry in the $(31,30)$ position became small rapidly. After nine iterations it had reached 10^{-21}, which is well below the unit round-off. After 43 iterations there were three tiny subdiagonal entries: $a_{27,26} \approx 10^{-25}$, $a_{30,29} \approx 10^{-18}$, and $a_{31,30} \approx 10^{-95}$. At this point the matrix had been deflated to 42×42. On subsequent iterations these entries became even smaller, and more eigenvalues were deflated from the matrix. Quadratic convergence was observed. After 62 iterations the entry in the $(31,30)$ position had reached 10^{-140}. On the next iteration it was deflated from the matrix.

5. Breakdown of forward stability. We have seen that a tiny $a_{i,i-1}$ entry will not normally cause the rotator Q_i to be inaccurate. Now let us see what conditions can cause inaccuracy. The accuracy of Q_i is entirely determined by the accuracy of the entries $a_{i,i-1}^{(i-1)}$ and $a_{i+1,i-1}^{(i-1)}$. We have

$$c_i = \frac{a_{i,i-1}^{(i-1)}}{m_i} \quad \text{and} \quad s_i = \frac{a_{i+1,i-1}^{(i-1)}}{m_i},$$

where

$$m_i = \sqrt{|a_{i,i-1}^{(i-1)}|^2 + |a_{i+1,i-1}^{(i-1)}|^2}.$$

Let us assume that $Q_1, \ldots, Q_{i-1}$ have been accurate, so that the errors in $a_{i,i-1}^{(i-1)}$ and $a_{i+1,i-1}^{(i-1)}$ are tiny relative to $\|A\|$. Generically both $a_{i,i-1}^{(i-1)}$ and $a_{i+1,i-1}^{(i-1)}$ will be $O(\|A\|)$; in this case they will have tiny relative errors, so Q_i will be accurate. Thus we need to look at the cases where one of the other of $a_{i,i-1}^{(i-1)}$ and $a_{i+1,i-1}^{(i-1)}$ is small.

Let us consider $a_{i+1,i-1}^{(i-1)}$ first. We have already noted that $a_{i+1,i-1}^{(i-1)} = s_{i-1}a_{i+1,i}$. There is no error in $a_{i+1,i}$, so the relative accuracy of $a_{i+1,i-1}^{(i-1)}$ depends entirely on the relative error in s_{i-1}. A reasonable simplifying assumption is that the latter relative error is tiny, as we now justify by induction on i. First of all, it is a simple matter to show that the relative error in s_1 is tiny. Now assume for the induction step that the relative error in s_{i-2} is tiny. By (6)

$$s_{i-1} = \frac{a_{i,i-2}^{(i-2)}}{m_{i-1}} = \frac{s_{i-2}a_{i,i-1}}{\sqrt{|a_{i-1,i-2}^{(i-2)}|^2 + |a_{i,i-2}^{(i-2)}|^2}}.$$

There is no error in $a_{i,i-1}$. Making the simplifying assumption that $a_{i-1,i-2}^{(i-2)}$ and $a_{i,i-2}^{(i-2)}$ are not both tiny, we can conclude that m_{i-1} has a tiny relative error. Thus s_{i-1} has a tiny relative error. We conclude that $a_{i+1,i-1}^{(i-1)}$ normally has a tiny relative error, regardless of whether it is big or small.

Thus the only way we can expect to get an inaccurate Q_i is by having an inaccurate $a_{i,i-1}^{(i-1)}$. It cannot be inaccurate unless it is small, so let us suppose that $a_{i,i-1}^{(i-1)}$ is tiny. We know from §3 that tininess alone does not imply inaccuracy. Whether $a_{i,i-1}^{(i-1)}$ is accurate or not depends on how it became tiny. In the previous section we assumed that $a_{i,i-1}$ was small. This caused s_{i-1} and $a'_{i,i-1}$ to be small, which implied that the smallness of $a_{i,i-1}^{(i-1)}$ was due not to cancellation but to the smallness of the two terms on the right-hand side of (9). Now suppose $a_{i,i-1}$ is not small. Then s_{i-1} is not small, and the terms on the right-hand side of (9) are normally both large. Thus the only way $a_{i,i-1}^{(i-1)}$ can become tiny is through cancellation in (9). When this happens, $a_{i,i-1}^{(i-1)}$ will normally have a large relative error.

Notice that once the transformations by Q_{i-1} are complete, the upper left-hand $i \times i$ submatrix has undergone a complete QR step with shift μ, for the bulge has now passed completely through that submatrix. In the course of this QR step, under the assumptions of the previous paragraph, the large entry $a_{i,i-1}$ has been replaced by the tiny entry $a_{i,i-1}^{(i-1)}$. This sudden appearance of a near zero in the $(i, i-1)$ position signals the sudden emergence (in position (i,i)) of an eigenvalue of the submatrix. The standard convergence theory [2], [5], [9], [10] shows that such sudden convergence can take place only if the shift μ is (almost exactly) equal to the emerging eigenvalue of the submatrix. This event is part of the syndrome that Parlett and Le [6] call premature deflation.

It turns out, remarkably, that even events of this type will not normally wash out a QR step, as we demonstrate in the next two sections.

6. Transmission of the shift in an implicit QR step. In an implicit QR step the shift μ is used only to generate the first rotator Q_1. This is done through the vector $\begin{bmatrix} a_{11}-\mu \\ a_{21} \end{bmatrix}$. Subsequent rotators are generated by using the elements of the intermediate matrices A_i, which contain the shift μ only implicitly. In this section we will see how the shift is transmitted through the matrix during a QR step and how we can extract it from any of the intermediate matrices. We assume exact arithmetic.

The main result is the first one, which establishes a fundamental relationship between the elements in the shifted matrix $A_i - \mu I$ in the vicinity of the bulge. As before, we denote by $a_{j,k}^{(i)}$ the (j,k) entry of A_i.

THEOREM 6.1. *For $i = 1, \ldots, n-2$ the matrix*

$$S_i = \begin{bmatrix} a_{i+1,i}^{(i)} & a_{i+1,i+1}^{(i)} - \mu \\ a_{i+2,i}^{(i)} & a_{i+2,i+1}^{(i)} \end{bmatrix}$$

is singular.

Proof. The proof is by induction on i. We have $A_0 = A$. If we define $a_{1,0}^{(0)} = a_{1,1} - \mu$ and $a_{2,0}^{(0)} = a_{21}$, the theorem holds trivially for $i = 0$. We start the induction from that point. Now let $i \geq 1$, and we show that S_i is singular if S_{i-1} is. Since the former is a submatrix of $A_i - \mu I$ and the latter is a submatrix of $A_{i-1} - \mu I$, we need to look at the transformation $A_{i-1} - \mu I \rightarrow A_i - \mu I = Q_i^*(A_{i-1} - \mu I)Q_i$. All of the action takes place in the 3×3 submatrix consisting of rows i through $i+2$ and columns $i-1$ through $i+1$. In $A_{i-1} - \mu I$ this submatrix looks like

$$\begin{bmatrix} a_{i,i-1}^{(i-1)} & a_{i,i}^{(i-1)} - \mu & * \\ a_{i+1,i-1}^{(i-1)} & a_{i+1,i}^{(i-1)} & * \\ 0 & 0 & * \end{bmatrix},$$

where the asterisks denote matrix entries whose values are not of immediate interest. The leading 2×2 submatrix is S_{i-1}, which is singular by the induction hypothesis. The transformation $A_{i-1} - \mu I \rightarrow Q_i^*(A_{i-1} - \mu I)$ acts on rows i and $i+1$. It is designed to annihilate the bulge, the nonzero element $a_{i+1,i-1}^{(i-1)}$. (This is so even when $i - 1 = 0$.) Since S_{i-1} is singular, this transformation must also annihilate $a_{i+1,i}^{(i-1)}$. Thus the active 3×3 submatrix of $Q_i^*(A_{i-1} - \mu I)$ has the form

$$\begin{bmatrix} * & * & * \\ 0 & 0 & b \\ 0 & 0 & d \end{bmatrix},$$

where $d = a_{i+1,i}^{(i-1)}$. We now complete the similarity transformation by multiplying by Q_i on the right. This operation transforms columns i and $i+1$. The active 3×3 submatrix of $A_i - \mu I$ has the form

$$\begin{bmatrix} * & * & * \\ 0 & a_{i+1,i}^{(i)} & a_{i+1,i+1}^{(i)} - \mu \\ 0 & a_{i+2,i}^{(i)} & a_{i+2,i+1}^{(i)} \end{bmatrix} = \begin{bmatrix} * & * & * \\ 0 & s_i b & \bar{c}_i b \\ 0 & s_i d & \bar{c}_i b \end{bmatrix}.$$

The trailing 2×2 submatrix is S_i, and this is clearly singular. □

The symmetric case of Theorem 6.1 was proved by Stewart [8], who used it to develop a version of the (symmetric) implicit QR algorithm that automatically restores the shift in cases where it has been lost through swamping by large main-diagonal entries in the matrix. This can be useful when the matrix is graded, having large entries at the top and much smaller entries at the bottom.

COROLLARY 6.2. *For $i = 1, \ldots, n-2$, the shift can be extracted from A_i by the formula*

$$\mu = a_{i+1,i+1}^{(i)} - \frac{a_{i+2,i+1}^{(i)} a_{i+1,i}^{(i)}}{a_{i+2,i}^{(i)}}.$$

Corollary 6.2 allows us to recover the shift at any point in the QR step, at least in principle. As we shall see, the formula generally works well in practice. The one situation in which we can clearly expect it to fail is when the subtraction results in severe cancellation. This happens exactly when $|\mu|$ is much smaller than $|a_{i+1,i+1}^{(i)}|$, that is, exactly when μ is swamped by the main diagonal entry.

For our next theorem we partition A_i into blocks,

$$A_i = \begin{bmatrix} A_{11}^{(i)} & A_{12}^{(i)} \\ A_{21}^{(i)} & A_{22}^{(i)} \end{bmatrix},$$

where $A_{11}^{(i)}$ is $i \times i$, and we partition $\hat{A} = A_{n-1}$ the same way:

$$\hat{A} = \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \end{bmatrix}.$$

THEOREM 6.3. *The transformation $A_{22}^{(i)} \rightarrow \hat{A}_{22}$ is a QR step with shift μ.*

Proof. The submatrix $A_{21}^{(i)}$ has only two nonzero entries, namely, $a_{i+1,i}^{(i)}$ and $a_{i+2,i}^{(i)}$, which lie in the upper right-hand corner. These entries determine the rotator Q_{i+1},

which continues the QR step. Left multiplication by Q_{i+1}^* annihilates $a_{i+2,i}^{(i)}$; that is, it transforms the vector

$$\begin{bmatrix} a_{i+1,i}^{(i)} \\ a_{i+2,i}^{(i)} \end{bmatrix}$$

to the form $\begin{bmatrix} * \\ 0 \end{bmatrix}$. Then right multiplication by Q_{i+1} creates a bulge in the $(2,2)$ submatrix. The subsequent transformations push this bulge through the $(2,2)$ submatrix until the Hessenberg matrix $\hat{A}_{22}$ is obtained. Each transformation is uniquely determined by the preceding one.

We wish to show that the entire transformation $A_{22}^{(i)} \to \hat{A}_{22}$ amounts to a QR step on $A_{22}^{(i)}$ with shift μ. To this end, consider how such a step is begun. The first rotator, which we briefly call $\tilde{Q}_{i+1}$, is constructed so that left multiplication by $\tilde{Q}_{i+1}^*$ would transform

$$\begin{bmatrix} a_{i+1,i+1}^{(i)} - \mu \\ a_{i+2,i+1}^{(i)} \end{bmatrix}$$

to the form $\begin{bmatrix} * \\ 0 \end{bmatrix}$. The similarity transformation $A_{22}^{(i)} \to \tilde{Q}_{i+1}^* A_{22}^{(i)} \tilde{Q}_{i+1}$ creates a bulge, which is then chased to the bottom of the matrix. Each transformation is uniquely determined by the preceding one.

Since S_i is singular, the vectors

$$\begin{bmatrix} a_{i+1,i}^{(i)} \\ a_{i+2,i}^{(i)} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a_{i+1,i+1}^{(i)} - \mu \\ a_{i+2,i+1}^{(i)} \end{bmatrix}$$

are proportional. Thus Q_{i+1} and $\tilde{Q}_{i+1}$ are identical. Since the subsequent transformations must then also be identical, we conclude that the transformation $A_{22}^{(i)} \to \hat{A}_{22}$ is exactly a QR step with shift μ. □

7. The effects of round-off errors on shift transmission. The theorems of the preceding section all assume exact arithmetic. Now let us see how well they hold up in the presence of round-off errors. First consider Theorem 6.1. Suppose S_{i-1} is numerically singular, by which we mean that its smaller singular value is of order $u\|A\|$. We would like to show, if possible, that S_i must then also be numerically singular. The transformation $A_{i-1} \to Q_i^* A_{i-1} Q_i = A_i$ will introduce some rounding errors, but these can perturb the entries of S_i by at most $O(u\|A\|)$. Thus the singular values of S_i will also be perturbed by at most $O(u\|A\|)$. Consequently it suffices to analyze the transformation as if it were exact. This allows us to consider the equivalent transformation $(A_{i-1} - \mu I) \to Q_i^*(A_{i-1} - \mu I)Q_i = A_i - \mu I$ instead. First consider the left multiplication by Q_i^*. The S_{i-1} part of $A_{i-1} - \mu I$ is transformed to

$$\begin{bmatrix} \bar{c}_i & \bar{s}_i \\ -s_i & c_i \end{bmatrix} S_{i-1} = \begin{bmatrix} g_i & * \\ 0 & h_i \end{bmatrix}, \tag{10}$$

where h_i would be zero if S_{i-1} were exactly singular. In practice h_i is merely small, or so we hope. The matrix on the right-hand side of (10) has the same singular values as S_{i-1}, but this does not absolutely guarantee that h_i is small. All we can say in general is that $|g_i h_i| = \sigma_1 \sigma_2$, where $\sigma_1 \geq \sigma_2$ are the singular values of S_{i-1}. Thus if g_i is tiny, h_i might not be. The active 3×3 submatrix of $Q_i^*(A_{i-1} - \mu I)$ has the form

$$\begin{bmatrix} * & * & * \\ 0 & h_i & b \\ 0 & 0 & d \end{bmatrix}.$$

Right multiplication by Q_i will transform the submatrix

$$\begin{bmatrix} h_i & b \\ 0 & d \end{bmatrix} \tag{11}$$

to S_i, so S_i has the same singular values as the matrix (11). Thus the smaller singular value of S_i is bounded above by $|h_i|$. We conclude that S_i is numerically singular if $h_i = O(u\|A\|)$. As we shall shortly see, there is a wide variety of conditions under which it can be guaranteed that $h_i = O(u\|A\|)$.

7.1. Cases where g_i is not small. If it happens that either $a_{i,i-1}^{(i-1)}$ or $a_{i+1,i-1}^{(i-1)}$ is of order $\|A\|$, then we have

$$|g_i| = \sqrt{|a_{i,i-1}^{(i-1)}|^2 + |a_{i+1,i-1}^{(i-1)}|^2} = O(\|A\|)$$

also, which implies

$$|h_i| = \frac{\sigma_1\sigma_2}{|g_i|} = O(u\|A\|), \tag{12}$$

because σ_1 is bounded above by $\|A\|$. We have proved the following theorem, or quasitheorem if you prefer.

THEOREM 7.1. *If S_{i-1} is numerically singular and the norm of its first column is of order $\|A\|$, then S_i is also numerically singular.*

This shows that under ordinary conditions, especially as seen in early iterations of the QR algorithm, the shift is transmitted accurately. As an illustration, consider the 100×100 symmetric, tridiagonal matrix given by $a_{i,i} = 2$ and $a_{i+1,i} = 1$ for all i. In the course of a QR step on this matrix, the smaller singular value of S_i never was greater than 3.3×10^{-15}. At no point did the estimate of the shift given by Corollary 6.2 differ from the true shift by more than 3.6×10^{-15}. The deviations did not show any significant tendency to grow as the bulge moved downward through the matrix. Subsequent QR iterations showed the same pattern. Similar results were obtained with several random nonsymmetric examples.

Even under certain nonordinary conditions, the shift is transmitted accurately. Parlett and Le [6] showed that for symmetric, tridiagonal matrices, forward stability breaks down when and only when premature deflation occurs. Let us take a look at this syndrome. Suppose μ is almost exactly an eigenvalue of both A and one of its leading principal submatrices, say the ith. Then, by the time we have computed A_{i-1}, we have the following situation:

$$\begin{bmatrix} a_{i-1,i-1}^{(i-1)} & \epsilon & a_{i+1,i-1}^{(i-1)} \\ \epsilon & \mu' & \beta \\ a_{i+1,i-1}^{(i-1)} & \beta & a_{i+1,i+1}^{(i-1)} \end{bmatrix}.$$

In the ith leading principal submatrix a QR step has just been completed. Since μ is almost exactly an eigenvalue of the submatrix, μ' is very close to μ, and ϵ is tiny. If β is also tiny, we have the situation known as premature deflation. If we were to stop the step at this point, we could deflate the eigenvalue μ' from the matrix simply by deleting the ith row and column. If we do not stop the step, the eigenvalue μ' tends to be shoved down through the matrix. To see this, consider the ideal case $\epsilon = \beta = 0$. Then, because $\epsilon = 0$, the rotator Q_i will perform a 90° rotation on rows and columns

i and $i+1$. Except for some sign changes, this rotator has the effect of just swapping these rows and columns. Thus the entry μ' is just moved to the $(i+1, i+1)$ entry, and zeros appear in positions $(i+1, i)$ and $(i+2, i+1)$. The next rotator then pushes the eigenvalue and the zeros down one more position, and so on.

In practice premature deflation tends to set in gradually. If μ is almost exactly an eigenvalue of the ith principal submatrix, it will usually also be a good approximation to an eigenvalue of each of the few preceding principal submatrices. Thus the premature deflation event seen in A_{i-1} will have been preceded by lesser premature deflations in A_{i-2}, A_{i-3}, and so on. For example, consider the 15×15 symmetric, tridiagonal matrix that was featured in Example 2.4 of [6]. It is defined by $a_{1,1} = a_{15,15} = 15$, $a_{i,i} = 0$ for $i = 2, \ldots, 14$, and $a_{i+1,i} = 1$ for $i = 1, \ldots, 14$. This has two very close eigenvalues that are both 15.0666666666667 to 15 decimal places. On the first QR step with Wilkinson shift, there is no premature deflation. The second step is more interesting. The Wilkinson shift is $\mu = 15.0666666666666$, which is nearly an eigenvalue of both A and its tenth leading principal submatrix. Thus a premature deflation occurs in A_9. However, it does not occur suddenly; it is preceded by lesser events in A_7 and A_8. In A_7 we have

$$\begin{bmatrix} a_{7,7}^{(7)} & 1.97 \times 10^{-5} & 1.00 \\ 1.97 \times 10^{-5} & 15.0666666608 & -2.96 \times 10^{-4} \\ 1.00 & -2.96 \times 10^{-4} & a_{9,9}^{(7)} \end{bmatrix},$$

in A_8 we have

$$\begin{bmatrix} a_{8,8}^{(8)} & -1.31 \times 10^{-6} & 1.00 \\ -1.31 \times 10^{-6} & 15.066666666641 & 1.97 \times 10^{-5} \\ 1.00 & 1.97 \times 10^{-5} & a_{10,10}^{(8)} \end{bmatrix},$$

and in A_9 we have

$$\begin{bmatrix} a_{9,9}^{(9)} & 1.01 \times 10^{-8} & 1.00 \\ 1.01 \times 10^{-8} & 15.0666666666666 & -1.31 \times 10^{-6} \\ 1.00 & -1.31 \times 10^{-6} & a_{11,11}^{(9)} \end{bmatrix}.$$

As we move downward, progressively better approximations to the shift appear on the main diagonal, and the adjacent subdiagonals become smaller. The rotations that perform the transformations $A_7 \rightarrow A_8$ and $A_8 \rightarrow A_9$ are very nearly 90° rotations, that is, they are nearly swapping operations; the effect is clear in the example.

Parlett and Le [6] have shown that premature deflation is accompanied by a loss of forward stability. The point that we wish to make here is that loss of forward stability does not imply loss of shift. In our example, the conditions of Theorem 7.1 are satisfied. We have

$$S_7 = \begin{bmatrix} 1.97 \times 10^{-5} & -5.83 \times 10^{-9} \\ 1.00 & -2.96 \times 10^{-4} \end{bmatrix}, \qquad S_8 = \begin{bmatrix} -1.31 \times 10^{-6} & -2.58 \times 10^{-11} \\ 1.00 & 1.97 \times 10^{-5} \end{bmatrix},$$

and

$$S_9 = \begin{bmatrix} 1.01 \times 10^{-8} & 0 \\ 1.00 & -1.31 \times 10^{-6} \end{bmatrix}.$$

In each case the norm of the first column is of order $\|A\|$, thanks to the entry 1.00 in the bulge, so the numerical singularity is passed from one submatrix to the next. The singular values of $S_7, \ldots, S_9$ are as in Table 4.

TABLE 4

	σ_1	σ_2
S_7	1.0	5.9×10^{-15}
S_8	1.0	1.1×10^{-14}
S_9	1.0	1.3×10^{-14}

In each case the smaller singular value is on the order of $u\|A\|$. The formula given by Corollary 6.2 returns the correct shift to fifteen decimal places. Indeed, under the conditions of premature deflation, the shift resides mainly in the term $a_{i+1,i+1}^{(i)}$. The contribution of the second term in the formula is small.

The fact that the first column of S_i has norm of order $u\|A\|$ in each case is a direct consequence of the near 90° rotations. If Q_{i-1} is near 90°, then $|a_{i+1,i-1}^{(i-1)}| \approx |a_{i,i-1}|$, so the first column of S_{i-1} will be of order $\|A\|$ as long as $a_{i,i-1}$ is. That is what happens in this example and the other examples discussed in [6]. Theorem 7.1 is applicable in all of these cases.

Referring back to Theorem 6.3, the forward instability associated with premature deflation implies that once a premature deflation has occurred, the computed $A_{22}^{(i)}$ need not be anywhere near what it would have been in the absence of round-off errors.[3] Theorem 6.3 is not exactly applicable now because S_i is not exactly singular. However, S_i is nearly singular, and this implies that Theorem 6.3 is nearly true. That is, the rest of the QR step will be almost exactly a QR step on $A_{22}^{(i)}$ with shift μ. Now let us suppose we are using some natural shifting strategy that chooses the shift from the lower right-hand corner of A. Since the lower right-hand corner of $A_{22}^{(i)}$ is identical to the lower right-hand corner of A, a shift that is good for A should normally be good for $A_{22}^{(i)}$, regardless of whether or not the rotators $Q_1, \ldots, Q_i$ were accurate. For example, the Wilkinson shift strategy applied to $A_{22}^{(i)}$ gives the same shift as the Wilkinson shift strategy applied to A. Consequently, the QR step is bound to be successful in spite of the forward instability. This explains why forward instability is not normally damaging to QR codes when applied to solving the eigenvalue problem.

It is worth pointing out that some of the most spectacular examples of forward instability in [6] were provoked by making an "unnatural" choice of shift; unnatural meaning not based on information from the lower right-hand corner. See, especially, [6, Examples 2.2 and 2.3]. In these cases the QR step will not normally be successful, because such a shift may well approximate an eigenvalue of A but not of $A_{22}^{(i)}$.

7.2. Cases where g_i is small. Theorem 7.1 is applicable in cases where g_i in (10) is not small. However, as we have seen, there are important situations in which g_i is tiny. Indeed, in §3 our attention was focused on matrices in which tiny g_i occur. As we saw in §3, a small g_i is not normally disastrous if it arose from a subdiagonal entry that was small to begin with. If $|a_{i,i-1}| \approx \epsilon$, then $a_{i,i-1}^{(i-1)}$ and $a_{i+1,i-1}^{(i-1)}$ are also of order ϵ, but they are fully accurate nevertheless. This means that the numerical singularity of S_{i-1} is not just an artifact of the smallness of the entries of its first column. If the first column were rescaled by $1/\epsilon$, the resulting matrix would still be numerically singular. Another way to put this is that the smaller singular value of S_{i-1} is of order $\epsilon u\|A\|$ rather than $u\|A\|$. If one does the computation (12) under these conditions, one gets $h_i = O(u\|A\|)$ as before, because the ϵ in g_i is cancelled by the ϵ in σ_2. Thus we conclude that S_i is numerically singular in this case as well.

[3] Actually $A_{22}^{(i)}$ differs from A_{22} only in the first row and column, but an inaccurate Q_i can cause these to be far from what they should be.

The situation is only slightly more complicated when A has two or more consecutive tiny subdiagonals. Suppose $a_{i,i-1} = O(\epsilon)$ and $a_{i+1,i} = O(\delta)$. Then $a_{i,i-1}^{(i-1)} = O(\epsilon)$, $a'_{i+1,i} = O(\delta)$, and $a_{i+1,i-1}^{(i-1)} = O(\epsilon\delta)$. These numbers are normally fully accurate. If we were to rescale S_{i-1} by multiplying the first column by $1/\epsilon$ and the second row by $1/\delta$, the resulting matrix would be numerically singular. In other words, the smaller singular value of S_{i-1} is $O(\delta\epsilon u\|A\|)$. Under these conditions (12) gives $h_i = O(\delta u\|A\|)$, which implies that the smaller singular value of S_i is $O(\delta u\|A\|)$.

As an example consider the 8×8 tridiagonal matrix

$$\text{tridiag}\left\{\begin{array}{cccccccc} 1 & 1 & \epsilon & \delta & \gamma & 1 & 1 & \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 1 & 1 & \epsilon & \delta & \gamma & 1 & 1 & \end{array}\right\},$$

where $\epsilon = 10^{-20}$, $\delta = 10^{-15}$, and $\gamma = 10^{-10}$. If we perform a QR step on this matrix with $\mu = 3.0$ (Wilkinson shift), the effect of the small numbers is not felt in S_1. The smaller singular value of S_1 is 6.3×10^{-16}, which is of order $u\|A\|$. The effect of ϵ is first felt in S_2, whose smaller singular value is 1.3×10^{-35}. The matrix S_3 is approximately

$$\begin{bmatrix} -1.414 \times 10^{-20} & -1.000 \\ -1.414 \times 10^{-35} & 1.000 \times 10^{-15} \end{bmatrix},$$

and its smaller singular value is 1.4×10^{-50}, which is of order $\epsilon\delta u\|A\|$. In the next submatrix, S_4, the influence of ϵ has disappeared, but γ has become a factor. The smaller singular value of S_4 is 9.0×10^{-41}, which is of order $\delta\gamma u\|A\|$. The smaller singular value of S_5 is 6.7×10^{-26}, which is of order $\gamma u\|A\|$. Finally, the smaller singular value of S_6 is 1.3×10^{-15}, which is of order $u\|A\|$. Thus numerical singularity is preserved for the entire QR step. At each stage the value of the shift given by the formula of Corollary 6.2 deviates from the true shift by less than 2.8×10^{-15}. Similar tests on larger matrices gave similar results.

Now let us consider an example of a different kind.[4] The matrix

$$A = \text{tridiag}\left\{\begin{array}{cccccccc} 1 & 1 & \epsilon & 1 & 1 & \\ 2 & 2 & 2 & 2 & 2 & 2 \\ 1 & 1 & \epsilon & 1 & 1 & \end{array}\right\},$$

where $\epsilon = 10^{-15}$, has three pairs of nearly identical eigenvalues 3.41421356237310, 2.00000000000000, and 0.58578643762690. These are essentially the eigenvalues of the submatrix

$$\begin{bmatrix} 2 & 1 & \\ 1 & 2 & 1 \\ & 1 & 2 \end{bmatrix}$$

repeated twice. A QR step with the shift $\mu = 3.41421356237310$ results in a premature deflation at A_2, since both A and its third leading principal submatrix have eigenvalues very near this shift. The matrix S_2 consists entirely of tiny numbers:

$$S_2 = \begin{bmatrix} a_{3,2}^{(2)} & a_{3,3}^{(2)} - \mu \\ a_{4,2}^{(2)} & a_{4,3}^{(2)} \end{bmatrix} = \begin{bmatrix} -9.99 \times 10^{-16} & 0 \\ 8.66 \times 10^{-16} & 5.00 \times 10^{-16} \end{bmatrix}.$$

[4] Again we look at a symmetric example. A nonsymmetric example with similar characteristics was run with similar results.

The number $a_{3,2}^{(2)}$ became small through cancellation and is therefore poorly determined. The rotator for the next step, which is determined by $a_{3,2}^{(2)}$ and $a_{4,2}^{(2)}$, will be inaccurate. Thus there is a breakdown of forward stability, which will affect all of the subsequent computations. This instance of premature deflation differs from the example discussed previously in that $a_{4,2}^{(2)}$ is tiny. Thus Theorem 7.1 is not directly applicable here. Nevertheless, the shift has not been lost. Obviously the singular values of S_2 are both tiny; in fact, $\sigma_1 = 1.4 \times 10^{-15}$ and $\sigma_2 = 3.7 \times 10^{-16}$. This forces h_3 in (10) to be tiny, in spite of the fact that g_3 is also tiny. Consequently S_3 must also be numerically singular; the shift is transmitted successfully. Of course, in this case, as in the case of forward instability that we examined previously, the shift resides mainly in the main diagonal. In S_2 we have $a_{3,3}^{(2)} - \mu = 0$, which implies $a_{3,3}^{(2)} = \mu$. Corollary 6.2 reproduces the shift accurately; the first term in the formula is the shift, and the second term is noise of order $u\|A\|$.

Because the shift is transmitted accurately through the region of forward instability, the QR step ends successfully with $\hat{a}_{6,5} = 1.4 \times 10^{-15}$ and $\hat{a}_{6,6} = \mu$, nearly allowing the eigenvalue μ to be deflated from the problem.

Notice, however, that the breakdown of forward stability is not completely without cost. The shifted matrix $A - \mu I$ has two eigenvalues near zero, so the standard convergence theory [9], [10] predicts that $\hat{a}_{5,4}$ will be (near) zero, and the bottom 2×2 submatrix will consist of two copies of the eigenvalue μ. Because of the forward instability, only one copy of μ appears in practice. The other copy will emerge on the second QR step.

Before leaving this example, it is perhaps useful to look at it from one more viewpoint. Part way through the QR step, the matrix has essentially the form

$$A_2 = \begin{bmatrix} * & * & & & & \\ * & * & \delta_1 & \delta_2 & & \\ & \delta_1 & \mu & \delta_3 & & \\ & \delta_2 & \delta_3 & 2 & 1 & \\ & & & 1 & 2 & 1 \\ & & & & 1 & 2 \end{bmatrix},$$

where $\delta_1, \delta_2, \delta_3$ are only slightly larger than u. The eigenvalue μ has just appeared in the $(3,3)$ position through premature deflation. Let B denote the lower right-hand 4×4 submatrix. Because δ_1 and δ_2 are tiny, B is essentially isolated from the upper left-hand 2×2 submatrix. B has μ as an eigenvalue (or very nearly so) with multiplicity 2. The remainder of the QR step is a similarity transformation on B; it does not mix B with the upper part of the matrix. Thus the double eigenvalue μ is preserved in the 4×4 trailing submatrix. The next rotator (Q_3) transforms B to the form

$$B' = \begin{bmatrix} * & * & * & \\ * & * & * & \\ * & * & 2 & 1 \\ & & 1 & 2 \end{bmatrix}.$$

Since Q_3 is the poorly determined rotator, the computed entries of B' will be far from the theoretical correct values. The subsequent rotators reduce B' to upper Hessenberg form, which we can call $\hat{B}$. But $\hat{B}$ has the same eigenvalues as B, so it has a double eigenvalue μ. An upper Hessenberg matrix with an eigenvalue of (geometric) multiplicity 2 must split, so the splitting we observe in this QR step is, from this viewpoint, inevitable.

8. Conclusions. Contrary to the conventional wisdom, tiny subdiagonal entries do not normally trigger forward instability in the QR algorithm, nor do they interfere with convergence in any way. Even in situations where forward instability does occur, the severe round-off errors do not normally result in degradation of the QR step; the shift is transmitted accurately through the region of instability.

Although our study has been neither rigorous nor exhaustive, we have looked at a diverse variety of situations. We do not see any way the shift can fail to be transmitted accurately during a single step of the QR algorithm on an ungraded matrix. The error in the effective shift will always be small relative to the norm of the matrix. In a graded matrix, a small shift can be lost through swamping by large entries at the top of the matrix, as Stewart [8] observed.

Acknowledgment. I tested the conventional wisdom at a conference at the Catholic University of Leuven, Belgium, in August, 1992. I would like to acknowledge the help and cooperation of my colleagues in this matter. Of the numerous experts whom I polled about the expected effect of an epsilon on the subdiagonal, almost all asserted that it would wash out the QR step. Only Pete Stewart knew what would really happen.

REFERENCES

[1] J. G. F. Francis, *The QR transformation, parts* I *and* II, Computer J., 4 (1961), pp. 265–272, 332–345.

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

[3] W. B. Gragg and W. J. Harrod, *The numerically stable reconstruction of Jacobi matrices from spectral data,* Numer. Math., 44 (1984), pp. 317–336.

[4] G. S. Miminis and C. C. Paige, *Implicit shifting in the QR algorithm and related algorithms,* SIAM J. Matrix Anal. Appl., 12 (1991), pp. 385–400.

[5] B. N. Parlett, *The Symmetric Eigenvalue Problem,* Prentice-Hall, Englewood Cliffs, NJ, 1980.

[6] B. N. Parlett and J. Le, *Forward instability of tridiagonal QR,* SIAM J. Matrix Anal. Appl., 14 (1993), pp. 279–316.

[7] B. T. Smith, J. M. Boyle, J. J. Dongarra, B. S. Garbow, Y. Ikebe, V. C. Klema, C. B. Moler, *Matrix Eigensystem Routines—EISPACK Guide,* 2nd ed., Springer-Verlag, Berlin, New York, 1976.

[8] G. W. Stewart, *Incorporating origin shifts into the QR algorithm for symmetric tridiagonal matrices,* Comm. Assoc. Comp. Mach., 13 (1970), pp. 365–367.

[9] D. S. Watkins, *Fundamentals of Matrix Computations,* John Wiley and Sons, New York, 1991.

[10] D. S. Watkins and L. Elsner, *Convergence of algorithms of decomposition type for the eigenvalue problem,* Linear Algebra Appl., 143 (1991), pp. 19–47.

[11] ———, *Theory of decomposition and bulge-chasing algorithms for the generalized eigenvalue problem,* SIAM J. Matrix Anal. Appl., 15 (1994), pp. 943–967.

[12] J. H. Wilkinson, *The Algebraic Eigenvalue Problem,* Clarendon Press, Oxford University, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 488–501, April 1995

012

LINEAR OPERATORS ON MATRICES: PRESERVING SPECTRUM AND DISPLACEMENT STRUCTURE*

KENNETH R. DRIESSEL† AND WASIN SO‡

Abstract. In this paper we characterize those linear operators on general matrices that preserve singular values and displacement rank. We also characterize those linear operators on Hermitian matrices that preserve eigenvalues and displacement inertia.

Key words. linear operator, displacement structure, Toeplitz

AMS subject classification. 15A04

1. Introduction.

We introduce some notation to facilitate our discussion.

$$\begin{aligned} C^{m\times n} &:= \text{ the set of all } m\times n \text{ complex matrices},\\ Gl(m) &:= \text{ the set of all nonsingular } m\times m \text{ matrices},\\ \mathrm{Herm}(m) &:= \text{ the set of all } m\times m \text{ Hermitian matrices},\\ U(m) &:= \text{ the set of all } m\times m \text{ unitary matrices}. \end{aligned}$$

For $1 \le i \le m, 1 \le j \le n$, let E^{ij} denote the $m \times n$ matrix with zero everywhere except one at the (i,j) position. Then $\{E^{ij}\}$ is a basis for $C^{m\times n}$. We also adopt the following notation.

$$\begin{aligned} \mathrm{sing}(A) &:= \text{the singular values of a matrix } A \text{ (including multiplicity)},\\ \mathrm{eigen}(A) &:= \text{the eigenvalues of a Hermitian matrix } A \text{ (including multiplicity)},\\ \mathrm{rank}(A) &:= \text{the rank of a matrix } A,\\ \mathrm{inertia}(A) &:= \text{the inertia of a Hermitian matrix } A. \end{aligned}$$

For $A \in C^{m\times n}, \mathrm{rank}(A) = k$ if and only if A has exactly k nonzero singular values. For $A \in \mathrm{Herm}(m), \mathrm{inertia}(A) = (p,n,z)$ if and only if A has p positive, n negative, and z zero eigenvalues, $m = p + n + z$.

We are interested in the spectral properties of matrices that are Toeplitz or nearly Toeplitz. As a consequence, we are interested in linear operators that preserve these properties. We know of only one previous result in this direction. It is the following theorem due to Chu [1]. Let J_m denote the $m \times m$ *exchange matrix* defined by

$$J_m(i,j) := \delta(i, m+1-j),$$

where δ denotes the Kronecker delta. For example, when $m = 3$,

$$J_3 = \begin{pmatrix} 0 & 0 & 1\\ 0 & 1 & 0\\ 1 & 0 & 0 \end{pmatrix}.$$

* Received by the editors July 17, 1992; accepted for publication (in revised form) by G. Cybenko, February 16, 1994.

† Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, Minnesota 55455 and Department of Mathematics, Iowa State University, Ames, Iowa 50011.

‡ Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, Minnesota 55455 and Division of Mathematical and Information Sciences, Sam Houston State University, Huntsville, Texas 77341 (mth_wso@shsu.edu).

We write J in place of J_m when m is easily determined from the context.

THEOREM 1.1. *Let Q be an $m \times m$ orthogonal matrix. Then the following conditions are equivalent.*

(*) *If A is an $m \times m$ symmetric Toeplitz matrix then so is QAQ^T.*

(**) *The matrix Q is one of the following:*

$$\pm I, \quad \pm J, \quad \pm I', \quad \pm I'J,$$

where I denotes the $m \times m$ identity matrix and $I' := \mathrm{Diag}(-1, (-1)^2, \ldots, (-1)^{m-1})$.

Chu's techniques can be used to characterize nonzero linear operators on Hermitian matrices that preserve both eigenvalues and Toeplitz structure.

In this paper, we study the nonzero linear operators that preserve spectra and displacement structure. We begin by recalling the relevant definitions. Let Z_m denote the $m \times m$ (*lower*) *shift matrix* defined by

$$Z_m(i,j) := \delta(i, j+1).$$

For example, when $m = 3$,

$$Z_3 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

We write Z in place of Z_m when m is easily determined from the context. Let ∇ be the linear operator defined by

$$\nabla := C^{m \times n} \longrightarrow C^{m \times n} : X \longrightarrow X - Z_m X Z_n^T.$$

For $A \in C^{m \times n}$, the *displacement rank* of A is defined as

$$\text{dis-rank } (A) := \text{rank } (\nabla.A).$$

In the case $m = n$, ∇ preserves Hermitian matrices. For $A \in \mathrm{Herm}(m)$, the *displacement inertia* of A is defined as

$$\text{dis-inertia } (A) := \text{inertia } (\nabla.A).$$

Kailath appears to be one of the first to emphasize the importance of the displacement structure of matrices. We recall a few of the major results in this area in order to illustrate the significance of these concepts. Note that Toeplitz matrices have displacement rank at most 2. Hence matrices with low displacement rank are regarded as being "nearly Toeplitz." The following result shows that displacement rank is preserved (loosely speaking) under inversion. The result is from Kailath, Kung, and Morf [8], but a different proof is given here. Recall that two matrices $A, B \in C^{m \times n}$ are *equivalent* if there exist $M \in Gl(m), N \in Gl(n)$ such that $B = MAN$. Note that A and B are equivalent if and only if $\mathrm{rank}(A) = \mathrm{rank}(B)$.

THEOREM 1.2. *For $A \in Gl(m)$,* $\text{dis-rank}(A^{-1}) = \text{dis-rank}(JAJ)$.

Proof. Note that

$$\begin{bmatrix} I & 0 \\ -Z^T A^{-1} & I \end{bmatrix} \begin{bmatrix} A & Z \\ Z^T & A^{-1} \end{bmatrix} \begin{bmatrix} I & -A^{-1}Z \\ 0 & I \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & A^{-1} - Z^T A^{-1} Z \end{bmatrix}$$

and

$$\begin{bmatrix} I & -ZA \\ 0 & I \end{bmatrix} \begin{bmatrix} A & Z \\ Z^T & A^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -AZ^T & I \end{bmatrix} = \begin{bmatrix} A - ZAZ^T & 0 \\ 0 & A^{-1} \end{bmatrix}.$$

Since rank is preserved under equivalence,

$$\begin{bmatrix} A & 0 \\ 0 & A^{-1} - Z^T A^{-1} Z \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} A - ZAZ^T & 0 \\ 0 & A^{-1} \end{bmatrix}$$

have the same rank. Moreover $\operatorname{rank}(A) = \operatorname{rank}(A^{-1})$ implies that $\operatorname{rank}(A^{-1} - Z^T A^{-1} Z) = \operatorname{rank}(A - ZAZ^T)$. Consequently,

$$\begin{aligned} \text{dis-rank}(A^{-1}) &= \operatorname{rank}(A^{-1} - Z^T A^{-1} Z) \\ &= \operatorname{rank}(A - ZAZ^T) \\ &= \operatorname{rank}(JAJ - JZAZ^T J) \\ &= \operatorname{rank}(JAJ - Z^T JAJZ) \\ &= \text{dis-rank}(JAJ). \qquad \square \end{aligned}$$

The following inequality, due to Comon [3], shows that if A has small displacement rank then so does its pseudoinverse A^+:

$$\text{dis-rank}(A^+) \leq 2 \text{ dis-rank}(JAJ).$$

Note that Hermitian Toeplitz matrices usually have displacement inertia $(1, 1, m-2)$. Hence Hermitian matrices with low displacement inertia are regarded as being "nearly Toeplitz." Similar to displacement rank, displacement inertia is preserved (loosely speaking) under inversion. We learned about this theorem from Tiberiu Constantinescu (Institute of Mathematics of the Romanian Academy of Sciences). Recall that two matrices $A, B \in \operatorname{Herm}(m)$ are *$*$-congruent* if there exists $S \in Gl(m)$ such that $B = SAS^*$ where S^* denotes the complex conjugated transpose of S. Note that A and B are $*$-congruent if and only if $\operatorname{inertia}(A) = \operatorname{inertia}(B)$.

THEOREM 1.3. *For $A \in Gl(m) \cap \operatorname{Herm}(m)$, $\text{dis-inertia}(A^{-1}) = \text{dis-inertia}(JAJ)$.*

Proof. If $X = -A^{-1}Z$ then

$$\begin{bmatrix} I & 0 \\ X^* & I \end{bmatrix} \begin{bmatrix} A & Z \\ Z^T & A^{-1} \end{bmatrix} \begin{bmatrix} I & X \\ 0 & I \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & A^{-1} - Z^T A^{-1} Z \end{bmatrix}.$$

If $Y = -AZ^T$ then

$$\begin{bmatrix} I & Y^* \\ 0 & I \end{bmatrix} \begin{bmatrix} A & Z \\ Z^T & A^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ Y & I \end{bmatrix} = \begin{bmatrix} A - ZAZ^T & 0 \\ 0 & A^{-1} \end{bmatrix}.$$

Since inertia is preserved under $*$-congruence,

$$\begin{bmatrix} A & 0 \\ 0 & A^{-1} - Z^T A^{-1} Z \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} A - ZAZ^T & 0 \\ 0 & A^{-1} \end{bmatrix}$$

have the same inertia. Moreover $\operatorname{inertia}(A) = \operatorname{inertia}(A^{-1})$ implies that $\operatorname{inertia}(A^{-1} - Z^T A^{-1} Z) = \operatorname{inertia}(A - ZAZ^T)$. Consequently,

$$\begin{aligned} \text{dis-inertia}(A^{-1}) &= \operatorname{inertia}(A^{-1} - Z^T A^{-1} Z) \\ &= \operatorname{inertia}(A - ZAZ^T) \\ &= \operatorname{inertia}(JAJ - JZAZ^T J) \\ &= \operatorname{inertia}(JAJ - Z^T JAJZ) \\ &= \text{dis-inertia}(JAJ) \qquad \square \end{aligned}$$

Other versions of displacement structure can be defined and theorems analogous to the last two can often be proved also. See Chun and Kailath [2] and Heinig and Rost [5].

The rest of this paper is organized as follows. In §2, we characterize those linear operators on general matrices that preserve both rank and displacement rank (Theorem 2.4). As a consequence, we obtain the characterization of those linear operators preserving singular values and displacement rank (Theorem 2.12). The aim of §3 is to characterize those linear operators on Hermitian matrices that preserve inertia and displacement inertia (Theorem 3.4). We also obtain the characterization of those linear operators preserving eigenvalues and displacement inertia (Theorem 3.5). Then we have concluding remarks in §4.

2. Preserving rank and displacement rank. In this section, we characterize those nonzero linear operators on $C^{m\times n}$ that preserve both rank and displacement rank and also those that preserve singular values and displacement rank. The following theorem appears in Horn, Li, and Tsing [7], and characterizes the linear operators preserving equivalence.

THEOREM 2.1. *Let $T : C^{m\times n} \longrightarrow C^{m\times n}$ be a nonzero linear operator. Then the following conditions are equivalent.*

(*) *$T.A$ is equivalent to $T.B$ whenever A is equivalent to B.*

(**) *There exist $M \in Gl(m), N \in Gl(n)$ such that either for all $X \in C^{m\times n}, T.X = MXN$ or $m=n$ and for all $X \in C^{m\times n}$, $T.X = MX^TN$.*

As a consequence of this result, we obtain the characterizations of linear operators preserving rank and those preserving singular values.

THEOREM 2.2. *Let $T : C^{m\times n} \longrightarrow C^{m\times n}$ be a nonzero linear operator. Then the following conditions are equivalent.*

(*) *For all $X \in C^{m\times n}$, $\mathrm{rank}(T.X) = \mathrm{rank}(X)$.*

(**) *There exist $M \in Gl(m), N \in Gl(n)$ such that either for all $X \in C^{m\times n}, T.X = MXN$ or $m = n$ and for all $X \in C^{m\times n}, T.X = MX^TN$.*

Proof. (**) $\Rightarrow$ (*). Direct verification. (*) $\Rightarrow$ (**). If T preserves rank then it also preserves equivalence. Hence T has the required forms by Theorem 2.1. □

THEOREM 2.3. *Let $T : C^{m\times n} \longrightarrow C^{m\times n}$ be a nonzero linear operator. Then the following conditions are equivalent.*

(*) *For all $X \in C^{m\times n}$, $\mathrm{sing}(T.X) = \mathrm{sing}(X)$.*

(**) *There exist $U \in U(m), V \in U(n)$ such that either for all $X \in C^{m\times n}, T.X = UXV$ or $m = n$ and for all $X \in C^{m\times n}, T.X = UX^TV$.*

Proof. (**) $\Rightarrow$ (*). Direct verification.

(*) $\Rightarrow$ (**). If T preserves singular values then it also preserves rank. By Theorem 2.2, there exist $M \in Gl(m), N \in Gl(n)$ such that either $T.X = MXN$ or $m = n$ and $T.X = MX^TN$. By the singular value decomposition, $M = U_1\Sigma_1U_2$ and $N = V_1\Sigma_2V_2$, where $U_i \in U(m), V_i \in U(n)$, $\Sigma_1 = \mathrm{Diag}(a_1, \ldots, a_m), \Sigma_2 = \mathrm{Diag}(b_1, \ldots, b_n)$. We consider the case when $T.X = MXN$. If $X = U_2^*E^{ij}V_1^*$ then $T.X = U_1\Sigma_1E^{ij}\Sigma_2V_2$. Since $\mathrm{sing}(X) = \mathrm{sing}(T.X)$, we have $a_ib_j = 1$. Consequently, $a_1 = \cdots = a_m =: a, b_1 = \cdots = b_n =: b$ and $ab = 1$. This implies that $T.X = UXV$ where $U := U_1U_2$ and $V := V_1V_2$. The proof is similar for the other case. □

Note that a linear operator $T : C^{m\times n} \longrightarrow C^{m\times n}$ preserves displacement rank if and only if $\nabla \circ T \circ \nabla^{-1}$ preserves rank. Hence we can use Theorem 2.2 to characterize the linear operators that preserve displacement rank. For nonzero T we obtain the following equivalent conditions.

(*) For all $X \in C^{m\times n}$, $\text{dis-rank}(T.X) = \text{dis-rank}(X)$.

(**) There exist $M \in Gl(m)$ and $N \in Gl(n)$ such that either for all $X \in C^{m\times n}, T.X = \nabla^{-1}.(M(\nabla.X)N)$ or $m = n$ and, for all $X \in C^{m\times n}, T.X = \nabla.(M(\nabla^{-1}.(X^T))N)$.

Now we characterize those linear operators on $C^{m\times n}$ preserving both rank and displacement rank. For $\lambda \in C$, we use $D_n(\lambda)$ to denote the $n \times n$ diagonal matrix with diagonal entries $1, \lambda, \ldots, \lambda^{n-1}$; in symbols

$$D_n(\lambda) := \mathrm{Diag}(1, \lambda, \ldots, \lambda^{n-1}).$$

THEOREM 2.4. *Let $T : C^{m\times n} \longrightarrow C^{m\times n}$ be a nonzero linear operator. Then the following conditions are equivalent.*

(*) *For all $X \in C^{m\times n}$, $\mathrm{rank}(T.X) = \mathrm{rank}(X)$ and $\text{dis-rank}(T.X) = \text{dis-rank}(X)$.*

(**) *There exist $\lambda \neq 0$ and lower triangular Toeplitz matrices $M \in Gl(m), N \in Gl(n)$ such that either for all $X \in C^{m\times n}, T.X = D_m(\lambda)MXN^TD_n(\lambda^{-1})$ or $m=n$ and for all $X \in C^{m\times n}, T.X = D_m(\lambda)MX^TN^TD_n(\lambda^{-1})$.*

Before we prove this theorem, we need some preliminary lemmas. The first one is a characterization of matrices that nearly commute with the shift matrix.

LEMMA 2.5. *Let $B \in C^{n\times n}$ and $\lambda \neq 0$. Then the following conditions are equivalent.*

(*) $BZ_n = \lambda Z_n B$.

(**) *There exists a lower triangular Toeplitz matrix L such that $B = D_n(\lambda)L$.*

Proof. First we observe that $D_n(\lambda)Z_n = \lambda Z_n D_n(\lambda)$. (*) $\Rightarrow$ (**). Let $L := D_n(\lambda^{-1})B$. Then $LZ_n = Z_nL$. By comparing entries, one deduces that L is a lower triangular Toeplitz matrix.

(**) $\Rightarrow$ (*). Since L is a lower triangular matrix, $LZ_n = Z_nL$. Hence $BZ_n = D_n(\lambda)LZ_n = D_n(\lambda)Z_nL = \lambda Z_n D_n(\lambda)L = \lambda Z_n B$. □

Next we collect some basic results about the Kronecker product. For $A \in C^{n\times n}$ and $B \in C^{m\times m}$, recall that $A \otimes B : C^{m\times n} \longrightarrow C^{m\times n}$ is a linear operator which may be defined by

$$(A \otimes B).X := BXA^T.$$

We prefer this "coordinate-free" definition to the usual one (compare to Horn and Johnson [6], Graham [4], or Lancaster and Tismenetsky [10]). A fundamental property (which is easy to verify using this definition) is that

$$(A \otimes B) \circ (C \otimes D) = (AC \otimes BD),$$

where $\circ$ denotes the composition of two operators. Moreover, it can be proved that $\mathrm{eigen}(A \otimes B) = \{\alpha_i\beta_j : 1 \leq i \leq n,\ 1 \leq j \leq m\}$, where $\mathrm{eigen}(A) = \{\alpha_i : 1 \leq i \leq n\}$ and $\mathrm{eigen}(B) = \{\beta_j : 1 \leq j \leq m\}$. Hence $\mathrm{tr}(A \otimes B) = (\mathrm{tr}\, A)(\mathrm{tr}\, B)$. The next result, which is taken from Marcus and Moyls [11], is a form of uniqueness for Kronecker product representations.

LEMMA 2.6. *Let $X_i, W_i \in C^{n\times n}$ and $Y_i, V_i \in C^{m\times m}$. If $\sum_{i=1}^r X_i \otimes Y_i = \sum_{i=1}^s W_i \otimes V_i$. and the X_i are linearly independent then each $Y_i \in \mathrm{Span}\{V_1, \ldots, V_s\}$.*

Proof. Since the X_j are linearly independent, for each i there exists P_i such that $\mathrm{tr}(P_iX_j) = \delta(i,j)$. Let K_p be the $m\times m$ matrix with all entries zero except the pth column with all entries one. Then $\mathrm{tr}[(P_i\otimes K_p)(A\otimes B)(I_n\otimes K_q^T)] = \mathrm{tr}(P_iA)\,\mathrm{tr}(K_pBK_q^T) = m\ \mathrm{tr}(P_iA)\ B(p,q)$, where A, B are matrices with appropiate dimensions and $B(p,q)$ is the (p,q) entry of B. Consequently, we have

$$m\sum_{j=1}^{r} tr(P_iX_j)Y_j(p,q) = m\sum_{j=1}^{s} tr(P_iW_j)\ V_j(p,q),$$

and so

$$Y_i(p,q) = \sum_{j=1}^{s} tr(P_i W_j)\, V_j(p,q). \qquad \square$$

COROLLARY 2.7. *Let $X_i \in C^{n\times n}$ and $Y_j \in C^{m\times m}$. If $\{X_i\}$ and $\{Y_j\}$ are linearly independent sets of matrices then $\{X_i \otimes Y_j\}$ is a linearly independent set.*

Proof. Assume $\sum_{i,j} a_{ij} X_i \otimes Y_j = 0$. We rewrite this equation as

$$\sum_i \left[X_i \otimes \left(\sum_j a_{ij} Y_j \right) \right] = 0.$$

Use Lemma 2.6 (with all $W_i = 0$ and $V_i = 0$) to conclude that for all i,

$$\sum_j a_{ij} Y_j = 0.$$

Since the Y_j are linearly independent, $a_{ij} = 0$ for all i, j. $\square$

COROLLARY 2.8. *Let $X \in Gl(n), Y \in Gl(m)$. Then $\{X \otimes Y, Z_n X \otimes Y, X \otimes Z_m Y, Z_n X \otimes Z_m Y\}$ is a linearly independent set.*

Proof. Since $X \in Gl(n)$, $\{X, Z_n X\}$ is a linearly independent set. Similarly $\{Y, Z_m Y\}$ is a linearly independent set. Apply Corollary 2.7 to obtain the required result. $\square$

The following result appears in Horn and Johnson [6] and Graham [4].

LEMMA 2.9. *Let* trans $:= C^{n\times n} \longrightarrow C^{n\times n} : X \longrightarrow X^T$ *denote the linear operator of taking transpose. Then* trans *has the following Kronecker product representation:*

$$\text{trans} = \sum_{i,j=1}^{n} E^{ij} \otimes E^{ji}.$$

Proof. For $X = (x_{ij}) \in C^{n\times n}$, note that $E^{ji} X E^{ji} = x_{ij} E^{ji}$. Now we have

$$\text{trans}.X = X^T = \sum_{i,j} x_{ij} E^{ji} = \sum_{i,j} E^{ji} X E^{ji} = \sum_{i,j} E^{ij} \otimes E^{ji}.X. \qquad \square$$

We adopt the convention that $E^{ij} = 0$ if $i > n, j > n, i < 1$, or $j < 1$. Then it is easy to verify that $Z E^{ij} = E^{(i+1)j}$, and $E^{ij} Z = E^{i(j-1)}$. With this observation, we are ready to prove the next lemma.

LEMMA 2.10. *If $P, Q, R, S \in C^{n\times n}$ are such that*

$$(I_n \otimes I_n - Z_n \otimes Z_n) \circ (Q \otimes P) = (S \otimes R) \circ \text{trans} \circ (I_n \otimes I_n - Z_n \otimes Z_n),$$

then at least one of $\{P, R, S\}$ is singular.

Proof. Assume that $P, R, S \in Gl(n)$. By Lemma 2.9, trans $= \sum_{i,j=1}^{n} E^{ij} \otimes E^{ji}$. Hence

$$(I_n \otimes I_n - Z_n \otimes Z_n)(Q \otimes P) = (S \otimes R) \left(\sum_{i,j=1}^{n} E^{ij} \otimes E^{ji} \right) (I_n \otimes I_n - Z_n \otimes Z_n).$$

Since

$$\begin{aligned}\sum_{i,j=1}^{n}(E^{ij}\otimes E^{ji})(I_n\otimes I_n - Z_n\otimes Z_n) &= \sum_{i,j=1}^{n} E^{ij}\otimes E^{ji} - E^{i(j-1)}\otimes E^{j(i-1)}\\ &= \sum_{i,j=1}^{n} E^{ij}\otimes E^{ji} - E^{ij}\otimes E^{(j+1)(i-1)}\\ &= \sum_{i,j=1}^{n} E^{ij}\otimes (E^{ji} - E^{(j+1)(i-1)}),\end{aligned}$$

we have

$$Q\otimes P - Z_nQ\otimes Z_nP = \sum_{i,j=1}^{n} SE^{ij}\otimes R(E^{ji} - E^{(j+1)(i-1)}).$$

Note that $\{SE^{ij}\}$ is a linearly independent set since S is nonsingular. Use Lemma 2.6 to conclude that, for all $i, j, R(E^{ji} - E^{(j+1)(i-1)}) \in \text{Span}\{P, Z_nP\}$. In particular, $RE^{11}, RE^{21}, R(E^{12} - E^{21}) \in \text{Span}\{P, Z_nP\}$. Hence, using the fact that R and P are nonsingular,

$$3 = \dim\ \text{Span}\{RE^{11}, RE^{21}, R(E^{12} - E^{21})\} \leq \dim\ \text{Span}\{P, Z_nP\} = 2.$$

This is a contraction. □

LEMMA 2.11. *If $P, R \in Gl(m)$ and $Q, S \in Gl(n)$ satisfy*

$$(I_n\otimes I_m - Z_n\otimes Z_m)\circ(Q\otimes P) = (S\otimes R)\circ(I_n\otimes I_m - Z_n\otimes Z_m),$$

then there exist $\lambda \neq 0$ and lower triangular Toeplitz matrices $N \in Gl(n), M \in Gl(m)$ such that $Q = D_n(\lambda^{-1})N$ and $P = D_m(\lambda)M$.

Proof. Note that we can rewrite the given equation as follows:

$$Q\otimes P - Z_nQ\otimes Z_mP = S\otimes R - SZ_n\otimes RZ_m. \tag{1}$$

By Lemma 2.6, $S, SZ_n \in \text{Span}\{Q, Z_nQ\}$, i.e., there exist $\alpha, \beta, \gamma, \delta \in C$ such that

$$S = \alpha Q + \beta Z_nQ \quad \text{and} \quad SZ_n = \gamma Q + \delta Z_nQ.$$

Note that $S = (\alpha I + \beta Z_n)Q$ has rank n; hence $\alpha \neq 0$. Also note that $SZ_n = (\gamma I + \delta Z_n)Q$ has rank $n-1$; hence $\gamma = 0$. Furthermore, $0 \neq SZ_n = \delta Z_nQ$ and hence $\delta \neq 0$. In summary, we have

$$S = \alpha Q + \beta Z_nQ \quad \text{and} \quad SZ_n = \delta Z_nQ,$$

where $\alpha \neq 0$ and $\delta \neq 0$. Similarly, we get

$$R = aP + bZ_mP \quad \text{and} \quad RZ_m = dZ_mP,$$

where $a \neq 0$ and $d \neq 0$. Substituting back into (1), we deduce that, by Corollary 2.8, $\beta = 0$, $b = 0$, and $\alpha a = \delta d = 1$. Thus

$$S = \alpha Q,\ SZ_n = \delta Z_nQ,$$

$$R = aP, \qquad RZ_m = dZ_mP,$$

and hence

$$\lambda^{-1}Z_nQ = QZ_n \quad \text{and} \quad \lambda Z_mP = PZ_m$$

where $\lambda := \alpha/\delta = d/a$. By Lemma 2.5,

$$Q = D_n(\lambda^{-1})N \quad \text{and} \quad P = D_m(\lambda)M,$$

where M, N are the lower triangular Toeplitz matrices of dimension m, n, respectively. □

We are now ready to prove Theorem 2.4.

Proof. $(**) \Rightarrow (*)$. It is clear that T preserves rank. It remains to show that T preserves displacement rank.

Case 1. $T.X = D_m(\lambda)MXN^TD_n(\lambda^{-1})$. By Lemma 2.5, we have

$$\begin{aligned} T.X - Z_m(T.X)Z_n^T &= D_m(\lambda)MXN^TD_n(\lambda^{-1}) - Z_mD_m(\lambda)MXN^TD(\lambda^{-1})Z_n^T \\ &= D_m(\lambda)MXN^TD_n(\lambda^{-1}) - \lambda^{-1}D_m(\lambda)MZ_mX\lambda Z_n^TN^TD(\lambda^{-1}) \\ &= D_m(\lambda)MXN^TD_n(\lambda^{-1}) - D_m(\lambda)MZ_mXZ_n^TN^TD(\lambda^{-1}) \\ &= D_m(\lambda)M(X - Z_mXZ_n^T)N^TD(\lambda^{-1}). \end{aligned}$$

Hence

$$\begin{aligned} \text{dis-rank}(T.X) &= \text{rank}(T.X - Z_m(T.X)Z_n^T) \\ &= \text{rank}(X - Z_mXZ_n^T) \\ &= \text{dis-rank}(X). \end{aligned}$$

Case 2. $T.X = D_m(\lambda)MX^TN^TD_n(\lambda^{-1})$. Using an argument like Case 1, we conclude that T preserves displacement rank.

$(*) \Rightarrow (**)$. We assume that T is a nonzero linear operator that preserves rank and displacement rank. We define $\hat{T} : C^{m\times n} \longrightarrow C^{m\times n}$ by

$$\hat{T} := (I_n \otimes I_m - Z_n \otimes Z_m) \circ T \circ (I_n \otimes I_m - Z_n \otimes Z_m)^{-1}.$$

Hence

$$(I_n \otimes I_m - Z_n \otimes Z_m) \circ T = \hat{T} \circ (I_n \otimes I_m - Z_n \otimes Z_m).$$

Since T preserves rank, by Theorem 2.2, there exist $P \in Gl(m), Q \in Gl(n)$ such that either $T = Q \otimes P$ or $m = n$ and $T = (Q \otimes P) \circ \text{trans}$. On the other hand, since T preserves displacement rank, it follows that $\hat{T}$ preserves rank. Then, by Theorem 2.2, there exist $R \in Gl(m), S \in Gl(n)$ such that either $\hat{T} = S \otimes R$ or $m = n$ and $\hat{T} = (S \otimes R) \circ \text{trans}$. We have four cases to consider.

Case 1. $T = Q \otimes P$ and $\hat{T} = S \otimes R$. Note that $(I_n \otimes I_m - Z_n \otimes Z_m) \circ (Q \otimes P) = (S \otimes R) \circ (I_n \otimes I_m - Z_n \otimes Z_m)$. Then, by Lemma 2.11, there exist $\lambda \neq 0$ and lower triangular Toeplitz matrices $N \in Gl(n), M \in Gl(m)$ such that $Q = D_n(\lambda^{-1})N$ and $P = D_m(\lambda)M$. Consequently, $T = D_n(\lambda^{-1})N \otimes D_m(\lambda)M$.

Case 2. $m = n$, $T = Q \otimes P$ and $\hat{T} = (S \otimes R) \circ \text{trans}$. Note that $(I_n \otimes I_m - Z_n \otimes Z_m) \circ (Q \otimes P) = (S \otimes R) \circ \text{trans} \circ (I_n \otimes I_m - Z_n \otimes Z_m)$. Then, by Lemma 2.10, one of $\{P, R, S\}$ is singular, a contradiction.

Case 3. $m = n$, $T = (Q \otimes P) \circ \text{trans}$ and $\hat{T} = S \otimes R$. Note that $(I_n \otimes I_m - Z_n \otimes Z_m) \circ (Q \otimes P) \circ \text{trans} = (S \otimes R) \circ (I_n \otimes I_m - Z_n \otimes Z_m)$. Using the fact that

$$(I_n \otimes I_m - Z_n \otimes Z_m) \circ \text{trans} = \text{trans} \circ (I_n \otimes I_m - Z_n \otimes Z_m),$$

we deduce that

$$(I_n \otimes I_m - Z_n \otimes Z_m) \circ (Q \otimes P) = (S \otimes R) \circ \text{trans} \circ (I_n \otimes I_m - Z_n \otimes Z_m).$$

Then, by Lemma 2.10, one of $\{P, R, S\}$ is singular, a contradiction.

Case 4. $m = n$, $T = (Q \otimes P) \circ \text{trans}$ and $\hat{T} = (S \otimes R) \circ \text{trans}$. Note that $(I_n \otimes I_m - Z_n \otimes Z_m) \circ (Q \otimes P) \circ \text{trans} = (S \otimes R) \circ \text{trans} \circ (I_n \otimes I_m - Z_n \otimes Z_m)$. Using the fact that

$$(I_n \otimes I_m - Z_n \otimes Z_m) \circ \text{trans} = \text{trans} \circ (I_n \otimes I_m - Z_n \otimes Z_m),$$

we deduce that

$$(I_n \otimes I_m - Z_n \otimes Z_m) \circ (Q \otimes P) = (S \otimes R) \circ (I_n \otimes I_m - Z_n \otimes Z_m).$$

Then, by Lemma 2.11, there exist $\lambda \neq 0$ and lower triangular Toeplitz matrices $N \in Gl(n), M \in Gl(m)$ such that $Q = D_n(\lambda^{-1})N$ and $P = D_m(\lambda)M$. Consequently, $T = D_n(\lambda^{-1})N \otimes D_m(\lambda)M$. □

Next we give the characterization of those linear operators on $C^{n\times n}$ preserving both singular values and displacement rank.

THEOREM 2.12. *Let $T : C^{m\times n} \longrightarrow C^{m\times n}$ be a nonzero linear operator. Then the following conditions are equivalent.*

(*) *For all $X \in C^{m\times n}$, $\text{sing}(T.X) = \text{sing}(X)$ and $\text{dis-rank}(T.X) = \text{dis-rank}(X)$.*

(**) *There exist $|\lambda| = |\mu| = 1$ such that either for all $X \in C^{m\times n}, T.X = \mu D_m(\lambda) X D_n(\lambda^{-1})$ or $m = n$ and for all $X \in C^{m\times n}, T.X = \mu D_m(\lambda) X^T D_n(\lambda^{-1})$.*

Proof. (**) $\Rightarrow$ (*). Since $|\lambda| = |\mu| = 1, \mu D_m(\lambda)$ and $D_n(\lambda^{-1})$ are unitary. Hence T preserves singular values. By Theorem 2.4, we know T also preserves displacement rank.

(*) $\Rightarrow$ (**). Since T preserves singular values, by Theorem 2.3, there exist $U \in U(m), V \in U(n)$ such that either $T.X = UXV$ or $m = n$ and $T.X = UX^TV$. On the other hand, since T preserves both rank and displacement rank, by Theorem 2.4, there exist $\lambda \neq 0$ and lower triangular Toeplitz matrices $M \in Gl(m), N \in Gl(n)$ such that either $T.X = D_m(\lambda)MXN^TD_n(\lambda^{-1})$ or $m = n$ and $T.X = D_m(\lambda)MX^TN^TD_n(\lambda^{-1})$. We consider the following four cases.

Case 1. $T.X = D_m(\lambda)MXN^TD_n(\lambda^{-1})$ and $T.X = UXV$. For all $X \in C^{m\times n}$, $D_m(\lambda)MXN^TD_n(\lambda^{-1}) = UXV$. Then there exists $\alpha \in C$ such that $\alpha D_m(\lambda)M = U$ and $\frac{1}{\alpha}N^TD_n(\lambda^{-1}) = V$. Therefore both $\alpha D_m(\lambda)M$ and $\frac{1}{\alpha}N^TD_n(\lambda^{-1})$ are diagonal and so $M = uI_m$ and $N = vI_n$ for some $u, v \in C$. Moreover $|\lambda| = |uv| = 1$. Consequently, $T.X = \mu D_m(\lambda)XD_n(\lambda^{-1})$ where $\mu := uv$.

Case 2. $m = n$, $T.X = D_m(\lambda)MXN^TD_n(\lambda^{-1})$ and $T.X = UX^TV$. For all $X \in C^{m\times n}$, $D_m(\lambda)MXN^TD_n(\lambda^{-1}) = UX^TV$. Evaluating at $X = I_n$, we get $D_m(\lambda)MN^TD_n(\lambda^{-1}) = UV$. Let $W := U^*D_m(\lambda)M = VD_n(\lambda)N^{-T}$. Then $WX = X^TW$ for all $X \in C^{m\times n}$. In particular, W commutes with every diagonal matix. Hence W is a diagonal matrix, and so $WX = (WX)^T$ for all $X \in C^{m\times n}$. Since W is invertible, it follows that $X{=}X^T$ for all $X \in C^{m\times n}$, a contradiction.

Case 3. $m = n$, $T.X = D_m(\lambda)MX^TN^TD_n(\lambda^{-1})$ and $T.X = UXV$. Using the same argument as in Case 2, we conclude that Case 3 is impossible.

Case 4. $m = n$, $T.X = D_m(\lambda)MX^TN^TD_n(\lambda^{-1})$ and $T.X = UX^TV$. Using the same argument as in Case 1, we conclude that $T.X = \mu D_m(\lambda)X^TD_n(\lambda^{-1})$. □

3. Preserving inertia and displacement inertia. In this section, we characterize those nonzero linear operators on Herm(n) that preserve both inertia and displacement inertia and those preserving eigenvalues and displacement inertia. The following theorem appears in Horn, Li, and Tsing [7]. It characterizes the linear operators preserving $*$-congruence. For simplicity, we write I_n as I, Z_n as Z, and $D_n(\lambda)$ as $D(\lambda)$ in the following theorem.

THEOREM 3.1. *Let* $T : \text{Herm}(n) \longrightarrow \text{Herm}(n)$ *be a nonzero linear operator. Then the following conditions are equivalent.*

(*) $T.A$ *is* $*$*-congruent to* $T.B$ *whenever* A *is* $*$*-congruent to* B.

(**) *There exists* $S \in Gl(n)$ *such that either for all* $X \in \text{Herm}(n), T.X = \pm SXS^*$ *or for all* $X \in \text{Herm}(n), T.X = \pm SX^TS^*$.

As a consequence of this result, we obtain the characterizations of linear operators preserving inertia and those preserving eigenvalues.

THEOREM 3.2. *Let* $T : \text{Herm}(n) \longrightarrow \text{Herm}(n)$ *be a nonzero linear operator. Then the following conditions are equivalent:*

(*) *For all* $X \in \text{Herm}(n), \text{inertia}(T.X) = \text{inertia}(X)$.

(**) *There exists* $S \in Gl(n)$ *such that either, for all* $X \in \text{Herm}(n), T.X = SXS^*$ *or, for all* $X \in \text{Herm}(n), T.X = SX^TS^*$.

Proof. (**) $\Rightarrow$ (*). Direct verification.

(*) $\Rightarrow$ (**). If T preserves inertia then it also preserves $*$-congruence. Hence, by Theorem 3.1, there exists $S \in Gl(n)$ such that either $T.X = \pm SXS^*$ or $T.X = \pm SX^TS^*$. However, the cases with minus signs are ruled out because T preserves inertia. □

THEOREM 3.3. *Let* $T : \text{Herm}(n) \longrightarrow \text{Herm}(n)$ *be a nonzero linear operator. Then the following conditions are equivalent.*

(*) *For all* $X \in \text{Herm}(n), \text{eigen}(T.X) = \text{eigen}(X)$.

(**) *There exists* $U \in U(n)$ *such that either for all* $X \in \text{Herm}(n), T.X = UXU^*$ *or for all* $X \in \text{Herm}(n), T.X = UX^TU^*$.

Proof. (**) $\Rightarrow$ (*). Direct verification.

(*) $\Rightarrow$ (**). If T preserves eigenvalues then it also preserves inertia. Hence, by Theorem 3.2, there exists $S \in Gl(n)$ such that either $T.X = SXS^*$ or $T.X = SX^TS^*$. Since T preserves eigenvalues, $\text{eigen}(I) = \text{eigen}(T.I) = \text{eigen}(SS^*)$. Therefore $SS^* = I$ and so $S \in U(n)$. □

Note that a linear operator $T : \text{Herm}(n) \longrightarrow \text{Herm}(n)$ preserves displacement inertia if and only if $\nabla \circ T \circ \nabla^{-1}$ preserves inertia. Hence we can use Theorem 3.2 to characterize the linear operators that preserve displacement inertia. For nonzero T we obtain the following equivalent conditions.

(*) For all $X \in \text{Herm}(n), \text{dis-inertia}(T.X) = \text{dis-inertia}(X)$.

(**) There exists $S \in Gl(n)$ such that either for all $X \in \text{Herm}(n), T.X = \nabla^{-1}.(S(\nabla.X)S^*)$ or for all $X \in \text{Herm}(n), T.X = \nabla^{-1}.(S(\nabla.X^T)S^*)$.

Now we characterize those linear operators on Herm(n) preserving both inertia and displacement inertia.

THEOREM 3.4. *Let* $T : \text{Herm}(n) \longrightarrow \text{Herm}(n)$ *be a nonzero linear operator. Then the following conditions are equivalent.*

(*) *For all* $X \in \text{Herm}(n), \text{inertia}(T.X) = \text{inertia}(X)$ *and* $\text{dis-inertia}(T.X) = \text{dis-inertia}(X)$.

(**) *There exists* $|\lambda| = 1$ *and a lower triangular Toeplitz* $N \in Gl(n)$ *such that either, for all* $X \in \text{Herm}(n), T.X = D(\lambda)NXN^*D(\lambda)^*$ *or for all* $X \in \text{Herm}(n), T.X = D(\lambda)NX^TN^*D(\lambda)^*$.

Proof. $(**) \Rightarrow (*)$. It is clear that T preserves inertia. It remains to show that T preserves displacement inertia.

Case 1. $T.X = D(\lambda)NXN^*D(\lambda)^*$. By Lemma 2.5, we have

$$\begin{aligned} T.X - Z(T.X)Z^T &= D(\lambda)NXN^*D(\lambda)^* - ZD(\lambda)NXN^*D(\lambda)^*Z^T \\ &= D(\lambda)NXN^*D(\lambda)^* - \lambda^{-1}D(\lambda)NZX(\lambda^*)^{-1}Z^TN^*D(\lambda)^* \\ &= D(\lambda)NXN^*D(\lambda)^* - D(\lambda)NZXZ^TN^*D(\lambda)^* \\ &= D(\lambda)N(X - ZXZ^T)N^*D(\lambda)^* \end{aligned}$$

Hence

$$\begin{aligned} \text{dis-inertia}(T.X) &= \text{inertia}(T.X - Z(T.X)Z^T) \\ &= \text{inertia}(X - ZXZ^T) \\ &= \text{dis-inertia}(X). \end{aligned}$$

Case 2. $T.X = D(\lambda)NX^TN^*D(\lambda)^*$. Using an argument like Case 1, we conclude that T preserves displacement inertia.

$(*) \Rightarrow (**)$. We assume that T is a nonzero linear operator that preserves inertia and displacement inertia. We define $\hat{T} : \text{Herm}(n) \longrightarrow \text{Herm}(n)$ by

$$\hat{T} := (I \otimes I - Z \otimes Z) \circ T \circ (I \otimes I - Z \otimes Z)^{-1}.$$

Hence

$$(I \otimes I - Z \otimes Z) \circ T = \hat{T} \circ (I \otimes I - Z \otimes Z).$$

Since T preserves inertia, by Theorem 3.2, there exist $S \in Gl(n)$ such that either $T = \bar{S} \otimes S$ or $T = (\bar{S} \otimes S) \circ \text{trans}$ where $\bar{S}$ denotes the complex conjugate of S. On the other hand, since T preserves displacement inertia, it follows that $\hat{T}$ preserves inertia. Then, by Theorem 3.2, there exist $R \in Gl(n)$ such that either $\hat{T} = \bar{R} \otimes R$ or $\hat{T} = (\bar{R} \otimes R) \circ \text{trans}$. We have four cases to consider.

Case 1. $T = \bar{S} \otimes S$ and $\hat{T} = \bar{R} \otimes R$. Note that $(I \otimes I - Z \otimes Z) \circ (\bar{S} \otimes S) = (\bar{R} \otimes R) \circ (I \otimes I - Z \otimes Z)$. Then, by Lemma 2.11, there exist $\lambda \neq 0$ and lower triangular Toeplitz matrices $N, M \in Gl(n)$ such that $\bar{S} = D(\lambda^{-1})M$ and $S = D(\lambda)N$. Hence $\bar{M} = D(|\lambda|^2)N$. On the other hand, $\bar{M} = N$ due to hermicity, and so $|\lambda| = 1$. Consequently, $T.X = D(\lambda)NX(D(\lambda)N)^*$ for all $X \in \text{Herm}(n)$.

Case 2. $T = \bar{S} \otimes S$ and $\hat{T} = (\bar{R} \otimes R) \circ \text{trans}$. Note that $(I \otimes I - Z \otimes Z) \circ (\bar{S} \otimes S) = (\bar{R} \otimes R) \circ \text{trans} \circ (I \otimes I - Z \otimes Z)$. By Lemma 2.10, we get a contradiction.

Case 3. $T = (\bar{S} \otimes S) \circ \text{trans}$ and $\hat{T} = \bar{R} \otimes R$. Note that $(I \otimes I - Z \otimes Z) \circ (\bar{S} \otimes S) \circ \text{trans} = (\bar{R} \otimes R) \circ (I \otimes I - Z \otimes Z)$. Using the fact that

$$(I \otimes I - Z \otimes Z) \circ \text{trans} = \text{trans} \circ (I \otimes I - Z \otimes Z),$$

we deduce that

$$(I \otimes I - Z \otimes Z) \circ (\bar{S} \otimes S) = (\bar{R} \otimes R) \circ \text{trans} \circ (I \otimes I - Z \otimes Z),$$

which leads to a contradiction by Lemma 2.10.

Case 4. $T = (\bar{S} \otimes S) \circ \text{trans}$ and $\hat{T} = (\bar{R} \otimes R) \circ \text{trans}$. Note that $(I \otimes I - Z \otimes Z) \circ (\bar{S} \otimes S) \circ \text{trans} = (\bar{R} \otimes R) \circ \text{trans} \circ (I \otimes I - Z \otimes Z)$. Using the fact that

$$(I \otimes I - Z \otimes Z) \circ \text{trans} = \text{trans} \circ (I \otimes I - Z \otimes Z),$$

we deduce that

$$(I \otimes I - Z \otimes Z) \circ (\bar{S} \otimes S) = (\bar{R} \otimes R) \circ (I \otimes I - Z \otimes Z).$$

Then we obtain the required result as in Case 1. □

Next we give the characterization of those linear operators on Herm(n) preserving both eigenvalues and displacement inertia.

THEOREM 3.5. *Let* T : Herm(n) $\longrightarrow$ Herm(n) *be a nonzero linear operator. Then the following conditions are equivalent.*

(*) *For all* $X \in$ Herm(n), eigen($T.X$) = eigen(X) *and* dis-inertia($T.X$) = dis-inertia(X).

(**) *There exists* $|\lambda| = 1$ *such that either for all* $X \in$ Herm(n), $T.X = D(\lambda)XD(\lambda)^*$ *or for all* $X \in$ Herm(n), $T.X = D(\lambda)X^TD(\lambda)^*$.

Proof. (**) $\Rightarrow$ (*). Since $|\lambda| = 1$, $D(\lambda)$ is unitary and hence T preserves eigenvalues. From Theorem 3.4, it is clear that T also preserves displacement inertia.

(*) $\Rightarrow$ (**). Since T preserves eigenvalues, from Theorem 3.3, there exists $U \in U(n)$ such that either $T.X = UXU^*$ or $T.X = UX^TU^*$. On the other hand, since T preserves both inertia and displacement inertia, from Theorem 3.4, there exists $|\lambda| = 1$ and a lower triangular Toeplitz $N \in Gl(n)$ such that either $T.X = D(\lambda)NXN^*D(\lambda)^*$ or $T.X = D(\lambda)NX^TN^*D(\lambda)^*$. We consider the following four cases.

Case 1. $T.X = D(\lambda)NXN^*D(\lambda)^*$ and $T.X = UXU^*$. Evaluating at $X = I$, we get $D(\lambda)NN^*D(\lambda)^* = UU^* = I$, i.e., $D(\lambda)N$ is unitary. Since $D(\lambda)N$ is lower triangular, it must be diagonal and so is N. This implies that $N = \mu I$ because N is Toeplitz. Moreover $|\mu| = 1$. Finally $T.X = D(\lambda)XD(\lambda)^*$.

Case 2. $T.X = D(\lambda)NXN^*D(\lambda)^*$ and $T.X = UX^TU^*$. For all $X \in$ Herm(n), $D(\lambda)NXN^*D(\lambda)^* = UX^TU^*$. Evaluating at $X = I$, we find that $D(\lambda)N$ is unitary. Let $V := U^*D(\lambda)N$. Then $VX = X^TV$ for all $X \in$ Herm(n). This implies that V must be a scalar, and so $X = X^T$ for all $X \in$ Herm(n) , a contradiction.

Case 3. $T.X = D(\lambda)NX^TN^*D(\lambda)^*$ and $T.X = UXU^*$. Using the same argument as in Case 2, we conclude that Case 3 is impossible.

Case 4. $T.X = D(\lambda)NX^TN^*D(\lambda)^*$ and $T.X = UX^TU^*$. Using the same argument as in Case 1, we conclude that $T.X = D(\lambda)X^TD(\lambda)^*$. □

4. Concluding remarks. There are many papers on rank preserving linear operators and inertia preserving linear operators, for example see Pierce et al. [12]. Some of these papers characterize linear preservers of one particular rank class or one particular inertia class (rather than characterizing preservers of all rank or inertia classes as was done in Theorems 2.2 and 3.2). These results probably make it possible to characterize linear preservers of one particular rank and displacement-rank class and linear preservers of one particular inertia- and displacement-inertia class. Many of the results in these references treat rank preservers or inertia preservers over the field of real numbers (rather than the field of complex numbers that we used in this paper). Some of the references even deal with more general fields of numbers. These preserver results over other fields probably make it possible to extend the results of this paper to other fields of numbers.

We mentioned earlier that there are definitions of displacement structure that are different than the ones we use in this paper. (See Chun and Kailath [2] and Heinig and Rost [5].) There linear preserver questions are analogous to the ones we studied here for the other definitions. We expect that the techniques that we have used here can be used to easily settle such analogous questions.

In the introduction of this paper we noted that we are interested in the spectral properties of matrices that are Toeplitz or nearly Toeplitz. In particular, we are interested in sets having the forms

$$\text{eigen}^{-1}(\lambda) \cap \text{Toep}(m)$$

or

$$\text{eigen}^{-1}(\lambda) \cap \text{dis-inertia}^{-1}(p, n, z),$$

where

$$\begin{aligned} \lambda &:= (\lambda_1, \ldots, \lambda_m) \in R^m, \\ \text{eigen}^{-1}(\lambda) &:= \{A \in \text{Herm}(m) : \text{eigen}(A) = \lambda\}, \\ \text{Toep}(m) &:= \{A \in C^{m \times m} : A \text{ is Toeplitz}\}, \\ \text{dis-inertia}^{-1}(p, n, z) &:= \{A \in \text{Herm}(m) : \text{dis-inertia}(A) = (p, n, z)\}. \end{aligned}$$

Now, by the spectral theorem, we have that

$$\text{eigen}^{-1}(\lambda) = \{Q \,\text{Diag}(\lambda) Q^* : Q \in U(m)\}.$$

From this we see that linear spectra preserving operators

$$\text{Herm}(m) \longrightarrow \text{Herm}(m) : X \longrightarrow QXQ^*$$

for $Q \in U(m)$ can be used to move around this isospectral surface $\text{eigen}^{-1}(\lambda)$. In more technical language, we see that $\text{eigen}^{-1}(\lambda)$ is the orbit of $\text{Diag}(\lambda)$ under the group action defined by

$$U(m) \times \text{Herm}(m) \longrightarrow \text{Herm}(m) : (Q, X) \longrightarrow QXQ^*.$$

We originally hoped that we could move around somewhat freely on the sets of the form $\text{eigen}^{-1}(\lambda) \cap \text{dis-inertia}^{-1}(p, n, z)$ by means of the linear preservers of such sets. This hope motivated our study of linear preservers. Unfortunately, our hope was too optimistic. Our results show that there are not enough such linear preservers.

Acknowledgment. This research was done during the 1991–1992 academic year which the authors spent at the Institute for Mathematics and Its Applications (IMA) at the University of Minnesota. We thank the members of the IMA for their hospitality. We thank Jack Conn (Department of Mathematics, University of Minnesota) for his suggestions that streamlined some of the proofs. We also thank Leiba Rodman (Department of Mathematics, College of William and Mary) for his comments.

REFERENCES

[1] M. T. Chu, *The stability group of symmetric Toeplitz matrices*, Linear Algebra Appl., 185 (1993), pp. 119–123.

[2] J. Chun and T. Kailath, *Displacement structure for Hankel, Vandermonde and related (derived) matrices*, Linear Algebra Appl., 151 (1991), pp. 199–227.

[3] P. Comon, *Displacement rank of pseudo-inverses*, IEEE Internat. Conf. Acoustics Speech and Signal Processing, Vol. 5, March 23–26, 1992, San Francisco, pp. 49–52.

[4] A. Graham, *Kronecker Products and Matrix Calculus: With Applications*, Wiley, New York, 1981.

[5] G. HEINIG AND K. ROST, *Algebraic Methods for Toeplitz-like Matrices and Operators*, Birkhauser, Berlin, New York, 1984.

[6] R. HORN AND C. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, MA, 1991.

[7] R. HORN, C. K. LI, AND N. K. TSING, *Linear operators preserving certain equivalence relations on matrices*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 195–204.

[8] T. KAILATH, S.-Y. KUNG, AND M. MORF, *Displacement ranks of matrices*, AMS Bull., 1 (1979), pp. 769–773.

[9] ———, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979), pp. 395–407.

[10] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices: With Applications*, Academic Press, New York, 1985.

[11] M. MARCUS AND B. MOYLS, *Transformations on tensor product spaces*, Pacific J. Math., 9 (1959), pp. 1215–1221.

[12] S. PIERCE, M. H. LIN, R. LOEWY, C. K. LI, N. K. TSING, B. K. MCDONALD, AND L. BEASLEY, *A survey of linear preserver problems*, Linear and Multilinear Algebra 33, number 1–2, Gordon Breach, New York, 1993.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 502–519, April 1995

013

RANK M WAVELETS WITH N VANISHING MOMENTS*

PETER NIELS HELLER†

Abstract. This work generalizes the rank 2 (scale factor of 2) orthogonal wavelet sequences of Daubechies to the case of a rank M *wavelet matrix*. Several equivalent definitions of Nth order vanishing moments for rank M wavelets are developed. These notions are used to find an explicit formula for rank M wavelet scaling sequences with N vanishing wavelet moments (of degree N in our terminology). A full wavelet matrix (scaling sequence and $M-1$ wavelet sequences) is constructed, with explicit examples.

Key words. wavelets, vanishing moments, multirate filter banks

AMS subject classifications. 15A57, 42A16, 94A12

1. Introduction.

In this paper we generalize Daubechies' discrete wavelets (which have a scale factor of 2) to the rank M case. The scale factor M is an integer greater than or equal to 2. Daubechies' wavelets are defined by a *scaling sequence* $\{a_k\}_{k=0}^{K-1}$ which satisfies

$$\sum_k a_k a_{k+2l} = 2\delta_{0,l} ,$$

$$\sum_k a_k = 2 .$$

Then the *wavelet sequence* $\{b_k\}_{k=0}^{K-1}$ defined by $b_k = (-1)^k a_{K-1-k}$ is orthogonal to the scaling sequence under shifts by 2, and together they form a rank 2 wavelet system; if we set $a_{0,k} = a_k$ and $a_{1,k} = b_k$ then the entries of the matrix $(a_{s,k})$ satisfy

$$\sum_k a_{s,k} a_{s',k+2l} = 2\delta_{s,s'}\delta_{0,l}$$

$$\text{and } \sum_k a_{s,k} = 2\delta_{s,0} ,$$

and they lead to the scaling and wavelet functions, which form an orthonormal basis of $L^2(\mathbf{R})$. Daubechies further defines the following notion: a rank 2 wavelet system *has N vanishing moments* if the scaling sequence satisfies the sum rules

$$\sum_k (-1)^k k^n a_k = 0 \text{ for } n = 0, 1, \ldots, N-1 . \tag{1}$$

This is simply the statement that the first N moments of the wavelet sequence vanish. Daubechies [1] develops explicit formulae for scaling sequences such that the associated wavelets have N vanishing moments, and uses them to construct arbitrarily differentiable wavelet functions on $\mathbf{R}$.

In the sections to follow, we develop the notion of a rank M wavelet system and define what it means to have N vanishing wavelet moments in the rank M setting. We

* Received by the editors March 11, 1993; accepted for publication (in revised form) by G. Strang, February 9, 1994.

† Aware, Inc., One Memorial Dr., Cambridge, Massachusetts 02142 (heller@aware.com). This research was supported in part by the Advanced Research Projects Agency of the Department of Defense and monitored by the Air Force Office of Scientific Research under contract F49620-92-C-0054.

then develop explicit formulas for rank M scaling sequences with N vanishing wavelet moments, and complete the construction of a full rank M wavelet matrix (one scaling sequence and $M-1$ wavelet sequences) given a scaling sequence and a Haar wavelet matrix. We conclude by giving several examples of our construction.

1.1. Rank M wavelet matrices. The notion of a discrete wavelet system has been generalized to the rank M case [3], [5], [15], [17], in which there is one scaling sequence and $M-1$ wavelet sequences. A real[1] rank M wavelet system is given by an $M \times K$ *wavelet matrix* $\mathbf{A}$ whose entries $a_{s,k}$ satisfy

$$\sum_k a_{s,k} a_{s',k+Ml} = M\delta_{s,s'}\delta_{0,l} \tag{2}$$

$$\text{and} \quad \sum_k a_{s,k} = M\delta_{s,0}\,. \tag{3}$$

The rows of $\mathbf{A}$ are orthogonal under shifts of M. In contrast to the rank 2 case, where the single wavelet sequence is determined by the scaling sequence, the rank M case has considerable freedom in the choice of the $M-1$ wavelet sequences. In §4 we describe a method for constructing a full wavelet matrix given its first row (the scaling sequence) and an $M \times M$ matrix that we call the characteristic Haar matrix [5].

A Haar wavelet matrix is an orthogonal matrix (up to scalar multiplication) whose first row is all ones; that is, its entries $h_{s,k}$ satisfy

$$\sum_k h_{s,k} h_{s',k} = M\delta_{s,s'}$$
$$\text{and} \quad h_{0,k} = 1 \quad \forall k\,.$$

Observe that every such Haar wavelet matrix is a rank M wavelet matrix (with $K = M$) and that every $M \times M$ wavelet matrix is a Haar matrix. Useful examples of Haar wavelet matrices include the M-point discrete Fourier transform (DFT), discrete cosine transform (DCT), and Hadamard matrix. The collection of rank M Haar matrices is isomorphic to the group of orthogonal matrices of rank $M-1$.

It will serve us to think of our wavelet matrices as being $M \times Mg$ for some integer g; we can always pad each row with zeros to bring the wavelet matrix into this form. We call g the *overlap* of the wavelet matrix. If we break up the $M \times Mg$ wavelet matrix $\mathbf{A}$ into its constituent $M \times M$ blocks

$$\mathbf{A} = (\mathbf{A}_0\ \mathbf{A}_1\ \ldots\ \mathbf{A}_{g-1})\,, \tag{4}$$

then the *characteristic Haar matrix associated with* $\mathbf{A}$ is given by

$$\mathbf{H}_0 = \mathbf{A}_0 + \mathbf{A}_1 + \cdots + \mathbf{A}_{g-1}\,.$$

It can be checked that $\mathbf{H}_0$ is in fact a Haar wavelet matrix.

The second step in generalizing Daubechies discrete wavelet systems is to develop a notion of vanishing wavelet moments in the rank M setting—now Mth roots of unity will play a role. For example, the sum rule (1) becomes

$$\sum_k \zeta^k k^n a_{0,k} = 0 \quad \text{for } n = 0, 1, \ldots, N-1\,,$$

[1] In this paper we consider the case of real rank M wavelets; the extension to the complex case is straightforward.

where $\zeta = e^{2\pi i/M}$ is the primitive Mth root of unity. In §2 we develop these ideas further, coining the name "wavelet system of degree N," and in §3 we use this set of conditions to derive explicit formulae for rank M scaling sequences with N vanishing wavelet moments. Using the results of §4, we are then able to construct full wavelet matrices of arbitrary rank and number of vanishing wavelet moments.

It is natural to take these rank M discrete wavelet systems and seek to construct compactly supported rank M wavelet orthonormal bases of $L^2(\mathbf{R})$; initial steps in this direction have been taken by Gopinath and Burrus [3]. These wavelet bases are derived from the *scaling function*, which is a solution to the scaling equation

$$\varphi(x) = \sum_k a_{0,k}\varphi(Mx - k) \,. \tag{5}$$

Thus there is a one-one correspondence between the scaling sequences discussed here and scaling functions on $\mathbf{R}$. The explicit formulae for scaling sequences developed below enable one to construct rank M wavelet bases with arbitrary smoothness (as measured by Sobolev differentiability); this is reported in [7].

1.2. The polyphase matrix representation. Matrices satisfying the orthogonality condition (2) have been extensively studied in the signal processing literature [12]–[14] under the name "M-band paraunitary perfect reconstruction filter banks." Engineers often restate (2) in the z-transform domain, as follows: form the $M \times M$ *polyphase matrix* $\mathbf{H}(z)$ with polynomial entries

$$h_{s,r}(z) = \sum_l a_{s,r+lM} z^l \,.$$

Observe that $\mathbf{H}$ and $\mathbf{A}$ are related by

$$\mathbf{H}(z) = \mathbf{A}_0 + z\mathbf{A}_1 + \cdots + z^{g-1}\mathbf{A}_{g-1} \,.$$

$\mathbf{H}(z)$ is said to be *paraunitary* if $\frac{1}{\sqrt{M}}\mathbf{H}(z)$ is unitary on the unit circle:

$$\mathbf{H}(z)\mathbf{H}^\dagger(z^{-1}) = M\mathbf{I} \quad \text{for } |z| = 1 \,. \tag{6}$$

Comparison of coefficients of powers of z shows that the paraunitarity of $\mathbf{H}$ is equivalent to the orthogonality under shifts (2) of the wavelet system. Paraunitary matrices and polyphase factorizations have been investigated in great detail [13]. We impose the additional linear condition (3) to form a wavelet matrix; this amounts to the requirement that the matrix $\mathbf{H}(z)|_{z=1} = \mathbf{H}_0$ be a Haar wavelet matrix. This proves essential for the later development of orthonormal bases of $L^2(\mathbf{R})$ (cf. [1], [3]).

The orthogonality condition (2) can also be stated in the Fourier domain; for each of the sequences a_s, consider its "symbol" or Fourier transform

$$A_s(e^{i\omega}) = \frac{1}{M}\sum_{k=0}^{K-1} a_{s,k}e^{ik\omega} \,, \quad 0 \le s < M \,.$$

Then (2) is equivalent to

$$\sum_{m=0}^{M-1} A_s(e^{i(\omega+2\pi m/M)})\overline{A_{s'}(e^{i(\omega+2\pi m/M)})} \equiv \delta_{s,s'}. \tag{7}$$

The paraunitarity (6) of $\mathbf{H}(z)$ implies

$$\mathbf{H}^{\dagger}(z^{-1})\mathbf{H}(z) = M\mathbf{I}, \tag{8}$$

and from this follows

$$\sum_{s=0}^{M-1} |A_s(e^{i\omega})|^2 \equiv 1\,. \tag{9}$$

1.3. Discrete wavelet bases. As observed in [5], [10], the wavelet matrix describes a basis for the space of functions on $\mathbf{Z}$. Specifically, a discrete function $f(k)$ may be expanded

$$f(k) = \sum_{s=0}^{M-1} \sum_{l=-\infty}^{\infty} c_{s,l} a_{s,Ml+k}, \tag{10}$$

where

$$c_{s,l} = \frac{1}{M} \sum_{k} f(k) a_{s,Ml+k}. \tag{11}$$

This discrete basis property is equivalent to (9); the wavelet matrix provides a set of *overlapping basis functions* for $\ell^2(\mathbf{Z})$. The additional "low-pass" condition (3) for wavelets confers the ability to develop orthonormal bases of $L^2(\mathbf{R})$.

2. Vanishing moments for rank M wavelets. In this section we generalize to the rank M setting several equivalent definitions for a wavelet matrix $\mathbf{A} = \{a_{s,k}\}$ to have N vanishing moments.

THEOREM 2.1. *A rank M scaling sequence a_0 is said to be* of degree N *if and only if one of the following equivalent conditions holds:*

(i) *the first N moments of the corresponding wavelet sequences vanish;*

(ii) *the symbol A_0 has a zero of order N at the Mth roots of unity $\zeta^m = e^{2\pi i m/M}$;*

(iii) *discrete polynomial sequences of degree $n < N$ are perfectly represented by shifts of the scaling sequence.*

First let us describe each of these conditions in more detail.

(i) The moments of the wavelet sequences vanish to order N if

$$\sum_{k} k^n a_{s,k} = 0, \quad s = 1, 2, \ldots, M-1, \quad \text{and } n = 0, 1, \ldots, N-1.$$

Equivalently,

$$A_s^{(n)}(e^{i\omega})\big|_{\omega=0} = 0 \quad \text{for } s = 1, 2, \ldots, M-1, \qquad n = 0, 1, \ldots, N-1.$$

(ii) A rank M scaling sequence has a zero (is flat) of order N at the roots of unity if its symbol A_0 satisfies

$$A_0^{(n)}(\zeta^m) = 0, \quad m = 1, 2, \ldots, M-1, \quad n = 0, 1, \ldots, N-1. \tag{12}$$

In other words, we can factor

$$A_0(e^{i\omega}) = \left(\prod_{m=1}^{M-1} \left(e^{i\omega} - \zeta^m\right)^N \right) Q(e^{i\omega}) = \left(\frac{1 - e^{iM\omega}}{1 - e^{i\omega}} \right)^N Q(e^{i\omega}), \tag{13}$$

where $Q(e^{i\omega})$ is some trigonometric polynomial. From the sum rule (3), every scaling sequence satisfies

$$A_0(1) = 1, \quad \text{and} \quad A_0(\zeta^m) = 0, \quad m = 1, 2, \ldots, M-1,$$

i.e., *every* scaling sequence is of degree 1. By definition of the symbol A_0, the flatness condition (12) is equivalent to the sum rules

$$\sum_k \zeta^{mk} k^n a_{0,k} = 0, \quad m = 1, 2, \ldots, M-1, \quad \text{and} \quad n = 0, 1, \ldots, N-1 .$$

Rewriting this as

$$\sum_{r=0}^{M-1} \sum_k (r + Mk)^n a_{0,r+Mk} \zeta^{mr} = 0,$$

we see that each of the *partial moments*

$$\mathcal{M}_r^n(a_0) = \sum_k (r + Mk)^n a_{0,r+Mk}$$

must be equal to a constant independent of r, for $n < N$.

(iii) A scaling sequence perfectly represents discrete polynomial sequences of degree $< N$ if, given a polynomial sequence

$$f(k) = \sum_{n=0}^{N-1} \alpha_n k^n ,$$

the discrete wavelet expansion (10) of $f(k)$ contains only shifts of the *scaling* sequence:

$$f(k) = \sum_l c_l a_{0,Ml+k} .$$

In signal processing terminology, putting the sequence f through a rank M wavelet filter bank with lowpass filter a_0 will produce zero outputs from all the wavelet or bandpass filters (and a lowpass output that is a polynomial in the index variable of degree $N-1$).

Proof of Theorem 2.1.

(ii) $\Rightarrow$ (i): Combining

$$A_0^{(n)}(\zeta^m) = 0, \qquad m = 1, 2, \ldots, M-1$$

with

$$\sum_{m=0}^{M-1} A_0(\zeta^m e^{i\omega}) \overline{A_s(\zeta^m e^{i\omega})} \equiv \delta_{0,s}$$

shows that

$$A_s^{(n)}(e^{i\omega})\big|_{\omega=0} = 0, \qquad n = 0, 1, \ldots, N-1,$$

which is (ii).

(i) $\Rightarrow$ (ii): Recall that

$$\sum_{s=0}^{M-1} |A_s(e^{i\omega})|^2 \equiv 1 .$$

If each of the symbols A_s, $s > 1$ vanish to order N at $\omega = 0$, then so must $A_0 - 1$.

(i) is clearly equivalent to (iii). □

Strang [11] gives additional formulations of degree N, relating it to approximation of order N for functions on $L^2(\mathbf{R})$.

3. An explicit formula for scaling sequences of degree N. We now derive an *explicit general formula* for rank M scaling sequences of degree N. This construction combines the flatness characterization (12) of degree N and the orthogonality condition (7) to obtain the modulus squared of the symbol (Fourier transform) of the scaling sequence. Following Daubechies, we then perform a Fejér factorization to obtain the scaling sequence itself. In fact, we are able to describe all possible rank M scaling sequences of degree N, not just the minimal length solutions. Our method generalizes a technique reported in [2], [13] for the $M = 2$ case.[2]

3.1. Minimal length solution. Our first goal is to find a minimal (finite) length rank M scaling sequence $\{a_{0,k}\}$ of degree N. In order to satisfy the flatness of order N condition (12), the symbol $A_0(e^{i\omega}) = \frac{1}{M}\sum_k a_{0,k}e^{ik\omega}$ must be factorizable as

$$A_0(e^{i\omega}) = \left(\frac{1 + e^{i\omega} + e^{i2\omega} + \cdots + e^{i(M-1)\omega}}{M}\right)^N Q(e^{i\omega}) ,$$

i.e., A_0 includes N powers of the rank M Haar trigonometric polynomial:[3]

$$\left(\frac{1 + e^{i\omega} + e^{i2\omega} + \cdots + e^{i(M-1)\omega}}{M}\right) = \frac{1 - e^{iM\omega}}{M(1 - e^{i\omega})} .$$

The modulus squared of A_0 is

$$P(e^{i\omega}) = A_0(e^{i\omega})\overline{A_0(e^{i\omega})} = H^N(e^{i\omega})R(e^{i\omega})$$

with

$$H(e^{i\omega}) = \left|\frac{1 + e^{i\omega} + \cdots + e^{i(M-1)\omega}}{M}\right|^2 . \tag{14}$$

Both $H(e^{i\omega})$ and the remainder $R(e^{i\omega}) = |Q(e^{i\omega})|^2$ are cosine polynomials. Henceforth we freely abuse notation by interchanging $x = \cos\omega$ and $e^{i\omega}$ as the independent variable for functions such as R.

The orthogonality condition (7) can be written

$$P(e^{i\omega}) + P(e^{i(\omega+2\pi/M)}) + \cdots + P(e^{i(\omega+2\pi\frac{M-1}{M})}) \equiv 1 . \tag{15}$$

[2] We have also developed [4] a set of purely algebraic formulae for the scaling sequences of degree $N = 2, 3, 4$, building upon work of Pollen [8]. These formulae yield *closed form* algebraic expressions for the sequences; in the $N = 4$ "$D8$" case this is a new result even for $M = 2$.

[3] Note that we have included a factor of $1/M$ inside the term in parentheses that was missing in (13). This new normalization means that each of the terms inside the parentheses is itself a valid wavelet symbol, that of the Haar scaling sequence.

However, since P vanishes to order $2N$ in ω at the frequencies $\omega = 2\pi m/M$, $m = 1, 2, \ldots, M-1$, we know

$$P(e^{i\omega}) = 1 + \mathcal{O}(|\omega|^{2N}) \tag{16}$$

at $\omega = 0$.

We can now find the minimal length solution $P = H^N R_N$. Since R_N is even, we can write

$$R_N(x) = \sum_{n=0}^{N-1} \rho_n \cos n\omega \, .$$

Writing $x = \cos\omega$, we can write this as a Taylor expansion of R_N about $x = 1$:

$$R_N(x) = \sum_{n=0}^{N-1} r_n (x-1)^n \text{ with } r_n = \frac{1}{n!} R_N^{(n)}(x)|_{x=1} \, .$$

However,

$$R_N(x) = P(x)\ [H(x)]^{-N} \ ,$$

so Leibniz' rule gives

$$R_N^{(n)}(x)|_{x=1} = \sum_{k=0}^{n} \binom{n}{k} \left[\left(\frac{d}{dx}\right)^k P(x)\right]_{x=1} \left[\left(\frac{d}{dx}\right)^{n-k} [H(x)]^{-N}\right]_{x=1} \, . \tag{17}$$

Here

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

is the usual binomial coefficient. Since $P - 1$ vanishes to order $2N$ at $x = 1$ by (16), the expression (17) simplifies to

$$R_N^{(n)}(x)|_{x=1} = \left[\left(\frac{d}{dx}\right)^n [H(x)]^{-N}\right]_{x=1} \, .$$

Thus

$$R_N(x) = \sum_{n=0}^{N-1} \left\{\frac{1}{n!}\left(\frac{d}{dx}\right)^n [H(x)]^{-N}\right\}_{x=1} (x-1)^n \, ,$$

with H given by (14). In other words, R_N is the first N terms in the Taylor expansion of H^{-N} about $x = 1$, and we can compute this! First rearrange the definition of H:

$$\begin{aligned} H(e^{i\omega}) &= \frac{1}{M^2}\left|1 + e^{i\omega} + \cdots + e^{i(M-1)\omega}\right|^2 \\ &= \frac{1}{M^2}\prod_{m=1}^{M-1}\left|1 - e^{i(\omega + 2\pi m/M)}\right|^2 . \end{aligned}$$

Now, for M even, with $M_1 = \frac{M}{2}$, this becomes

$$H(e^{i\omega}) = \frac{1}{M^2}\left|1+e^{i\omega}\right|^2 \prod_{m=1}^{M_1-1}\left|(1-e^{i(\omega+2\pi m/M)})(1-e^{i(\omega-2\pi m/M)})\right|^2$$

$$= \frac{1}{M^2}2(1+\cos\omega)\prod_{m=1}^{M_1-1} 4\left(\cos\omega - \cos\frac{2\pi m}{M}\right)^2, \quad \text{or}$$

$$H(x) = \frac{2^{M-1}}{M^2}(x+1)\prod_{m=1}^{M_1-1}\left(x-\cos\frac{2\pi m}{M}\right)^2;$$

while for M odd, with $M_1 = \frac{M-1}{2}$, we obtain

$$H(e^{i\omega}) = \frac{1}{M^2}\prod_{m=1}^{M_1}\left|(1-e^{i(\omega+2\pi m/M)})(1-e^{i(\omega-2\pi m/M)})\right|^2$$

$$= \frac{1}{M^2}\prod_{m=1}^{M_1} 4\left(\cos\omega - \cos\frac{2\pi m}{M}\right)^2, \quad \text{or}$$

$$H(x) = \frac{2^{M-1}}{M^2}\prod_{m=1}^{M_1}\left(x-\cos\frac{2\pi m}{M}\right)^2.$$

Let us consider the power series expansion of each of the factors of $H(x)$. The Taylor expansion of $\left(x-\cos\frac{2\pi m}{M}\right)^{-2N}$ about $x=1$ is

$$\sum_{n=0}^{\infty}\binom{2N+n-1}{2N-1}(-1)^n\left(1-\cos\frac{2\pi m}{M}\right)^{-2N-n}(x-1)^n$$

and the Taylor expansion of $(x+1)^{-N}$ about $x=1$ is

$$\sum_{n=0}^{\infty}\binom{N+n-1}{N-1}(-1)^n(1-\cos\pi)^{-N-n}(x-1)^n.$$

The Taylor expansion of H^{-N} about $x=1$ is a product of these expansions,

$$[H(x)]^{-N} = \sum_{n=0}^{\infty} r_n(1-x)^n.$$

For M even, we find:

$$r_n = \left(\frac{M^2}{2^{M-1}}\right)^N \sum_{k_1+k_2+\cdots+k_{M_1}=n}\left\{\prod_{m=1}^{M_1-1}\binom{2N+k_m-1}{2N-1}\left(1-\cos\frac{2\pi m}{M}\right)^{-2N-k_m}\right\}$$

$$\times\binom{N+k_{M_1}-1}{N-1}(1-\cos\pi)^{-N-k_{M_1}}$$

or

$$r_n = \sum_{k_1+k_2+\cdots+k_{M_1}=n}\left\{\prod_{m=1}^{M_1-1}\binom{2N+k_m-1}{2N-1}\left(1-\cos\frac{2\pi m}{M}\right)^{-k_m}\right\}$$

$$\times\binom{N+k_{M_1}-1}{N-1}(1-\cos\pi)^{-k_{M_1}}. \tag{18}$$

For M odd,

$$r_n = \left(\frac{M^2}{2^{M-1}}\right)^N \sum_{k_1+k_2+\cdots+k_{M_1}=n} \left\{ \prod_{m=1}^{M_1} \binom{2N+k_m-1}{2N-1} \left(1-\cos\frac{2\pi m}{M}\right)^{-2N-k_m} \right\}$$

or

$$r_n = \sum_{k_1+k_2+\cdots+k_{M_1}=n} \left\{ \prod_{m=1}^{M_1} \binom{2N+k_m-1}{2N-1} \left(1-\cos\frac{2\pi m}{M}\right)^{-k_m} \right\}. \tag{19}$$

Thus R_N is the finite trigonometric polynomial

$$R_N(e^{i\omega}) = \sum_{n=0}^{N-1} r_n(1-\cos\omega)^n \tag{20}$$

with r_n given by (18) for M even and (19) for M odd. Observe that since the r_n are visibly positive, and $\cos\omega < 1$ for all $\omega \neq 0$, R_N is a positive trigonometric polynomial. This will be important later.

LEMMA 3.1. *The solution $P = H^N R_N$, with R_N given by* (20) *and* (18) *or* (19), *which is guaranteed to have the desired flatness properties, satisfies the orthogonality condition* (15).

Proof. Define

$$\Phi(e^{i\omega}) = P(e^{i\omega}) + P(e^{(i\omega+2\pi/M)}) + \cdots + P(e^{i(\omega+2\pi\frac{M-1}{M})}) - 1\,;$$

then $\Phi + 1$ is the periodization of P to the interval $[0, 2\pi/M]$. Since Φ is real, even, and periodic with period $2\pi/M$, it must have the trigonometric polynomial expansion

$$\Phi(e^{i\omega}) = \sum_{k=0}^{N-1} c_k(e^{iMk\omega} + e^{-iMk\omega})\,. \tag{21}$$

By construction, Φ is flat of order N in $x = \cos\omega$ at $x = 1$, or

$$\Phi(x) \approx (x-1)^N \text{ for } x \approx 1\,.$$

Thus

$$\Phi(e^{i\omega}) \approx \omega^{2N} \text{ for } \omega \approx 0\,,$$

and

$$\left[\left(\frac{d}{d\omega}\right)^n \Phi(e^{i\omega})\right]_{\omega=0} = 0, \qquad n = 0, 1, \ldots, 2N-1. \tag{22}$$

However, from (21)

$$\left(\frac{d}{d\omega}\right)^n \Phi(e^{i\omega})|_{\omega=0} = \sum_k c_k\left[(ikM)^n + (-ikM)^n\right]$$

$$= \begin{cases} (-M^2)^{n/2} \sum_{k=0}^{N-1} c_k\, k^n & \text{if } n \text{ is even, } n \leq 2N-2\,, \\ 0 & \text{if } n \text{ is odd.} \end{cases}$$

Thus we have an $N \times N$ Vandermonde system for the c_k:

$$\begin{bmatrix} 1 & 1 & 1 & \dots & 1 & 1 & 1 \\ 0 & 1 & 4 & \dots & k^2 & \dots & (N-1)^2 \\ 0 & 1 & 16 & \dots & k^4 & \dots & (N-1)^4 \\ 0 & 1 & & & \dots & & \\ 0 & 1 & & & \dots & & \\ 0 & 1 & & & \dots & & \\ 0 & 1 & 2^{2N-2} & \dots & k^{2N-2} & \dots & (N-1)^{2N-2} \end{bmatrix} \begin{bmatrix} c_o \\ c_1 \\ c_2 \\ . \\ . \\ . \\ c_{N-1} \end{bmatrix} = 0 .$$

The invertibility of this matrix, combined with the flatness of Φ (22), guarantees that each of the $c_k = 0$, i.e.,

$$\Phi(e^{i\omega}) \equiv 0. \qquad \square$$

Notice that if P has the trigonometric polynomial expansion

$$P(e^{i\omega}) = \sum_{k=0}^{NM-1} p_k(e^{ik\omega} + e^{-ik\omega}) ,$$

then since $\Phi + 1$ is the periodization of P to the interval $[0, 2\pi/M]$, Φ has the trigonometric polynomial expansion

$$\Phi(e^{i\omega}) = -1 + \sum_{k=0}^{N-1} p_{Mk}(e^{iMk\omega} + e^{-iMk\omega}) .$$

Since $\Phi \equiv 0$,

$$p_{Mk} = 0 \text{ for } k \neq 0 ,$$

i.e., the Fourier coefficients of P vanish at every Mth index other than zero.

With Lemma 3.1, we have completed the proof of the following theorem.

THEOREM 3.2. *The minimal length solution for P, the modulus squared of the symbol of a rank M scaling sequence of degree N, is*

$$P(e^{i\omega}) = \left| \frac{1 + e^{i\omega} + \cdots + e^{i(M-1)\omega}}{M} \right|^{2N} R_N(e^{i\omega})$$

with R_N given by (20) *and* (18) *or* (19).

The symbol A_0 will be a spectral factor of P, of the form

$$A_0(e^{i\omega}) = \left(\frac{1 + e^{i\omega} + \cdots + e^{i(M-1)\omega}}{M} \right)^N Q_N(e^{i\omega}),$$

where the trigonometric polynomial Q_N is a spectral factor of R_N. As in Daubechies [1], we compute this via the method of Fejér–Riesz [9], finding

$$Q_N(e^{i\omega}) = \sum_{n=0}^{N-1} c_n e^{in\omega} \quad \text{such that}$$

$$R_N(e^{i\omega}) = \sum_{n=0}^{N-1} b_n \cos n\omega = Q_N(e^{i\omega})\overline{Q_N(e^{i\omega})}.$$

This factorization depends on the fact that $R_N(e^{i\omega}) \geq 0$ for $\omega \in [0, 2\pi]$, which we observed previously. When determining the spectral factor Q_N, one has a degree of choice over which roots of R_N to put into $Q_N(e^{i\omega})$ and which to put into $\overline{Q_N(e^{i\omega})}$; this can lead to minimum phase, midphase, and maximum phase scaling sequences, and complex sequences as well as real ones.

3.2. Arbitrary length solutions. The minimal length solution $P = H^N R_N$ we have found is a *particular solution* to the equation

$$P(e^{i\omega}) + P(e^{i(\omega+2\pi/M)}) + \cdots + P(e^{i(\omega+2\pi\frac{M-1}{M})}) \equiv 1 .$$

The first N coefficients r_n in the expansion of R_N are determined by the degree N constraint. We can generate *arbitrary* length solutions with degree N by augmenting R_N with a higher order cosine polynomial

$$\tilde{R}(x) = (x-1)^N \sum_{n=0}^{N_2} \tilde{r}_n (x-1)^n$$

that satisfies the homogeneous equation

$$(H^N\tilde{R})(e^{i\omega}) + (H^N\tilde{R})(e^{i(\omega+2\pi/M)}) + \cdots + (H^N\tilde{R})(e^{i(\omega+2\pi\frac{M-1}{M})}) \equiv 0 . \tag{23}$$

In other words, the M-fold periodization of $H^N(e^{i\omega})\tilde{R}(e^{i\omega})$ vanishes. If $H^N(e^{i\omega})\tilde{R}(e^{i\omega})$ has the Fourier expansion $\sum_k c_k e^{ik\omega}$ then its M-fold periodization has the expansion $\sum_k c_{Mk} e^{iMk\omega}$. The homogenenous equation (23) then holds if and only if $c_{Mk} = 0$ for all k. However,

$$(x-1)^N H^N(x) = |e^{iM\omega} - 1|^{2N}$$

is $2\pi/M$-periodic and therefore has a trigonometric polynomial expansion of pure Mth harmonics, $\sum_k h_{Mk} e^{iMk\omega}$. Thus $\tilde{R}(e^{i\omega})$ will solve (23) if and only if it has a trigonometric expansion

$$\tilde{R}(e^{i\omega}) = (\cos\omega - 1)^N \sum_n \tilde{r}_n \cos n\omega, \quad \text{with } \tilde{r}_n = 0 \quad \text{for } n = Mk .$$

Furthermore, in order to use the Fejér–Riesz algorithm on the result, we must have

$$R_N(e^{i\omega}) + \tilde{R}(e^{i\omega}) \geq 0 \text{ for } \omega \in [0, \pi] .$$

Summarizing this information, we have Theorem 3.3.

THEOREM 3.3. *The general solution $P(e^{i\omega})$ to*

$$P(e^{i\omega}) + P(e^{i(\omega+2\pi/M)}) + \cdots + P(e^{i(\omega+2\pi(M-1)/M)}) \equiv 1 ,$$

subject to the degree N constraint

$$P(e^{i\omega}) = 1 + \mathcal{O}(|\omega|^{2N}) \text{ at } \omega = 0$$

is given by

$$P(e^{i\omega}) = H^N(e^{i\omega}) \left(R_N(e^{i\omega}) + \tilde{R}(e^{i\omega}) \right) \tag{24}$$

with H as in (14), R_N *given by* (20) *and* (18) *or* (19), *and*

$$\tilde{R}(e^{i\omega}) = (\cos\omega - 1)^N \sum_{n \neq Mk} \tilde{r}_n \cos n\omega \,,$$

such that

$$R_N(e^{i\omega}) + \tilde{R}(e^{i\omega}) \geq 0 \text{ for } \omega \in [0, \pi] \,.$$

As in the minimal length case, we can spectrally factor $R = R_N + \tilde{R}$ to arrive at the general scaling sequence of degree N. This generalizes the $M = 2$ results of Daubechies [1] and Wells [16].

Figure 1 compares the graph of $|A|$ for an $M = 3$, $N = 2$ minimal length (6-coefficient) scaling sequence obtained from Theorem 3.2 with the graph of $|A|$ for a sequence obtained from the more general Theorem 3.3, with $M = 3$, $N = 2$, and

$$\tilde{R}(e^{i\omega}) = r_1 (\cos\omega - 1)^N \cos\omega \,,$$

i.e., having one additional parameter and resulting in an 8-coefficient sequence. In this example we have chosen the parameter r_1 to force a zero in the symbol at $\omega = \pi$. This is useful both for signal processing (since it reduces the filter sidelobe) and for creating a smoother scaling function [7]. Table 1 tabulates the two sequences.

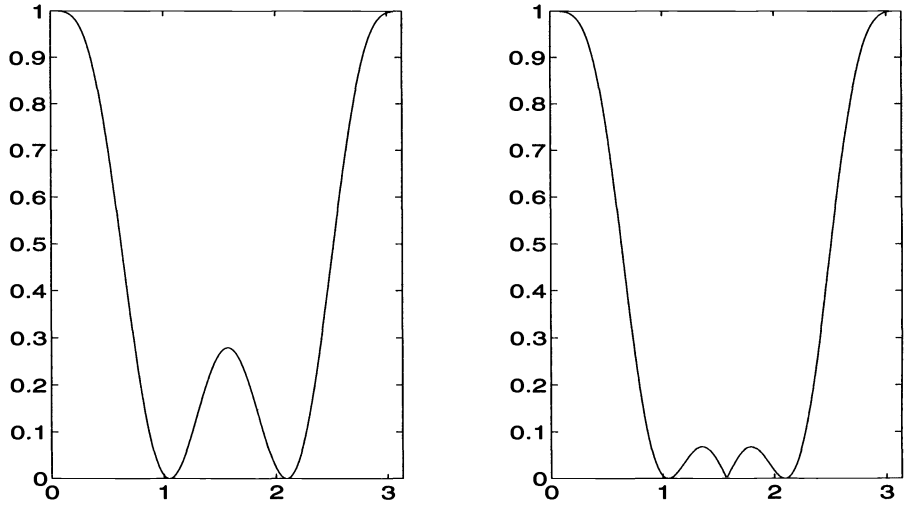


FIG. 1. *Moduli of the symbols for* 6*-coefficient* (*minimal length*) *and* 8*-coefficient* $M = 3$ *scaling sequences of degree* $N = 2$.

3.3. Examples. We now use the methods developed above to construct examples of rank M, degree N scaling sequences. In the $M = 3$ case, the square modulus of the symbol of the degree N minimal length scaling sequence is

$$P(e^{i\omega}) = \left(\frac{1 + 2\cos\omega}{3}\right)^{2N} \sum_{n=0}^{N-1} \binom{2N + k - 1}{2N - 1} \left(\frac{2}{3}\right)^n (1 - \cos\omega)^n \,.$$

TABLE 1
Minimal and nonminimal length scaling sequences for $M = 3$ and $N = 2$.

	k	$a_{0,k}$		k	$a_{0,k}$
Minimal	0	.58610191	8 coeffs	0	.40120209
Length	1	.91943525		1	.91425445
(6 coeffs)	2	1.2527686		2	1.1927766
	3	.41389809		3	.72369175
	4	.080564754		4	.030934374
	5	-.25276858		5	-.19277662
				6	-.12490085
				7	.054811171

When $M = 3$, $N = 2$ (a sequence that perfectly interpolates linear polynomials, the generalization of Daubechies 4-coefficient scaling sequence), we carry through the spectral factorization to find

$$Q_2(e^{i\omega}) = \frac{1}{2}\left\{1 \pm \frac{\sqrt{57}}{3} + \left(1 \mp \frac{\sqrt{57}}{3}\right) e^{i\omega}\right\}$$

and

$$\{a_{0,k}\} = \left\{\frac{3 \pm \sqrt{57}}{18}, \frac{9 \pm \sqrt{57}}{18}, \frac{15 \pm \sqrt{57}}{18}, \frac{15 \mp \sqrt{57}}{18}, \frac{9 \mp \sqrt{57}}{18}, \frac{3 \mp \sqrt{57}}{18}\right\}.$$

These numbers appear in Table 1. Table 2 below displays the $M = 3$ minimal length scaling sequences of degree N for $N = 3$, 4, and 5.

TABLE 2
Scaling sequences for $M = 3$ and $N = 3, 4, 5$.

	k	$a_{0,k}$		k	$a_{0,k}$		k	$a_{0,k}$
$N = 3$	0	0.35184039	$N = 4$	0	0.21374716	$N = 5$	0	0.13078303
	1	0.73291789		1	0.55061140		1	0.39986248
	2	1.2251065		2	1.0761524		2	0.88862061
	3	0.77288503		3	0.97241848		3	1.0265830
	4	0.34406337		4	0.63896711		4	0.87196905
	5	-0.30698050		5	-0.14940338		5	0.12599947
	6	-0.12472542		6	-0.22132111		6	-0.20022120
	7	-0.076981260		7	-0.23167774		7	-0.37766755
	8	0.081874011		8	0.10177363		8	0.011991962
				9	0.035155465		9	0.050460123
				10	0.0420992319		10	0.12621021
				11	-0.028522632		11	-0.036899302
							12	-0.0076049461
							13	-0.020374182
							14	0.010287257

When $M = 4$, the symbol P of the autocorrelation of the minimal length scaling sequence of degree N is

$$P(e^{i\omega}) = \left(\frac{\cos^3\omega + \cos^2\omega}{2}\right)^N \times \ldots$$
$$\sum_{n=0}^{N-1}\sum_{k=0}^{n}\binom{2N+k-1}{2N-1}\binom{N+n-k-1}{N-1}2^{k-n}(1-\cos\omega)^n,$$

Again, we carry through the spectral factorization in the $N = 2$ case to find the $M = 4$ generalization of Daubechies 4-coefficient scaling sequence:

$$\{a_{0,k}\} = \left\{ \frac{1\pm\sqrt{11}}{8}, \frac{3\pm\sqrt{11}}{8}, \frac{5\pm\sqrt{11}}{8}, \frac{7\pm\sqrt{11}}{8}, \frac{7\mp\sqrt{11}}{8}, \frac{5\mp\sqrt{11}}{8}, \frac{3\mp\sqrt{11}}{8}, \frac{1\mp\sqrt{11}}{8} \right\} . \tag{25}$$

Finally, we compute R_N for arbitrary M and the first three values of N:

$$\begin{aligned} R_1(e^{i\omega}) &= 1 , \\ R_2(e^{i\omega}) &= \frac{M^2+2}{3} + \frac{(1-M^2)\cos\omega}{3} , \\ R_3(e^{i\omega}) &= \frac{4M^4+5M^2+11}{20} - \frac{8M^4+5M^2-13}{30}\cos\omega \\ &\quad + \frac{4M^4-5M^2+1}{60}\cos 2\omega. \end{aligned}$$

4. Construction of the full wavelet matrix. Having constructed rank M scaling sequences of degree N and arbitrary length, we now turn to the construction of the corresponding wavelet sequences. In the rank M case, there is considerable freedom in this construction of a full wavelet matrix given its first row. In [5] we solved the following problem: given a Haar wavelet matrix $\mathbf{H}_0$ and a scaling sequence a_0, construct a full wavelet matrix $\mathbf{A}$ whose first row is a_0 and whose characteristic Haar matrix is $\mathbf{H}_0$. Here we clarify and refine that explicit construction using Vaidyanathan's paraunitary factorization technique. In particular, this section shows that a parametrization of the choice of the $M-1$ wavelets is given by the choice of the characteristic Haar matrix, which is equivalent to the choice of an $(M-1)\times(M-1)$ orthogonal matrix (or unitary matrix in the case of complex wavelets).

Working in the z-transform domain, Vaidyanathan [13] has proven that *every* paraunitary polyphase matrix $\mathbf{H}(z)$ of McMillan degree[4] K can be factored into the form

$$\mathbf{H}(z) = \left(\prod_{k=0}^{K-1} (\mathbf{I} - \mathbf{v}_k\mathbf{v}_k^\dagger + z\mathbf{v}_k\mathbf{v}_k^\dagger) \right) \mathbf{H}_0, \tag{26}$$

where each $\mathbf{v}_k$ is a unit M-vector. $\mathbf{H}(z)$ will be the polyphase matrix of a wavelet matrix if and only if $\mathbf{H}_0$ is a Haar wavelet matrix, so (26) provides a factorization of all wavelet matrices with polyphase matrix of McMillan degree K. We refer to the term

$$\mathbf{I} - \mathbf{v}_k\mathbf{v}_k^\dagger + z\mathbf{v}_k\mathbf{v}_k^\dagger$$

as a *prime factor* of the polyphase matrix.

THEOREM 4.1. *Given a scaling sequence a_0 of overlap g and a characteristic Haar matrix $\mathbf{H}_0$, there exists a unique wavelet matrix of McMillan degree $g-1$ (i.e.,*

[4] A polyphase matrix of McMillan degree K will correspond to a wavelet matrix of overlap $K+1$, while a wavelet matrix of overlap $K+1$ has a polyphase matrix with McMillan degree at least K. However, there exist wavelet matrices of overlap $K+1$ and McMillan degree strictly greater than K; for examples see [6]. The construction presented here describes a unique wavelet matrix with first row a_0 and characteristic Haar $\mathbf{H}_0$ *and* having a polyphase matrix of McMillan degree K.

with $g-1$ prime factors) whose first row is a_0 and with characteristic Haar $\mathbf{H}_0$. Furthermore, we can explicitly construct the prime factors $\mathbf{I}-\mathbf{v}_k\mathbf{v}_k^\dagger+z\mathbf{v}_k\mathbf{v}_k^\dagger$ (and thus the wavelet matrix) from a_0 and $\mathbf{H}_0$.

Proof. We wish to obtain vectors $\mathbf{v}_k$ such that the relationship (26) holds with $K=g-1$. Recall the $M\times M$ submatrices $\mathbf{A}_k$ of $\mathbf{A}$ defined in (4). We know the first row of each of these matrices to be the kth length-M subvector of the scaling sequence a_0, but the remaining $M-1$ rows of each $\mathbf{A}_k$ are undetermined. Right-multiplying by $\mathbf{H}_0^{-1}$, we seek

$$\mathbf{B}_0^0+z\mathbf{B}_1^0+\cdots+z^{g-1}\mathbf{B}_{g-1}^0=\prod_{k=0}^{g-2}(\mathbf{I}-\mathbf{v}_k\mathbf{v}_k^\dagger+z\mathbf{v}_k\mathbf{v}_k^\dagger)\;; \tag{27}$$

again the first row of each $\mathbf{B}_k^0$ is known but the remaining $M-1$ rows of each submatrix are undetermined. We write β_k^j for the first row of the $M\times M$ matrix $\mathbf{B}_k^j$. If we write α_k for the length-M subvectors of the given scaling sequence a_0, then the fact that a_0 is a scaling sequence can be written as

$$\sum_{k=0}^{g-1-l}\alpha_{k+l}\alpha_k^\dagger=M\delta_{0,l} \tag{28}$$

and

$$\sum_{k=0}^{g-1}\alpha_k=(1,\,1,\,\ldots,\,1)\;. \tag{29}$$

Right-multiplication of the α_k by $\mathbf{H}_0^{-1}$ to get the β_k^0 renormalizes (28) and simply rotates the vector of ones in (29):

$$\sum_{k=0}^{g-1-l}\beta_{k+l}^0\beta_k^{0\dagger}=\frac{1}{M}\delta_{0,l} \tag{30}$$

and

$$\sum_{k=0}^{g-1}\beta_k^0=(1,\,0,\,0,\,\ldots,\,0)\;. \tag{31}$$

This will be useful shortly.

Now compute the product on the right-hand side of (27) to find that the coefficient of z^{g-1} is

$$\mathbf{v}_0\mathbf{v}_0^\dagger\mathbf{v}_1\mathbf{v}_1^\dagger\ldots\mathbf{v}_{g-2}\mathbf{v}_{g-2}^\dagger\;,$$

a rank 1 matrix, each of whose rows is proportional to $\mathbf{v}_{g-2}^\dagger$. Since we have specified the first row β_{g-1}^0 of $\mathbf{B}_{g-1}^0$, equating coefficients of z^{g-1} in (27) requires that

$$\mathbf{v}_{g-2}^\dagger=\frac{\beta_{g-1}^0}{\|\,\beta_{g-1}^0\,\|}$$

and $\mathbf{B}_{g-1}^0$ must have rank 1; each of its rows must be a multiple of the first. In fact $\mathbf{v}_{g-2}^\dagger$ is only determined up to a complex number of modulus 1; however, this phase

factor does not matter because we only care about $\mathbf{v}_{g-2}^\dagger$ insofar as it determines the prime factor

$$\mathbf{I} - \mathbf{v}_{g-2}\mathbf{v}_{g-2}^\dagger + z\mathbf{v}_{g-2}\mathbf{v}_{g-2}^\dagger \,.$$

Right-multiply (27) by

$$\mathbf{I} - \mathbf{v}_{g-2}\mathbf{v}_{g-2}^\dagger + z^{-1}\mathbf{v}_{g-2}\mathbf{v}_{g-2}^\dagger \,,$$

the inverse of the newly determined prime factor, to obtain

$$\mathbf{B}_0^1 + z\mathbf{B}_1^1 + \cdots + z^{g-2}\mathbf{B}_{g-2}^1 = \prod_{k=0}^{g-3}(\mathbf{I} - \mathbf{v}_k\mathbf{v}_k^\dagger + z\mathbf{v}_k\mathbf{v}_k^\dagger) \,.$$

Again, since we are given each of the first rows β_k^0, we know each of the new first rows β_k^1, and right-multiplication by a paraunitary prime factor preserves (30) and (31):

$$\sum_{k=0}^{g-2-l} \beta_{k+l}^1 \beta_k^{1\dagger} = \frac{1}{M}\delta_{0,l}$$

and

$$\sum_{k=0}^{g-2} \beta_k^1 = (1, 0, 0, \ldots, 0) \,.$$

We iterate this procedure to determine $\mathbf{v}_{g-3}, \ldots, \mathbf{v}_1$, arriving at the point where we wish to establish

$$\mathbf{B}_0^{g-2} + z\mathbf{B}_1^{g-2} = \mathbf{I} - \mathbf{v}_0\mathbf{v}_0^\dagger + z\mathbf{v}_0\mathbf{v}_0^\dagger \tag{32}$$

given the knowledge of the first rows β_0^{g-2} and β_1^{g-2}. We know that

$$\beta_0^{g-2} + \beta_1^{g-2} = (1, 0, 0, \ldots, 0), \quad \text{and} \quad (\beta_0^{g-2})^\dagger \beta_1^{g-2} = 0.$$

Set

$$\mathbf{v}_0^\dagger = \frac{\beta_1^{g-2}}{\| \beta_1^{g-2} \|};$$

and (32) will be satisfied; one can verify that

$$\beta_1^{g-2} = v_{0,0}\mathbf{v}_0^\dagger \,, \quad \text{where } \mathbf{v}_0^\dagger = (v_{0,0}, v_{0,1}, \ldots, v_{0,M-1}) \,.$$

This fully determines the matrices $\mathbf{B}_0^{g-2}$ and $\mathbf{B}_1^{g-2}$.

We now form the product

$$\left(\prod_{k=0}^{g-2}(\mathbf{I} - \mathbf{v}_k\mathbf{v}_k^\dagger + z\mathbf{v}_k\mathbf{v}_k^\dagger)\right)\mathbf{H}_0$$

and by construction it will produce a polyphase matrix

$$\mathbf{H}(z) = \mathbf{A}_0 + z\mathbf{A}_1 + \cdots + z^{g-1}\mathbf{A}_{g-1},$$

whose corresponding wavelet matrix has characteristic Haar matrix $\mathbf{H}_0$ and first row a_0. □

As a closing example, we use the methods of §§3 and 4 to construct a minimal length wavelet matrix with $M = 4$ and $N = g = 2$. The minimal length scaling sequence for this case was given in (26). The full wavelet matrix with this sequence for its first row and the rank-4 DCT for its characteristic Haar matrix $\mathbf{H}_0$ is

$$\begin{pmatrix} 0.5396 & 0.7896 & 1.0396 & 1.2896 & 0.4604 & 0.2104 & -0.0396 & -0.2896 \\ -0.1962 & -0.1456 & -0.4120 & -0.3614 & 1.5028 & 0.6868 & -0.1292 & -0.9451 \\ 1.0 & -1.0 & -1.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.4344 & -1.3554 & 1.3157 & -0.4740 & 0.1068 & 0.0488 & -0.0092 & -0.0672 \end{pmatrix}.$$

Plots of the symbols of the scaling sequence and the three wavelet sequences (i.e., the four rows of the matrix) appear in Fig. 2.

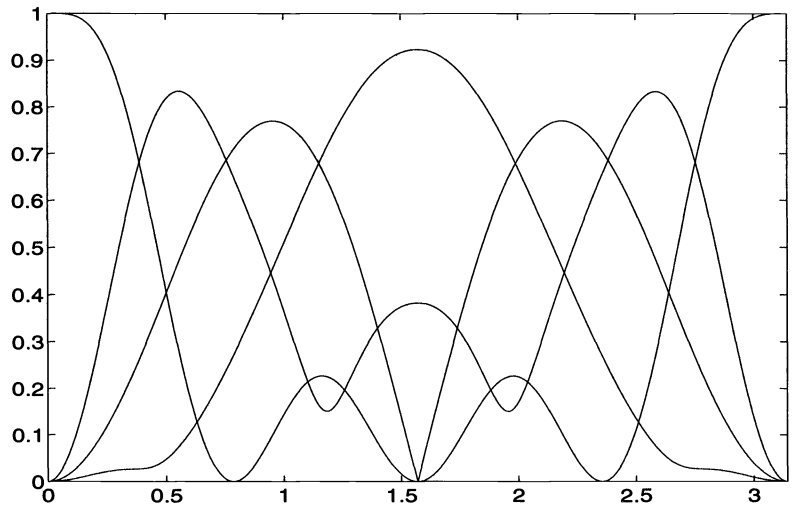


FIG. 2. *Moduli of the symbols for minimal length* $M = 4$, $N = 2$ *wavelet matrix based on a* DCT *characteristic Haar matrix.*

5. Conclusions. In this paper we have generalized the discrete wavelets of Daubechies to the rank M case. By describing the "N vanishing moments" property in terms of maximal flatness, we have been able to obtain an explicit formula for the square modulus of the symbol of a rank M scaling sequence with N vanishing wavelet moments, and subsequently the scaling sequences themselves. This yields a collection of discrete bases distinguished by their polynomial interpolation properties. We have explicitly constructed a full wavelet matrix given a scaling sequence and the desired characteristic Haar matrix. In a separate work [7] we explore the implications of these discrete constructions for differentiability of the associated wavelet scaling functions, i.e., solutions of the scaling equation (5).

Acknowledgments. The author would like to thank his colleagues at Aware, Inc., particularly J. Weiss, H. Resnikoff, and R. Tolimieri, as well as Professors G. Strang and R. O. Wells, Jr. for useful conversations.

REFERENCES

[1] I. DAUBECHIES, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[2] ———, *Ten Lectures on Wavelets*, CBMS Conference Series 61, Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[3] R. A. GOPINATH AND C. S. BURRUS, *Wavelet transforms and filter banks*, in Wavelets: A Tutorial in Theory and Applications, C. K. Chui, ed., Academic Press, San Diego, 1992.

[4] P. N. HELLER, *Higher rank Daubechies wavelets—preliminary report*, Aware Tech. Report AD911204, Cambridge, MA, 1991.

[5] P. N. HELLER, H. L. RESNIKOFF, AND R. O. WELLS, JR., *Wavelet matrices and the representation of discrete functions*, in Wavelets: A Tutorial in Theory and Applications, C. K. Chui, ed., Academic Press, San Diego, 1992.

[6] P. N. HELLER AND R. TOLIMIERI, *A general construction of $M \times 2M$ perfect reconstruction filter banks*, Aware Tech. Report AD921202, Cambridge, MA, 1992.

[7] P. N. HELLER AND R. O. WELLS, JR., *Sobolev regularity for rank M wavelets*, manuscript.

[8] D. POLLEN, *Linear one-dimensional scaling functions*, Aware Tech. Report AD900101, Cambridge, MA, 1990.

[9] G. POLYA AND G. SZEGÖ, *Problems and Theorems of Analysis*, Springer-Verlag, New York, 1972.

[10] A. K. SOMAN AND P. P. VAIDYANATHAN, *On orthonormal wavelets and paraunitary filter banks*, IEEE Trans. Signal Processing, 41 (1993), pp. 1170–1183.

[11] G. STRANG, *Wavelets and dilation equations: a brief introduction*, SIAM Rev., 31 (1989), pp. 614–627.

[12] P. P. VAIDYANATHAN, *Quadrature mirror filter banks, M-band extensions and perfect-reconstruction techniques*, IEEE ASSP Magazine, 4 (1987), pp. 4–20.

[13] P. P. VAIDYANATHAN, *Multirate Systems and Filter Banks*, Prentice Hall, Englewood Cliffs, NJ, 1992.

[14] M. VETTERLI AND D. LEGALL, *Perfect reconstruction* FIR *filter banks*, IEEE Trans. ASSP, 37 (1989), pp. 1057–1071.

[15] G. WELLAND AND M. LUNDBERG, *Construction of compact p-wavelets*, Constructive Approximation, 9 (1993), pp. 347–370.

[16] R. O. WELLS, JR., *Parametrizing smooth compactly supported wavelets*, Trans. Amer. Math. Soc., 338 (1993), pp. 919–931.

[17] H. ZOU AND A. H. TEWFIK, *Discrete orthogonal M-band wavelet decompositions*, in Proc. IEEE ICASSP-San Francisco, March 1992.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 520–535, April 1995

014

ANALYSIS OF A QR ALGORITHM FOR COMPUTING SINGULAR VALUES*

S. CHANDRASEKARAN† AND I.C.F. IPSEN‡

Abstract. We extend the Golub–Kahan algorithm for computing the singular value decomposition of bidiagonal matrices to triangular matrices R. Our algorithm avoids the explicit formation of R^TR or RR^T.

We derive a relation between left and right singular vectors of triangular matrices and use it to prove monotonic convergence of singular values and singular vectors. The convergence rate for singular values equals the square of the convergence rate for singular vectors. The convergence behaviour explains the occurrence of deflation in the interior of the matrix.

We analyse the relationship between our algorithm and rank-revealing QR and URV decompositions. As a consequence, we obtain an algorithm for computing the URV decomposition, as well as a divide-and-conquer algorithm that computes singular values of dense matrices and may be beneficial on a parallel architecture. Our perturbation result for the smallest singular values of a triangular matrix is stronger than the traditional results because it guarantees high *relative* accuracy in the smallest singular values after an off-diagonal block of the matrix has been set to zero.

Key words. singular value decomposition, eigenvalue decomposition, QR decomposition, rank revealing QR decomposition, URV decomposition, deflation

AMS subject classifications. 15A18, 15A23, 15A42, 65F15, 65F25, 65W05

1. Introduction.

We present an algorithm for computing the singular value decomposition (SVD) of a real upper triangular matrix R that is based on the repeated QR decomposition of R.

1.1. The algorithm.

In 1965 Golub and Kahan [21] introduced an algorithm for the computation of the singular values and vectors of a real upper bidiagonal matrix B. The algorithm is based on the QR algorithm for computing eigenvalues but avoids the explicit formation of the tridiagonal matrix B^TB. An Algol implementation of this algorithm was proposed by Golub and Reinsch in 1970 [22].

The following extension of the unshifted Golub–Kahan algorithm from bidiagonal matrices to triangular matrices was proposed in [28]. It determines a new iterate from the QR decomposition of the transpose of the old iterate,

$$(*) \qquad R^{(0)} = R, \qquad [R^{(i)}]^T = Q^{(i+1)}R^{(i+1)}, \qquad i \geq 0,$$

and so avoids the explicit formation of R^TR or RR^T. Since the iterates $R^{(i)}$ are related to each other by orthogonal equivalence transformations, they all have the same singular values. We show in §3 that this algorithm computes the singular values of R.

The repeated transformation from lower to upper triangular form by means of orthogonal transformations was motivated by an algorithm for computing partial correlation coefficients [11], [12]. F. Chatelin and A. Ruhe pointed out to us that (*) had already been proposed by Fadeev, Kublanovskaya, and Fadeeva in 1966 [17],

* Received by the editors August 31, 1992; accepted for publication (in revised form) by F. T. Luk, February 25, 1994. The work presented in this paper was supported by National Science Foundation grant CCR-9102853.

† Department of Electrical and Computer Engineering, University of California, Santa Barbara, California 93106-9560 (shir@ece.ucsb.edu).

‡ Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205 (ipsen@math.ncsu.edu).

where it is formulated as applying an LQ iteration to $R^{(2i)}$ and a QR iteration to $R^{(2i+1)}$.

Fernando and Parlett [19] derive a version of Rutishauser's differential QD algorithm based on (*) that computes singular values of bidiagonal matrices to high relative accuracy. Mathias and Stewart [31] use (*) to update rank-revealing URV and ULV decompositions, while Dowling, Ammann, and DeGroat [16] use it to develop a systolic real-time algorithm for computing the SVD. Moonen, Van Dooren, and Vanpouke [32] insert permutations to turn (*) into a Jacobi-type algorithm. In a second paper, Fernando and Parlett [18] incorporate shifts into (*) to compute singular values and vectors and to derive lower bounds on the smallest singular value.

This paper concentrates on the unshifted algorithm. At this point we do not advocate (*) as a practical method for computing singular values of dense matrices. Our motivation is to obtain insight into the behaviour of the Golub–Kahan algorithm for bidiagonal matrices [21] and into the unshifted QR algorithm [23], [33], [38], [41] for computing eigenvalues of symmetric matrices.

1.2. Overview. In §2 we derive a relation between left and right singular vectors of triangular matrices. It provides the basis for a simple analysis in §3 of the monotonic convergence of (*). In particular, we show that the tangent of the angle between certain canonical spaces and the singular vector subspaces of the iterates $R^{(i)}$ decreases monotonically at the usual rate; and that the convergence rate of the singular values is equal to the square of that of the singular vectors. These results explain the occurrence of deflation in the interior of the matrix.

Our analysis helps to understand the relation between algorithms that produce a complete SVD and those that produce a partial SVD, such as rank-revaling QR (RRQR) decompositions [8] and URV decompositions [24], [30], [36]. In §4 we show that with respect to a particular block partitioning of the matrix R, (*) proceeds in two phases: a rank-revealing phase where the large singular values are separated from the small ones, and a monotonic phase, where the iterates converge monotonically to block-diagonal form. Hence, preceding (*) with a rank-revealing algorithm accomplishes two things: it reverses the grading of inappropriately graded matrices and so enhances subsequent convergence; and, it forces premature deflation of a particular off-diagonal block and thus amounts to the computation of a URV decomposition. Based on this observation, we sketch a divide-and-conquer algorithm for computing singular values of dense matrices, which may be advantageous on a parallel architecture.

Section 5 derives a simple perturbation result for the smallest singular values of a triangular matrix. It is stronger than the traditional results because it guarantees high *relative* accuracy in the smallest singular values after an off-diagonal block of the matrix has been set to zero.

Some of the material in §§2 and 5 has appeared in preliminary form in [6], [7].

1.3. Relation to other algorithms. Two successive iterations of (*) are mathematically equivalent to one iteration of the unshifted QR algorithm for computing eigenvalues [23], [33], [38], [41] applied to both $R^{(i)}[R^{(i)}]^T$ and $[R^{(i)}]^T R^{(i)}$, as

$$R^{(i+2)}[R^{(i+2)}]^T = [Q^{(i+2)}]^T \left(R^{(i)}[R^{(i)}]^T \right) Q^{(i+2)}$$

and

$$[R^{(i+2)}]^T R^{(i+2)} = [Q^{(i+1)}]^T \left([R^{(i)}]^T R^{(i)} \right) Q^{(i+1)}.$$

This is also observed in [31]. If $R^{(0)}$ is upper bidiagonal, so are all iterates $R^{(i)}$, and two successive iterations amount to applying one iteration of the Golub–Kahan algorithm [21].

Fadeev, Kublanoskaya, and Fadeeva [17] and Fernando and Parlett [19] observe that one iteration of (*) is mathematically equivalent to one iteration of the Cholesky LR algorithm [41] applied to $A^{(i)} \equiv R^{(i)}[R^{(i)}]^T$. This is because $A^{(i)}$ has the upper-lower Cholesky factorisation $A^{(i)} = [R^{(i+1)}]^T R^{(i+1)}$. A subsequent multiplication of the Cholesky factors in reverse order gives the next iterate $A^{(i+1)} \equiv R^{(i+1)}[R^{(i+1)}]^T$.

From

$$A^{(i)} = R^{(i)}[R^{(i)}]^T = [R^{(i+1)}]^T R^{(i+1)}$$

it follows that $R^{(i)}$ is the factor from the upper-lower Cholesky factorisation of $A^{(i)}$, while $R^{(i+1)}$ is the factor from its lower-upper Cholesky factorisation. Hence the two factors are related through the orthogonal transformation $Q^{(i+1)}$. The fact that the two Cholesky factors of a matrix are related by an orthogonal transformation is used in [11], [12] to compute partial correlation coefficients. It is a consequence of the more general result that $M = M_1^T M_1 = M_2^T M_2$ for a positive-definite matrix M implies the existence of an orthogonal matrix W with $M_2 = W M_1$, cf. the exercise beneath [27, Coro. 7.2.8] and [19, §3].

Notation. The norm $\|\cdot\|$ represents the Euclidean two-norm. The identity matrix of order k is denoted by I_k and its ith column by e_i.

2. SVD of triangular matrices. To understand why (*) makes progress in every iteration we establish a relation between left and right singular vectors of triangular matrices. Let $R = U\Sigma V^T$ be the SVD of the upper triangular matrix

$$R = \begin{array}{c} \\ k \\ n-k \end{array} \begin{array}{c} \begin{array}{cc} k & n-k \end{array} \\ \begin{pmatrix} R_{11} & R_{12} \\ & R_{22} \end{pmatrix} \end{array},$$

where

$$U = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix}, \qquad V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}$$

are orthogonal matrices, and

$$\Sigma = \begin{pmatrix} \Sigma_1 & \\ & \Sigma_2 \end{pmatrix}, \qquad \text{with} \qquad \Sigma_1 = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} \sigma_{k+1} & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix},$$

is a diagonal matrix whose diagonal contains the singular values of R in descending order,

$$\sigma_1 \geq \cdots \geq \sigma_k \geq \sigma_{k+1} \geq \cdots \geq \sigma_n.$$

The following theorem implies that if the singular values are well separated, then the left singular vectors are almost always closer to canonical form than the right singular vectors. By "canonical form" we mean a matrix $\binom{Z}{0}$ with Z orthogonal. The columns of such a canonical form span what we casually call the "canonical space $\binom{I_k}{0}$," that is, the column space of $\binom{I_k}{0}$.

THEOREM 2.1. *If R_{11} is nonsingular, and if U_{11} or V_{11} is also nonsingular, then*

$$\tan\theta_{u,k} \le \frac{\sigma_{k+1}}{\sigma_k} \tan\theta_{v,k},$$

where $\theta_{u,k}$ is the largest principal angle between $\binom{U_{11}}{U_{21}}$ and $\binom{I_k}{0}$; and $\theta_{v,k}$ is the largest principal angle between $\binom{V_{11}}{V_{21}}$ and $\binom{I_k}{0}$.

Proof. From the SVD $U^TR = \Sigma V^T$ one gets $U_{11}^TR_{11} = \Sigma_1 V_{11}^T$. Hence the nonsingularity of R_{11} implies that U_{11} is nonsingular whenever V_{11} is nonsingular. According to the CS decomposition [23, Thm. 2.6.1] for orthogonal matrices, V_{22} and U_{22} must also be nonsingular. Furthermore, the $(2,1)$ block in $R = U\Sigma V^T$ yields

$$U_{21}\Sigma_1 V_{11}^T + U_{22}\Sigma_2 V_{12}^T = 0$$

and

$$\|U_{22}^{-1}U_{21}\| \le \frac{\sigma_{k+1}}{\sigma_k} \|V_{11}^{-1}V_{12}\|.$$

Again, from the CS decomposition,

$$\sin\theta_{u,k} = \|U_{12}\| = \|U_{21}\| = \sqrt{1 - \frac{1}{\|U_{11}^{-1}\|^2}} = \sqrt{1 - \sigma_{\min}^2(U_{11}^T I_k)}.$$

Since the square root term represents the distance between the column space of $\binom{U_{11}}{U_{21}}$ and the canonical space $\binom{I_k}{0}$ [23, Coro. 2.6.2], the angle $\theta_{u,k}$ must be the largest principal angle [23, §12.4.3] between these two spaces. Moreover,

$$\|U_{11}^{-1}U_{12}\| = \|U_{22}^{-1}U_{21}\| = \tan\theta_{u,k}.$$

Substituting this in the above inequality gives

$$\tan\theta_{u,k} \le \frac{\sigma_{k+1}}{\sigma_k} \tan\theta_{v,k},$$

where $\theta_{v,k}$ is the analogous angle for V. □

COROLLARY 2.2. *If R is nonsingular, and if U_{11} or V_{11} is also nonsingular, then*

$$\frac{\sigma_k}{\sigma_{k+1}} \tan\theta_{u,k} \le \tan\theta_{v,k} \le \frac{\sigma_1}{\sigma_n} \tan\theta_{u,k}.$$

3. Monotonic convergence results. We determine some of the quantities that undergo monotonic change during one iteration of (*). Partition the iterates as in §2

$$R^{(i)} = \begin{array}{c} \\ k \\ n-k \end{array} \begin{array}{c} \begin{array}{cc} k & n-k \end{array} \\ \begin{pmatrix} R_{11}^{(i)} & R_{12}^{(i)} \\ & R_{22}^{(i)} \end{pmatrix} \end{array},$$

and denote their SVDs by $R^{(i)} = U^{(i)}\Sigma {V^{(i)}}^T$, where

$$U^{(i)} = \begin{pmatrix} U_{11}^{(i)} & U_{12}^{(i)} \\ U_{21}^{(i)} & U_{22}^{(i)} \end{pmatrix}, \qquad V^{(i)} = \begin{pmatrix} V_{11}^{(i)} & V_{12}^{(i)} \\ V_{21}^{(i)} & V_{22}^{(i)} \end{pmatrix}$$

are orthogonal.

For simplicity we assume that the initial matrix R is nonsingular. If this is not the case then the zero singular values can be extracted in two iterations. First perform a QR decomposition with column pivoting [5], [20], [23] that moves the zeros to the bottom of the matrix:

$$R = Q_P \begin{pmatrix} \hat{R}_{11} & \hat{R}_{12} \\ 0 & 0 \end{pmatrix} P^T,$$

where Q_P is orthogonal, P is a permutation matrix, and $\hat{R}_{11}$ is nonsingular upper triangular. In the next iteration eliminate the off-diagonal block,

$$\begin{pmatrix} \hat{R}_{11}^T & 0 \\ \hat{R}_{12}^T & 0 \end{pmatrix} = Q \begin{pmatrix} R_{11} & 0 \\ 0 & 0 \end{pmatrix}.$$

Our algorithm (*) can now be applied to the nonsingular triangular matrix R_{11}.

The convergence properties of the unshifted QR algorithm are well known [23], [33], [34], [38], [40], [41]. They are usually derived from the fact that one iteration of the QR algorithm is mathematically equivalent to one nested subspace iteration, applied to particular starting spaces, cf. in particular [34], [38], [40]. The subspace iterates converge linearly to eigenspaces with an asymptotic convergence rate equal to a ratio of adjacent singular values (in fact, the distance between the iterates and the eigenspace decreases from the start [40]). In [39] these results are extended to the computation of the SVD of R from R^TR and RR^T. The monotonic convergence of the eigenvalues during nested subspace iteration is proved through the connection to Toda flows [29].

3.1. Convergence of the singular vectors. We show that the angles between the invariant subspaces and the canonical spaces almost always decrease monotonically during (*). First we prove that the convergence rate for the leading k columns of the singular vector matrices depends on the gap between the kth and $(k+1)$st singular values.

THEOREM 3.1. *Let $R^{(0)}$ be nonsingular; $U_{11}^{(0)}$ or $V_{11}^{(0)}$ be nonsingular; and $\sigma_k > \sigma_{k+1}$.*

Then $U_{11}^{(i)}$ and $V_{11}^{(i)}$ are nonsingular for all $i \geq 0$; and convergence is monotonic in the sense that

$$\tan\theta_{v,k}^{(i+1)} \leq \frac{\sigma_{k+1}}{\sigma_k} \tan\theta_{v,k}^{(i)}, \qquad \tan\theta_{u,k}^{(i+1)} \leq \frac{\sigma_{k+1}}{\sigma_k} \tan\theta_{u,k}^{(i)},$$

where $\theta_{v,k}^{(i)}$ is the largest principal angle between the canonical space $\binom{I_k}{0}$ and the space spanned by the leading k columns of $V^{(i)}$; and $\theta_{u,k}^{(i)}$ is the analogous angle for $U^{(i)}$.

Proof. Consider one iteration $R^T = Q\hat{R}$. From the SVD $U^TR = \Sigma V^T$ follows $U_{11}^T R_{11} = \Sigma_1 V_{11}^T$. The nonsingularity of $R^{(0)}$ implies that U_{11} is nonsingular whenever V_{11} is. Applying Theorem 2.1 to R gives

$$\tan\theta_{u,k} \leq \frac{\sigma_{k+1}}{\sigma_k} \tan\theta_{v,k}.$$

Let $\hat{R} = \hat{U}\Sigma\hat{V}^T$ be the SVD of $\hat{R}$, where $\hat{U} = Q^TV$ and $\hat{V} = U$. The analogous relation for $\hat{R}$ is $\hat{U}_{11}^T\hat{R}_{11} = \Sigma_1\hat{V}_{11}^T$. Another application of Theorem 2.1, this time to $\hat{R}$, yields

$$\tan\theta_{\hat{u},k} \leq \frac{\sigma_{k+1}}{\sigma_k} \tan\theta_{\hat{v},k}.$$

Putting the two inequalities together via $\hat{V} = U$ results in

$$\tan\theta_{\hat{u},k} \leq \frac{\sigma_{k+1}}{\sigma_k} \tan\theta_{u,k}. \qquad \square$$

Now we prove that the rate of convergence for interior columns of the singular vector matrices depends on the gaps with the adjacent distinct singular values. This result holds if the initial singular vector matrices are strongly nonsingular.[1] The fact that the strong nonsingularity is preserved throughout the iteration (*) follows already from the convergence results of the eigenvalue QR algorithms [34], [38]–[41].

THEOREM 3.2. *Let $R^{(0)}$ be nonsingular, $U^{(0)}$ or $V^{(0)}$ be strongly nonsingular, and*

$$\sigma_k > \sigma_{k+1} = \cdots = \sigma_{k+m} > \sigma_{k+m+1}.$$

Then $U^{(i)}$ and $V^{(i)}$ are strongly nonsingular, for all $i \geq 0$, and columns $k+1, \ldots, k+m$ of $U^{(i)}$ and $V^{(i)}$ converge to a $n \times m$ matrix of the form

$$\begin{array}{c} \\ k \\ m \end{array} \begin{array}{c} \begin{array}{c} m \end{array} \\ \begin{pmatrix} 0 \\ Z \\ 0 \end{pmatrix} \end{array},$$

where Z is orthogonal, at the rate

$$\rho_k = \max\left\{\frac{\sigma_{k+1}}{\sigma_k}, \frac{\sigma_{k+m+1}}{\sigma_{k+1}}\right\}.$$

Proof. The strong nonsingularity of $U^{(i)}$ and $V^{(i)}$ can be proved as in Theorem 3.1. According to the convergence results for distinct singular values in Theorem 3.1, the singular vector matrices converge at the rate σ_{k+1}/σ_k to the canonical form

$$\begin{array}{c} \begin{array}{cc} k & \end{array} \\ \begin{pmatrix} X & X & & & & \\ X & X & & & & \\ & & X & X & X & X \\ & & X & X & X & X \\ & & X & X & X & X \\ & & X & X & X & X \end{pmatrix} \end{array} \begin{array}{c} k \\ \\ \\ \\ \\ \\ \end{array},$$

while they converge at the rate $\sigma_{k+m+1}/\sigma_{k+m}$ to the canonical form

$$\begin{array}{c} \begin{array}{cccc} k & + & m & \end{array} \\ \begin{pmatrix} X & X & X & X & & \\ X & X & X & X & & \\ X & X & X & X & & \\ X & X & X & X & & \\ & & & & X & X \\ & & & & X & X \end{pmatrix} \end{array} \begin{array}{c} k \\ + \\ m \\ \\ \\ \\ \end{array}.$$

[1] A square matrix is called "strongly nonsingular" if all its leading principal submatrices are nonsingular.

Here X represents a matrix element that may be nonzero. Thus the singular vector matrices converge to the form

$$\begin{array}{c} \\ k \\ \\ m \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cc} k & \quad m \end{array} \\ \begin{pmatrix} X & X & & & & \\ X & X & & & & \\ & & X & X & & \\ & & X & X & & \\ & & & & X & X \\ & & & & X & X \end{pmatrix} \end{array}$$

at the rate $\max\{\sigma_{k+1}/\sigma_k,\ \sigma_{k+m+1}/\sigma_{k+m}\}$. □

Therefore, if all singular values of the matrix $R^{(0)}$ are distinct and if the singular vector matrices are strongly nonsingular, then the singular vector matrices converge to the identity matrix monotonically at the rate $\max_k \sigma_{k+1}/\sigma_k$. In general, the singular vector matrices converge to a block-diagonal matrix whose diagonal blocks are orthogonal. The convergence rate is equal to the largest ratio of adjacent distinct singular values. The size of the kth diagonal block equals the multiplicity of the kth distinct singular value, and the columns making up the block represent an orthogonal basis for the associated invariant subspace.

3.2. Convergence of the singular values. From the convergence rate of the singular vector matrices we can in turn estimate the convergence rate for the singular values. First we show that the singular values converge monotonically. The inequalities in the lemma below are also derived in [31, Thm. 2.1]. They are special cases of the monotonicity properties of eigenvalues during subspace iteration [29].

LEMMA 3.3. *If $R^{(0)}$ is nonsingular then*

$$\|R_{11}^{(i+1)^{-1}}\| \le \|R_{11}^{(i)^{-1}}\|, \qquad \|R_{22}^{(i+1)}\| \le \|R_{22}^{(i)}\|.$$

Proof. From one iteration $R^T = Q\hat{R}$ follows that $R_{11}^T = Q_{11}\hat{R}_{11}$ and $\hat{R}_{22} = Q_{22}^T R_{22}^T$. Hence

$$\|\hat{R}_{11}^{-1}\| \le \|R_{11}^{-1}\|, \qquad \|\hat{R}_{22}\| \le \|R_{22}\|. \qquad \square$$

Now we derive the rate of convergence of the extreme singular values of the leading and trailing principal submatrices.

THEOREM 3.4. *Let $R^{(0)}$ be nonsingular, $U_{11}^{(0)}$ or $V_{11}^{(0)}$ be nonsingular, and $\sigma_k > \sigma_{k+1}$.*

Then convergence of the singular values is monotonic in the sense that

$$\frac{\|R_{22}^{(i)}\| - \sigma_{k+1}}{\sigma_{k+1}} \le \frac{\sigma_1}{\sigma_k} \tan^2 \theta_{v,k}^{(i)}, \qquad \frac{\|R_{11}^{(i)^{-1}}\| - \frac{1}{\sigma_k}}{\frac{1}{\sigma_k}} \le \frac{\sigma_{k+1}}{\sigma_n} \tan^2 \theta_{v,k}^{(i)}.$$

Proof. Consider one iteration $R^T = Q\hat{R}$. The SVD $R = U\Sigma V^T$ gives

$$R_{22} = U_{21}\Sigma_1 V_{21}^T + U_{22}\Sigma_2 V_{22}^T = U_{22}(\Sigma_2 + U_{22}^{-1}U_{21}\Sigma_1 V_{21}V_{22}^{-T})V_{22}^T.$$

Following the proof of Theorem 2.1,

$$\|R_{22}\| \le \sigma_{k+1} + \sigma_1 \tan\theta_{u,k} \tan\theta_{v,k}$$

and substituting

$$\tan\theta_{u,k} \leq \frac{\sigma_{k+1}}{\sigma_k} \tan\theta_{v,k}$$

for $\tan\theta_{u,k}$ gives

$$\frac{\|R_{22}\| - \sigma_{k+1}}{\sigma_{k+1}} \leq \frac{\sigma_1}{\sigma_k} \tan^2\theta_{v,k}.$$

The second inequality is derived analogously from $R^{-1} = V\Sigma^{-1}U^T$. □

Theorem 3.4 implies that the relative distance of $\|R_{11}^{-1}\|$ from $1/\sigma_k$ is bounded above by the condition number of Σ_2, as well as the square of the angle between the leading k columns of the right singular vector matrix and the corresponding canonical space. Similarly, the relative distance of $\|R_{22}\|$ from σ_{k+1} is bounded above by the condition number of Σ_1 and the square of the same angle. Hence if V_{11} is well conditioned and the spread of singular values in Σ_1 is small then $\|R_{22}\|$ is close to σ_{k+1}.

Furthermore, the rate of convergence of the singular values is approximately the square of that of the associated singular vectors. B. Parlett pointed out that this is a result of Rayleigh's principle.

COROLLARY 3.5. *The following convergence estimates hold:*

$$\frac{\|\hat{R}_{22}\| - \sigma_{k+1}}{\|R_{22}\| - \sigma_{k+1}} \approx \left(\frac{\tan\theta_{\hat{v},k}}{\tan\theta_{v,k}}\right)^2 \leq \left(\frac{\sigma_{k+1}}{\sigma_k}\right)^2$$

and

$$\frac{\|\hat{R}_{11}^{-1}\| - \frac{1}{\sigma_k}}{\|R_{11}^{-1}\| - \frac{1}{\sigma_k}} \approx \left(\frac{\tan\theta_{\hat{v},k}}{\tan\theta_{v,k}}\right)^2 \leq \left(\frac{\sigma_{k+1}}{\sigma_k}\right)^2.$$

Now we estimate the convergence of an interior principal submatrix. The theorem below implies that the iterates $R^{(i)}$ converge to a diagonal matrix with the singular values in sorted order along the diagonal.

THEOREM 3.6. *Let $R^{(0)}$ be nonsingular, $V^{(0)}$ or $U^{(0)}$ be strongly nonsingular, and*

$$\sigma_k > \sigma_{k+1} = \cdots = \sigma_{k+m} > \sigma_{k+m+1}.$$

Then the principal submatrix of order m of $R^{(i)}$,

$$\begin{pmatrix} R^{(i)}_{k+1,k+1} & \cdots & R^{(i)}_{k,k+m} \\ & \ddots & \vdots \\ & & R^{(i)}_{k+m,k+m} \end{pmatrix},$$

converges to $\sigma_{k+1}I_m$ at approximately the rate ρ_k^2, where

$$\rho_k = \max\left\{\frac{\sigma_{k+1}}{\sigma_k}, \frac{\sigma_{k+m+1}}{\sigma_{k+1}}\right\}.$$

Proof. Partition the iterates as in the proof of Theorem 3.2,

$$R^{(i)} = \begin{pmatrix} R^{(i)}_{11} & X & X \\ & R^{(i)}_{22} & X \\ & & R^{(i)}_{33} \end{pmatrix},$$

where $R_{11}^{(i)}$ is of order k and $R_{22}^{(i)}$ is of order m.

Corollary 3.5 implies that the convergence of $\|{R_{11}^{(i)}}^{-1}\|$ to $1/\sigma_k$ and the convergence of

$$\left\| \begin{pmatrix} R_{22}^{(i)} & X \\ & R_{33}^{(i)} \end{pmatrix} \right\|$$

to σ_{k+1} occur at approximately the rate $\sigma_{k+1}^2/\sigma_k^2$, while the convergence of

$$\left\| \begin{pmatrix} R_{11}^{(i)} & X \\ & R_{22}^{(i)} \end{pmatrix}^{-1} \right\|$$

to $1/\sigma_{k+1}$ and the convergence of $\|R_{33}^{(i)}\|$ to σ_{k+m+1} occurs at approximately the rate $\sigma_{k+1}^2/\sigma_{k+m+1}^2$.

Consider the essential limit of the iterates, which we define as $R^{(\infty)} = U^{(\infty)}\Sigma V^{(\infty)}$, and partition their singular vector matrices like $R^{(i)}$:

$$U^{(\infty)} = \begin{pmatrix} U_{11}^{(\infty)} & & \\ & U_{22}^{(\infty)} & \\ & & U_{33}^{(\infty)} \end{pmatrix}, \qquad V^{(\infty)} = \begin{pmatrix} V_{11}^{(\infty)} & & \\ & V_{22}^{(\infty)} & \\ & & V_{33}^{(\infty)} \end{pmatrix},$$

where the diagonal blocks $U_{ii}^{(\infty)}$ and $V_{ii}^{(\infty)}$ are orthogonal and

$$\Sigma = \begin{pmatrix} \Sigma_1 & & \\ & \sigma_{k+1} I_m & \\ & & \Sigma_3 \end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{pmatrix}, \quad \Sigma_3 = \begin{pmatrix} \sigma_{k+m+1} & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix}.$$

Then $R_{22}^{(\infty)} V_{22}^{(\infty)} = \sigma_{k+1} U_{22}^{(\infty)}$, so that $R_{22}^{(\infty)} = \sigma_{k+1} U_{22}^{(\infty)} {V_{22}^{(\infty)}}^T$ is a multiple of an orthogonal matrix. But $R_{22}^{(\infty)}$ is also upper triangular. Therefore $R_{22}^{(\infty)} = \sigma_{k+1} I_m$ is a scalar matrix, and $U_{22}^{(\infty)} = V_{22}^{(\infty)}$ (where we have assumed that $R^{(\infty)}$ has positive diagonal elements). Hence a principal submatrix $R_{22}^{(i)}$ associated with a singular value σ_{k+1} of multiplicity m converges to $\sigma_{k+1} I_m$ at approximately the rate $\max\{\sigma_{k+1}^2/\sigma_k^2, \sigma_{k+m+1}^2/\sigma_{k+1}^2\}$. □

3.3. Consequences. Our upper bounds on the relative distance between $\|{R_{11}^{(i)}}^{-1}\|$ and $\|R_{22}^{(i)}\|$ to the respective singular values depend on the spreads $\sigma_1 \ldots \sigma_k$ and $\sigma_{k+1} \ldots \sigma_n$, and the conditioning of the leading principal submatrices of order k of $V^{(i)}$. The number of iterations required to reduce the relative distance between $\|R_{22}\|$ and σ_{k+1} to ϵ can thus be estimated as

$$\frac{\log \sigma_1/\sigma_k - \log \epsilon + \log \tan \theta_{v,k}^{(0)}}{\log \sigma_k/\sigma_{k+1}}.$$

An analogous estimate can be made for $\|R_{11}^{-1}\|$.

According to [34, §2.2], the QR algorithm tends to converge to the small eigenvalues first. According to our analysis, though, there is no preference of (*) for small singular values over larger ones. However, such a preference may be enforced by a suitable choice of shifts [33], [41].

The bounds on the relative distance also explain why (*) and the QR algorithm have such a hard time with graded matrices whose elements increase in size towards the bottom, cf. [14, §5] and [19, Thm. 5]. These matrices have a large spread in the spectrum and very ill-conditioned leading principal submatrices. One of the simplest examples of a graded matrix is

$$R^{(0)} = \begin{pmatrix} 1 & \epsilon \\ & \alpha \end{pmatrix},$$

where $\epsilon \ll 1 \ll \alpha$. One iteration of (*) gives

$$R^{(1)} = \frac{1}{\sqrt{1+\epsilon^2}} \begin{pmatrix} 1+\epsilon^2 & \alpha\epsilon \\ & \alpha \end{pmatrix},$$

whose off-diagonal element has increased from ϵ to about $\alpha\epsilon$. But the diagonal elements have only changed marginally, and it is obvious that many iterations are needed to arrive at a diagonal matrix with diagonal elements in descending order (a similar example was used in [41, §8.7] to illustrate slow convergence of the LR algorithm). Section 4.2 illustrates how to force fast convergence on such graded matrices without the need to decide between QR- and QL-type algorithms as in [14], [19].

The leading principal submatrices of the singular vector matrices are almost always nonsingular [38, p. 430] but may be very ill conditioned, in which case the convergence is slow.

4. RRQR and URV decompositions. We discuss the connections between (*) on the one hand and RRQR decompositions [8] and URV decompositions [36] on the other hand.

4.1. Two phases in the algorithm. For each partitioning index k of the nonsingular matrix $R^{(0)}$ define

$$\gamma_k^{(i)} \equiv \|{R_{11}^{(i)}}^{-1}\| \, \|R_{22}^{(i)}\|.$$

When $\gamma_k^{(i)} < 1$ then $\|R_{22}^{(i)}\| < 1/\|{R_{11}^{(i)}}^{-1}\|$, which means that all singular values of $R_{11}^{(i)}$ are larger than the singular values of $R_{22}^{(i)}$, and a partial ordering of the singular values of $R^{(i)}$ has occurred. Lemma 3.3 implies that

$$\gamma_k^{(i+1)} \leq \gamma_k^{(i)},$$

so the separation between singular values of $R_{11}^{(i)}$ and $R_{22}^{(i)}$ never decreases throughout the iterations (*). Furthermore, if $\sigma_{k+1}/\sigma_k < 1$ then $\gamma_k^{(i)} \to \sigma_{k+1}/\sigma_k$ as $i \to \infty$, provided $U^{(0)}$ and $V^{(0)}$ are strongly nonsingular. Because the convergence of $\gamma_k^{(i)}$ to $\sigma_{k+1}/\sigma_k < 1$ is monotone, there exists a number i_k such that $\gamma_k^{(i)} < 1$ for all $i \geq i_k$. It makes sense therefore to distinguish, for each k, two phases of (*) depending on the value of $\gamma_k^{(i)}$.

1. A rank-revealing phase, where $\gamma_k^{(i)} > 1$, during which the singular values of $R_{11}^{(i)}$ and $R_{22}^{(i)}$ are in the process of separating.

2. A monotonic phase, where $\gamma_k^{(i)} \leq 1$, during which *all* quantities of interest converge monotonically.

4.2. The rank-revealing phase. The name for the first phase comes from its resemblance to RRQR decompositions. Given a matrix R and a specific k (usually determined by the number of singular values of R that are smaller than a certain threshold), RRQR algorithms try to find a permutation matrix P so that the triangular matrix $\bar{R}$ in the QR decomposition $RP = Q\bar{R}$ has a $(1,1)$ block with maximal smallest singular value, and/or a $(2,2)$ block with minimal largest singular value [8]. The existence of RRQR decompositions was proved in [26], and one of the most accurate RRQR algorithms is Hybrid III(k) [8], which finds a permutation matrix P so that $RP = Q\bar{R}$ with

$$\bar{\gamma}_k \equiv \|\bar{R}_{11}^{-1}\| \, \|\bar{R}_{22}\| \leq (k+1)(n-k+1) \, \frac{\sigma_{k+1}}{\sigma_k}.$$

In practice, though, the cheaper and possibly less accurate forms of column pivoting, such as QR with column pivoting [5], [20], [23], tend to work quite well (an attempt at explaining the practical effectiveness of the simple column pivoting strategies, regardless of their potential failures, is made in [8]).

Therefore, if the singular values σ_k and σ_{k+1} are well separated then one can try to enforce the onset of the monotonic phase for a particular k by preceding $(*)$ with an RRQR decomposition. This also reverses the grading in a matrix all of whose large elements are at the bottom, thus obviating the need for a decision between an algorithm of QR or of QL type [14], [19].

The idea of permuting rows or columns of the iterates during eigenvalue computations is not new. Pivoting, in the form of row exchanges, has been suggested for the LR algorithm, [41, §8.13] and [34, §2.7], to enhance numerical stability in those cases where the orthodox LR algorithm fails to converge. A preliminary pivoting step has also been suggested for Jacobi methods: Hari and Veselič use QR with column pivoting [25] and Cholesky decomposition with symmetric pivoting [37], while Demmel and Veselič [15, Algorithm 4.4] propose to compute the eigendecomposition of a symmetric positive-definite matrix A by first determining the Cholesky factor R of A with complete pivoting, followed by the application of a one-sided Jacobi method to R.

4.3. The monotonic phase. Since $\|R_{12}^{(i+1)}\| \leq \gamma_k^{(i)} \|R_{12}^{(i)}\|$, this implies for the monotonic phase $\|R_{12}^{(i+1)}\| < \|R_{12}^{(i)}\|$. Hence the off-diagonal blocks $R_{12}^{(i)}$ decrease monotonically; and convergence to block-diagonal form is fast once the monotonic phase has been reached. Since $\gamma_k^{(i)} \to \sigma_{k+1}/\sigma_k$, the blocks corresponding to well-separated singular values may decrease faster and deflation[2] is likely to set in earlier.

4.4. A divide and conquer algorithm. The previous sections showed that once the rank-revealing phase has been completed for some k, the iterates converge rapidly to block diagonal form. Hence preceding $(*)$ with an RRQR algorithm tends to force completion of the rank-revealing phase and the start of deflation for that k. This observation leads to a divide and conquer algorithm for computing singular values of dense or banded matrices A, which may be advantageous on a parallel architecture. Below is a rough sketch.

1. Select a k and apply an RRQR algorithm to $AP = Q\bar{R}$ so that $\bar{\gamma}_k < 1$.
2. Set $R^{(0)} = \bar{R}$ and iterate $(*)$ until $\|R_{12}^{(i)}\|$ is small enough.
3. Apply Steps 1 and 2 recursively to $R_{11}^{(i)}$ and to $R_{22}^{(i)}$.

[2] The splitting of a matrix into two or more independent diagonal blocks due to almost zero off-diagonal blocks is called "deflation."

There are several ways to determine the index k in Step 1 where the matrix is to be split. The simplest option is to set $k = n/2$ and choose Hybrid III($n/2$) as the RRQR algorithm to break the matrix into equally sized blocks and ensure load balance with regard to parallel execution. But the separation of the singular values $\sigma_{n/2}$ and $\sigma_{n/2+1}$ may not be large enough. Alternatively one can apply QR with column pivoting and select as k that index for which $|\bar{r}_{k+1,k+1}|/|\bar{r}_{kk}| \approx \bar{\gamma}_k$ is smallest. A third possibility is to estimate the norm of $\|\bar{R}_{11}^{-1}\|$ by an incremental condition estimator [1]–[3]. We have not yet gathered enough computational experience to judge whether the algorithm presents a viable alternative to other methods that operate on dense matrices, such as Jacobi methods [4], [9], for instance.

4.5. Computation of the URV decomposition. We show how to compute a URV decomposition by means of (*). The URV decomposition was introduced by Hanson and Lawson, [24] and [30, Thm. (3.19)], to solve (rank deficient) least squares problems. Stewart [36] emphasizes its use for computing the null space of a matrix that is repeatedly updated. If R has rank $k < n$ then there exist orthogonal matrices U and V and a nonsingular upper triangular matrix $\bar{R}$ of order k such that

$$R = U \begin{pmatrix} \bar{R} & 0 \\ 0 & 0 \end{pmatrix} V^T.$$

In practice, R is often only of numerical rank k, where the singular values $\sigma_{k+1}, \ldots, \sigma_n$ are small. In this case one would like to find a decomposition

$$R = U \begin{pmatrix} \bar{R}_{11} & \bar{R}_{12} \\ & \bar{R}_{22} \end{pmatrix} V^T,$$

where $\|\bar{R}_{11}^{-1}\| \approx 1/\sigma_k$ is large and where $\|\bar{R}_{22}\| \approx \sigma_{k+1}$ and $\|\bar{R}_{12}\| \approx \sigma_{k+1}$ are small.

Hanson and Lawson, [24] and [30, §14], as well as Stewart and Mathias [31], [35], [36], compute a URV decomposition by determining orthogonal matrices P and Q such that $RP = Q\bar{R}$ where $\|\bar{R}_{11}^{-1}\| \approx 1/\sigma_k$ and $\|\begin{pmatrix} \bar{R}_{12}^T & \bar{R}_{22}^T \end{pmatrix}\| \approx \sigma_{k+1}$. Stewart and Mathias [31], [35] then perform several of the following "refinement steps" on $R^{(0)} = \bar{R}$ to further decrease the size of the $(1,2)$ block: first determine an orthogonal matrix $Q^{(1)}$ so that ${R^{(1)}}^T = {R^{(0)}}{Q^{(1)}}$ is lower triangular and, second, determine an orthogonal matrix $Q^{(2)}$ so that $R^{(2)} = {Q^{(2)}}^T {R^{(1)}}^T$ is upper triangular. In [36] Stewart proposes an incomplete version of these refinement steps: reduce only the last column of $R^{(0)}$ to e_n, and in this resulting matrix in turn reduce only the last row to e_n^T.

Note that in the beginning these algorithms accomplish more than an RRQR decomposition. Due to the rotations performed on both sides of the matrix the off-diagonal block also ends up being small. Hence the following result from [35] applies. If

$$\|R_{12}^{(0)}\| + \|R_{22}^{(0)}\| < \sigma_k$$

then the first part of the refinement steps in [35], [36] causes a monotonic decrease $\|R_{12}^{(1)}\| < \|R_{12}^{(0)}\|$ in the (1,2) block, and so does, of course, the second part of the refinement step. The refinement step in [35] represents two iterations of (*)

$$[R^{(0)}]^T = Q^{(1)} R^{(1)}, \qquad [R^{(1)}]^T = Q^{(2)} R^{(2)},$$

while the refinement step in [36] amounts to one incomplete iteration of (*) where $R_{11}^{(1)}$ of order $n-1$ remains lower triangular.

Section 4.3 showed that generally no assumption on the (1,2) block is necessary to ensure monotonic decrease provided the singular values of $R_{11}^{(i)}$ and $R_{22}^{(i)}$ are well separated: if $\gamma_k^{(i)} < 1$ then $\|R_{12}^{(i+1)}\| < \|R_{12}^{(i)}\|$. This is true regardless of whether σ_{k+1} is small or not. However, if $\|R_{22}^{(i)}\|$ is small then $\|R_{12}^{(i+1)}\|$ is as small—regardless of the relation between $R_{11}^{(i)}$ and $R_{22}^{(i)}$—because $R_{12}^{(i+1)} = Q_{21}^{(i+1)} {R_{22}^{(i)}}^T$, so $\|R_{12}^{(i+1)}\| \leq \|R_{22}^{(i)}\|$.

Therefore, one can compute a URV decomposition of R by determining an RRQR decomposition $RP = Q\bar{R}$ and then applying several iterations of $(*)$ to $\bar{R}$, which then converges monotonically to the desired URV decomposition.

5. Deflation criteria. We extend some of the existing convergence and deflation critera for computing singular values of bidiagonal matrices to triangular matrices.

Demmel and Kahan [14] and Deift et al. [10] have shown that, in floating point arithmetic, a particular implementation of the Golub–Kahan algorithm for bidiagonal matrices computes small singular values to high relative accuracy. This implementation is based on deflation and convergence criteria that preserve high relative accuracy of the computed singular values.

Fernando and Parlett [19] introduce a modification of Rutishauser's differential QD algorithm for bidiagonal matrices that is faster than the current implementations of the Golub–Kahan algorithm. Their deflation criterion for shifted matrices continues to preserve high relative accuracy for the singular values. Demmel and Gragg [13] extend the criterion from [10] to biacyclic matrices.

In [35] Stewart proves a deflation criterion that bounds the relative accuracy of the smallest singular value and can be considered an extension of Criterion 2a in [14] to triangular matrices: If the off-diagonal block is small enough with regard to the singular value separation,

$$\|R_{12}\| < \sigma_k - \|R_{22}\|,$$

then

$$\frac{|\sigma_{k+i} - \sigma_i(R_{22})|}{\sigma_1(R_{22})} \leq \frac{\|R_{12}\|^2}{\delta^2 - \|R_{22}\|^2}, \qquad \delta = \sigma_k - \|R_{12}\|.$$

In [31] Mathias and Stewart prove a deflation criterion for the eigenvalues of RR^T that bounds the relative accuracy of the $n-k$ smallest eigenvalues,

$$\frac{\sigma_i^2(R_{22}) - \sigma_{k+i}^2}{\sigma_i^2(R_{22})} \leq \frac{\|R_{12}\|^2}{\mathrm{gap}_k(\sigma_{\min}(R_{11}) + \|R_{22}\|)}, \qquad \mathrm{gap}_k = 6\mathrm{min}(R_{11}) - \|R_{22}\|$$

(a similar theorem in [31] also bounds the relative accuracy of the largest k eigenvalues of R^TR). It implies a first-order bound on the relative accuracy of the $n-k$ smallest singular values of R,

$$\frac{\sigma_i(R_{22}) - \sigma_{k+i}}{\sigma_i(R_{22})} \leq \frac{\|R_{12}\|^2}{2\mathrm{gap}_k(\sigma_{\min}(R_{11}) + \|R_{22}\|)} + O\left(\frac{\|R_{12}\|^4}{\mathrm{gap}_k^2}\right).$$

If $\|R_{12}\|$ is small then this criterion permits earlier deflation than the one from [35].

Our deflation criterion below guarantees high relative accuracy in $\sigma_1(R_{22})$ and holds without any assumptions on the size of $\|R_{12}\|$.

THEOREM 5.1. *If R is nonsingular, V is strongly nonsingular and* $\mathrm{gap}_k = \sigma_{\min}(R_{11}) - \|R_{22}\| > 0$ *then*

$$\frac{|\sigma_{k+j} - \sigma_j(R_{22})|}{\sigma_1(R_{22})} \leq \frac{\|R_{12}\|}{\mathrm{gap}_k}.$$

Proof. From $R^T = Q\hat{R}$ it follows that

$$\begin{pmatrix} 0 \\ R_{22}^T \end{pmatrix} = Q \begin{pmatrix} 0 \\ \hat{R}_{22} \end{pmatrix} + Q \begin{pmatrix} \hat{R}_{12} \\ 0 \end{pmatrix}.$$

Using $|\sigma_j(A+E) - \sigma_j(A)| \leq \|E\|$ [23, Cor. 8.3.2] with

$$A = Q \begin{pmatrix} 0 \\ \hat{R}_{22} \end{pmatrix}, \qquad A + E = Q \begin{pmatrix} 0 \\ \hat{R}_{22} \end{pmatrix} + Q \begin{pmatrix} \hat{R}_{12} \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ R_{22}^T \end{pmatrix}$$

yields

$$|\sigma_j(R_{22}) - \sigma_j(\hat{R}_{22})| \leq \|\hat{R}_{12}\|.$$

This implies, together with

$$\|\hat{R}_{12}\| \leq \gamma_k \|R_{12}\|, \qquad \gamma_k \equiv \|R_{11}^{-1}\| \, \|R_{22}\|$$

from §4.3, that

$$|\sigma_j(R_{22}) - \sigma_j(\hat{R}_{22})| \leq \gamma_k \|R_{12}\|.$$

Now set $R^{(0)} \equiv R$ and apply $(*)$ to $R^{(0)}$. According to §4.1 $\gamma_k^{(i+1)} \leq \gamma_k^{(i)}$, so the difference between two successive iterations is

$$|\sigma_j(R_{22}^{(i+1)}) - \sigma_j(R_{22}^{(i)})| \leq \|R_{12}^{(i+1)}\| \leq \gamma_k^{(i)} \|R_{12}^{(i)}\| \leq \gamma_k^i \|R_{12}\|.$$

As for the difference between iteration $i+2$ and i, we make use of Stewart's idea [36],

$$\begin{aligned} |\sigma_j(R_{22}^{(i+2)}) - \sigma_j(R_{22}^{(i)})| &\leq |\sigma_j(R_{22}^{(i+2)}) - \sigma_j(R_{22}^{(i+1)})| + |\sigma_j(R_{22}^{(i+1)}) - \sigma_j(R_{22}^{(i)})| \\ &\leq \left(\gamma_k^{i+1} + \gamma_k^i\right) \|R_{12}\|. \end{aligned}$$

Because V is strongly nonsingular, Theorem 3.6 implies that the singular values of $R_{22}^{(i)}$ converge to the singular values $\sigma_{k+1}, \ldots, \sigma_n$ as $i \to \infty$. The assumption $\gamma_k < 1$ allows extrapolation to the limit

$$|\sigma_{k+j} - \sigma_j(R_{22})| \leq \|R_{12}\| \sum_{l=1}^{\infty} \gamma_k^l = \|R_{12}\| \, \frac{\gamma_k}{1-\gamma_k}$$

as $\sum_{l=1}^{\infty} \gamma_k^l = \frac{1}{1-\gamma_k} - 1$. Hence

$$\frac{|\sigma_{k+j} - \sigma_j(R_{22})|}{\sigma_1(R_{22})} \leq \frac{\|R_{11}^{-1}\|}{1-\gamma_k} \|R_{12}\| = \frac{\|R_{12}\|}{\sigma_{\min}(R_{11}) - \|R_{22}\|}. \qquad \square$$

Theorem 5.1 is most valuable for the case $k = n-m$, where m is the multiplicity of the smallest singular value σ_n, because it assures that $\|R_{22}\|$ approximates σ_n to high relative accuracy whenever the norm of the off-diagonal block is small and the singular values of R_{11} are much larger than those of R_{22}. Since the requirement $\text{gap}_k > 0$ is equivalent to $\gamma_k < 1$, Theorem 5.1 can be applied as a deflation criterion for $(*)$ only once the monotonic phase for k has set in. Note that $\text{gap}_k > 0$ is not satisfied for a graded matrix whose elements increase in size towards the bottom, regardless of how small R_{12} is.

In general, Theorem 5.1 suggests using the simple deflation criterion

$$\|R_{12}\| \leq \eta \frac{\text{gap}_k}{\|R_{22}\|}$$

to guarantee absolute accuracy η for all singular values of R_{22}. If $\|R_{22}\|$ is small and if the singular values of R_{11} and R_{22} are well separated then this criterion permits earlier deflation than the traditional criterion [23, Coro. 8.3.2]

$$\|R_{12}\| \leq \eta.$$

Relative accuracy η for *all* singular values is achieved if

$$\|R_{12}\| \leq \eta \frac{\text{gap}_k}{\kappa(R_{22})},$$

where $\kappa(R_{22}) = \|R_{22}\| \, \|R_{22}^{-1}\|$ is the condition number of R_{22}.

Acknowledgments. We thank Stan Eisenstat, Beresford Parlett, and David Watkins for helpful discussions, as well as Françoise Chatelin, Axel Ruhe, and a referee for references to the literature.

REFERENCES

[1] J. Barlow and U. Vemulapati, *Rank detection methods for sparse matrices*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1279–1297.

[2] C. Bischof, *Incremental condition estimation*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 312–322.

[3] C. Bischof, J. Lewis, and D. Pierce, *Incremental condition estimation for sparse matrices*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 644–659.

[4] R. Brent and F. Luk, *The solution of singular-value and symmetric eigenvalue problems on multiprocessor arrays*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 69–84.

[5] P. Businger and G. Golub, *Linear least squares solutions by Householder transformations*, Numer. Math., 7 (1965), pp. 269–276.

[6] S. Chandrasekaran and I. Ipsen, *On the singular value decomposition of triangular matrices*, in Proceedings of the 1992 Shanghai International Numerical Algebra and its Applications Conference.

[7] ———, *A divide and conquer algorithm for computing singular values*, Z. Angew. Math. Mech., 74 (1994), pp. T 532–534.

[8] ———, *On rank-revealing QR factorisations*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 592–622.

[9] J. Charlier, M. Vanbegin, and P. Van Dooren, *On efficient implementations of Kogbetliantz's algorithm for computing the singular value decomposition*, Numer. Math., 52 (1988), pp. 279–300.

[10] P. Deift, J. Demmel, L. Li, and C. Tomei, *The bidiagonal singular value decomposition and Hamiltonian mechanics*, SIAM J. Numer. Anal., 28 (1991), pp. 1463–1516.

[11] J. Delosme and I. Ipsen, *Computing partial correlations from the data matrix*, Research Report 541, Department of Computer Science, Yale University, New Haven, CT, 1987.

[12] ———, *From Bareiss' algorithm to the stable computation of partial correlations*, J. Comput. Appl. Math., 27 (1989), pp. 53–91; also *Parallel Algorithms for Numerical Linear Algebra* (*Advances in Parallel Computing*, 1), H. van der Vorst and P. Van Dooren, eds., North Holland, Amsterdam, 1990.

[13] J. Demmel and W. Gragg, *On computing accurate singular values and eigenvalues of matrices with acyclic graphs*, Linear Algebra Appl., 185 (1993), pp. 203–217.

[14] J. Demmel and W. Kahan, *Accurate singular values of bidiagonal matrices*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 873–912.

[15] J. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.

[16] E. Dowling, L. Ammann, and R. DeGroat, *A TQR-iteration based adaptive SVD for real time angle and frequency tracking*, IEEE Trans. Signal Processing, 42 (1994), pp. 914–926.

[17] D. FADEEV, V. KUBLANOVSKAYA, AND V. FADEEVA, *Sur les systèmes linéaires algébraiques de matrices rectangulaires et mal-conditionnées*, Colloq. Int. du C.N.R.S. Besançon 1966, No. 165 (1968), pp. 161–170.

[18] K. FERNANDO AND B. PARLETT, *Implicit Cholesky algorithm for singular values and vectors*, Tech. Report PAM-587, Center for Pure and Applied Mathematics, University of California, Berkeley, 1993.

[19] ———, *Accurate singular values and differential QD algorithms*, Numer. Math., 67 (1994), pp. 191–229.

[20] G. GOLUB, *Numerical methods for solving linear least squares problems*, Numer. Math., 7 (1965), pp. 206–216.

[21] G. GOLUB AND W. KAHAN, *Calculating the singular values and pseudo-inverse of a matrix*, SIAM J. Numer. Anal. Ser., B2 (1965), pp. 205–224.

[22] G. GOLUB AND C. REINSCH, *Singular value decomposition and least squares solutions*, Numer. Math., 14 (1970), pp. 403–420.

[23] G. GOLUB AND C. VAN LOAN, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, 1989.

[24] R. HANSON AND C. LAWSON, *Extensions and applications of the Householder algorithms for solving least squares problems*, Math. Comp., 23 (1969), pp. 787–812.

[25] V. HARI AND K. VESELIĆ, *On Jacobi methods for singular value decompositions*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 741–754.

[26] H. HONG AND C. PAN, *The rank-revealing QR decomposition and* SVD, Math. Comp., 58 (1992), pp. 213–232.

[27] R. HORN AND C. JOHNSON, *Matrix Analysis*, Cambridge University Press, London, 1985.

[28] I. IPSEN, *A close look at an extended Golub–Reinsch algorithm for computing singular values (abstract)*, in The Householder Symposium XI Meeting on Numerical Linear Algebra, Technical Report, Department of Mathematics, Linköping University, Sweden, 1990.

[29] J. LAGARIAS, *Monotonicity properties of the Toda flow, the QR flow, and subspace iteration*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 449–462.

[30] C. LAWSON AND R. HANSON, *Solving Least Squares Problems*, Prentice-Hall, New York, 1974.

[31] R. MATHIAS AND G. STEWART, *A block QR algorithm and the singular value decomposition*, Linear Algebra Appl., 182 (1993), pp. 91–100.

[32] M. MOONEN, P. VAN DOOREN, AND F. VANPOUKE, *On the QR algorithm and updating the* SVD *and* URV *decomposition in parallel*, Linear Algebra Appl., 188/189 (1993), pp. 549–568.

[33] B. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, New York, 1980.

[34] B. PARLETT AND W. POOLE, *A geometric theory for the QR, LU and power iterations*, SIAM J. Numer. Anal., 10 (1973), pp. 389–412.

[35] G. STEWART, *On an algorithm for refining a rank-revealing* URV *factorization and a perturbation theorem for singular values*, Tech. Report UMIACS-TR-91-38, Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, 1991.

[36] ———, *An updating algorithm for subspace tracking*, IEEE Transactions on Signal Processing, SP-40 (1992), pp. 1535–1541.

[37] K. VESELIĆ AND V. HARI, *A note on a one-sided Jacobi algorithm*, Numer. Math, 56 (1989), pp. 627–633.

[38] D. WATKINS, *Understanding the QR algorithm*, SIAM Rev., 24 (1982), pp. 427–440.

[39] D. WATKINS AND L. ELSNER, *Self-equivalent flows associated with the singular value decomposition*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 244–258.

[40] ———, *Convergence of algorithms of decomposition type for the eigenvalue problem*, Linear Algebra Appl., 143 (1991), pp. 19–47.

[41] J. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 536–564, April 1995

015

DISPLACEMENT STRUCTURES OF COVARIANCE MATRICES, LOSSLESS SYSTEMS, AND NUMERICAL ALGORITHM DESIGN*

PHILLIP A. REGALIA† AND FRANÇOIS DESBOUVRIES†

Abstract. Low displacement rank theory underlies many fast algorithms designed for structured covariance matrices. Some of these have gained notoriety for their numerical instability problems, particularly fast least-squares algorithms. Recent studies have shown that instability is not inherent to fast algorithms, but rather comes from the violation of backward consistency constraints. This paper thus details the connection between covariance matrices of a given displacement inertia and lossless rational matrices, as well as the role of this connection in numerically consistent algorithms. This basic connection allows displacement structures to be parametrized via a sequence of rotation angles obtained from a lossless system. The utility of this approach is that, irrespective of errors in the rotation parameter set, they remain consistent with a positive definite matrix of a prescribed displacement inertia. This property in turn may be rephrased as meaningful forms of backward consistency in numerical algorithms. The rotation parameters then take the form of Givens or Jacobi angles applied to data, in contrast to classical approaches which directly manipulate dyadic decompositions of the displacement structure. The concepts are illustrated in popular signal processing applications. In particular, these connections lend clear insight into the stable computation of reflection coefficients of Toeplitz systems, and also serve to resolve the numerical instability problem of fast least-squares algorithms.

Key words. displacement ranks, covariance matrices, Schur reduction, lossless systems, fast least-squares algorithms

AMS subject classifications. 15A21, 15A24, 93A25, 30C45

1. Introduction.

Displacement rank theory plays a fundamental role in the development of fast computational algorithms [1]–[7], and is closely connected to problems in inverse scattering and interpolation theory [8]–[10], providing a rich algebraic link between modern results of analytic functions and concrete applications in matrix computation. This paper focuses on the intimate connection between displacement structures of covariance matrices and lossless transfer matrices, and the implications of this connection in numerical algorithm design.

One of the greatest impacts of displacement rank theory has been the development of fast recursive least-squares algorithms. The successful exploitation of a certain shift property of the data allows an order of magnitude reduction in the computational complexity, compared to conventional recursive least-squares algorithms. Many of these algorithms suffer numerical instability problems in the form of unstable error propagation: round-off errors in the computed quantities are amplified by successive time update recursions, leading to numerical divergence. This has led some skeptics to conjecture (incorrectly) that some form of numerical instability should be inherited by all fast algorithms developed from low displacement rank theory.

The philosophy of this paper evolved during the numerical stability study of such fast algorithms, and the isolation of structural features underlying instability problems. It was first recognized in [38], [39] that backward consistency gives a sufficient criterion for stable error propagation in any fast least-squares algorithm. Backward consistency means that the quantities computed in finite precision arithmetic are indistinguishable

*Received by the editors October 19, 1992; accepted for publication (in revised form) by G. Cybenko, March 2, 1994.

†Département Signal et Image, Institut National des Télécommunications, 9 rue Charles Fourier, 91011 Evry Cedex France (regalia@galaxie.int-evry.fr).

from those obtained by first perturbing the input data to the filtering algorithm, and then running the same algorithm in exact arithmetic, a familiar concept in numerical analysis [11], [12]. This idea, in the context of fast least-squares algorithms, was considerably amplified in [36]–[40], which established that numerical divergence effects are necessarily preceded by the violation of backward consistency constraints. A key development in this approach was advanced by Slock [36], who deduced a manifold characterizing the set of "reachable" variables in exact arithmetic for the so-called fast transversal equations, which are notorious for numerical divergence problems. Soon thereafter, this manifold was recognized [37] to be a special case of more general results known from lossless inverse scattering [8]. This connection proved crucial towards establishing the numerical stability of alternate fast least-squares algorithms, such as those obtained from fast QR decomposition approaches [40].

The essential lessons from the study of fast least-squares algorithms all reduce to backward consistency constraints. Backward consistency is a necessary first step in any backward error analysis, and despite many papers touting fast algorithms obtained from low displacement rank theory, questions of existence and applicability of backward error analyses are by and large absent. Thus a paper devoted to displacement structures, lossless systems, and numerical algorithm design would seem appropriate, particularly if the features underlying backward consistency can be brought to the forefront.

Many of the supporting results of this paper have previously appeared in very specialized contexts, some rather advanced. Recognizing that the reader may not have references [1]–[47] at his/her fingertips, proofs or verifications of many supporting arguments are included to enhance the tutorial aspects of the paper.

The organization is as follows. Section 2 begins with a selected overview of known results in displacement rank theory. This section leads into the important equivalence between positive definite matrices of a given displacement inertia and lossless rational matrices (Theorem 2.1). This result has very natural implications concerning backward consistency in numerical algorithm design, which will be emphasized throughout the paper. For the benefit of the nonexpert, §3 gives background information on Schur recursions that underlie Theorem 2.1 and that lead to alternate parametrizations of displacement structures. Sections 4–6 address computational aspects and backward consistency notions in light of the previous sections. To keep the presentation tractable, in §§4–6 attention is restricted to the so-called displacement inertia $(1,1)$ and $(2,1)$ cases. The former is intimately connected with the inverses of Toeplitz matrices [29] and orthogonal polynomials [27] as widely used in linear prediction theory [30]. The latter underlies fast least-squares algorithms, as well as certain forms of digital filter synthesis [22] and model reduction strategies in linear system theory [33]. Sections 5 and 6 are partly tutorial, but provide complete proofs of key propositions concerning fast least-squares algorithms, particularly Slock's manifold [39] for which the supporting arguments were previously incomplete.

Although the examples we consider are the simplest occurrences of low displacement rank theory, they have had immense impact on modern signal processing, and also serve as excellent examples illustrating the interplay between the topics of the paper's title.

2. Review of displacement structures. We suppose that $\mathbf{P}$ is an $M \times M$ covariance matrix, meaning symmetric ($\mathbf{P} = \mathbf{P}^t$) and positive definite ($\mathbf{P} > \bigcirc$). The

classical matrix displacement residue takes the form [1]

$$\mathbf{P} - \mathbf{Z}\mathbf{P}\mathbf{Z}^t = \sum_{k=1}^{p} \mathbf{a}_k \mathbf{a}_k^t - \sum_{k=1}^{q} \mathbf{b}_k \mathbf{b}_k^t \qquad (M \times M), \tag{2.1}$$

where $\mathbf{Z}$ is the $M \times M$ "down-shift" matrix consisting of ones along the subdiagonal and zeros elsewhere. We suppose moreover that the column vectors $\{\mathbf{a}_k\}$ and $\{\mathbf{b}_k\}$ (of length M) are linearly independent; for if not, they may always be reduced to a linearly independent set. The matrix $\mathbf{P}$ is then said to have displacement inertia (p, q) with respect to the displacement residue (2.1),[1] and its displacement rank is then $p + q$.

Displacement residues are of interest when the structure of $\mathbf{P}$ leads to $p+q < M$ (particularly when $p + q << M$), since the displacement residue may then provide a more compact parametric representation than that available from the elements of $\mathbf{P}$ themselves. Computational algorithms using structured covariance matrices may then often be rephrased in terms of new algorithms using the displacement generators, with a decrease in the computational load (e.g., [1]–[4], [6], [7], [46]).

Equation (2.1) expresses the displacement generators $\{\mathbf{a}_k\}$ and $\{\mathbf{b}_k\}$ in terms of $\mathbf{P}$, and a direct expression for $\mathbf{P}$ in terms of $\{\mathbf{a}_k\}$ and $\{\mathbf{b}_k\}$ may be obtained from the representation theorem [1]. Specifically, from a vector $\mathbf{v} = [v_1 \ \cdots \ v_M]^t$ one may define a lower triangular Toeplitz matrix $\mathcal{L}(\mathbf{v})$ via

$$\mathcal{L}(\mathbf{v}) = \begin{bmatrix} v_1 & 0 & \cdots & 0 \\ v_2 & v_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ v_M & \cdots & v_2 & v_1 \end{bmatrix} \qquad (M \times M).$$

With this, $\mathbf{P}$ may be recovered as [1]

$$\mathbf{P} = \sum_{k=1}^{p} \mathcal{L}(\mathbf{a}_k)\, \mathcal{L}^t(\mathbf{a}_k) - \sum_{k=1}^{q} \mathcal{L}(\mathbf{b}_k)\, \mathcal{L}^t(\mathbf{b}_k). \tag{2.2}$$

As a consequence, one may always choose the generator vectors freely (subject to a linear independence constraint); the symmetric matrix $\mathbf{P}$ determined from (2.2) then fulfills the displacement equation (2.1). Characterizations of positivity are reviewed below.

Many other forms of displacement residues are also used in different contexts, as evidenced by [29] in the study of Hankel and Toeplitz forms, or [10] in the study of classical interpolation theory. Our interest focuses on a slight variant of (2.1), taken from definition 1.1 in [29]:

$$\begin{bmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{0}^t & 0 \end{bmatrix} - \begin{bmatrix} 0 & \mathbf{0}^t \\ \mathbf{0} & \mathbf{P} \end{bmatrix} = \sum_{k=1}^{p} \mathbf{a}_k \mathbf{a}_k^t - \sum_{k=1}^{q} \mathbf{b}_k \mathbf{b}_k^t \qquad [(M+1) \times (M+1)]. \tag{2.3}$$

[1]The inertia of a symmetric matrix is properly the ordered triple (π, ν, δ) where π is the number of positive eigenvalues, ν is the number of negative eigenvalues, and δ is the number of zero eigenvalues. For simplicity, we only consider the ordered couple (π, ν), since the number of zero eigenvalues is then obtained by counting those that remain. Thus in (2.1) we have $(\pi, \nu) = (p, q)$.

Here $\mathbf{0}$ is the zero column vector of M elements and the generator vectors now of length $M+1$. One may observe that the residue appearing in (2.1) is the $M \times M$ principal submatrix of (2.3), and similarly, the so-called "up-shifted" displacement residue $\mathbf{Z}^t\mathbf{P}\mathbf{Z} - \mathbf{P}$ corresponds to the lower right $M \times M$ subblock of (2.3). This yields an appealing balance, and if $\mathbf{P}$ has displacement rank r (say) in (2.3), its displacement rank in (2.1) will be r or less.

Suppose now that $\mathbf{P}_\infty = \begin{bmatrix} P_1 & \cdots \\ \vdots & \ddots \end{bmatrix}$ is a doubly infinite matrix and extend the dimensions of the shift matrix $\mathbf{Z}$ accordingly. This gives rise to an operator displacement residue of the form

$$\mathbf{P}_\infty - \mathbf{Z}\mathbf{P}_\infty\mathbf{Z}^t = \sum_{k=1}^{p} \mathbf{a}_k\,\mathbf{a}_k^t - \sum_{k=1}^{q} \mathbf{b}_k\,\mathbf{b}_k^t, \tag{2.4}$$

where the generator vectors are now of infinite length. If $\mathbf{P}$ is the $M \times M$ principal submatrix of $\mathbf{P}_\infty$, then clearly the residues (2.1) and (2.3) are both particular instances of the infinite form (2.4). Specifically, (2.1) is obtained by displacing the operator $\mathbf{P}_\infty$ and then truncating the result, while (2.3) is obtained by first truncating the operator and then displacing the result. Thus any result applicable to (2.4) must specialize to (2.3) upon setting $\mathbf{P}_\infty = \begin{bmatrix} \mathbf{P} & 0 \\ 0 & 0 \end{bmatrix}$.

Characterizations of positivity of $\mathbf{P}$ first surfaced with respect to the infinite form (2.4) in Lev-Ari and Kailath [5]. Specifically, partition $\mathbf{P}_\infty$ as

$$\mathbf{P}_\infty = \begin{bmatrix} P_1 & \mathbf{p}^t \\ \mathbf{p} & \mathbf{P}_2 \end{bmatrix},$$

where P_1 is a scalar. Then a standard test attributed to Schur [14] asserts that $\mathbf{P}_\infty$ is positive definite if and only if $P_1 > 0$ and the Schur complement

$$\mathbf{P}_2 - \mathbf{p}\mathbf{p}^t/P_1$$

is positive definite. Once the Schur complement is similarly partitioned, a recursive test for positivity is obtained. It turns out [5] that this test may be rephrased in terms of the generator vectors appearing in (2.4), in the form of an infinite lattice synthesis procedure. Since this test involves an infinite number of steps, convergence is not immediately clear.

A more complete characterization of positivity was obtained by Alpay, Dewilde, and Dym [8]. Multiply the form (2.4) from the right by the row vector $[1 \;\; z \;\; z^2 \;\; \cdots]$ and from the left by the column vector $[1 \;\; w \;\; w^2 \;\; \cdots]^t$, where z and w are two complex variables. One obtains

$$(1 - zw)\,P(z,w) = \sum_{k=1}^{p} A(z)\,A(w) - \sum_{k=1}^{q} B(z)\,B(w), \tag{2.5}$$

where

$$P(z,w) = [1 \;\; z \;\; z^2 \;\; \cdots]\,\mathbf{P}_\infty \begin{bmatrix} 1 \\ w \\ w^2 \\ \vdots \end{bmatrix},$$

$$A_k(z) = [1 \;\; z \;\; z^2 \;\; \cdots]\,\mathbf{a}_k, \qquad k = 1, 2, \ldots, p;$$

$$B_k(z) = [1 \;\; z \;\; z^2 \;\; \cdots]\,\mathbf{b}_k, \qquad k = 1, 2, \ldots, q.$$

The terms $\{A_k(z)\}$ and $\{B_k(z)\}$ are called the generator functions [5]. The results of [8] show that $\mathbf{P}_\infty$ is positive (semi) definite if and only if there exists a $q \times p$ matrix Schur function $\mathbf{S}(z)$, which reconciles the generator functions as

$$\begin{bmatrix} B_1(z) \\ \vdots \\ B_q(z) \end{bmatrix} = \mathbf{S}(z) \begin{bmatrix} A_1(z) \\ \vdots \\ A_p(z) \end{bmatrix}. \tag{2.6}$$

By Schur function, we mean that $\mathbf{S}(z)$ is analytic in $|z| \leq 1$, and contractive there:

$$\mathbf{S}^*(z)\,\mathbf{S}(z) \leq \mathbf{I}_p, \quad \text{and} \quad \mathbf{S}(z)\,\mathbf{S}^*(z) \leq \mathbf{I}_q \quad \text{for all } |z| \leq 1,$$

where the superscript asterisk denotes Hermitian transpose. Extensions to this result, involving displacement residues in which $\mathbf{Z}$ is replaced by more general triangular structures, may be found in Theorem 3.2 of Sayed [10].

With respect to the matrix displacement residue (2.3), the two-variable form (2.5) still holds, but now simplifies to a polynomial equation, where

$$P(z,w) = [1 \;\; z \;\; \cdots \;\; z^{M-1}]\,\mathbf{P} \begin{bmatrix} 1 \\ w \\ \vdots \\ w^{M-1} \end{bmatrix}, \tag{2.7a}$$

$$A_k(z) = [1 \;\; z \;\; \cdots \;\; z^M]\,\mathbf{a}_k, \qquad k = 1, 2, \ldots, p, \tag{2.7b}$$

$$B_k(z) = [1 \;\; z \;\; \cdots \;\; z^M]\,\mathbf{b}_k, \qquad k = 1, 2, \ldots, q. \tag{2.7c}$$

The generator functions $\{A_k(z)\}$ and $\{B_k(z)\}$ are now polynomials of degree not exceeding M.

A simple property may be observed at this point. Set $w = z^{-1}$ in (2.5) and note that the left-hand side vanishes. Thus the polynomial generator functions must be related as

$$\sum_{k=1}^{p} A_k(z)\,A_k(z^{-1}) = \sum_{k=1}^{q} B_k(z)\,B_k(z^{-1}) \quad \text{for all } z, \tag{2.8}$$

if they are indeed obtained from the generator vectors in residue (2.3). This condition is sufficient as well. For if (2.8) holds, then the polynomial $\sum_k A(z)\,A(w) - \sum_k B_k(z)\,B_k(w)$ contains a factor $(1 - zw)$. As such,

$$\Big(\sum_k A_k(z)\,A_k(w) - \sum_k B_k(z)\,B_k(w)\Big)/(1 - zw)$$

remains a polynomial of degree $M - 1$ in both z and w. It may thus be written in the form (2.7a) for some $M \times M$ matrix $\mathbf{P}$.

This puts forth a simple lesson: If arithmetic operations are applied to the generator vectors, the locally independent errors in these vectors will lead to constraint (2.8) being violated. In this case, the perturbed vectors cannot be associated with any matrix $\mathbf{P}$ via the displacement residue (2.3), which is to say that consistency is lost.

The structural constraint (2.8) turns out to be quite exploitable: it says, in effect, that the vectors $[A_1(z), \ldots, A_p(z)]$ and $[B_1(z), \ldots, B_p(z)]$ are different spectral factors of the same function. By a standard result [13], any two spectral factors of a given

function may always be reconciled by a matrix function that is unitary on the unit circle. Indeed, the result of Alpay, Dewilde, and Dym [8] concerning positivity of $\mathbf{P}_\infty$ may be specialized to the matrix case as follows.

THEOREM 2.1. *Let* $\{A_k(z)\}_{k=1}^p$ *and* $\{B_k(z)\}_{k=1}^q$ *be polynomials obtained from candidate generator vectors in* (2.7). *There exists a positive definite* $M \times M$ *matrix* $\mathbf{P}$ *fulfilling the displacement equation* (2.3), *if and only if there exists a* $p \times q$ *lossless function* $\mathbf{U}(z)$ *of McMillan degree* M *which fulfills*

$$\begin{bmatrix} A_1(z) \\ \vdots \\ A_p(z) \end{bmatrix} = \mathbf{U}(z) \begin{bmatrix} B_1(z) \\ \vdots \\ B_q(z) \end{bmatrix}. \tag{2.9}$$

The matrix $\mathbf{P}$ *may be recovered from* $\mathbf{U}(z)$, *to within a scale factor ambiguity.*

By lossless, we mean that $\mathbf{U}(z)$ is analytic in $|z| \geq 1$, and paraunitary:

$$\mathbf{U}^t(z^{-1})\,\mathbf{U}(z) = \mathbf{I}_q \qquad \text{if } p \geq q$$

or

$$\mathbf{U}(z)\,\mathbf{U}^t(z^{-1}) = \mathbf{I}_p \qquad \text{if } p \leq q.$$

The first (respectively, second) equality says that the column (respectively, row) vectors of $\mathbf{U}(e^{i\omega})$ are orthonormal for any $-\pi \leq \omega \leq \pi$. Lossless systems play a fundamental role in many signal processing applications [19]–[26], to which the interested reader is referred for more background information. This paper is concerned principally with the case $p \geq q$, as the case $p < q$ may be treated in a parallel fashion. For the case $p \geq q$, we have $\mathbf{U}^t(z^{-1})\,\mathbf{U}(z) = \mathbf{I}$, so that (2.9) may be rewritten as

$$\mathbf{U}^t(z^{-1}) \begin{bmatrix} A_1(z) \\ \vdots \\ A_p(z) \end{bmatrix} = \begin{bmatrix} B_1(z) \\ \vdots \\ B_q(z) \end{bmatrix}.$$

It is easy to show that $\mathbf{U}^t(z^{-1})$ is analytic and contractive in $|z| \leq 1$, thereby yielding a Schur function, and thus a particular instance of (2.6). The modified form (2.9), however, proves more convenient for the purposes of this paper.

The numerical consequences of this result must be emphasized. By the bounded real lemma [26], a matrix function $\mathbf{U}(z)$ is lossless if and only if may be expressed as

$$\mathbf{U}(z) = \mathbf{D} + \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B},$$

where $\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$ has orthonormal columns (if $p \geq q$) or rows (if $p \leq q$). In network synthesis [19]–[23], the matrix $\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$ is split into elementary rotation angles $\{\theta_k\}$, such that $\mathbf{U}(z)$ is lossless irrespective of the exact values of these angles. To any perturbation of the parameters $\{\theta_k\}$ proper to $\mathbf{U}(z)$, there thus corresponds a perturbed matrix $\mathbf{P}$ of the same displacement inertia which remains positive definite, by Theorem 2.1. This connection gives rise to backward consistency: any error in the rotation parameter set could have been obtained by first perturbing the initial data that build $\mathbf{P}$ and then running the computations in exact arithmetic. We will return to this interpretation in §§4 and 6.

3. Schur reduction.

3.1. Construction of Theorem 2.1. The construction underlying Theorem 2.1 is a consequence of Schur reductions [5], [10], [14]–[17], [22], [28] whose review may be helpful to some readers. We begin by partitioning $\mathbf{P}$ as

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_1 & \mathbf{p} \\ \mathbf{p}^t & P_2 \end{bmatrix},$$

where P_2 is a scalar. Then $\mathbf{P}$ is positive definite if and only if $P_2 > 0$ and the Schur complement

$$\mathbf{P}_1 - \mathbf{p}\mathbf{p}^t/P_2 \tag{3.1}$$

is positive definite. Our approach follows [5] closely, to extract the Schur complement (3.1) in terms of the generator vectors of $\mathbf{P}$. The procedure then continues on this Schur complement, yielding M successful steps if and only if $\mathbf{P}$ is positive definite. This test will naturally take the form of a synthesis of the lossless function $\mathbf{U}(z)$ fulfilling Theorem 2.1.

We begin by setting

$$\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_p], \qquad \mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_q],$$

so that the displacement equation (2.3) may be written

$$\begin{bmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{0}^t & 0 \end{bmatrix} - \begin{bmatrix} 0 & \mathbf{0}^t \\ \mathbf{0} & \mathbf{P} \end{bmatrix} = [\mathbf{A}\ \mathbf{B}] \underbrace{\begin{bmatrix} \mathbf{I}_p & \\ & -\mathbf{I}_q \end{bmatrix}}_{\triangleq \mathbf{J}} \begin{bmatrix} \mathbf{A}^t \\ \mathbf{B}^t \end{bmatrix}. \tag{3.2}$$

Let Σ be any $(p+q) \times (p+q)$ $\mathbf{J}$-orthogonal matrix, i.e., satisfying

$$\Sigma^t \mathbf{J} \Sigma = \mathbf{J}.$$

As is well known, the row vectors of the matrix

$$\begin{matrix} p\{ \\ q\{ \end{matrix} \begin{bmatrix} \mathbf{A}_\Sigma^t \\ \mathbf{B}_\Sigma^t \end{bmatrix} = \Sigma \begin{bmatrix} \mathbf{A}^t \\ \mathbf{B}^t \end{bmatrix}$$

are also generator vectors of the same matrix $\mathbf{P}$, as may be verified by direct substitution. The Schur complement (3.1) may be deduced by elementary operations involving strategic choices of Σ [5].

For convenience, let $\begin{matrix} \} p \\ \} q \end{matrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$ be the final column of the matrix $\begin{bmatrix} \mathbf{A}^t \\ \mathbf{B}^t \end{bmatrix}$. We may then rewrite the displacement equation (3.2) in the form

$$\begin{bmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{0}^t & 0 \end{bmatrix} - \begin{bmatrix} 0 & \mathbf{0}^t \\ \mathbf{0} & \mathbf{P} \end{bmatrix} = \begin{bmatrix} \times_{\mathbf{A}}^t & \times_{\mathbf{B}}^t \\ \mathbf{x}^t & \mathbf{y}^t \end{bmatrix} \begin{bmatrix} \mathbf{I}_p & \\ & -\mathbf{I}_q \end{bmatrix} \begin{bmatrix} \times_{\mathbf{A}} & \mathbf{x} \\ \times_{\mathbf{B}} & \mathbf{y} \end{bmatrix}.$$

Inspection of the lower right entry of both sides reveals

$$P_2 = \|\mathbf{y}\|^2 - \|\mathbf{x}\|^2.$$

This observation has a simple interpretation: There exists a $\mathbf{J}$-orthogonal matrix Σ fulfilling

$$\Sigma \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_p \\ \mathbf{0}_{q-1} \\ \sqrt{\|\mathbf{y}\|^2 - \|\mathbf{x}\|^2} \end{bmatrix} \tag{3.3}$$

if and only if $P_2 > 0$. Such a Σ may be chosen as a hyperbolic Householder transformation, for example. Note that in the limiting case $\|\mathbf{x}\|^2 = \|\mathbf{y}\|^2$, the entries of Σ would not necessarily be bounded.

For illustration purposes, suppose that $M = 4$ and that $(p, q) = (2, 2)$. We would then have

$$\begin{bmatrix} \mathbf{A}^t \\ \mathbf{B}^t \end{bmatrix} = \begin{bmatrix} \times & \times & \times & \times & \otimes \\ \times & \times & \times & \times & \otimes \\ \times & \times & \times & \times & \otimes \\ \times & \times & \times & \times & \otimes \end{bmatrix}.$$

The distinguished elements $\otimes$ of the last column are the vector $\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$. Applying the transformation Σ from (3.3), we find

$$\begin{aligned} \begin{bmatrix} \mathbf{A}_\Sigma^t \\ \mathbf{B}_\Sigma^t \end{bmatrix} &= \Sigma \begin{bmatrix} \times & \times & \times & \times & \otimes \\ \times & \times & \times & \times & \otimes \\ \times & \times & \times & \times & \otimes \\ \times & \times & \times & \times & \otimes \end{bmatrix} \\ &= \begin{bmatrix} \times & \times & \times & \times & 0 \\ \times & \times & \times & \times & 0 \\ \times & \times & \times & \times & 0 \\ 0 & \times & \times & \times & \sqrt{\|\mathbf{y}\|^2 - \|\mathbf{x}\|^2} \end{bmatrix}. \end{aligned} \tag{3.4}$$

Note the "free" zero that is obtained in the lower left-hand entry. This is a consequence of a well-known degree-reduction property of the Schur reduction procedure that is explicit, for instance, in [22]–[24].

For our purposes, this free zero may be explained as follows. Let

$$\begin{bmatrix} \mathbf{A}(z) \\ \mathbf{B}(z) \end{bmatrix} = \begin{bmatrix} \mathbf{A}^t \\ \mathbf{B}^t \end{bmatrix} \begin{bmatrix} 1 \\ z \\ \vdots \\ z^M \end{bmatrix},$$

which is simply the vector built from the polynomial generator functions $\{A_k(z)\}_{k=1}^p$ and $\{B_k(z)\}_{k=1}^q$. Similarly, set

$$\begin{bmatrix} \mathbf{A}_\Sigma(z) \\ \mathbf{B}_\Sigma^t(z) \end{bmatrix} = \begin{bmatrix} \mathbf{A}_\Sigma^t \\ \mathbf{B}_\Sigma^t \end{bmatrix} \begin{bmatrix} 1 \\ z \\ \vdots \\ z^M \end{bmatrix},$$

which is the vector obtained from the transformed generator functions $\{A_{\Sigma,k}(z)\}_{k=1}^p$ and $\{B_{\Sigma,k}(z)\}_{k=1}^q$. Because Σ is $\mathbf{J}$-orthogonal, one has

$$\begin{aligned} &\sum_{k=1}^{p} A_{\Sigma,k}(z)\, A_{\Sigma,k}(z^{-1}) - \sum_{k=1}^{q} B_{\Sigma,k}(z)\, B_{\Sigma,k}(z^{-1}) \\ &= \sum_{k=1}^{p} A_k(z)\, A_k(z^{-1}) - \sum_{k=1}^{q} B_k(z)\, B_k(z^{-1}) \quad (=0) \end{aligned}$$

or

$$B_{\Sigma,q}(z)\, B_{\Sigma,q}(z^{-1}) = \sum_{k=1}^{p} A_{\Sigma,k}(z)\, A_{\Sigma,k}(z^{-1}) - \sum_{k=1}^{q-1} B_{\Sigma,k}(z)\, B_{\Sigma,k}(z^{-1}).$$

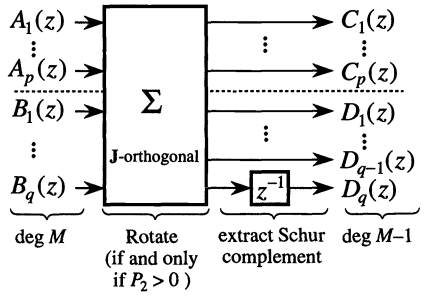


FIG. 1. *One step in the Schur reduction procedure.*

By construction, the right-hand side is a polynomial of degree $\pm(M-1)$. Thus the coefficients of $z^{\pm M}$ on the left-hand side must vanish. If β_0 and β_M denote the extreme coefficients of the polynomial $B_{\Sigma,q}(z)$, the coefficient of $z^{\pm M}$ on the left-hand side is $\beta_0 \beta_M$. But $\beta_M = \sqrt{\|\mathbf{y}\|^2 - \|\mathbf{x}\|^2}$ exists and is nonzero if and only if $P_2 > 0$, whence $\beta_0 = B_{\Sigma,q}(0) = 0$ must result. This accounts for the free zero in (3.4).

At this point we may set

$$\begin{aligned} C_k(z) &= A_{\Sigma,k}(z), \qquad k = 1, 2, \ldots, p, \\ D_k(z) &= B_{\Sigma,k}(z), \qquad k = 1, 2, \ldots, q-1, \\ D_q(z) &= z^{-1} B_{\Sigma,q}(z). \end{aligned}$$

The polynomials $\{C_k(z)\}$ and $\{D_k(z)\}$ now have degrees not exceeding $M-1$. It is easy to verify that

$$\sum_{k=1}^{p} C_k(z)\, C_k(z^{-1}) = \sum_{k=1}^{q} D_k(z)\, D_k(z^{-1}).$$

Thus $\{C_k(z)\}$ and $\{D_k(z)\}$ are generator functions of some $(M-1) \times (M-1)$ matrix. More precisely, we have the following property.

PROPERTY 3.1. Let $\{C_k(z)\}_{k=1}^{p}$ and $\{D_k(z)\}_{k=1}^{q}$ be the polynomials obtained from the above procedure. These are generator functions of the Schur complement $\mathbf{P}_1 - \mathbf{p}\mathbf{p}^t/P_2$.

This result was first established in [5] (the Schur invariance theorem) with respect to the operator residue (2.4). The result specializes to our case upon setting $\mathbf{P}_\infty = \begin{bmatrix} \mathbf{P} & 0 \\ 0 & 0 \end{bmatrix}$.

The operations performed thus far admit the flowgraph interpretation of Fig. 1. The polynomials $\{C_k(z)\}$ and $\{D_k(z)\}$ now correspond to the Schur complement $\mathbf{R} \triangleq \mathbf{P}_1 - \mathbf{p}\mathbf{p}^t/P_1$. If we partition $\mathbf{R}$ as

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_1 & \mathbf{r} \\ \mathbf{r}^t & R_2 \end{bmatrix},$$

where R_2 is a scalar, then the next Schur reduction step will succeed in forcing a further degree reduction in the polynomial generator functions if and only if $R_2 > 0$. By induction, M successive Schur reduction steps will reduce the generator functions to constants if and only if $\mathbf{P}$ is positive definite.

The resulting flowgraph interpretation appears in Fig. 2. The overall transformation may be written as

$$\Sigma(z) = \mathbf{F}(z)\,\Sigma_1\mathbf{F}(z)\,\Sigma_2 \cdots \mathbf{F}(z)\,\Sigma_M, \tag{3.5}$$

where each Σ_k is a constant **J**-orthogonal matrix and $\mathbf{F}(z) = \mathrm{diag}[\mathbf{I}_{p+q-1}, z^{-1}]$. It is straightforward to verify that $\mathbf{F}(z)$ is **J**-inner [9] in the sense that

$$\mathbf{J} - \mathbf{F}^*(z)\,\mathbf{J}\,\mathbf{F}(z) \begin{cases} \geq \bigcirc, & |z| < 1, \\ = \bigcirc, & |z| = 1, \\ \leq \bigcirc, & |z| \geq 1, \end{cases}$$

where the asterisk denotes Hermitian transpose. It follows that $\Sigma(z)$ from (3.5) is also **J**-inner, since the product of **J**-inner and **J**-orthogonal factors will always yield a **J**-inner matrix [9].

Let **c** and **d** be the constant vectors obtained from the end result in Fig. 2:

$$\begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = \Sigma(z) \begin{bmatrix} \mathbf{A}(z) \\ \mathbf{B}(z) \end{bmatrix}.$$

Since $\Sigma(z)$ is **J**-inner, it is easily verified that

$$\|\mathbf{c}\|^2 - \|\mathbf{d}\|^2 = \mathbf{A}^t(z^{-1})\,\mathbf{A}(z) - \mathbf{B}^t(z^{-1})\,\mathbf{B}(z) = 0.$$

Thus, **c** and **d** have the same Euclidean norm. By elementary considerations, there exists a constant matrix **V** fulfilling $\mathbf{V}^t\mathbf{V} = \mathbf{I}$ and $\mathbf{c} = \mathbf{V}\,\mathbf{d}$.

Partition now the **J**-inner function $\Sigma(z)$ as

$$\Sigma(z) = \begin{bmatrix} \Sigma_{11}(z) & \Sigma_{12}(z) \\ \Sigma_{21}(z) & \Sigma_{22}(z) \end{bmatrix}$$

with $\Sigma_{11}(z)$ of dimensions $p \times p$. Since $\Sigma(z)$ is **J**-inner, the matrix $\mathbf{U}(z)$ defined via

$$\mathbf{U}(z) = [\Sigma_{11}(z) - \mathbf{V}\,\Sigma_{21}(z)]^{-1}[\mathbf{V}\,\Sigma_{22}(z) - \Sigma_{12}(z)], \qquad (p \times q)$$

is now lossless of McMillan degree M [9], [19], [21]–[23], and by construction fulfills $\mathbf{A}(z) = \mathbf{U}(z)\,\mathbf{B}(z)$. This gives the "only if" clause of Theorem 2.1.

A flowgraph of $\mathbf{U}(z)$ appears in Fig. 3; it is obtained by reversing the flow directions of the upper branches of Fig. 2. In doing so, each **J**-orthogonal matrix Σ_k is converted to an orthogonal matrix Θ_k.

To conclude the "if" clause, suppose the generator functions are related as $\mathbf{A}(z) = \mathbf{U}(z)\,\mathbf{B}(z)$, with $\mathbf{U}(z)$ lossless of degree M. Then the relation $\mathbf{U}^t(z^{-1})\,\mathbf{U}(z) = \mathbf{I}$ implies $\mathbf{A}^t(z^{-1})\,\mathbf{A}(z) = \mathbf{B}^t(z^{-1})\,\mathbf{B}(z)$. Thus the functions are indeed generators obtained from some matrix **P** via the displacement residue (3.2). To show that **P** must be positive definite, we recall [22], [23] that every lossless rational function of degree M may, by Schur recursions, be synthesized in the form of Fig. 3. This may be converted into Fig. 2 by flowgraph manipulations. Figure 2 shows that such Schur recursions will have reduced the generator functions to constant functions in precisely M steps, which implies that **P** is positive definite.

3.2. Illustrative examples. For the benefit of the nonexpert, we illustrate the above Schur reduction steps for two concrete cases: the displacement inertia $(1, 1)$

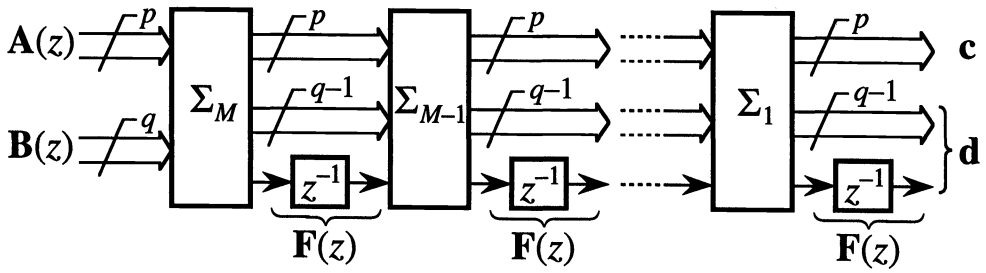


FIG. 2. *The Schur test for positive definiteness in terms of the generator polynomials* $\mathbf{A}(z)$ *and* $\mathbf{B}(z)$.

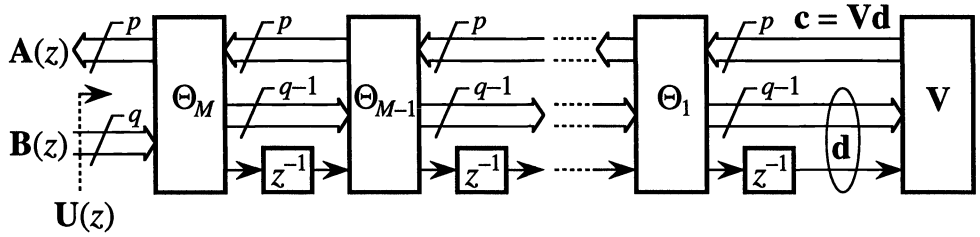


FIG. 3 *Lossless system* $\mathbf{U}(z)$ *fulfilling* $\mathbf{A}(z) = \mathbf{U}(z)\mathbf{B}(z)$, *obtained from Fig. 2 via flowgraph manipulation.*

case that is adapted from [28], and the displacement inertia (2, 1) case that is adapted from [22]. These cases are further examined with respect to computational aspects in §§4–6. We emphasize that these procedures are intended to illustrate candidate parametrizations of the displacement structure and are not recommended as computational algorithms.

The displacement inertia (1, 1) case is motivated by the following result

THEOREM 3.2. *Let* $\mathbf{P}$ *be positive definite. Then* $\mathbf{P}$ *has displacement inertia* (1, 1), *i.e.,*

$$\begin{bmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{0}^t & 0 \end{bmatrix} - \begin{bmatrix} 0 & \mathbf{0}^t \\ \mathbf{0} & \mathbf{P} \end{bmatrix} = \mathbf{a}\,\mathbf{a}^t - \mathbf{b}\,\mathbf{b}^t$$

if and only if $\mathbf{P}^{-1}$ *is Toeplitz.*

This result may be found in Theorem 2.1 of [29] and may also be inferred from [27]. Since $A(z)$ and $B(z)$ are both scalar, the lossless function $\mathbf{U}(z)$ is easily found as

$$\mathbf{U}(z) = \frac{A(z)}{B(z)}.$$

The identity $\mathbf{U}(z^{-1})\,\mathbf{U}(z) = 1$ is a restatement of the constraint $A(z)\,A(z^{-1}) = B(z)\,B(z^{-1})$.

Without loss of generality, suppose $M = 3$. We then have

$$\begin{bmatrix} A(z) \\ B(z) \end{bmatrix} = \begin{bmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \end{bmatrix} \begin{bmatrix} 1 \\ z \\ z^2 \\ z^3 \end{bmatrix}.$$

Let Σ_3 be a hyperbolic rotation that annihilates the upper right-hand entry of the above array. Then in array form we have

$$\begin{bmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \end{bmatrix} \overset{\Sigma_3}{\rightarrow} \begin{bmatrix} \times & \times & \times & 0 \\ 0 & \times & \times & \times \end{bmatrix}.$$

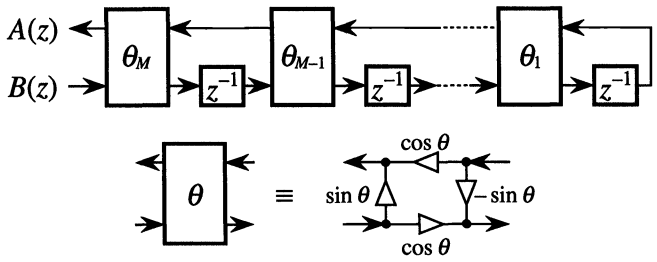


FIG. 4. *Lossless system* $\mathbf{U}(z)$ *for displacement inertia* $(p, q) = (1, 1)$ *giving the normalized lattice.*

The lower left-hand entry is, of course, also annihilated for free. Let z^{-1} denote the right shift operation applied to the bottom row:

$$\begin{bmatrix} \times & \times & \times & 0 \\ 0 & \times & \times & \times \end{bmatrix} \overset{z^{-1}}{\rightarrow} \begin{bmatrix} \times & \times & \times & 0 \\ \times & \times & \times & 0 \end{bmatrix}.$$

Delete the final zero column, and reiterate the above procedure twice to reduce the generator vectors to constants.

If $C(z)$ and $D(z)$ are the generator functions obtained after the first reduction step, then

$$\begin{bmatrix} C(z) \\ z\,D(z) \end{bmatrix} = \frac{1}{\cos\theta_3} \begin{bmatrix} 1 & -\sin\theta_3 \\ -\sin\theta_3 & 1 \end{bmatrix} \begin{bmatrix} A(z) \\ B(z) \end{bmatrix}, \tag{3.6}$$

where

$$\sin\theta_3 = \frac{A(\infty)}{B(\infty)} = \mathbf{U}(\infty).$$

This of course is a classical formula advanced by Schur [14]. The matrix form (3.6) can be rearranged as

$$\begin{bmatrix} A(z) \\ D(z) \end{bmatrix} = \begin{bmatrix} 1 & \\ & z^{-1} \end{bmatrix} \begin{bmatrix} \cos\theta_3 & \sin\theta_3 \\ -\sin\theta_3 & \cos\theta_3 \end{bmatrix} \begin{bmatrix} C(z) \\ B(z) \end{bmatrix}.$$

The interconnection of this system is sketched in Fig. 4, and in prediction theory is known as the normalized lattice filter [30]. The parameters $\sin\theta_k$ are known as the reflection coefficients associated to the Toeplitz matrix $\mathbf{P}^{-1}$. As is well known, the function $\mathbf{U}(z)$ realized in Fig. 4 is lossless of degree M, if and only if $|\sin\theta_k| < 1$ for all k. Theorem 2.1 then reduces to a classical result: The reflection coefficients obtained from a Toeplitz matrix are all upper bounded by unit magnitude if and only if the Toeplitz matrix is positive definite. Reliable numerical methods for obtaining the reflection coefficients are presented in §4.

For the displacement inertia $(2, 1)$ case, we now have

$$\begin{bmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{0}^t & 0 \end{bmatrix} - \begin{bmatrix} 0 & \mathbf{0}^t \\ \mathbf{0} & \mathbf{P} \end{bmatrix} = \mathbf{a}_1\,\mathbf{a}_1^t + \mathbf{a}_2\,\mathbf{a}_2^t - \mathbf{b}_1\,\mathbf{b}_1^t.$$

The structure of the associated $\mathbf{P}$ matrix is made explicit in §5.

The array form appears as

$$\begin{bmatrix} \mathbf{a}_1^t \\ \mathbf{a}_2^t \\ \mathbf{b}_1^t \end{bmatrix} = \begin{bmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \end{bmatrix}.$$

Let Σ_1' be a hyperbolic rotation operating in the $(2,3)$ plane, chosen such that

$$\begin{bmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \end{bmatrix} \overset{\Sigma_1'}{\longrightarrow} \begin{bmatrix} \times & \times & \times & \times \\ \times & \times & \times & 0 \\ \times & \times & \times & \times \end{bmatrix}.$$

Let next Σ_2' be a hyperbolic rotation operating in the $(1,3)$ plane, chosen such that

$$\begin{bmatrix} \times & \times & \times & \times \\ \times & \times & \times & 0 \\ \times & \times & \times & \times \end{bmatrix} \overset{\Sigma_2'}{\longrightarrow} \begin{bmatrix} \times & \times & \times & 0 \\ \times & \times & \times & 0 \\ 0 & \times & \times & \times \end{bmatrix}.$$

At this point, a free zero is obtained in the lower left entry. Finally, let Σ_3' be a planar rotation operating in the $(1,2)$ plane:

$$\begin{bmatrix} \times & \times & \times & 0 \\ \times & \times & \times & 0 \\ 0 & \times & \times & \times \end{bmatrix} \overset{\Sigma_3'}{\longrightarrow} \begin{bmatrix} \times & \times & \times & 0 \\ 0 & \times & \times & 0 \\ 0 & \times & \times & \times \end{bmatrix}. \tag{3.7}$$

The product $\Sigma_3' \Sigma_2' \Sigma_1'$ is $\mathbf{J}$-orthogonal, and hence the array form (3.7) yields generator vectors for the same matrix $\mathbf{P}$. The point is that we may always choose the generator vectors with zeros distributed as in (3.7), if $\mathbf{P}$ is positive definite. Thus the array form (3.7) is taken as our starting point, i.e., the generator vectors are redefined as

$$\begin{bmatrix} \mathbf{a}_1^t \\ \mathbf{a}_2^t \\ \mathbf{b}_1^t \end{bmatrix} = \begin{bmatrix} \times & \times & \times & 0 \\ 0 & \times & \times & 0 \\ 0 & \times & \times & \times \end{bmatrix}.$$

Shift the final row to obtain

$$\begin{bmatrix} \times & \times & \times & 0 \\ 0 & \times & \times & 0 \\ 0 & \times & \times & \times \end{bmatrix} \overset{z^{-1}}{\longrightarrow} \begin{bmatrix} \times & \times & \times & 0 \\ 0 & \times & \times & 0 \\ \times & \times & \times & 0 \end{bmatrix}.$$

Delete the final zero column, and choose now a hyperbolic rotation $\Sigma_{2(M-1)} = \Sigma_4$ operating in the $(1,3)$ plane to yield

$$\begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ \times & \times & \times \end{bmatrix} \overset{\Sigma_4}{\longrightarrow} \begin{bmatrix} \times & \times & 0 \\ 0 & \times & \times \\ 0 & \times & \times \end{bmatrix}.$$

Both the upper right and lower left entries are annihilated. Choose Σ_3 as a hyperbolic rotation operating in the $(2,3)$ plane:

$$\begin{bmatrix} \times & \times & 0 \\ 0 & \times & \times \\ 0 & \times & \times \end{bmatrix} \overset{\Sigma_3}{\longrightarrow} \begin{bmatrix} \times & \times & 0 \\ 0 & \times & 0 \\ 0 & \times & \times \end{bmatrix}.$$

Now reiterate the above procedure once more to reduce the generator functions to constants. The lossless function $\mathbf{U}(z)$ that reconciles the generator vectors from (3.7) appears as Fig. 5. This structure first appeared in [22] in the context of digital filter synthesis. It may be shown that the $\mathbf{U}(z)$ so realized is lossless of degree M if and

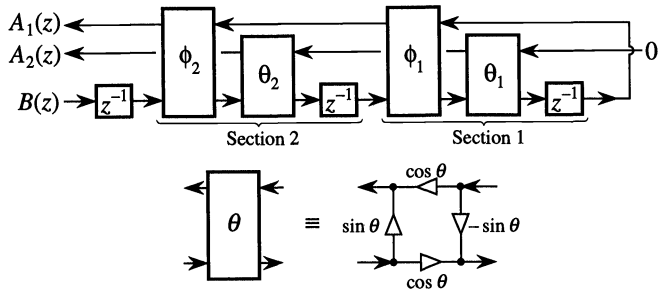


FIG. 5. *Lossless system* $\mathbf{U}(z)$ *for displacement inertia* $(p, q) = (2, 1)$.

only if $|\sin\theta_k| < 1$ and $|\sin\phi_k| < 1$ for all k [22]. Hence, for any rotation angle set satisfying this simple constraint, the implied generator vectors are consistent with a positive definite $\mathbf{P}$ of displacement inertia $(2, 1)$. The role of this property in the numerical stability of fast least-squares algorithms is developed in §6.

4. Calculation of reflection coefficients. An important problem in linear prediction is to determine the reflection coefficients $\sin\theta_k$ that appear in the lattice filter of Fig. 4, given empirical estimates of an autocorrelation sequence that constructs a Toeplitz matrix. Here we address their computation.

Specifically, let $\{x(k)\}_{k=0}^{n}$ be some data sequence, collected into the vector

$$\mathbf{x} = \begin{bmatrix} x(0) \\ \vdots \\ x(n) \\ \mathbf{0}_{M-1} \end{bmatrix}$$

with $M-1$ zeros appended at the bottom. Construct the M-column data matrix $\mathbf{X}_M$ according to

$$\mathbf{X}_M = \begin{bmatrix} \mathbf{x} & \mathbf{Z}\mathbf{x} & \cdots & \mathbf{Z}^{M-1}\mathbf{x} \end{bmatrix},$$

where $\mathbf{Z}$ is the shift matrix with ones on the subdiagonal. Then $\mathbf{P}^{-1} \triangleq \mathbf{X}_M^t \mathbf{X}_M$ is a positive definite Toeplitz matrix. From $\mathbf{P}^{-1}$ (or $\mathbf{P}$) one could in principle apply a Levinson or Schur algorithm to calculate the reflection coefficients, but in finite precision these algorithms sometimes return computed reflection coefficients with magnitudes exceeding unity. We show in this section an orthogonal algorithm that operates on the data $\{x(k)\}$ themselves to extract computed reflection coefficients whose magnitudes will always be bounded by one.

Note first that $\mathbf{X}_M$ is itself a Toeplitz matrix (the so-called prewindowed and postwindowed case), and admits the simultaneous partitioning

$$\mathbf{X}_M = \left[\begin{array}{c|c} \mathbf{x} & \begin{matrix} 0 & \cdots & 0 \end{matrix} \\ \hline & \\ & \mathbf{X}_{M-1} \\ & \end{array}\right] = \left[\begin{array}{c|c} \mathbf{X}_{M-1} & \\ \hline \begin{matrix} 0 & \cdots & 0 \end{matrix} & \mathbf{Z}^{M-1}\mathbf{x} \end{array}\right].$$

Let $\mathbf{Q}$ be an orthogonal matrix that rotates $\mathbf{X}_{M-1}$ into an upper triangular matrix $\mathbf{C}_{M-1}$:

$$\mathbf{Q}\mathbf{X}_{M-1} = \begin{bmatrix} \mathbf{C}_{M-1} \\ \bigcirc \end{bmatrix}. \tag{4.1}$$

We assume that the diagonal elements of $\mathbf{C}_{M-1}$ are positive. Shifting $\mathbf{Q}$ diagonally one slot and applying to $\mathbf{X}_M$, we have

$$\begin{bmatrix} 1 & \\ & \mathbf{Q} \end{bmatrix} \left[\begin{array}{c|c} \mathbf{x} & \begin{matrix} 0 & \mathbf{0}_{M-1}^t \\ \mathbf{0}_{M-1} & \mathbf{X}_{M-1} \end{matrix} \end{array}\right] = \begin{bmatrix} x(0) & \mathbf{0}_{M-1}^t \\ \mathbf{x}_q & \mathbf{C}_{M-1} \\ \mathbf{z} & \bigcirc \end{bmatrix} \begin{matrix} \} 1 \\ \} M-1 . \end{matrix} \tag{4.2}$$

Let now $\widehat{\mathbf{Q}}$ be an orthogonal matrix that eliminates $\mathbf{z}$ from the first column, while leaving $\mathbf{x}_q$ unaltered:

$$\widehat{\mathbf{Q}} \begin{bmatrix} x(0) & \mathbf{0}_{M-1}^t \\ \mathbf{x}_q & \mathbf{C}_{M-1} \\ \mathbf{z} & \bigcirc \end{bmatrix} = \begin{bmatrix} \alpha_{M-1} & \mathbf{0}_{M-1}^t \\ \mathbf{x}_q & \mathbf{C}_{M-1} \\ \mathbf{0} & \bigcirc \end{bmatrix}, \qquad \alpha_{M-1} > 0. \tag{4.3}$$

(Note that only the first column is affected). We then have the algorithm that follows.

ALGORITHM. Let $\mathbf{x}_q = [x_{q,1}, \ldots, x_{q,M-1}]^t$, and choose the rotation angles $\{\theta_k\}$ to fulfill the annihilations

$$\begin{bmatrix} \alpha_{k-1} \\ 0 \end{bmatrix} = \begin{bmatrix} \cos\theta_k & -\sin\theta_k \\ \sin\theta_k & \cos\theta_k \end{bmatrix} \begin{bmatrix} \alpha_k \\ x_{q,k} \end{bmatrix}, \qquad k = M-1, \ldots, 2, 1. \tag{4.4}$$

Then $\sin\theta_k$ is the kth reflection coefficient.

The verification is very similar to that presented in [47] and is omitted for brevity. Irrespective of the numerical errors accumulated in the calculation of $\begin{bmatrix} \alpha_{M-1} \\ \mathbf{x}_q \end{bmatrix}$, there always exists a sequence of rotation angles to fulfill the annihilation step (4.4). It is also easy to check that they will satisfy $|\sin\theta_k| < 1$ for all k, if the computed α_{M-1} ($= \sqrt{|x(0)|^2 + \|\mathbf{z}\|^2}$) is positive. Thus they are the exact reflection coefficients of a nearby Toeplitz matrix that remains positive definite. The above algorithm is thus a reliable numerical alternative to the Schur reduction steps outlined in §3.2.

Remark. If $\widetilde{\mathbf{Q}}$ is the product of successive rotations from (4.4), then one may observe that $\widetilde{\mathbf{Q}}$ completes the orthogonal triangularization of (4.3):

$$\widetilde{\mathbf{Q}} \begin{bmatrix} \alpha_{M-1} & \mathbf{0}_{M-1}^t \\ \mathbf{x}_q & \mathbf{C}_{M-1} \end{bmatrix} = \mathbf{C}_M. \tag{4.5}$$

The algorithm may then be compared with some previous square-root versions.

1. Delosme and Ipsen [31] first rotate $\mathbf{X}_M$ into upper triangular form $\mathbf{C}_M$, and then find an orthogonal matrix $\widetilde{\mathbf{Q}}^t$, which eliminates all but the first element of the first row:

$$\widetilde{\mathbf{Q}}^t \mathbf{C}_M = \begin{bmatrix} \alpha_{M-1} & \mathbf{0}_{M-1}^t \\ \mathbf{x}_q & \mathbf{C}_{M-1} \end{bmatrix}. \tag{4.6}$$

This is simply the inverse of (4.5), whence $\widetilde{\mathbf{Q}}^t$ splits into $M-1$ elementary rotations furnishing the reflection coefficients. Whereas (4.6) requires $M(M-1)/2$ rotations/annihilations, the step (4.4) requires only $M-1$ annihilations, and thus is faster. For both methods though, the computational load is dominated by the initial orthogonal transformation steps, and one may show that the initial steps of both algorithms yield identical computational counts.

2. One may also convert the triangularization step into a fast order-recursive algorithm that directly extracts the reflection coefficients; see Rialan and Scharf [32].

Their order recursions, however, use hyperbolic rotations, and hence the error growth factors may not be as favorable as for the above two algorithms that use exclusively orthogonal transformations.

5. Displacement inertia $(2,1)$ and Hankel grammians. We begin this section with an interpolation problem. Suppose we are given data $h_0, h_1, \ldots, h_{M-1}$ and $r_0, r_1, \ldots, r_{M-1}$. The problem is to find an l_2 sequence whose first M terms begin with $h_0, \ldots, h_{M-1}$ and whose tail end is chosen to be compatible with the data r_k according to

$$r_k = \sum_{l=0}^{\infty} h_l h_{k+l}, \qquad k = 0, 1, \ldots, M-1. \tag{5.1}$$

This problem was studied by Mullis and Roberts [33] in the context of linear system approximation and model reduction; a solution was shown to exist if and only if the $M \times M$ matrix

$$\begin{bmatrix} r_0 & r_1 & \cdots & r_{M-1} \\ r_1 & r_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & r_1 \\ r_{M-1} & \cdots & r_1 & r_0 \end{bmatrix} - \begin{bmatrix} h_0 & 0 & \cdots & 0 \\ h_1 & h_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ h_{M-1} & \cdots & h_1 & h_0 \end{bmatrix} \begin{bmatrix} h_0 & h_1 & \cdots & h_{M-1} \\ 0 & h_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & h_1 \\ 0 & \cdots & 0 & h_0 \end{bmatrix} \tag{5.2}$$

is positive (semi) definite. Specifically, we have the following:

(i) If the matrix (5.2) is positive semidefinite but rank deficient, then the l_2 sequence $\{h_k\}$ is uniquely determined, and the function $H(z) = \sum_{k=0}^{\infty} h_k z^{-k}$ has McMillan degree equal to the rank of (5.2).

(ii) If the matrix (5.2) is positive definite, then infinitely many l_2 sequences exist that interpolate the given data. All of them yield a function $H(z)$ of degree at least M.

Suppose then that (5.2) is positive (semi) definite, so that a solution exists. The constraint (5.1) may be written in matrix form as

$$\begin{bmatrix} r_0 & r_1 & \cdots & r_{M-1} \\ r_1 & r_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & r_1 \\ r_{M-1} & \cdots & r_1 & r_0 \end{bmatrix} = \begin{bmatrix} 0 & \cdots & 0 & h_0 & h_1 & \cdots \\ \vdots & \iddots & h_0 & h_1 & h_2 & \cdots \\ 0 & \iddots & \iddots & \vdots & \vdots & \iddots \\ h_0 & h_1 & \cdots & h_{M-1} & h_M & \cdots \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 & h_0 \\ \vdots & \iddots & h_0 & h_1 \\ 0 & \iddots & \iddots & \vdots \\ h_0 & h_1 & \cdots & h_{M-1} \\ h_1 & h_2 & \cdots & h_M \\ \vdots & \vdots & \iddots & \vdots \end{bmatrix}.$$

Upon noting that

$$\begin{bmatrix} h_0 & 0 & \cdots & 0 \\ h_1 & h_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ h_{M-1} & \cdots & h_1 & h_0 \end{bmatrix} \begin{bmatrix} h_0 & h_1 & \cdots & h_{M-1} \\ 0 & h_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & h_1 \\ 0 & \cdots & 0 & h_0 \end{bmatrix} = \begin{bmatrix} 0 & \cdots & 0 & h_0 \\ \vdots & \iddots & h_0 & h_1 \\ 0 & \iddots & \iddots & \vdots \\ h_0 & h_1 & \cdots & h_{M-1} \end{bmatrix}^2,$$

a straightforward calculation shows

$$(5.2) = \begin{bmatrix} h_1 & h_2 & h_3 & \cdots \\ h_2 & h_3 & h_4 & \iddots \\ \vdots & \vdots & \vdots & \iddots \\ h_M & h_{M+1} & h_{M+2} & \cdots \end{bmatrix} \begin{bmatrix} h_1 & h_2 & \cdots & h_M \\ h_2 & h_3 & \cdots & h_{M+1} \\ h_3 & h_4 & \cdots & h_{M+2} \\ \vdots & \iddots & \iddots & \vdots \end{bmatrix}, \tag{5.3}$$

which yields the grammian of a Hankel form. We now summarize in the property that follows.

PROPERTY 5.1. Let the data $h_0, \ldots, h_{M-1}$ and $r_0, \ldots, r_{M-1}$ be given. The matrix (5.2) is positive (semi) definite if and only if it may be written as the Grammian of a Hankel form as in (5.3), where $\{h_k\}$ is any l_2 sequence that satisfies the given interpolation problem.

Remark. A particular instance is obtained when $\{h_k\}$ happens to be a finite length sequence, in which case

$$(5.3) = \begin{bmatrix} h_1 & \cdots & h_{n-M+1} & \cdots & h_{n-1} & h_n \\ h_2 & \cdots & \vdots & \iddots & h_n & 0 \\ \vdots & \cdots & h_{n-1} & \iddots & \iddots & \vdots \\ h_M & \cdots & h_n & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} h_1 & h_2 & \cdots & h_M \\ \vdots & \vdots & \vdots & \vdots \\ h_{n-M+1} & \cdots & h_{n-1} & h_n \\ \vdots & \iddots & \iddots & 0 \\ h_{n-1} & h_n & \iddots & \vdots \\ h_n & 0 & \cdots & 0 \end{bmatrix}.$$

By a simple permutation operation, this may be written as

$$(5.3) = \begin{bmatrix} h_n & h_{n-1} & \cdots & h_{n-M+1} & \cdots & h_1 \\ 0 & h_n & \ddots & \vdots & \cdots & h_2 \\ \vdots & \ddots & \ddots & h_{n-1} & \ddots & \vdots \\ 0 & \cdots & 0 & h_n & \cdots & h_M \end{bmatrix} \begin{bmatrix} h_n & 0 & \cdots & 0 \\ h_{n-1} & h_n & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ h_{n-M+1} & \cdots & h_{n-1} & h_n \\ \vdots & \ddots & \ddots & \vdots \\ h_1 & h_2 & \cdots & h_M \end{bmatrix}.$$

This in turn yields the grammian formed from a prewindowed Toeplitz matrix. This latter form underlies the development of most fast recursive least-squares filtering algorithms. Irrespective of the number of rows of this prewindowed Toeplitz matrix, its grammian is always of the form (5.2).

The basic result we show is Theorem 5.2.

THEOREM 5.2. *Let* $\mathbf{P}$ *be symmetric and positive definite. Then* $\mathbf{P}$ *has displacement inertia* $(2,1)$, *i.e.,*

$$(5.4) \qquad \begin{bmatrix} \mathbf{P} & \mathbf{0}_M \\ \mathbf{0}_M^t & 0 \end{bmatrix} - \begin{bmatrix} 0 & \mathbf{0}_M^t \\ \mathbf{0}_M & \mathbf{P} \end{bmatrix} = \mathbf{a}_1 \mathbf{a}_1^t + \mathbf{a}_2 \mathbf{a}_2^t - \mathbf{b}_1 \mathbf{b}_1^t$$

if and only if $\mathbf{P}^{-1}$ *is the grammian of a Hankel form.*

Although the "if" clause has perhaps been known for some time, concrete statements are difficult to trace prior to Slock [36].[2] The "only if" clause was later claimed as well [39], but the supporting arguments were overly complicated and incomplete. We give direct proofs of both clauses.

For convenience, we recall the following Schur complement inversion formulas.

IDENTITY 5.3. Let $\mathbf{P}$ be partitioned as

$$\mathbf{P} = \begin{bmatrix} P_1 & \mathbf{p}^t \\ \mathbf{p} & \mathbf{P}_2 \end{bmatrix},$$

[2]Actually, the "if" clause is well known with respect to the classical displacement residue $\mathbf{P} - \mathbf{Z}\,\mathbf{P}\,\mathbf{Z}^t$. The set of positive definite $\mathbf{P}$ having displacement inertia $(2,1)$ with respect to the residue (5.4) is, of course, a subset of those having displacement inertia at most $(2,1)$ with respect to the classical residue. Thus the "only if" clause does not apply to the classical residue.

where P_1 is scalar. Then

$$\mathbf{P} = \begin{bmatrix} P_1 \\ \mathbf{p} \end{bmatrix} \frac{[P_1 \;\; \mathbf{p}^t]}{P_1} + \begin{bmatrix} 0 & \mathbf{0}_{M-1}^t \\ \mathbf{0}_{M-1} & \mathbf{P}_2 - \mathbf{p}\mathbf{p}^t/P_1 \end{bmatrix}.$$

Let $\mathbf{S} \triangleq \mathbf{P}^{-1}$ be partitioned as

$$\mathbf{S} = \begin{bmatrix} S_1 & \mathbf{s}^t \\ \mathbf{s} & \mathbf{S}_2 \end{bmatrix},$$

where S_1 is scalar. Then

$$\mathbf{S}_2^{-1} = \mathbf{P}_2 - \mathbf{p}\mathbf{p}^t/P_1.$$

For the "if" clause of Theorem 5.2, suppose $\mathbf{S} \triangleq \mathbf{P}^{-1}$ is the grammian of a Hankel form, as in (5.3). Partition $\mathbf{S}$ as

$$\mathbf{S} = \begin{bmatrix} & & * \\ & \mathbf{S}_1 & \vdots \\ * & \cdots & * \end{bmatrix} = \begin{bmatrix} * & \cdots & * \\ \vdots & \mathbf{S}_2 & \\ * & & \end{bmatrix}, \tag{5.5}$$

where $\mathbf{S}_1$ and $\mathbf{S}_2$ are, respectively, the $(M-1) \times (M-1)$ upper-left and lower-right submatrices of $\mathbf{S}$. Because $\mathbf{S}$ is the grammian of a Hankel form, it is easy to verify that

$$\mathbf{S}_1 = \mathbf{S}_2 + \begin{bmatrix} h_1 \\ \vdots \\ h_{M-1} \end{bmatrix} [h_1 \;\; \cdots \;\; h_{M-1}].$$

By applying the matrix inversion lemma, it follows that $\mathbf{S}_1^{-1}$ and $\mathbf{S}_2^{-1}$ differ by a rank one term:

$$\mathbf{S}_2^{-1} - \mathbf{S}_1^{-1} = \mathbf{c}\,\mathbf{c}^t. \tag{5.6}$$

Here the $M-1$-element vector $\mathbf{c}$ is given by

$$\mathbf{c} = \frac{\mathbf{S}_2^{-1}\mathbf{h}}{\sqrt{1 + \mathbf{h}^t \mathbf{S}_2^{-1} \mathbf{h}}}, \qquad \mathbf{h} = \begin{bmatrix} h_1 \\ \vdots \\ h_{M-1} \end{bmatrix}.$$

Let now $\begin{bmatrix} 1 \\ \mathbf{a} \end{bmatrix}$ be proportional to the first column of $\mathbf{P} = \mathbf{S}^{-1}$:

$$\mathbf{S} \begin{bmatrix} 1 \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \alpha \\ \mathbf{0}_{M-1} \end{bmatrix}.$$

The scalar $\alpha = [1 \;\; \mathbf{a}^t]\,\mathbf{S} \begin{bmatrix} 1 \\ \mathbf{a} \end{bmatrix}$ is clearly positive. By Identity 5.3, we may write

$$\mathbf{S}^{-1} = \mathbf{P} = \begin{bmatrix} 1 \\ \mathbf{a} \end{bmatrix} \frac{[1 \;\; \mathbf{a}^t]}{\alpha} + \begin{bmatrix} 0 & \mathbf{0}_{M-1}^t \\ \mathbf{0}_{M-1} & \mathbf{S}_2^{-1} \end{bmatrix}. \tag{5.7}$$

In the same way, if we let $\begin{bmatrix} \mathbf{b} \\ 1 \end{bmatrix}$ be proportional to the last column of $\mathbf{S} = \mathbf{P}^{-1}$, i.e.,

$$\mathbf{S} \begin{bmatrix} \mathbf{b} \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{M-1} \\ \beta \end{bmatrix},$$

then β is positive and we may write

$$\mathbf{P} = \begin{bmatrix} 1 \\ \mathbf{b} \end{bmatrix} \frac{[\mathbf{b}^t \ 1]}{\beta} + \begin{bmatrix} \mathbf{S}_1^{-1} & \mathbf{0}_{M-1} \\ \mathbf{0}_{M-1}^t & 0 \end{bmatrix}. \tag{5.8}$$

Upon combining (5.6), (5.7), and (5.8), we find

(5.9)

$$\begin{aligned} \begin{bmatrix} \mathbf{P} & \mathbf{0}_M \\ \mathbf{0}_M^t & 0 \end{bmatrix} - \begin{bmatrix} 0 & \mathbf{0}_M^t \\ \mathbf{0}_M & \mathbf{P} \end{bmatrix} &= \begin{bmatrix} 1 \\ \mathbf{a} \\ 0 \end{bmatrix} \frac{[1 \ \mathbf{a}^t \ 0]}{\alpha} - \begin{bmatrix} 0 \\ \mathbf{b} \\ 1 \end{bmatrix} \frac{[0 \ \mathbf{b}^t \ 1]}{\beta} + \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \mathbf{S}_2^{-1} - \mathbf{S}_1^{-1} & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ \mathbf{a} \\ 0 \end{bmatrix} \frac{[1 \ \mathbf{a}^t \ 0]}{\alpha} + \begin{bmatrix} 0 \\ \mathbf{c} \\ 0 \end{bmatrix} [0 \ \mathbf{c}^t \ 0] - \begin{bmatrix} 0 \\ \mathbf{b} \\ 1 \end{bmatrix} \frac{[0 \ \mathbf{b}^t \ 1]}{\beta}, \end{aligned}$$

which shows that $\mathbf{P}$ has displacement inertia $(2,1)$.

For the "only if" clause, suppose (5.4) holds. Since $\mathbf{P}$ is positive definite by assumption, one may always apply a $\mathbf{J}$-orthogonal transformation to the generator vectors to distribute zeros as in the expression (5.9); cf. §3.2. The resulting $\begin{bmatrix} 1 \\ \mathbf{a} \end{bmatrix}$ is proportional to the first column of $\mathbf{P}$, while the resulting $\begin{bmatrix} \mathbf{b} \\ 1 \end{bmatrix}$ is proportional to the last column of $\mathbf{P}$. Equation (5.6) then still holds, i.e., the upper-left and lower-right $(M-1)\times(M-1)$ submatrices of $\mathbf{P}^{-1}$ necessarily differ by a rank-one term. Considering the partition of $\mathbf{S} = \mathbf{P}^{-1}$ as in (5.5), one obtains by the matrix inversion lemma

$$\mathbf{S}_2 - \mathbf{S}_1 = -\begin{bmatrix} h_1 \\ \vdots \\ h_{M-1} \end{bmatrix} [h_1 \ \cdots \ h_{M-1}], \qquad \text{with} \quad \begin{bmatrix} h_1 \\ \vdots \\ h_{M-1} \end{bmatrix} = \frac{\mathbf{S}_1 \mathbf{c}}{\sqrt{1 + \mathbf{c}^t \mathbf{S}_1 \mathbf{c}}}.$$

Let now $r_0, \ldots, r_{M-1}$ be the elements on the first row of $\mathbf{S}$. Consideration of the classical residue gives

$$\begin{aligned} \mathbf{S} - \mathbf{Z}\mathbf{S}\mathbf{Z}^t &= \begin{bmatrix} r_0 & \cdots & r_{M-1} \\ \vdots & \bigcirc & \\ r_{M-1} & & \end{bmatrix} + \begin{bmatrix} 0 & \mathbf{0}_{M-1}^t \\ \mathbf{0}_{M-1} & \mathbf{S}_2 - \mathbf{S}_1 \end{bmatrix} \\ &= \begin{bmatrix} r_0/\sqrt{r_0} & 0 & 0 \\ r_1/\sqrt{r_0} & r_1/\sqrt{r_0} & h_1 \\ \vdots & \vdots & \vdots \\ r_{M-1}/\sqrt{r_0} & r_{M-1}/\sqrt{r_0} & h_{M-1} \end{bmatrix} \\ &\quad \times \begin{bmatrix} 1 & \\ & -\mathbf{I}_2 \end{bmatrix} \begin{bmatrix} r_0/\sqrt{r_0} & r_1/\sqrt{r_0} & \cdots & r_{M-1}/\sqrt{r_0} \\ 0 & r_1/\sqrt{r_0} & \cdots & r_{M-1}/\sqrt{r_0} \\ 0 & h_1 & \cdots & h_{M-1} \end{bmatrix}. \end{aligned}$$

By applying finally the representation theorem (2.2), the matrix $\mathbf{S} = \mathbf{P}^{-1}$ may be written as

$$\mathbf{S} = \mathbf{P}^{-1} = \begin{bmatrix} r_0 & r_1 & \cdots & r_{M-1} \\ r_1 & r_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & r_1 \\ r_{M-1} & \cdots & r_1 & r_0 \end{bmatrix} - \begin{bmatrix} 0 & 0 & \cdots & 0 \\ h_1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ h_{M-1} & \cdots & h_1 & 0 \end{bmatrix} \begin{bmatrix} 0 & h_1 & \cdots & h_{M-1} \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & h_1 \\ 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Since $\mathbf{P}^{-1}$ is positive definite by assumption, Property 5.1 now shows that $\mathbf{P}^{-1}$ is the grammian of a Hankel form, which completes the proof.

The following corollary is essential to the stability study of the next section.

COROLLARY 5.4. *Let* $\mathbf{P}$ *be positive definite and let* $0 < \lambda \leq 1$. *Then* $\mathbf{P}$ *verifies*

$$\begin{bmatrix} \mathbf{P} & \mathbf{0}_M \\ \mathbf{0}_M^t & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & \mathbf{0}_M^t \\ \mathbf{0}_M & \mathbf{P} \end{bmatrix} = \mathbf{a}_1 \mathbf{a}_1^t + \mathbf{a}_2 \mathbf{a}_2^t - \mathbf{b}_1 \mathbf{b}_1^t \tag{5.10}$$

if and only if $\mathbf{P}^{-1}$ *is an exponentially weighted grammian of a Hankel form*

$$\mathbf{P}^{-1} = \begin{bmatrix} h_1 & h_2 & h_3 & \cdots \\ h_2 & h_3 & h_4 & \cdots \\ \vdots & \iddots & \iddots & \cdots \\ h_M & h_{M+1} & h_{M+2} & \cdots \end{bmatrix} \begin{bmatrix} 1 & & & \\ & \lambda & & \\ & & \lambda^2 & \\ & & & \ddots \end{bmatrix} \begin{bmatrix} h_1 & h_2 & \cdots & h_M \\ h_2 & h_3 & \cdots & h_{M+1} \\ h_3 & h_4 & \cdots & h_{M+2} \\ \vdots & \iddots & \iddots & \vdots \end{bmatrix}. \tag{5.11}$$

Proof. Observe that

$$\begin{aligned} &[1 \;\; (z/\sqrt{\lambda}) \;\; \cdots \;\; (z/\sqrt{\lambda})^M]\,(5.10) \begin{bmatrix} 1 \\ (w/\sqrt{\lambda}) \\ \vdots \\ (w/\sqrt{\lambda})^M \end{bmatrix} = (1 - zw)\, P(z/\sqrt{\lambda}, w/\sqrt{\lambda}) \\ &= A_1(z/\sqrt{\lambda})\, A_1(w/\sqrt{\lambda}) + A_2(z/\sqrt{\lambda})\, A_2(w/\sqrt{\lambda}) - B_1(z/\sqrt{\lambda})\, B_1(w/\sqrt{\lambda}). \end{aligned} \tag{5.12}$$

Thus upon setting

$$\begin{aligned} \overline{\mathbf{P}} &= \operatorname{diag}[1, 1/\sqrt{\lambda}, \ldots, 1/(\sqrt{\lambda})^{M-1}]\, \mathbf{P}\, \operatorname{diag}[1, 1/\sqrt{\lambda}, \ldots, 1/(\sqrt{\lambda})^{M-1}], \\ \overline{\mathbf{a}}_k &= \operatorname{diag}[1, 1/\sqrt{\lambda}, \ldots, 1/(\sqrt{\lambda})^{M}]\, \mathbf{a}_k, \\ \overline{\mathbf{b}}_1 &= \operatorname{diag}[1, 1/\sqrt{\lambda}, \ldots, 1/(\sqrt{\lambda})^{M}]\, \mathbf{b}_1, \end{aligned}$$

we find that $\overline{\mathbf{P}}$ remains positive definite and verifies

$$\begin{bmatrix} \overline{\mathbf{P}} & \mathbf{0}_M \\ \mathbf{0}_M^t & 0 \end{bmatrix} - \begin{bmatrix} 0 & \mathbf{0}_M^t \\ \mathbf{0}_M & \overline{\mathbf{P}} \end{bmatrix} = \overline{\mathbf{a}}_1 \overline{\mathbf{a}}_1^t + \overline{\mathbf{a}}_2 \overline{\mathbf{a}}_2^t - \overline{\mathbf{b}}_1 \overline{\mathbf{b}}_1^t.$$

This is equivalent to saying that $\overline{\mathbf{P}}^{-1}$ is the grammian of a Hankel form, so that

$$\begin{aligned} \mathbf{P}^{-1} &= \operatorname{diag}[1, 1/\sqrt{\lambda}, \ldots, 1/(\sqrt{\lambda})^{M-1}]\, \overline{\mathbf{P}}^{-1}\, \operatorname{diag}[1, 1/\sqrt{\lambda}, \ldots, 1/(\sqrt{\lambda})^{M-1}] \\ &= \begin{bmatrix} \overline{h}_1 & \overline{h}_2 & \overline{h}_3 & \cdots \\ \overline{h}_2/\lambda^{1/2} & \overline{h}_3/\lambda^{1/2} & \overline{h}_4/\lambda^{1/2} & \iddots \\ \vdots & \vdots & \vdots & \iddots \\ \overline{h}_M/\lambda^{(M-1)/2} & \overline{h}_{M+1}/\lambda^{(M-1)/2} & \overline{h}_{M+2}/\lambda^{(M-1)/2} & \cdots \end{bmatrix} \\ &\quad \cdot \begin{bmatrix} \overline{h}_1 & \overline{h}_2/\lambda^{1/2} & \cdots & \overline{h}_M/\lambda^{(M-1)/2} \\ \overline{h}_2 & \overline{h}_3/\lambda^{1/2} & \cdots & \overline{h}_{M+1}/\lambda^{(M-1)/2} \\ \overline{h}_3 & \overline{h}_4/\lambda^{1/2} & \cdots & \overline{h}_{M+2}/\lambda^{(M-1)/2} \\ \vdots & \iddots & \iddots & \vdots \end{bmatrix} \end{aligned}$$

for some l_2 sequence $\{\overline{h}_k\}$. If we set

$$h_k = \overline{h}_k/\lambda^{(k-1)/2}, \qquad k = 1, 2, 3, \ldots,$$

the above expression for $\mathbf{P}^{-1}$ is equivalent to

$$\mathbf{P}^{-1} = \begin{bmatrix} h_1 & h_2 & h_3 & \cdots \\ h_2 & h_3 & h_4 & \iddots \\ \vdots & \vdots & \vdots & \iddots \\ h_M & h_{M+1} & h_{M+2} & \cdots \end{bmatrix} \begin{bmatrix} 1 & & & \\ & \lambda & & \\ & & \lambda^2 & \\ & & & \ddots \end{bmatrix} \begin{bmatrix} h_1 & h_2 & \cdots & h_M \\ h_2 & h_3 & \cdots & h_{M+1} \\ h_3 & h_4 & \cdots & h_{M+2} \\ \vdots & \iddots & \iddots & \vdots \end{bmatrix}$$

to show the result.

6. Numerical stability of fast least-squares algorithms. We begin this section with a standard recursive least-squares estimation problem. After a review of some numerical considerations, we then move to fast algorithms to show how the results of §5 serve to identify the constraints behind backward consistency and stable error propagation.

For the standard recursive least-squares problem, one is given a sequence of column vectors $\{\mathbf{x}(k)\}_{k=0}^{n}$ and a scalar sequence $\{y(k)\}_{k=0}^{n}$. The problem is to determine the weight vector $\mathbf{w}$, which minimizes

$$\sum_{k=0}^{n} \lambda^{n-k} [y(k) - \mathbf{x}^t(k)\,\mathbf{w}]^2.$$

Of interest are solutions computed recursively in time. Thus suppose the data matrix $\mathbf{X}(n+1)$ and the "reference" vector $\mathbf{y}(n+1)$ may be obtained as

$$\mathbf{X}(n+1) = \begin{bmatrix} \lambda^{1/2}\,\mathbf{X}(n) \\ \mathbf{x}^t(n+1) \end{bmatrix}, \qquad \mathbf{y}(n+1) = \begin{bmatrix} \lambda^{1/2}\,\mathbf{y}(n) \\ y(n+1) \end{bmatrix}. \tag{6.1}$$

Suppose we have the vector $\mathbf{w}(n)$, which minimizes $\|\Lambda(n)\,[\mathbf{y}(n) - \mathbf{X}(n)\,\mathbf{w}(n)]\|^2$, with $\Lambda(n) = \operatorname{diag}[\lambda^{n/2}, \ldots, \lambda^{1/2}, 1]$. Upon appending new data as in (6.1), the updated $\mathbf{w}(n+1)$ is, of course, available from the recursive least-squares algorithm:

$$\begin{aligned} \mathbf{g}(n+1) &= \frac{\mathbf{P}(n)\,\mathbf{x}(n+1)}{\lambda + \mathbf{x}^t(n+1)\,\mathbf{P}(n)\,\mathbf{x}(n+1)}, \\ \epsilon(n+1) &= y(n+1) - \mathbf{x}^t(n+1)\,\mathbf{w}(n), \\ \mathbf{w}(n+1) &= \mathbf{w}(n) + \mathbf{g}(n+1)\,\epsilon(n+1), \\ \mathbf{P}(n+1) &= \lambda^{-1}\,[\mathbf{I} - \mathbf{g}(n+1)\,\mathbf{x}^t(n+1)]\,\mathbf{P}(n) \\ &= [\mathbf{X}^t(n+1)\,\mathbf{X}(n+1)]^{-1}. \end{aligned} \tag{6.2}$$

The vector $\mathbf{g}(n+1)$ is known as the Kalman gain vector; its computation involves matrix operations requiring order M^2 operations. Fast least-squares algorithms exploit the low displacement rank of $\mathbf{P}$ (when applicable) to reduce the computational complexity.

The numerical behavior of these recursions is well documented [34], [39], [42]; we recall some basic notions that are easy to establish, to provide a smooth transition into the study of fast algorithms.

First note that the critical link is the matrix recursion from (6.2). Its numerical properties were investigated by Ljung and Ljung [34] using the error propagation approach. For this approach, suppose at time n (say) a perturbation has been introduced into $\mathbf{P}(n)$, giving $\widetilde{\mathbf{P}}(n)$. One then examines how such a perturbation influences the future evolution of the system. Specifically, set

$$\begin{aligned} \tilde{\mathbf{g}}(k+1) &= \frac{\widetilde{\mathbf{P}}(k)\,\mathbf{x}(k+1)}{\lambda + \mathbf{x}^t(k+1)\,\widetilde{\mathbf{P}}(k)\,\mathbf{x}(k+1)} \qquad k > n \\ \widetilde{\mathbf{P}}(k+1) &= \lambda^{-1}\,[\mathbf{I}_M - \tilde{\mathbf{g}}(k+1)\,\mathbf{x}^t(k+1)]\,\widetilde{\mathbf{P}}(k) \end{aligned} \tag{6.3}$$

and compare $\widetilde{\mathbf{P}}(k)$ with the "true" matrix $\mathbf{P}(k)$ that would have been obtained from (6.2). Note that both recursions use the same driving sequence $\mathbf{x}(n+1)$, $\mathbf{x}(n+2)$, ..., so that the comparison has some meaning. The following result is standard [34], [42].

PROPERTY 6.1. The difference $\mathbf{P}(k) - \widetilde{\mathbf{P}}(k)$ tends exponentially fast to the zero matrix as $k \to \infty$, with base given by $\lambda < 1$, provided:

(i) The perturbed matrix $\widetilde{\mathbf{P}}(n)$ remains symmetric and positive definite; and

(ii) the driving sequence $\mathbf{x}(\cdot)$ is persistently exciting in the sense that there exists some integer N and positive constants a and b such that

$$a\mathbf{I} \le \mathbf{x}(k)\,\mathbf{x}^t(k) + \cdots + \mathbf{x}(k+N)\,\mathbf{x}^t(k+N) \le b\mathbf{I} \quad \text{for all } k. \tag{6.4}$$

The verification is straightforward and instructive. First, note that $\widetilde{\mathbf{P}}(n)$ is symmetric and positive definite if and only if we may write $\widetilde{\mathbf{P}}^{-1}(n) = \widetilde{\mathbf{X}}(n)\,\widetilde{\mathbf{X}}(n)$, using some "perturbed" data matrix $\widetilde{\mathbf{X}}(n)$. Next, note that the two recursions (6.2) and (6.3) may be written as

$$\begin{aligned} \mathbf{P}^{-1}(k+1) &= \lambda\,\mathbf{P}^{-1}(k) + \mathbf{x}(k+1)\,\mathbf{x}^t(k+1) \\ \widetilde{\mathbf{P}}^{-1}(k+1) &= \lambda\,\widetilde{\mathbf{P}}^{-1}(k) + \mathbf{x}(k+1)\,\mathbf{x}^t(k+1) \end{aligned} \qquad k > n. \tag{6.5}$$

Since the perturbation is introduced at time n, it follows easily that

$$\mathbf{P}^{-1}(k) - \widetilde{\mathbf{P}}^{-1}(k) = \lambda^{k-n}\,[\mathbf{X}^t(n)\,\mathbf{X}(n) - \widetilde{\mathbf{X}}^t(n)\,\widetilde{\mathbf{X}}(n)] \quad \text{for all } k \ge n,$$

which decays exponentially fast to zero as $k \to \infty$, provided $\lambda < 1$. This shows that error propagation in the recursion (6.5) is exponentially stable. Under the persistence of the excitation condition (6.4), it may then be shown [42] that $\mathbf{P}(k) - \widetilde{\mathbf{P}}(k)$ also tends to zero exponentially fast as $k \to \infty$. Thus the iteration (6.2) is exponentially stable with respect to perturbations that preserve symmetry and positive definiteness.

Conversely, if the perturbations destroy either symmetry or positive definiteness, or if the driving sequence $\{\mathbf{x}(\cdot)\}$ is not persistently exciting, the recursion (6.2) can approach exponentially unstable error propagation. See [39] or [42] for more detail on this point.

We should remark that condition (i) of Property 6.1 is a restatement of the principle of backward consistency: $\widetilde{\mathbf{P}}(n)$ remains symmetric and positive definite if and only if this same $\widetilde{\mathbf{P}}(n)$ could have been obtained by first perturbing $\mathbf{X}(n)$ to $\widetilde{\mathbf{X}}(n)$ and then running the identical algorithm in exact arithmetic. We shall see that this same characterization carries over to the family of fast least-squares algorithms.

Note finally that in practice, perturbations are introduced at each iteration. It is a standard result that exponentially stable error propagation implies bounded error accumulation with respect to successive perturbations. Analyses showing such bounded error accumulation can be found in [43], [44].

Fast least-squares algorithms derive from the low displacement rank of $\mathbf{P}(\cdot)$. Specifically, suppose that at each iteration n, the vector $\mathbf{x}(n)$ contains delayed versions of a scalar sequence $\{x(\cdot)\}$

$$\mathbf{x}(n) = [x(n) \;\; x(n-1) \;\; \cdots \;\; x(n-M+1)]^t \quad \text{for all } n,$$

and that $x(n) = 0$ for $n < 0$. This implies that $\mathbf{P}^{-1}(n)$ is the grammian of an exponentially weighted prewindowed Toeplitz matrix:

$$\mathbf{P}^{-1}(n) = [\cdot]^t \begin{bmatrix} \lambda^n & & & \\ & \ddots & & \\ & & \lambda & \\ & & & 1 \end{bmatrix} \begin{bmatrix} x(0) & 0 & \cdots & 0 \\ x(1) & x(0) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ x(M-1) & \cdots & x(1) & x(0) \\ \vdots & \ddots & \ddots & \vdots \\ x(n) & x(n-1) & \cdots & x(n-M+1) \end{bmatrix}. \tag{6.6}$$

By permutation operations, this may be rearranged as the grammian of an exponentially weighted Hankel form as per (5.11). Thus by Corollary 5.4, $\mathbf{P}(n)$ has displacement inertia $(2,1)$, i.e.,

$$\begin{bmatrix} \mathbf{P}(n) & \mathbf{0} \\ \mathbf{0}^t & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & \mathbf{0}^t \\ \mathbf{0} & \mathbf{P}(n) \end{bmatrix} = \mathbf{a}_1 \mathbf{a}_1^t + \mathbf{a}_2 \mathbf{a}_2^t - \mathbf{b}_1 \mathbf{b}_1^t. \tag{6.7}$$

Conversely, if $\mathbf{P}(\cdot)$ satisfies (6.7), then $\mathbf{P}^{-1}(\cdot)$ may be written in the form (6.6), allowing possibly for an infinite number of rows in $\mathbf{X}(n)$.

It proves convenient to partition the generator vectors according to

$$\mathbf{a}_1 = \frac{1}{\sqrt{\alpha(n)}} \begin{bmatrix} 1 \\ \mathbf{a}(n) \\ 0 \end{bmatrix}, \quad \mathbf{a}_2 = \frac{\sqrt{\lambda}}{\sqrt{\gamma(n)}} \begin{bmatrix} 0 \\ \mathbf{c}(n) \\ 0 \end{bmatrix}, \quad \mathbf{b}_1 = \frac{\sqrt{\lambda}}{\sqrt{\beta(n)}} \begin{bmatrix} 0 \\ \mathbf{b}(n) \\ 1 \end{bmatrix},$$

which is generically possible; cf. §3.2. The formulas for the first and last generator vectors may then be summarized as

$$\begin{bmatrix} 1 \\ \mathbf{a}(n) \end{bmatrix} = \mathbf{P}(n) \begin{bmatrix} \alpha(n) \\ \mathbf{0} \end{bmatrix}, \quad \begin{bmatrix} \mathbf{b}(n) \\ 1 \end{bmatrix} = \mathbf{P}(n) \begin{bmatrix} \mathbf{0} \\ \beta(n) \end{bmatrix}. \tag{6.8a}$$

For the middle vector, consider the partitioning $\mathbf{P}(n) = \begin{bmatrix} \mathbf{P}_1 & \mathbf{p} \\ \mathbf{p}^t & P_2 \end{bmatrix}$, where P_2 is scalar. Then

$$\begin{aligned} \mathbf{c}(n) &= [\mathbf{P}_1 - \mathbf{p}\mathbf{p}^t/P_2] \begin{bmatrix} x(n) \\ \vdots \\ x(n-M+2) \end{bmatrix}, \\ \gamma(n) &= 1 - [x(n) \;\; \cdots \;\; x(n-M+2)]\, \mathbf{c}(n). \end{aligned} \tag{6.8b}$$

The scalar γ is called a conversion factor or likelihood variable. A readable account of these formulas may be found in Haykin [35].

The trick behind fast least-squares algorithms is as follows: Upon adding the next datum sample $x(n+1)$, the updated data matrix $\mathbf{X}(n+1)$ remains in prewindowed Toeplitz form, and hence the updated grammian $\mathbf{P}(n+1)$ continues to have

displacement inertia $(2,1)$. Thus $\mathbf{P}(n+1)$ satisfies an equation of the form (6.7), and the generator vectors relate to $\mathbf{P}(n+1)$ via the formulas (6.8) with the time index incremented to $n+1$. Thus, rather than directly updating $\mathbf{P}(n) \rightarrow \mathbf{P}(n+1)$, requiring order M^2 operations, it is possible to directly update the generator vectors $[\mathbf{a}_1(n), \mathbf{a}_2(n), \mathbf{b}_1(n)] \rightarrow [\mathbf{a}_1(n+1), \mathbf{a}_2(n+1), \mathbf{b}_1(n+1)]$ in order M operations. Algorithms that directly update the generator vectors are called fast transversal filters [4], [36], [45]. They are algebraically equivalent to the matrix recursion (6.2) provided $\mathbf{P}(\cdot)$ has displacement inertia $(2,1)$.

Many of these algorithms suffer unstable error propagation.

To understand the origins of this problem, consider again the error propagation question. Suppose that $\mathbf{P}(n)$ is perturbed to $\widetilde{\mathbf{P}}(n)$, which remains symmetric and positive definite. Suppose also that the displacement inertia of $\widetilde{\mathbf{P}}(n)$ remains $(2,1)$, a nontrivial assumption. Then it is completely described by three generator vectors $[\tilde{\mathbf{a}}_1(n), \tilde{\mathbf{a}}_2(n), \tilde{\mathbf{b}}_1(n)]$ by formulas akin to (6.8) above. The property $\mathbf{P}(k) - \widetilde{\mathbf{P}}(k) \overset{k\rightarrow\infty}{\rightarrow} \bigcirc$ then implies that $[\mathbf{a}_1(k), \mathbf{a}_2(k), \mathbf{b}_1(k)] - [\tilde{\mathbf{a}}_1(k), \tilde{\mathbf{a}}_2(k), \tilde{\mathbf{b}}_1(k)] \overset{k\rightarrow\infty}{\rightarrow} \mathbf{0}$.[3]

The key assumption here is that $\widetilde{\mathbf{P}}(n)$ retains its displacement inertia. This is equivalent to saying that its generator vectors could have been obtained by first perturbing the scalar sequence $\{x(\cdot)\}$ to some new scalar sequence $\{\tilde{x}(\cdot)\}$ and then running the computations in exact arithmetic. It follows then that backward consistency plus persistent excitation is sufficient for stable error propagation, whether we consider fast or full least-squares algorithms.

It thus suffices to examine consistency in greater detail. At any iteration n, introduce the generator functions

$$A_1(z) = \frac{[1 \;\; (z/\sqrt{\lambda}) \;\; \cdots \;\; (z/\sqrt{\lambda})^M]}{\sqrt{\alpha(n)}} \begin{bmatrix} 1 \\ \mathbf{a}(n) \\ 0 \end{bmatrix},$$

$$A_2(z) = \frac{[1 \;\; (z/\sqrt{\lambda}) \;\; \cdots \;\; (z/\sqrt{\lambda})^M]}{\sqrt{\gamma(n)}} \begin{bmatrix} 0 \\ \mathbf{c}(n) \\ 0 \end{bmatrix} \times \sqrt{\lambda},$$

$$B_1(z) = \frac{[1 \;\; (z/\sqrt{\lambda}) \;\; \cdots \;\; (z/\sqrt{\lambda})^M]}{\sqrt{\beta(n)}} \begin{bmatrix} 0 \\ \mathbf{b}(n) \\ 1 \end{bmatrix} \times \sqrt{\lambda}.$$

For convenience, the exponential weighting is absorbed directly. The displacement equation (6.7) is then equivalent to the generator equation

$$(1 - zw)\, P(z/\sqrt{\lambda}, w/\sqrt{\lambda}) = A_1(z)\, A_1(w) + A_2(z)\, A_2(w) - B_1(z)\, B_1(w)$$

(cf. (5.12)). From Theorem 2.1, $\mathbf{P}(n)$ is positive definite if and only if we may find a lossless function $\mathbf{U}(z)$ of degree M, which fulfills

$$\begin{bmatrix} A_1(z) \\ A_2(z) \end{bmatrix} = \mathbf{U}(z)\, B_1(z).$$

---

[3] Actually, this implies only that $[\mathbf{a}_1(k), \mathbf{a}_2(k), \mathbf{b}_1(k)]$ and $[\tilde{\mathbf{a}}_1(k), \tilde{\mathbf{a}}_2(k), \tilde{\mathbf{b}}_1(k)]$ converge to generator vectors of the same matrix, i.e., that they are asymptotically related by a J-orthogonal transformation. If both sets of generator vectors have zeros distributed as in (6.8), then their difference must converge to zero.

This gives $\mathbf{U}(z) = \frac{1}{B_1(z)} \begin{bmatrix} A_1(z) \\ A_2(z) \end{bmatrix}$, and the constraint that $\mathbf{U}(z)$ be lossless of degree M may be rephrased as

$$A_1(z)\,A_1(z^{-1}) + A_2(z)\,A_2(z^{-1}) = B_1(z)\,B_1(z^{-1}) \quad \text{for all } z, \tag{6.9a}$$

$$B_1(z) \text{ has no zeros in } |z| \geq 1, \tag{6.9b}$$

$$A_1(z) \text{ and } A_2(z) \text{ have no common zeros in } |z| \leq 1. \tag{6.9c}$$

The first two of these constraints were deduced by Slock [36], who dubbed them the "FTF manifold"; the third constraint appears to have been overlooked. This manifold characterizes the set of "reachable" generator vectors in exact arithmetic, as the scalar sequence $\{x(\cdot)\}$ varies over all possibilities. Thus, as long as the computed generators obey these constraints, they are indistinguishable from the exact generators obtained by first perturbing the sequence $\{x(\cdot)\}$, and then running all computations in exact arithmetic.[4]

It is easy to see that locally independent errors in the computed generator vectors lead to constraint (6.9a) being violated. To any such error, backward consistency no longer applies. As a result of (6.9a) being generically violable, all fast transversal filter algorithms have parasitic dynamics in the time update formulas [38], [39]. These parasitic dynamics theoretically vanish in exact arithmetic, but are generically present once finite precision arithmetic is introduced and account for the numerical instability of such algorithms. See [38] and [39] for more detail on this point.

To see how consistency can be enforced, recall that an equivalent parametrization of the displacement structure is available from the rotation angles of the lossless system $\mathbf{U}(z)$ of Fig. 5. We address now the computation of such rotation angles. Let $\mathbf{R}(n)$ be the $(M-1) \times (M-1)$ principal submatrix of the Cholesky factor of $\mathbf{P}^{-1}(n)$. Then define two vectors as

$$\mathbf{x}_f(n) = \begin{bmatrix} x_{f,0}(n) \\ \vdots \\ x_{f,M-2}(n) \end{bmatrix} = \mathbf{R}(n-1)\,\mathbf{a}(n)$$

and

$$\epsilon_b(n) = \begin{bmatrix} \epsilon_{b,0}(n) \\ \vdots \\ \epsilon_{b,M-2}(n), \end{bmatrix} = \mathbf{R}(n)\,\mathbf{c}(n),$$

where $\mathbf{a}(n)$ and $\mathbf{c}(n)$ are taken from the first two generator vectors via (6.8). It may be shown that the vector $\epsilon_b(n)$ has norm less than one, and that the vector

$$\begin{bmatrix} \epsilon_b(n) \\ \gamma^{1/2}(n) \end{bmatrix}$$

has unit norm.

IDENTITY 6.2. Set $\alpha_{M-1}(n) = \alpha(n)$ and $\gamma^{1/2}_{M-1}(n) = \gamma^{1/2}(n)$. Then determine a sequence of rotation angles $\{\phi_k\}$ and $\{\theta_k\}$ from the annihilations

$$\begin{bmatrix} \alpha^{1/2}_{k-1}(n) \\ 0 \end{bmatrix} = \begin{bmatrix} \cos\phi_{k-1} & \sin\phi_{k-1} \\ -\sin\phi_{k-1} & \cos\phi_{k-1} \end{bmatrix} \begin{bmatrix} \alpha^{1/2}_k(n) \\ x_{f,k-1}(n) \end{bmatrix},$$

$$\begin{bmatrix} \gamma^{1/2}_{k-1}(n) \\ 0 \end{bmatrix} = \begin{bmatrix} \cos\theta_{k-1} & \sin\theta_{k-1} \\ -\sin\theta_{k-1} & \cos\theta_{k-1} \end{bmatrix} \begin{bmatrix} \gamma^{1/2}_k(n) \\ \epsilon_{b,k-1}(n) \end{bmatrix}, \qquad k = M-1, \ldots, 1. \tag{6.10}$$

[4] One subtlety does arise here: the "starting" time for the perturbed sequence $\{\tilde{x}(\cdot)\}$ may have to be readjusted for the perturbed generators to be reached. It may be shown [40, §5.b] that the property of stable error propagation is insensitive to such a change in the starting time.

These rotation angles are precisely those of the lossless system $\mathbf{U}(z)$ of Fig. 5 obtained by applying the Schur reduction steps to $\mathbf{P}(n)$.

This identity was established in [40], based on order recursions that appeared earlier in [4]. We recall that $\mathbf{U}(z)$ of Fig. 5 is lossless of degree M for all rotation angles fulfilling $|\sin\theta_k| < 1$ and $|\sin\phi_k| < 1$. This in turn may be rephrased as $\alpha_{M-1} > 0$ and $\gamma_{M-1} > 0$. As such, this parametrization is inherently consistent with a positive definite $\mathbf{P}(n)$ of displacement inertia $(2,1)$. Thus numerical errors in their determination via (6.10) are indistinguishable from having first perturbed the scalar sequence $\{x(\cdot)\}$.

It turns out that the vectors $\mathbf{x}_f(n)$ and $\epsilon_b(n)$, plus the scalar $\alpha^{1/2}(n)$, are all that is needed to perform time updates. With the arrival of $x(n+1)$, the variables $\mathbf{x}_f(n)$, $\epsilon_b(n)$, and $\alpha^{1/2}(n)$ may be time-updated to $\mathbf{x}_f(n+1)$, $\epsilon_b(n+1)$, and $\alpha^{1/2}(n+1)$ using orthogonal transformations determined from the annihilation steps (6.10). Such an algorithm was first obtained by Proudler, McWhirter, and Shepherd [41] from a fast QR decomposition approach. That the rotation angles of this algorithm coincide with a Schur parametrization of the displacement structure came as a surprise. The numerical stability properties of such an algorithm are detailed in [40], based on the above identity.

We close this section with some comments on accuracy aspects. Considerable attention has been drawn to this question in the context of fast least-squares algorithms. Most analyses derive expressions for the accumulated error variances in the time-propagated variables. Comparisons of different algorithms are complicated when the propagated variables are not the same, as happens when comparing stabilized transversal filters, lattice filters, or fast QR algorithms. Among those algorithms that are backward consistent [37]–[40], a meaningful approach is to guage accuracy in terms of the required equivalent perturbation in the input sequence.

To this end, let us review how to reconstruct a perturbed input sequence from the perturbed $\widetilde{\mathbf{P}}(\cdot)$. If $\widetilde{\mathbf{P}}(\cdot)$ remains positive definite of displacement inertia $(2,1)$, the matrix

$$\overline{\mathbf{P}}^{-1} = \operatorname{diag}[1, \sqrt{\lambda}, \ldots, (\sqrt{\lambda})^{M-1}]\, \widetilde{\mathbf{P}}^{-1}\, \operatorname{diag}[1, \sqrt{\lambda}, \ldots, (\sqrt{\lambda})^{M-1}]$$

may be written in the form (5.2) for some perturbed data set $\tilde{h}_1, \ldots, \tilde{h}_{M-1}$ and $\tilde{r}_0, \ldots, \tilde{r}_{M-1}$. Append further data $\tilde{h}_M, \ldots, \tilde{h}_{M+k}$ and $\tilde{r}_M, \ldots, \tilde{r}_{M+k}$ until the augmented matrix

$$\begin{bmatrix} \tilde{r}_0 & \tilde{r}_1 & \cdots & \tilde{r}_{M+k} \\ \tilde{r}_1 & \tilde{r}_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & r_1 \\ \tilde{r}_{M+k} & \cdots & \tilde{r}_1 & \tilde{r}_0 \end{bmatrix} - \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \tilde{h}_1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \tilde{h}_{M+k} & \cdots & \tilde{h}_1 & 0 \end{bmatrix} \begin{bmatrix} 0 & \tilde{h}_1 & \cdots & \tilde{h}_{M+k} \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \tilde{h}_1 \\ 0 & \cdots & 0 & 0 \end{bmatrix} \tag{5.11}$$

is positive semidefinite with precisely one zero eigenvalue. Such a "singular extension" is generically possible [33]. Let $\mathbf{d} = [1, d_1, \ldots, d_{M+k}]^t$ lie in the nullspace of matrix (5.11). It may be shown that the polynomial $D(z) = 1 + d_1\, z^{-1} + \cdots + d_{M+k}\, z^{-(M+k)}$ has all zeros in $|z| < 1$ [33]. Then determine $N(z) = n_1\, z^{-1} + \cdots + n_{M+k}\, z^{-(M+k)}$ according to

$$\begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_{M+k} \end{bmatrix} = \begin{bmatrix} \tilde{h}_1 & 0 & \cdots & 0 \\ \tilde{h}_2 & \tilde{h}_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \tilde{h}_{M+k-1} & \cdots & h_2 & h_1 \end{bmatrix} \begin{bmatrix} 1 \\ d_1 \\ \vdots \\ d_{M+k-1} \end{bmatrix}.$$

The function

$$\tilde{H}(z) = \frac{N(z)}{D(z)} = \sum_{k=1}^{\infty} \tilde{h}_k z^{-k}$$

yields an l_2 sequence $\{\tilde{h}_k\}$ whose Hankel form gives a grammian equal to $\overline{\mathbf{P}}^{-1}$. It may be shown that all such l_2 sequences may be placed in one-to-one correspondence with all singular extensions from procedure [33].

At the same time, set

$$h_1 = x(n), \quad h_2 = \lambda^{1/2} x(n-1), \quad h_3 = \lambda x(n-2), \quad \cdots$$

corresponding to the true data. Then natural distance measures may be taken as follows:

(i) The l_2-norm of the error, as given by

$$\left(\sum_{k=1}^{\infty} (h_k - \tilde{h}_k)^2 \right)^{1/2}.$$

(ii) The Hankel norm of the error, which is the largest singular value of the matrix

$$\begin{bmatrix} h_1 & h_2 & h_3 & \cdots \\ h_2 & h_3 & h_4 & \cdots \\ h_3 & h_4 & h_5 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} - \begin{bmatrix} \tilde{h}_1 & \tilde{h}_2 & \tilde{h}_3 & \cdots \\ \tilde{h}_2 & \tilde{h}_3 & \tilde{h}_4 & \cdots \\ \tilde{h}_3 & \tilde{h}_4 & \tilde{h}_5 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

In either case, bounds on the error norm in terms of the accumulated errors in the time update formulas are difficult to establish in closed form. Such a study would be desirable to draw more meaningful accuracy comparisons between backward consistent fast least-squares algorithms. This is a topic for future work.

7. Concluding remarks. Our principal theme has been the interplay between displacement structures, lossless systems, and numerical algorithm design. From these connections, it is easily seen how consistency is destroyed in structured algorithms deriving from low displacement rank: if the generator vectors are manipulated directly, then numerical errors destroy the structural constraint of (2.8). The equivalence of Theorem 2.1, which comes from lossless inverse scattering theory [8], leads to a feasible method of imposing consistency: The lossless function $\mathbf{U}(z)$ can be parametrized by a sequence of rotation angles, such that consistency holds irrespective of the exact values of these rotation angles.

The examples of §§4 and 6 illustrate how the rotation angles of $\mathbf{U}(z)$ can be calculated from square-root algorithms operating directly on the data given to the problem. The role of these relations in resolving the numerical instability problem of fast least-squares algorithms [37]–[40] suggests that similar progress should be feasible with respect to other numerical algorithms based on low displacement rank theory [16], [46]. Numerical algorithm design from this perspective is rarely elementary, and our attention has accordingly been restricted to the simplest occurrences of low displacement rank. It is hoped that the concepts exposed here will encourage a revised look at low displacement rank algorithms, their consistency constraints, and the development of backward error analyses.

Acknowledgment. The authors are indebted to an anonymous reviewer for many constructive criticisms on the first draft.

REFERENCES

[1] T. KAILATH, S. Y. KUNG, AND M. MORF, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979), pp. 395–407.

[2] ———*Displacement ranks of a matrix*, Bull. Amer. Math. Soc., 1 (1979), pp. 769–773.

[3] B. FRIEDLANDER, M. MORF, T. KAILATH, AND L. JUNG, *New inversion formulas for matrices classified in terms of their distance from Toeplitz matrices*, Linear Algebra Appl., 27 (1979), pp. 31–60.

[4] H. LEV-ARI, T. KAILATH, AND J. CIOFFI, *Least-squares adaptive lattice and transversal filters: A unified geometric theory*, IEEE Trans. Information Theory, 30 (1984), pp. 222–236.

[5] H. LEV-ARI AND T. KAILATH, *Lattice filter parametrization and modeling of nonstationary processes*, IEEE Trans. Information Theory, 30 (1984), pp. 2–16, p. 878.

[6] C. GUEGUEN, *An introduction to displacement ranks and fast related algorithms*, USGM NATO Summer Course on Digital Signal Processing, course 12, in Signal Processing, 2, Lacoume et al., eds., North Holland, Amsterdam, 1987.

[7] F. DESBOUVRIES, *Rangs de Déplacement et Algorithmes Rapides*, Ph.D. thesis, Ecole Nationale Supérieure des Télécommunications, Paris, January 1991.

[8] D. ALPAY, P. DEWILDE, AND H. DYM, *On the existence and construction of solutions to the partial lossless inverse scattering problem, with applications to estimation theory*, IEEE Trans. Information Theory, 35 (1989), pp. 1184–1205.

[9] H. DYM, *J-Contractive Matrix Functions, Reproducing Kernel Hilbert Spaces, and Interpolation*, American Mathematical Society, Regional Conferences Series in Mathematics 71, Providence, RI, 1989.

[10] A. H. SAYED, *Displacement Structure in Signal Processing and Mathematics*, Ph.D. thesis, Stanford University, Stanford, CA, August 1992.

[11] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.

[12] B. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[13] D. C. YOULA, *On the factorization of rational matrices*, IRE Trans. Information Theory, 7 (1961), pp. 172–189.

[14] I. SCHUR, *Über Potenzreihen die im Innern des Einheitskreises beschränkt sind*, J. für die Reine und Angewandte Mathematik, 147 (1917), pp. 205–232, 148 (1918), pp. 122–145; English translation: Operator Theory: Advances and Applications, 18, pp. 31–59, 61–87, Birkhäuser-Verlag, Basel, 1986.

[15] P. DEWILDE AND E. DEPRETTERE, *The generalized Schur algorithm: Approximation and hierarchy*, in Topics in Operator Theory and Interpolation, OT-29, I. Gohberg, ed., Birkhäuser-Verlag, Basel, 1988, pp. 97–116.

[16] K. DIEPOLD AND R. PAULI, *Schur parametrization of symmetric indefinite matrices based on a circuit theoretic model*, Archiv für Elektronic und Übertragungstechnik, 45 (1991), pp. 375–385.

[17] V. M. POTAPOV, *The multiplicative structure of J-contractive matrix functions*, Amer. Math. Soc., Translation Series 2, 15 (1960), pp. 131–243.

[18] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[19] V. BELEVITCH, *Classical Network Theory*, Holden-Day, San Francisco, 1968.

[20] A. FETTWEIS, *Stability and pseudo-passivity in wave digital filters*, IEEE Trans. Circuit Theory, 19 (1972), pp. 668–673.

[21] E. F. DEPRETTERE AND P. DEWILDE, *Orthogonal cascade realization of real multiport digital filters*, Internat. J. Circuit Theory Appl., 8 (1980), pp. 245–257.

[22] S. K. RAO AND T. KAILATH, *Orthogonal digital filters for* VLSI *implementation*, IEEE Trans. Circuits and Systems, 31 (1984), pp. 933–945.

[23] P. P. VAIDYANATHAN AND S. K. MITRA, *A general family of multivariable digital lattice filters*, IEEE Trans. Circuits and Systems, 32 (1985), pp. 1234–1245.

[24] ———, *A unified structural interpretation of some well-known stability test procedures for linear systems*, Proc. IEEE, 75 (1987), pp. 478–497.

[25] P. P. VAIDYANATHAN, *Theory and design of M-channel maximally decimated quadrature mirror filter banks, having the perfect reconstruction property*, IEEE Trans. Acoustics, Speech, and Signal Processing, 35 (1987), pp. 476–492.

[26] ———, *The discrete-time bounded real lemma in digital filtering*, IEEE Trans. Circuits and Systems, 32 (1985), pp. 918–924.

[27] T. KAILATH, A. VIEIRA, AND M. MORF, *Inverses of Toeplitz operators, innovations, and orthogonal polynomials*, SIAM Rev., 20 (1978) pp. 106–119.

[28] T. KAILATH, *A theorem of I. Schur and its impact on modern signal processing*, in I. Schur Methods in Operator Theory and Signal Processing, Operator Theory: Advances and Applications 18 (1976), I. Gohberg, ed., Birkhäuser-Verlag, Basel, pp. 9–30.

[29] G. HEINIG AND K. ROST, *Algebraic Methods for Toeplitz-like Matrices and Operators*, Akademie-Verlag, Berlin, 1984.

[30] A. H. GRAY, JR. AND J. D. MARKEL, *Linear Prediction of Speech*, Springer-Verlag, Berlin, 1976.

[31] J.-M. DELOSME AND C. F. IPSEN, *From Bareiss's algorithm to the stable computation of partial correlations*, J. Comput. Appl. Math., 27 (1989), pp. 53–91.

[32] C. P. RIALAN AND L. L. SCHARF, *Fast algorithms for computing QR and Cholesky factors of Toeplitz operators*, IEEE Trans. Acoustics, Speech, and Signal Processing, 36 (1988), pp. 1740–1748.

[33] C. T. MULLIS AND R. A. ROBERTS, *The use of second order information in the approximation of discrete-time linear systems*, IEEE Trans. Acoustics, Speech, and Signal Processing, 24 (1976), pp. 226–238.

[34] S. LJUNG AND L. LJUNG, *Error propagation properties of recursive least-squares adaptation algorithms*, Automatica, 21 (1985), pp. 157–167.

[35] S. HAYKIN, *Adaptive Filter Theory*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1991.

[36] D. T. M. SLOCK, *An overview of some recent advances in fast* RLS *algorithms*, in Algorithms and Parallel VLSI Architectures, E. Deprettere and A.-J. van der Veen, eds., Elsevier, Amsterdam, 1990.

[37] P. A. REGALIA, *Fast recursive least-squares filters: Error propagation, backward consistency, and lossless inverse scattering*, Proc. IEEE Benelux Pro-RISC workshop, Houthalen, Belgium, March 1993, pp. 15–21.

[38] P. A. REGALIA, *Numerical stability issues in fast least-squares adaptation algorithms*, Optical Engrg., 31 (1992), pp. 1144–1152.

[39] D. T. M. SLOCK, *The backward consistency concept and round-off error propagation dynamics in recursive least-squares algorithms*, Optical Engrg., 31, (1992), pp. 1153–1169.

[40] P. A. REGALIA, *Numerical stability properties of a QR-based fast least-squares algorithm*, IEEE Trans. Signal Processing, 41 (1993), pp. 2096–2109.

[41] I. K. PROUDLER, J. G. MCWHIRTER, AND T. J. SHEPHERD, *Fast* QRD*-based algorithms for least-squares linear prediction*, in Mathematics in Signal Processing, 2 (1990), J. G. McWhirter, ed., Clarendon Press, Oxford, pp. 465–488.

[42] M. H. VERHAEGEN, *Round-off error propagation in four generally applicable recursive least-squares estimation schemes*, Automatica, 25 (1989), pp. 437–444.

[43] G. W. STEWART, *Error analysis of QR updating with exponential windowing*, Report CS-TR 2685, University of Maryland, College Park, 1991.

[44] H. LEUNG AND S. HAYKIN, *Stability of recursive* QRD*-LS algorithms using finite-precision systolic array implementation*, IEEE Trans. Acoustics, Speech, Signal Processing, 37 (1989), pp. 760–763.

[45] J. CIOFFI AND T. KAILATH, *Fast, recursive least-squares filters for adaptive filtering*, IEEE Trans. Acoustics, Speech, and Signal Processing 32 (1984), pp. 304–337.

[46] H. LEV-ARI AND T. KAILATH, *Triangular factorization of structured Hermitian matrices*, in I. Schur Methods in Operator Theory and Signal Processing, Operator Theory: Advances and Applications, 18 (1986), Birkhäuser-Verlag, Basel, pp. 301–324.

[47] P. A. REGALIA AND M. G. BELLANGER, *On the duality between fast QR methods and lattice methods in least-squares adaptive filtering*, IEEE Trans. Signal Processing, 39 (1991), pp. 375–388.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 565–578, April 1995

016

CONDITION ESTIMATION FOR MATRIX FUNCTIONS VIA THE SCHUR DECOMPOSITION*

ROY MATHIAS†

Abstract. We show how to cheaply estimate the Fréchet derivative and the condition number for a general class of matrix functions (the class includes the matrix sign function and functions that can be expressed as power series) via the Schur decomposition. In the case of the matrix sign function we also give a method to compute the Fréchet derivative exactly. We also show that often this general method, based on the Schur decomposition, when applied the matrix sign function and the matrix exponential, enables one to compute the function and estimate its condition number more cheaply than the various special techniques that exploit special properties of these two functions.

Key words. matrix function, condition estimation, matrix sign function, matrix exponential, Fréchet derivative, Sylvester equation, primary matrix function

AMS subject classifications. 15A12, 65F35, 65F99, 15A99

1. Introduction.

One way to compute $f(A)$ is via the Schur decomposition of A. We show that once one has the Schur decomposition one can cheaply obtain a condition estimate of $f(A)$ (roughly speaking, a bound on $\|f(A+E)-f(A)\|/\|E\|$ for $\|E\|$ small). There are other ways to compute $f(A)$ if f has special properties (for example, $f(A)=\exp(A), A^{1/2}, \operatorname{sgn}(A)$). These are sometimes computationally less expensive. However, we show that if one also wants a condition estimate then the Schur method is often cheaper. The direct solution of $Ax=b$ requires $O(n^3)$ flops, while a condition estimate can be obtained in an additional $O(n^2)$ flops. However, for the methods presented here (and indeed all known methods) the cost of computing $f(A)$ and the additional cost of a condition estimate are both $O(n^3)$.

In the rest of this section we discuss primary matrix functions (a general class of matrix function that includes both the matrix sign function and matrix functions defined by power series), mention the relationship between the Fréchet derivative, the directional derivative, and the condition number, and give a simple lemma on which our results are based.

In §2 we give a way to compute the Fréchet derivative of the matrix sign function and show that the cost of the evaluation of $\operatorname{sgn}(A)$ and estimation of the condition number is a little less by the Schur method than by the iterative methods proposed in [7, §3]. We also give an improvement of one of the methods in [7, §3].

In §3 we consider general (primary) matrix functions and show how one can obtain an estimate of the condition number for $f(A)$ for a fraction of the cost of computing $f(A)$. We then compare the Schur method for $\exp(A)$ with the scaling and squaring methods discussed in [7], [12]. The Schur method is cheaper if the spectral norm of A is at least 16, and more expensive otherwise.

Let $M_{m,n}$ denote the space of $m \times n$ complex matrices and define $M_n \equiv M_{n,n}$. Given D, a subset of the complex plane, let $D_n \subset M_n$ denote the set of matrices with spectrum contained in D. Let $\|\cdot\|$ denote the spectral norm ($\|A\| \equiv \sqrt{\lambda_{\max}(A^*A)}$),

* Received by the editors February 16, 1993; accepted for publication (in revised form) by N. J. Higham, March 4, 1994.

† Department of Mathematics, College of William & Mary, Williamsburg, Virginia 23187 (na.mathias@na-net.ornl.gov). This research was supported in part by National Science Foundation grant DMS-9201586 and was done while the author was visiting the Institute for Mathematics and Its Applications at the University of Minnesota.

and let $\|\cdot\|_F$ denote the Frobenius norm

$$\|A\|_F \equiv \sqrt{\sum_{ij} |a_{ij}|^2}.$$

Let f be analytic on a domain $D \subset C$ (D need not be simply connected). We define the *primary matrix function* associated with f on D_n as follows. If $A \in D_n$ is diagonalizable, $A = S\mathrm{diag}(\lambda_1, \ldots, \lambda_n)S^{-1}$, then $f(A) \equiv S\mathrm{diag}(f(\lambda_1), \ldots, f(\lambda_n))S^{-1}$, and if A is not diagonalizable then define $f(A)$ by continuity. The differentiability condition on f ensures that $f(A)$ is well defined. One can see from the definition that $f(A)A = Af(A)$. For a further discussion of primary matrix functions and an alternate definition see [5, §6.6].

We define the (*relative*) *condition number* of f at $A \in M_n$ by

$$\kappa_f(A) \equiv \left(\lim_{\delta \downarrow 0} \max_{\|E\|_F \leq \delta} \frac{\|f(A+E) - f(A)\|_F}{\delta} \right) \frac{\|A\|_F}{\|f(A)\|_F}. \tag{1.1}$$

This is equivalent to the (asymptotic) relative condition defined in [13, Definition 2]. We use $\|\cdot\|_F$ in this definition (although we could use any other norm) so that later we will be able to exploit the fact that $\|\cdot\|_F$ is derived from an inner product to estimate $\kappa_f(\cdot)$ by a power method. If one takes $f(A) = A^{-1}$ and uses the spectral norm rather than the Frobenius norm in (1.1), then one can show that the resulting relative condition number is $\kappa_f(A) = \|A\|\,\|A^{-1}\|$, i.e., what is usually meant by the condition number of A.

If we take $D = \{|z| < R\}$ and $f(z) = \sum_{i=1}^{\infty} a_i z^i$, where the series is convergent on D then $f(A) = \sum_{i=1}^{\infty} a_i A^i$, which is what is usually meant by a matrix function. The matrix exponential, $\exp(A) = \sum_{i=1}^{\infty} A^i/i!$ is one such function. However, there are matrix functions that cannot be expressed as a power series. For example, if we take $D = \{z : \mathrm{Re}(z) \neq 0\}$ and $f(z) = \mathrm{sign}(\mathrm{Re}(z))$, then $f(A) = \mathrm{sgn}(A)$ is the *sign* of the matrix A. Another way to define $\mathrm{sgn}(A)$ for $A \in D_n$ is to write $A = S(P \oplus N)S^{-1}$ where P and $-N$ have spectra in the open right half-plane, and define $\mathrm{sgn}(A) \equiv S[I \oplus (-I)]S^{-1}$. The matrix sign function separates the positive and negative invariant subspaces of a matrix and has various applications in the solution of Lyapunov and Ricatti equations [2], and recently has been used in parallel algorithms for the nonsymmetric eigenvalue problem; see, for example, [10].

It can be shown from [5, Thm. 6.6.14(3)], that if f is analytic then

$$\left.\frac{d}{dt} f(A + tE)\right|_{t=0} = p_A(A, E) \quad \text{for all} \quad A \in D_n, E \in M_n,$$

where p_A is a polynomial in A and E and is linear in E. Note that the condition in [5, Thm. 6.6.14] that D be simply connected is not necessary. Note also that the coefficients of p_A may depend on A, but are independent of E. From [5, Thm. 6.6.20(3)], one can show that that

$$\left\| \left.\frac{d^2}{dt^2} f(A + tE)\right|_{t=0} \right\| = O(\|E\|^2).$$

By combining these two facts one can show that the Fréchet derivative of f at A exists, and so is equal to the directional derivative. Let $L_f(A, E)$ denote the Fréchet derivative of f at A in the matrix direction E, then by the above

$$f(A + E) = f(A) + L_f(A, E) + O(\|E\|^2). \tag{1.2}$$

Define

$$\|L_f(A,\cdot)\| \equiv \max_{\|E\|_F \le 1} \|L_f(A,E)\|_F. \tag{1.3}$$

From this definition and (1.2) one can easily show that

$$\kappa_f(A) = \frac{\|A\|}{\|f(A)\|}\|L_f(A,\cdot)\|.$$

So to estimate, $\kappa_f(A)$ we need only estimate $\|A\|$, $\|f(A)\|$ and the induced norm $L_f(A,\cdot)$ with respect to the Frobenius norm on M_n. Because the Frobenius norm is derived from an inner product one can estimate $\kappa_f(A) = \|L_f(A,\cdot)\|$ by a power method ([7, (1.8-9)], or the slight improvement given in [12, (3.1)]) or a Lanczos method [12, §4]. The key point of all these algorithms is that one can estimate $\kappa_f(A)$ by evaluating $L_f(A,\cdot)$ a few times—usually twice is sufficient to get within an order of magnitude of $\kappa_f(A)$. In this paper we do not consider the details of these algorithms, rather we assume that two evaluations of $L_f(A,\cdot)$ are sufficient to estimate $\kappa_f(A)$. (This assumption is based on the satisfactory results obtained for the matrix exponential and logarithm in [6].)

In general there are no convenient formulae for $\|L_f(A,\cdot)\|$. Kenney and Laub showed [6, Lem. 2.1] that if f is given by a power series and if A is normal with eigenvalues λ_i, then

$$\|L_f(A,\cdot)\| = \max_{i,j=1,\dots,n} \frac{f(\lambda_i) - f(\lambda_j)}{\lambda_i - \lambda_j}, \tag{1.4}$$

where we take $[f(\lambda_i) - f(\lambda_j)]/[\lambda_i - \lambda_j] = f'(\lambda_i)$ if $\lambda_i = \lambda_j$. They showed, using a separate argument in [7, Thm. 3.2], that (1.4) is still valid for $f(z) = \operatorname{sign}(\operatorname{Re}(z))$ (on $\{\operatorname{Re}(z) \ne 0\}$) and normal A. Using [5, Thm. 6.6.14 3] one can show that (1.4) is true for any primary function and any normal matrix A. Thus the problem of condition estimation for primary matrix functions of normal matrices is solved. The numerical techniques proposed in this paper are for the nonnormal case.

We define the separation between $A \in M_m$ and $B \in M_n$ by

$$\operatorname{sep}(A,B) = \min_{X \in M_{m,n}, \|X\| \le 1} \|AX - XB\|.$$

Note that we use the spectral norm rather that the Frobenius norm. It is well known that $\operatorname{sep}(A,B) > 0$ if and only if A and B have no common eigenvalues [5, Thm. 4.4.6]. The most common way to solve a Sylvester equation is to first reduce the matrices on the left-hand side to quasitriangular form. One can check that if $A \in M_n$ and $B \in M_m$ are in quasitriangular form then, considering only the highest order terms, one can solve $AX + XB = C$ in

$$\min\left\{mn^2 + \frac{9}{2}m^2n, m^2n + \frac{9}{2}mn^2\right\} \tag{1.5}$$

flops. See [11] for the details. One can check that the quantity in (1.5) is maximized over $m, n \ge 0$ and $m + n = k$ at $m = n = k/2$ and that the corresponding value of (1.5) is $11k^3/16$.

Our results in the rest of the paper are based on the following simple lemma.

LEMMA 1.1. *Let*

$$W = \begin{pmatrix} A & C \\ 0 & B \end{pmatrix} \quad \text{and} \quad E = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix}$$

be conformally partitioned. Assume that A and B have disjoint spectra. Let P be the unique solution to

$$PA - BP = E_{21} \tag{1.6}$$

and set

$$S = \begin{pmatrix} I & -P^* \\ P & I \end{pmatrix}, \qquad \eta = \operatorname{sep}^{-1}(A, B)\|E\|. \tag{1.7}$$

Assume that $\eta \leq \frac{1}{3}$. Then

$$\|I - S\| \leq \eta, \tag{1.8}$$

$$\|S^* - S^{-1}\| \leq \frac{\eta^2}{1-\eta}, \tag{1.9}$$

and

$$\|S^{-1}(W+E)S - \hat{W}\| \leq 2\eta^2\|W\| + 3\eta\|E\|, \tag{1.10}$$

where

$$\hat{W} = \begin{pmatrix} A + CP + E_{11} & C + P^*B - AP^* + E_{12} \\ 0 & B - PC + E_{22} \end{pmatrix}. \tag{1.11}$$

Notice that if W is fixed then $\operatorname{sep}^{-1}(A, B)$ is a constant, though possibly very large, and so $\eta = O(\|E\|)$. In the next section where we derive an expression for the Fréchet derivative of the matrix sign function the fact that $\eta = O(\|E\|)$ is sufficient. However, in §4, where we use forward differences, it is necessary to remember that η contains a factor of $\operatorname{sep}^{-1}(A, B)$. The proof of (1.10) with η replaced by $O(\|E\|)$ is somewhat simpler than the proof of the more precise bound.

Let W be block upper triangular and let E be small. This lemma shows how to find S such that the 2,1 block of $S^{-1}(W+E)S$ is $O(\|E\|^2)$. Iterating this one can find T such that $T(W+E)T^{-1}$ is block upper triangular and $\|T - I\| \leq 2\eta$. This is precisely what was done in [14, §V.2.1 and Thm. V.2.1] (or see [15]).

Proof. The first inequality follows from the fact that P is linear in E:

$$\|I - S\| = \|P\| \leq \operatorname{sep}^{-1}(A, B)\|E_{12}\| \leq \operatorname{sep}^{-1}(A, B)\|E\| = \eta.$$

Let $\hat{P} = I - S$. Then $\hat{P} = -\hat{P}^*$, and so provided that $\|P\| < 1$, we have

$$\begin{aligned} \|S^* - S^{-1}\| &\leq \|(I - \hat{P})^* - \sum_{i=0}^{\infty} \hat{P}^i\| \\ &= \|\sum_{i=2}^{\infty} \hat{P}^i\| \\ &\leq \sum_{i=2}^{\infty} \|\hat{P}^i\| \end{aligned}$$

$$\leq \sum_{i=2}^{\infty} \eta^i$$
$$= \frac{\eta^2}{1-\eta}.$$

Now let us consider (1.10).

$$\|S^{-1}(W+E)S - \hat{W}\| \leq \|(S^{-1}-S^*)(W+E)S\| + \|S^*WS + E - \hat{W}\| + \|S^*ES - E\|.$$

The first term is bounded by

$$\|S^{-1}-S^*\|\,\|W+E\|\,\|S\| \leq \frac{\eta^2}{1-\eta}[1+\eta^2]^{1/2}(\|W\|+\|E\|).$$

For the second term, one can check that

$$\|S^*WS + E - \hat{W}\| = \left\| \begin{pmatrix} 0 & P^* \\ -P & 0 \end{pmatrix} W \begin{pmatrix} 0 & -P^* \\ P & 0 \end{pmatrix} \right\| \leq \eta^2\|W\|.$$

Finally, writing $S = I + (I-S)$, one can bound the last term by

$$2\|I-S\|\,\|S\|\,\|E\| + \|I-S\|^2\,\|E\| \leq (2\eta(1+\eta^2)^{1/2} + \eta^2)\|E\|.$$

Inequality (1.10) follows from the three bounds and the assumption that $\eta \leq \frac{1}{3}$. □

Note that although the matrix S defined in (1.7) is not unitary, if we take

$$\hat{S} = \begin{pmatrix} I & -P^* \\ P & I \end{pmatrix} \begin{pmatrix} C_1^{-1} & 0 \\ 0 & C_2^{-1} \end{pmatrix}, \tag{1.12}$$

where $C_1^*C_1 = (I + P^*P)$ and $C_2^*C_2 = (I + PP^*)$, then $\hat{S}$ is unitary and (1.10) still holds, (in fact, a slightly stronger inequality is true). The cost of computing $\hat{S}$ by (1.12) is not much greater than by (1.7) since we can take C_1 to be the Cholesky factor of $(I + P^*P)$ and similarly for C_2. From a numerical point of view it would be better to compute $\hat{S}$ from a QR factorization of $\binom{I}{P}$. See [1] for a discussion of this point.

2. Matrix sign function. In this section we give a method of computing the Fréchet derivative of the matrix sign function, and compare the cost of this method with the cost of computing it by an iterative method (Newton's method), and the cost of estimating it by a forward difference.

THEOREM 2.1. *Let*

$$W = \begin{pmatrix} A & C \\ 0 & B \end{pmatrix} \quad \text{and} \quad E = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix}. \tag{2.1}$$

Assume that A and $-B$ have spectra in the open right half-plane. Then

$$L_{\text{sgn}}(W,E) = \begin{pmatrix} -XP & 2P^* - \delta X \\ 2P & PX \end{pmatrix}, \tag{2.2}$$

where P, X, and δX are the solutions to the Sylvester equations

$$PA - BP = E_{21}, \tag{2.3}$$
$$AX - XB = 2C, \tag{2.4}$$
$$A\delta X - \delta X B = 2\delta C - \delta A X + X\delta B \tag{2.5}$$

and

$$\delta A = E_{11} + CP, \tag{2.6}$$

$$\delta B = E_{22} - PC, \tag{2.7}$$

$$\delta C = E_{12} - AP^* + P^*B. \tag{2.8}$$

Note that if W is given by (2.1) and A and $-B$ have spectra in the open right half-plane then

$$\operatorname{sgn}(W) = \begin{pmatrix} I & X \\ 0 & -I \end{pmatrix},$$

where X is the solution of $AX - XB = 2C$.

Proof. Notice that (2.3), (2.6), and (2.7) imply that $\|\delta A\|$ and $\|\delta B\|$ are $O(\|E\|)$. Thus, for $\|E\|$ sufficiently small, $(A + \delta A)$ and $-(B + \delta B)$ will have spectra in the open right half-plane.

We prove the result by showing that whenever $\|E\|$ is sufficiently small,

$$\operatorname{sgn}(W + E) - \operatorname{sgn}(W) = \begin{pmatrix} -XP & 2P^* - \delta X \\ 2P & PX \end{pmatrix} + O(\|E\|^2).$$

Set

$$S = I + \begin{pmatrix} 0 & -P^* \\ P & 0 \end{pmatrix}.$$

Then

$$\begin{aligned} \operatorname{sgn}(W + E) &= S \operatorname{sgn}(S^{-1}(W + E)S)S^{-1} \\ &= S \operatorname{sgn}(S^*(W + E)S)S^* + O(\|E\|^2) \qquad (2.9) \\ &= S \operatorname{sgn} \begin{pmatrix} A + \delta A & C + \delta C \\ 0 & B + \delta B \end{pmatrix} S^{-1} + O(\|E\|^2) \\ &= S \begin{pmatrix} I & X + \delta X \\ 0 & -I \end{pmatrix} S^T + O(\|E\|^2) \qquad (2.10) \\ &= \begin{pmatrix} I - XP & X + \delta X + 2P^* \\ 2P & -I + PX \end{pmatrix} + O(\|E\|^2) \qquad (2.11) \\ &= \operatorname{sgn}(W) + \begin{pmatrix} -XP & \delta X + 2P^* \\ 2P & PX \end{pmatrix} + O(\|E\|^2). \end{aligned}$$

We have used (1.9) and the fact that $\operatorname{sgn}(\cdot)$ is twice differentiable at W for (2.9), the remark preceding this proof and Lemma 1.1 for (2.10), and have multiplied out and collected second order terms for (2.11). □

This result is superficially similar to Byers' result [3, Thm. 2] since both involve the Schur decomposition and $\operatorname{sep}(A, B)$. However, Byers was interested in the backward stability of Newton's Method (which is independent of the conditioning of the matrix sign function) for computing the positive and negative invariant subspaces of a matrix (which is closely related to computing the matrix sign function).

We now estimate the computational cost of various methods of computing $\operatorname{sgn}(A)$ and estimating its condition. These are summarized in Table 1. The point is that if

TABLE 2.1
Estimated flop counts for various methods of computing $\operatorname{sgn}(A)$ *and estimating* $\kappa_{\operatorname{sgn}}(A)$.

	$\operatorname{sgn}(A)$	Preprocessing	$L_{\operatorname{sgn}}(A,E)$	$\operatorname{sgn}(A)$ and $L_{\operatorname{sgn}}(A,E)$ twice
Schur	$29\frac{2}{3}n^3$	$2\frac{11}{16}n^3$	$52n^3/16$	$39n^3$
Newton (2.12) and Fwd. diff. (2.13)	14–$20n^3$	—	14–$20n^3$	42-$60n^3$
Newton (2.12) and (2.14)	14–$20n^3$	—	28–$40n^3$	70–$100n^3$
Newton (2.12) and (2.14) iterated to convergence	14–$20n^3$	—	14–$20n^3$	42–$60n^3$

one requires only $\operatorname{sgn}(A)$ then one can see, from column 1, that it would be cheaper to use Newton's method. On the other hand, if one also requires a condition estimate for the sign function at A then, from column 4, it is cheaper to compute the entire Schur decomposition (using the QR algorithm) and compute the sign function and a condition estimate using this.

Note that all the methods under consideration are iterative and that the flop counts depend heavily on the number of iterations required for convergence. So the totals given in Table 1 are merely estimates. For this reason we have rounded the final total for the Schur method. One should also note that the table merely gives flop counts—it does not differentiate between flops associated with matrix multiplication, matrix inversion, or applying the QR iteration. To get a true idea of the relative efficiency of the different methods one should consider these factors and the computer architecture.

Theorem 2.1 outlines a method to compute the Fréchet derivative of the sign function at a matrix W of the form given in (2.1). The following algorithm presents this in algorithmic form.

ALGORITHM 2.2. Given W, E as in (2.1) and X the solution to $AX - XB = 2C$
1. solve $PA - BP = E_{21}$ for P.
2. Set $\delta A = E_{11} + CP$, $\delta B = E_{22} - PC$, $\delta C = E_{12} - AP^* + P^*B$.
3. Solve $A\delta X - \delta X B = 2\delta C - \delta A X + X \delta B$ for δX.
4. Compute PX and XP. L_{sgn} is now given by (2.2).

Let us determine the computational cost, assuming that A and B are quasi-upper triangular, as is generally be the case, and that X, the solution to (2.4) is known. Since X is independent of E it need only be computed once and we include this cost in the cost preprocessing to be discussed later. Assume also that A is $m \times m$, B is $l \times l$, and that $l + m = n$.

The cost of forming the products $CP, PC, AP^*, P^*B, \delta AX, X\delta B, XP$, and PX is $8(l^2m + m^2l) = 8lmn \leq 2n^3$ flops (since $l + m = n$). It is also necessary to solve a Sylvester equation of the form $AY - YB = Z$ twice. The cost of this, using for example a variant of the Bartels–Stewart algorithm [4, Algorithm 7.6.3] for quasi-upper triangular A and B, is at most $11n^3/16$ flops per Sylvester equation from the discussion after (1.5). Thus the cost of each evaluation of $L_{\operatorname{sgn}}(W, E)$ is at most $52n^3/16$, after the necessary preprocessing. If m and l are not equal, then the cost

will be less than this.

Now we consider the cost of computing $\operatorname{sgn}(M)$ via the Schur decomposition and the cost of the preprocessing for the computation of L_{sgn}. Computation of orthogonal U and upper triangular T such that $M = UTU^*$ requires approximately $25n^3$ flops [4, Algorithm 7.5.2]. Then $\operatorname{sgn}(T)$ requires a further $2n^3/3$ flops (using a variant of Parlett's algorithm for computing a function of an upper triangular matrix). Finally computing $\operatorname{sgn}(M) = U\operatorname{sgn}(T)U^*$ uses a further $4n^3$ flops. To estimate the condition number for $\operatorname{sgn}(T)$ (which is the same as that for $\operatorname{sgn}(M)$) we must find an orthogonal V such that $W = VTV^*$ is upper triangular as well as being of the form required by Theorem 2.1. This can be done in at most $2n^3$ flops. For example, by using the algorithm for ordering the eigenvalues of a triangular matrix given in [1]. Finally, solving $AX - XB = 2C$ takes $11n^3/16$ flops.

Let us compare the cost of our method with that of a method based on the scaled Newton iteration

$$S_{k+1} = [\gamma_k S_k + (\gamma_k S_k)^{-1}]/2, \quad S_0 = M. \tag{2.12}$$

If the γ_k are suitably chosen then S_k converges quadratically to $\operatorname{sgn}(M)$. See [8] for a discussion of several simple and effective choices of γ_k. Each iteration of (2.12) requires $2n^3$ flops for the computation of an inverse. The cost of computing γ_k is generally negligible.

The cost of this method and the associated condition estimation scheme is strongly dependent on the number of iterations of (2.12) required for convergence. This number depends on the matrix M and is least if M is normal and has clustered eigenvalues. Based on the results in [8] and our own experiments, at least seven, and frequently more, iterations are required if M is nonnormal and has widely scattered eigenvalues. We use 7–10 iterations in our comparisons. Note that in [8, Example 1] there are examples where 15 iterations are required, even with the optimal choice of the γ_k. In these cases (2.12) will be much more expensive than the Schur approach. Another factor that will affect the cost comparison is the number of times L_{sgn} needs to be evaluated to obtain an acceptable estimate of $\kappa_{\operatorname{sgn}}(M)$. We take this number to be 2, but it may be larger, and this would greatly favor the Schur method, which can evaluate $L_{\operatorname{sgn}}(M, \cdot)$ cheaply after the preprocessing. It is rarely less than 2.

One can estimate $L_{\operatorname{sgn}}(M, E)$ by a forward difference

$$L_{\operatorname{sgn}}(M, E) \approx \frac{\operatorname{sgn}(M + tE) - \operatorname{sgn}(M)}{t}, \tag{2.13}$$

for some suitably chosen value of t. Given $\operatorname{sgn}(M)$ this requires one extra evaluation of $\operatorname{sgn}(\cdot)$. Thus, under our assumptions, this method of condition estimation combined with (2.12) requires $42 - 60n^3$ flops to compute $\operatorname{sgn}(M)$ and estimate $\kappa_{\operatorname{sgn}}(M)$. The cost of the Schur method for computing $\operatorname{sgn}(M)$ and estimating $\kappa_{\operatorname{sgn}}(M) = \kappa_{\operatorname{sgn}}(W)$ is about $39n^3$ flops, which is a little less.

One can also compute $L_{\operatorname{sgn}}(M, E)$ by a Newton iteration without the use of forward differences [7, Thm. 3.3]

$$\delta S_{k+1} = [\gamma_k\, \delta S_k - \gamma_k^{-1}\, S_k^{-1} \delta S_k S_k^{-1}]/2, \quad \delta S_0 = E, \tag{2.14}$$

where S_k^{-1} are given by (2.12). This iteration is also quadratically convergent and takes about as many iterations to converge as (2.12), but because two matrix multiplications are required at each step, it is twice as expensive as (2.13) for computing

L_{sgn}. However, we can approximate the iteration (2.14) by

$$\delta S_k - S_k^{-1}\delta S_k S_k^{-1} \approx \delta S_k + [(S_k + t\delta S_k)^{-1} - S_k^{-1}]/t. \tag{2.15}$$

Since S_k^{-1} has already been calculated in (2.12), this iteration requires only one extra matrix inversion per step, and so is half the cost of (2.14). Using the Neumann series, one can check that the error in the approximation (2.15) is $O(t)$, so if t is suitably chosen then not too much error is incurred at each step. The discussion in [7, p. 501] shows that an error incurred at a given step of iteration (2.14) does not grow as the iteration proceeds. Thus if t is chosen chosen appropriately, one can obtain a good estimate of $L_{\mathrm{sgn}}(M, E)$ by the iteration (2.15) for the same cost as (2.13). If one only wants an order of magnitude estimate of $\|L_{\mathrm{sgn}}(M, \cdot)\|$ it may not be necessary to iterate (2.15) to convergence. This idea needs to be tested in practice, especially in the case than $\kappa_{\mathrm{sgn}}(M)$ is large, but it may prove to be less expensive than the Schur method.

Note that in *exact* arithmetic, the method in Theorem 2.1 gives the exact value of $L_{\mathrm{sgn}}(M, E)$, the iteration (2.14) converges to the exact value, (2.13) gives an $O(t)$ approximation, and (2.15) combined with a suitable stopping criterion will terminate at an $O(t)$ approximation.

3. Primary matrix functions. In this section we consider estimating $\kappa_f(T)$ for a primary matrix function (defined in §1) and a block upper triangular matrix T with well-separated main diagonal blocks. Such matrices arise when $f(M)$ is computed via the Schur decomposition as described in the next paragraph.

A popular way to compute $f(M)$ is to first find an orthogonal U such that

$$UMU^* = T = [T_{ij}]_{i,j=1}^k,$$

where T is block upper triangular and the main diagonal blocks of T are well separated, i.e.,

$$\delta_0 = \min_{i \neq j} \operatorname{sep}(T_{ii}, T_{jj}) \tag{3.1}$$

is "not too small." Then compute $f(T) = [F_{ij}]_{i,j=1}^k$, which will again be block upper triangular, by (block) diagonals. Start with the main diagonal, where one uses $F_{ii} = f(T_{ii})$, and compute each superdiagonal in turn using the fact that $Tf(T) = f(T)T$. This requires the solution of a Sylvester equation of the form

$$T_{ii}F_{ij} - F_{ij}T_{jj} = R_{ij}$$

for each block F_{ij} with $i < j$. This computation is well conditioned because, by assumption, T_{ii} and T_{jj} are well separated. See [4, §11.1.4] for further details of this method.

If T is block upper triangular, let $T_{i,[i]}$ denote the block row to the right of T_{ii} and let $T_{[i,i]}$ denote the square block in the bottom right-hand corner below T_{ii}. That is

$$T = \begin{pmatrix} * & * & * \\ 0 & T_{ii} & T_{i,[i]} \\ 0 & 0 & T_{[i,i]} \end{pmatrix}. \tag{3.2}$$

Define

$$\delta = \min_{i=1,\ldots,(k-1)} \operatorname{sep}(T_{ii}, T_{[i,i]}). \tag{3.3}$$

One can check that

$$\operatorname{sep}(T_{ii}, T_{[i,i]}) \leq \min_{j<i} \operatorname{sep}(T_{ii}, T_{jj})$$

and so $\delta \leq \delta_0$. We assume that δ is not too small in relation to $\|E\|$. Then by applying Lemma 1.1 $(k-1)$ times, we can compute a matrix U that is orthogonal (up to $O(\|E\|^2)$) such that $U^*(T+E)U = \hat{T} + O(\|E\|^2)$. Note, there is no need to actually form U, we can keep it as a product of matrices of the form (1.7). We can then compute $f(\hat{T})$ cheaply by the method outlined in the previous paragraph. (This computation will be stable because, by assumption, $\|E\|$ is small in comparison to $\delta_0 \geq \delta$.) We can then form $f_{\text{est}}(T+E) \equiv Uf(\hat{T})U^*$, which, by arguments similar to those in the proof of Theorem 2.1, is an $O(\|E\|^2)$ approximation to $f(T+E)$. So given E, we have the forward difference estimate of $L_f(T,E)$:

$$L_f(T,E) = \frac{f_{\text{est}}(T+tE) - f(T)}{t} + O(t).$$

In Appendix I we give a more precise discussion of the error analysis, and, in particular, show how to choose t. In Appendix II we give a listing of a MATLAB .m file that implements this idea to compute $f_{\text{est}}(T+tE)$.

The cost of transforming $T+E$ into upper triangular form is approximately $3n^3$ flops, the cost of computing $f(\hat{T})$ is approximately $2n^3/3$ flops, and the cost of forming $f_{\text{est}}(T+E) = Uf(\hat{T})U^*$ is another $8n^3/3$ flops, if one uses the factored form of U. The total cost of each estimate of $L_f(T,E)$ is $19n^3/3$ flops as compared to approximately $29\frac{2}{3}n^3$ flops (see the previous section for the justification for this figure) for the computation of $f(M)$ alone. Thus, the cost of computing $f(M)$ and $L_f(T,E)$ twice (which should be sufficient for a reasonable estimate of $\kappa_f(M) = \kappa_f(T)$) is about $43n^3$ flops.

Let us compare this with the cost of computing $\exp(M)$ by scaling and squaring [4, Algorithm 11.3.1] and estimating $\kappa_{\exp}(M)$ by the trapezoidal rule [6], [12]. The cost of scaling and squaring depends on q, the order of Pade approximation used—we take $q=8$, which gives accuracy of 10^{-16}, and use the Horner multiplication scheme described after [4, Algorithm 11.3.1]. It also depends on $j = \max\{0, 1 + \lfloor \|M\| \rfloor\}$. The trapezoidal rule calls for computing terms of the form

$$\frac{d}{dt}(e^{X/2^j} + tW)^2\Big|_{t=0} = \exp(X/2^j)W + W\exp(X/2^j), \tag{3.4}$$

which requires two matrix multiplications. Following [9, (100)], we estimate (3.4) by a forward difference approximation

$$[(\exp(X/2^j) + tW)^2 - \exp(X/2^{j-1})]/t \tag{3.5}$$

for suitably chosen small t. This halves the cost of evaluating the trapezoidal rule estimate of $L_{\exp}(M,E)$ since the matrices $\exp(X/2^k), k = 0,1,\ldots,j$ have already been computed in the scaling and squaring. The cost is now $2(j+1)n^3$ flops. Using the flop count from [4, Algorithm 11.3.1], we see that this method requires $2(6+j+1/3+2(j+1))n^3 \approx (17+6j)n^3$ flops to compute $\exp(M)$ and $L_{\exp}(T,E)$ twice. That is $47n^3$ flops if $\|M\| = 16$ and $41n^3$ flops if $\|M\| = 8$. So the Schur method is faster if $\|M\|$ is at least 16.

Sometimes one wishes to compute e^{tA} for several values of t. In this case the Schur method is more efficient, both for the computation of e^{tA} and for condition estimation.

4. Appendix I. Now we give a more precise analysis of the errors that occur in the forward difference approximation $L_f(T,E)$. This enables us to choose t appropriately. We use the notation of §3 and assume that $\|E\| = 1$.

The first step in estimating $f(T+tE)$ is to restore $T+tE$ to block upper triangular form one block column at a time. At the first stage we have

$$T + tE = \begin{pmatrix} (T+tE)_{11} & (T+tE)_{1,[1]} \\ tE_{[1],1} & (T+tE)_{[1,1]} \end{pmatrix} \tag{4.1}$$

in the notation of (3.2). After applying the transformation in Lemma 1.1 to make the $2,1$ block of this $O(\|E\|^2)$, the resulting $2,2$ block is

$$P(T+tE)_{11}P^* - P(T+tE)_{1,[1]} + (T+tE)_{[1,1]}.$$

Let L denote the strictly (block) lower triangular part of this. Let

$$\gamma = \frac{\max_{i=1,\ldots,k-1} \|T_{i,[i]}\|}{\delta} \quad \text{and} \quad \eta = \frac{t}{\delta}. \tag{4.2}$$

Then the spectral norm of any block column of L is bounded by

$$\eta^2(\|T_{11}\| + t) + \eta(\|T_{1,[1]}\| + t) + t. \tag{4.3}$$

Under the further assumptions that $t << \|T\|$ and $\eta << 1$, (4.3) is bounded by $c(1+\gamma)\|E\|$ where $c \geq 1$ and $c \approx 1$. Thus the size of the subdiagonal has grown by a factor of $c(1+\gamma)$. We would like γ to be small. This is the case if δ is large or if T is almost block diagonal. For simplicity we assume that $[c(1+\gamma)]^{k-1}$, the growth factor for the whole process, is at most 2. Recall that η also depends on t so η is also increasing, but by no more than a factor of 2 in total. Let U be the product of the matrices S defined in Lemma 1.1. Then, by Lemma 1.1,

$$U^*(T+tE)U = \hat{T} + \hat{E},$$

where $\hat{T}$ is block upper triangular and, using the bound on the growth of the subdiagonal of L and the growth of η, we have

$$\|\hat{E}\| \leq (k-1)[8\eta^2\|T\| + 12\eta t \leq 20(k-1)\eta t.$$

We have used the fact $\gamma \leq 1$, which is implied by $[c(1+\gamma)]^{k-1} \leq 2$. We can also obtain the bound $\|U - I\| \leq 20(k-1)\eta$. Note that U is not unitary. Let $\hat{\eta} = 20(k-1)\eta$ and assume that $\hat{\eta} < \frac{1}{3}$.

Now let us estimate the accuracy of $Uf(\hat{T})U^*$ as an estimate of $f(T+tE)$.

$$\begin{aligned}
\|f(T+E) - Uf(\hat{T})U^*\| &\leq \|Uf(\hat{T})(U^* - U^{-1})\| \\
&\quad + \|U\,[f(U^{-1}(T+tE)U) - f(\hat{T})]\,U^*\| \\
&\leq 2\hat{\eta}^2\|f(\hat{T})\| \\
&\quad + (1+\hat{\eta}^2)\,\|L_f(T,\cdot)\|\,\|U^{-1}(T+tE)U - \hat{T}\| \\
&\quad + O(\|U^{-1}(T+tE)U - \hat{T}\|^2) \\
&\approx 2\hat{\eta}^2\|f(T)\| + (1+\hat{\eta}^2)\|L_f(T,\cdot)\|\,\|U^{-1}(T+tE)U - \hat{T}\| \\
&\leq 2\hat{\eta}^2\|f(T)\| + 2\hat{\eta}\|L_f(T,\cdot)\|\,t \\
&= 2\hat{\eta}[20(k-1)\delta^{-1}\|f(T)\| + \|L_f(T,\cdot)\|]t \\
&\equiv \tilde{\eta}t.
\end{aligned}$$

The first of these two terms is due to the fact that U is not unitary and can be eliminated by using unitary transformations in Lemma 1.1; that is use (1.12) rather than (1.7). However, since when one forms the product $Uf(\hat{T})U^*$, there will be an error of norm about $n\epsilon\|f(T)\|$ where ϵ is machine precision, the extra term $2\hat{\eta}^2\|f(T)\|$ is unlikely to cause a serious problem. Using a unitary U will not reduce the second term. The only way to reduce the second term is by applying Lemma 1.1 to $\hat{T}+\hat{E}$ to further reduce the (block) subdiagonal part of $\hat{T}+\hat{E}$. This would potentially double the cost of condition estimation.

One can find a constant M, based on the norm of the second derivative of f, such that

$$\left\| L_f(T,E) - \frac{f(T+tE)-f(T)}{t} \right\| \le Mt$$

for t sufficiently small.

Combining these bounds we have

$$\left\| L_f(T,E) - fl\left(\frac{Uf(\hat{T})U^* - f(T)}{t}\right)\right\| \le n\epsilon\frac{\|f(T)\|}{t} + \tilde{\eta} + Mt, \tag{4.4}$$

where $fl(\cdot)$ denotes the quantity actually computed.

One approach is to minimize the right-hand side of (4.4). All the quantities on the right-hand side are known except for $\|L_f(T,\cdot)\|$, which can be estimated after one step of the procedure, and M. One could estimate M if $f(z)=\sum_i a_i z^i$ by considering $f_{\text{abs}}(z)=\sum_i |a_i| z^i$. This is likely to give a gross overestimate of M.

It is not necessary to estimate $L_f(T,E)$ very accurately since we are only interested in obtaining an order of magnitude estimate of $\|L_f(T,E)\|$ for condition estimation. So another approach is to find an acceptable value of t rather than the best. Since the last two terms of (4.4) are increasing functions of t (recall that $\tilde{\eta}$ contains a factor of t), we choose t as small as possible, without making the first term too large. In particular, take

$$t = 2n\epsilon\frac{\|f(T)\|}{\|L_f(T,\cdot)\|_{\text{est}}}, \tag{4.5}$$

where $\|L_f(T,\cdot)\|_{\text{est}}$ is an estimate of $\|L_f(T,\cdot)\|_{\text{est}}$, and may be changed at each iteration. If the sum of the last two terms in (4.4) is less than $.4\|L_f(T,\cdot)\|$, then our choice of t in (4.5) ensures that the right-hand side of (4.4) is less than $.9\|L_f(T,\cdot)\|$ and hence that our forward difference estimate of $\|L_f(T,\cdot)\|$ is correct to within a factor of 10. On the other hand, if the sum of the last two terms in (4.4) is greater than $.9^2\|L_f(T,\cdot)\|$, then even the optimal choice of t will not ensure that the right-hand side of (4.4) is less than $.9\|L_f(T,\cdot)\|$. Thus it is unlikely that our choice of t in (4.5) will give an unacceptable estimate of $\|L_f(T,\cdot)\|$ and that the optimal choice of t, had we been able to determine it, would have given an acceptable estimate.

5. Appendix II. The following listing implements the ideas in §§3 and 4. Both when block triangularizing $W+E$ and when applying the inverse transformation, we ignore terms of size $O(\|E\|^2)$ to save computation.

```
function [wt] = blocktri(w, blksz, e)
% function to compute func(w+e) using the ideas in Sections 3 and 4
```

```
%
% input
%       w         block upper triangular
%       blksz     vector of block sizes
%       e         perturbation -- small enough that one can ignore
%                       terms of norm O(||e||^2)
% return
%       wt        estimate of func(w+e) that is within O(||e||^2) of true
%                       value
% calls
%       func.m    evaluates the function one wishes to estimate
%                       (this .m file should exploit the fact that its
%                       argument is block upper triangular)
%       lyap.m    solves a Sylvester equation
%
k = length(blksz);
blks = cumsum(blksz);
n = max(size(w));
pstore  = zeros(n);
%
% zero sub diagonal of w+e and accumulate changes in e
for i = 1:k-1,
        cols = blks(i)-blksz(i)+1:blks(i);
        row1 = 1:blks(i);
        rows = blks(i)+1:n;
        p = lyap( -w(rows,rows), w(cols,cols), -e(rows,cols));
        pstore(rows, cols) = p;
        e(rows, cols) = zeros(length(rows), length(cols));
        e(rows, rows) = e(rows, rows) - p*w(cols, rows);
        e(row1, cols) = e(row1, cols) + w(row1, rows)*p;
        e(row1, rows) = e(row1, rows) - w(row1, cols)*p';
        e(cols, rows) = e(cols, rows) + p'*w(rows, rows);
end
%
% evaluate func at the block triangular matrix w+e
wt = func(w + e);
%
% apply the inverse transformation
for i = k-1:-1:1,
        cols = blks(i)-blksz(i)+1:blks(i);
        row1 = 1:blks(i);
        rows = blks(i)+1:n;
        p = pstore(rows, cols);
        wt(rows, cols) = p*wt(cols,cols) - wt(rows,rows)*p;
        wt(rows, rows) = wt(rows, rows) + p*wt(cols, rows);
        wt(row1, cols) = wt(row1, cols) - wt(row1, rows)*p;
        wt(row1, rows) = wt(row1, rows) + wt(row1, cols)*p';
        wt(cols, rows) = wt(cols, rows) - p'*wt(rows, rows);
end
```

REFERENCES

[1] Z. Bai and J. Demmel, *On swapping diagonal blocks in the real Schur form*, Linear Algebra Appl., 186 (1993), pp. 73–96.

[2] R. Byers, *Solving the algebraic Ricatti equation with matrix sign function*, Linear Algebra Appl., 85 (1987), pp. 267–279.

[3] ———, *Numerical stability and instability in matrix sign function based algorithms*, in Computational and Combinatorial Methods in Systems Theory, C. Byrnes and A. Lindquist, eds., Elsevier Science Publishers B.V., North-Holland, 1990.

[4] G. Golub and C. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, 1989.

[5] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.

[6] C. Kenney and A. J. Laub, *Condition estimates for matrix functions*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 191–209.

[7] ———, *Polar decomposition and matrix sign function condition estimates*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 488–504.

[8] ———, *On scaling Newton's method for the polar decomposition and the matrix sign function*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 688–706.

[9] ———, *Small-sample statistical condition estimates for general matrix functions*, SIAM J. Sci. Comput., 15 (1994), pp. 36–61.

[10] C.-C. Lin and E. Zmijewski, *A parallel algorithm for computing the eigenvalues of an unsymmetric matrix on an* SIMD *mesh of processors*, Tech. Report TRCS 91-15, Dept. of Computer Science, University of California, Santa Barbara, 1991.

[11] R. Mathias, *Operation counts for the solution of the Sylvester equation*, manuscript.

[12] R. Mathias, *Evaluating the Fréchet derivative of the matrix exponential*, Numer. Math., 63 (1992), pp. 213–226.

[13] J. R. Rice, *A theory of condition*, SIAM J. Numer. Anal., 3 (1966), pp. 287–310.

[14] G. Stewart and G.-S. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.

[15] G. W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 579–601, April 1995

017

THE p-PRODUCT AND ITS APPLICATIONS IN SIGNAL PROCESSING *

HUIXIA ZHU† AND GERHARD X. RITTER†

Abstract. This paper introduces a new matrix product that proves useful in the representation of a large class of orthogonal transforms. It is shown that transform representation in terms of this product provides novel computational methods for computing the fast Fourier transform, the fast Walsh transform, a fast generalized Walsh transform, as well as a fast wavelet transform. Uniqueness and the advantages of this new formalism over traditional methods are also discussed.

Key words. Fourier transform, matrix product, p-product, signal processing, tensor product, Walsh transform, wavelet transform

AMS subject classifications. 15A69, 68Q40, 68U10, 68U30

1. Introduction.

The generalized matrix or p-product was first defined in [16]. This new matrix operation includes the matrix and vector products of linear algebra, the matrix product of minimax algebra [8], as well as generalized convolutions as special cases [16]. It provides for a transformation that combines the same or different types of values (or objects) into values of a possibly different type from those initially used in the combining operation. It has been shown that the p-product can be applied to express various image processing transforms in computing form [17]. In this paper, after briefly defining the p-product, we discuss its applications to a large class of transforms used in signal processing.

One of the most fundamental transforms in signal analysis and signal processing is the Fourier transform. With the rapid advances in digital computers, the discrete version of the Fourier transform gained in importance and various efficient algorithms for its computation, known as fast Fourier transforms (FFTs), were developed [21], [6], [20]. Most of these algorithms are based on, or are variations of, the Cooley–Tukey method [7] and are required to have a reordering process. Even though this process can be done without arithmetic operations and only by moving some elements in storage, it is still an important part of the overall cost of an FFT computation on most computers. As pointed out by Stone [19], Cvetanovic [9] showed that for any typical implementation of the FFT that uses $\log N$ *butterfly* operations on N-vectors, followed or preceded by one reverse-binary operation, these two kinds of operations are incompatible on most parallel machines. If the butterfly operation is conflict free, then the reverse-binary results in a maximum conflict in the network and vice versa. This means that at least one of the two types of operations will cause some problems. The operations required for the reverse-binary can be up to $O(N)$ for any N-vectors. However, the p-product FFT, which is presented in §3, is designed to compute the FFT in a proper order without requiring the reordering process either after or before computational stages. This means that only one type of operation is used in the p-product FFT. Therefore, the conflict problem no longer exists on parallel machines, and the extra $O(N)$ operations required for the reverse process have been

* Received by the editors February 11, 1993; accepted for publication (in revised form) by C. Van Loan, March 18, 1994.

† Center for Computer Vision and Visualization, University of Florida, Gainesville, Florida 32611 (ritter@cis.ufl.edu).

eliminated. Furthermore, all of the stages of the p-product FFT are identical and the only computations required at each stage are the p-product and Hadamard matrix product, which makes this algorithm very fast on either parallel or sequential machines.

The Walsh transform and the generalized Walsh transform are similar to the Fourier transform, but simpler. The Walsh transform was first defined in 1923 by J. L. Walsh [22]. In 1931, R. E. A. C. Paley [14] gave an entirely different definition of the Walsh transform, which is the one we discuss here. His definition was based on finite products of Rademacher functions and the order obtained was quite different from that of Walsh. In 1955, Chrestenson extended Paley's idea and formed a class of the generalized Walsh transforms based on finite products of Rademacher functions of order α, which takes the Walsh transform as its special case [3]. The Walsh transform can be computed by a fast algorithm identical in form to the successive-doubling method given for the FFT. Due to the requirement of the reordering process, this fast method is restricted to the base-two format only. It is probably for this reason that the technical literature is devoid of fast algorithms for computing the generalized Walsh transform. The p-product provides the key mathematical language in which to describe and analyze, in a unified format, similarities and differences between these transformations. In §4, we derive a new algorithm in terms of the p-product language, which offers a fast computation not only used for *both* transforms, but also in proper order without requiring the reverse process. This is especially important on a supercomputer where the data flow is usually the major time-consuming part of the computation.

It is well known that the Fourier transform decomposes a signal into individual frequency components but does not provide information as to *when* the frequencies occurred. When the signal to be analyzed is nonstationary, a relevant analysis calls for keeping the time information to exhibit its time-varying spectral properties. The most straightforward solution is, therefore, to split the signal into fractions within which the stationary assumptions apply. The Gabor transform (or the short time Fourier transform) is commonly used to perform this decomposition. It introduces a time-localization *window function*, $g(t-b)$, where the parameter b is used to translate the window to cover the whole time domain for extracting local information of the Fourier transform of the signal. The principal problem here is that any one choice of $g(t)$ results in windows that are too wide to capture all nonstationary behavior and too narrow to capture low-frequency information. The recently introduced wavelet transform is an alternative tool that deals with nonstationary signals. The decomposition is carried out by means of a special analysis function ψ, called the *basic wavelet*, which is translated in time (for selecting the part of the signal to be analyzed), then dilated or contracted using a scale parameter (to focus on a given range of oscillations). It is different from the Gabor transform in that it simultaneously localizes a signal and its Fourier transform with zoom-in and zoom-out capability. The wavelet transform has drawn a great deal of attention from mathematicians and scientists in various disciplines. Its numerous applications can be found in [2], [5], [11], [12].

As mentioned earlier, the p-product provides novel algorithms for computing and expressing the Fourier transform, the Walsh transform, the generalized Walsh transform, and the wavelet transform. A well-known fact of linear algebra is that linear transforms can be represented in terms of matrix–vector products. However, since the wavelet transform is a function of two variables (both time and frequency), it is difficult to express it in matrix product form. Even though Heller et al. [5] have defined wavelet matrices, they are only used to prove certain mathematical properties. In contrast, the generalized matrix product lends itself well to expressing the

wavelet transform in matrix form. In this paper, following a definition of the wavelet matrices given in [5], we show how to express the wavelet transform and its inverse in terms of the p-product and, in addition, provide a simple and fast wavelet transform algorithm using the p-product and parallelism. The principle of the new algorithm is to decompose a long summation into several short ones and then use the p-product to carry out the computation. When executing this algorithm on parallel machines, the computing time is g times less than the computing time of the standard method, where g denotes the *genus* of the wavelet matrix **a**.

2. The generalized matrix product. We reserve the symbols $\mathbb{Z}, \mathbb{R}$, and $\mathbb{C}$ to denote the set of integers, real numbers, and complex numbers, respectively. The set $\mathbb{Z}_n^+$ is defined by $\mathbb{Z}_n^+ = \{1, 2, \ldots, n\}$. This distinguishes $\mathbb{Z}_n^+$ from the commonly used notation $\mathbb{Z}_n = \{0, 1, \ldots, n-1\}$.

An arbitrary field is denoted by $\mathbb{F}$. In our discussion of the wavelet transform, the field $\mathbb{F}$ is usually $\mathbb{R}$ or $\mathbb{C}$. For a given set $\mathbf{X}$, the set of all functions $\mathbf{X} \to \mathbb{F}$ is denoted by $\mathbb{F}^{\mathbf{X}}$, while the set of all $m \times n$ matrices with entries from $\mathbb{F}$ is denoted by $\mathbb{F}_{m\times n}$. We follow the usual convention of setting $\mathbb{F}^n = \mathbb{F}_{1\times n}$ and view $\mathbb{F}^n$ as the set of all n-dimensional row vectors with entries from $\mathbb{F}$. Similarly, the set of all m-dimensional column vectors with entries from $\mathbb{F}$ is given by $(\mathbb{F}^m)' = [\mathbb{F}_{1\times m}]' = \mathbb{F}_{m\times 1}$.

In the subsequent discussion, let m, n, and p be positive integers with p dividing both m and n. Define the following correspondences:

$$\begin{aligned} &c_p\colon \mathbb{Z}_p^+ \times \mathbb{Z}_{n/p}^+ \to \mathbb{Z}_n^+ \\ &\text{by} \quad c_p(k,j) = (k-1)\frac{n}{p} + j, \\ &\text{where} \quad 1 \le j \le \frac{n}{p} \quad \text{and} \quad 1 \le k \le p \end{aligned} \tag{1}$$

and

$$\begin{aligned} &r_p\colon \mathbb{Z}_{m/p}^+ \times \mathbb{Z}_p^+ \to \mathbb{Z}_m^+ \\ &\text{by} \quad r_p(i,k) = (i-1)p + k, \\ &\text{where} \quad 1 \le k \le p \quad \text{and} \quad 1 \le i \le \frac{m}{p}. \end{aligned} \tag{2}$$

Now let $A = (a_{sj'}) \in \mathbb{F}_{l\times m}$ and $B = (b_{i',t}) \in \mathbb{F}_{n\times q}$. Using the maps r_p and c_p, A and B can be rewritten as

$$\begin{aligned} &A = (a_{s,(i,k)})_{l\times m}, \quad \text{where} \quad 1 \le s \le l, 1 \le r_p(i,k) = j' \le m, \quad \text{and} \\ &B = (b_{k,j),t})_{n\times q}, \quad \text{where} \quad 1 \le c_p(k,j) = i' \le n \quad \text{and} \quad 1 \le t \le q. \end{aligned} \tag{3}$$

The *p-product* or *generalized matrix product* of A and B is denoted by $A \oplus_p B$, and is the matrix

$$C = A \oplus_p B \in \mathbb{F}_{l(n/p)\times(m/p)q} \tag{4}$$

defined by

$$c_{(s,j)(i,t)} = \sum_{k=1}^{p} (a_{s,(i,k)} b_{(k,j),t}) = (a_{s,(i,1)} b_{(1,j),t}) + \cdots + (a_{s,(i,p)} b_{(p,j),t}), \tag{5}$$

where $c_{(s,j)(i,t)}$ denotes the (s,j)th row and (i,t)th column entry of C. Here we use the lexicographical order $(s,j) < (s',j') \Leftrightarrow s < s'$ or if $s = s', j < j'$. Thus, matrix C has the following form.

(6)

$$\begin{bmatrix}
c_{(1,1)(1,1)} & \cdots & c_{(1,1)(1,q)} & c_{(1,1)(2,1)} & \cdots & c_{(1,1)(2,q)} & \cdots & c_{(1,1)(i,t)} & \cdots & c_{(1,1)(m/p,q)} \\
c_{(1,2)(1,1)} & \cdots & c_{(1,2)(1,q)} & c_{(1,2)(2,1)} & \cdots & c_{(1,2)(2,q)} & \cdots & c_{(1,2)(i,t)} & \cdots & c_{(1,2)(m/p,q)} \\
\vdots & & \vdots & & & \vdots & & \vdots & & \vdots \\
c_{(1,n/p)(1,1)} & \cdots & c_{(1,n/p)(1,q)} & c_{(1,n/p)(2,1)} & \cdots & c_{(1,n/p)(2,q)} & \cdots & c_{(1,n/p)(i,t)} & \cdots & c_{(1,n/p)(m/p,q)} \\
c_{(2,1)(1,1)} & \cdots & c_{(2,1)(1,q)} & c_{(2,1)(2,1)} & \cdots & c_{(2,1)(2,q)} & \cdots & c_{(2,1)(i,t)} & \cdots & c_{(2,1)(m/p,q)} \\
\vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\
c_{(2,n/p)(1,1)} & \cdots & c_{(2,n(p))(1,q)} & c_{(2,n/p)(2,1)} & \cdots & c_{(2,n/p)(2,q)} & \cdots & c_{(2,n/p)(i,t)} & \cdots & c_{(2,n/p)(m/p,q)} \\
\vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\
c_{(s,j)(1,1)} & \cdots & c_{(s,j)(1,q)} & c_{(s,j)(2,1)} & \cdots & c_{(s,j)(2,q)} & \cdots & \underline{c_{(s,j)(i,t)}} & \cdots & c_{(s,j)(m/p,q)} \\
\vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\
c_{(l,1)(1,1)} & \cdots & c_{(l,1)(1,q)} & c_{(l,1)(2,1)} & \cdots & c_{(l,1)(2,q)} & \cdots & c_{(l,1)(i,t)} & \cdots & c_{(l,1)(m/p,q)} \\
\vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\
c_{(l,n/p)(1,1)} & \cdots & c_{(l,n/p)(1,q)} & c_{(l,n/p)(2,1)} & \cdots & c_{(l,n/p)(2,q)} & \cdots & c_{(l,n/p)(i,t)} & \cdots & c_{(l,n/p)(m/p,q)}
\end{bmatrix}$$

The entry $c_{(s,j)(i,t)}$ in the (s,j)-row and (i,t)-column is underlined for emphasis.

To provide an example, suppose that $l = 2, m = 6, n = 4$, and $q = 3$. Then for $p = 2$, one obtains $m/p = 3, n/p = 2$, and $1 \leq k \leq 2$. Now let

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} \end{pmatrix} \in M_{2\times 6}(\mathbb{R}) \tag{7}$$

and

$$B = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \\ b_{41} & b_{42} & b_{43} \end{pmatrix} \in M_{4\times 3}(\mathbb{R}). \tag{8}$$

Then the (2, 1)-row and (2, 3)-column element $c_{(2,1)(2,3)}$ of the matrix

$$C = A \oplus_2 B \in M_{l(n/p)\times(m/p)q}(\mathbb{R}) = M_{4\times 9}(\mathbb{R}) \tag{9}$$

is given by

$$\begin{aligned} c_{(2,1)(2,3)} &= \sum_{k=1}^{2} a_{2,r_2(2,k)} \cdot b_{c_2(k,1),3} \\ &= a_{2,r_2(2,1)} \cdot b_{c_2(1,1),3} + a_{2,r_2(2,2)} \cdot b_{c_2(2,1),3} \\ &= a_{23} \cdot b_{13} + a_{24} \cdot b_{33}. \end{aligned} \tag{10}$$

Thus, to compute $c_{(2,1)(2,3)}$, the two underlined elements of A are combined with the

two underlined elements of B as illustrated.

$$
\begin{aligned}
&\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} \\ a_{21} & a_{22} & \underline{a_{23}} & \underline{a_{24}} & a_{25} & a_{26} \end{pmatrix} \oplus_2 \begin{pmatrix} b_{11} & b_{12} & \underline{b_{13}} \\ b_{21} & b_{22} & \underline{b_{23}} \\ b_{31} & b_{32} & \underline{b_{33}} \\ b_{41} & b_{42} & b_{43} \end{pmatrix} \\
&\quad = \begin{pmatrix} a_{1,r_2(1,1)} & a_{1,r_2(1,2)} & a_{1,r_2(2,1)} & a_{1,r_2(2,2)} & a_{1,r_2(3,1)} & a_{1,r_2(3,2)} \\ a_{2,r_2(1,1)} & a_{2,r_2(1,2)} & \underline{a_{2,r_2(2,1)}} & \underline{a_{2,r_2(2,2)}} & a_{2,r_2(3,1)} & a_{2,r_2(3,2)} \end{pmatrix} \\
&\quad \oplus_2 \begin{pmatrix} b_{c_2(1,1),1} & b_{c_2(1,1),2} & \underline{b_{c_2(1,1),3}} \\ b_{c_2(1,2),1} & b_{c_2(1,2),2} & \underline{b_{c_2(1,2),3}} \\ b_{c_2(2,1),1} & b_{c_2(2,1),2} & \underline{b_{c_2(2,1),3}} \\ b_{c_2(2,2),1} & b_{c_2(2,2),2} & \underline{b_{c_2(2,2),3}} \end{pmatrix} \\
&\quad = \begin{pmatrix} c_{(1,1)(1,1)} & c_{(1,1)(1,2)} & \cdots & c_{(1,1)(2,3)} & \cdots & c_{(1,1)(3,3)} \\ c_{(1,2)(1,1)} & c_{(1,2)(1,2)} & \cdots & c_{(1,2)(2,3)} & \cdots & c_{(1,2)(3,3)} \\ c_{(2,1)(1,1)} & c_{(2,1)(1,2)} & \cdots & \underline{c_{(2,1)(2,3)}} & \cdots & c_{(2,1)(3,3)} \\ c_{(2,2)(1,1)} & c_{(2,2)(1,2)} & \cdots & c_{(2,2)(2,3)} & \cdots & c_{(2,2)(3,3)} \end{pmatrix} \\
&\quad = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{16} & \cdots & c_{19} \\ c_{21} & c_{22} & \cdots & c_{26} & \cdots & c_{29} \\ c_{31} & c_{32} & \cdots & \underline{c_{36}} & \cdots & c_{39} \\ c_{41} & c_{42} & \cdots & c_{46} & \cdots & c_{49} \end{pmatrix}.
\end{aligned} \tag{11}
$$

In particular,

$$
\begin{aligned}
&\begin{pmatrix} 1 & 2 & 0 & 5 & 4 & 3 \\ 7 & 3 & 4 & 1 & 0 & 6 \end{pmatrix} \oplus_2 \begin{pmatrix} 2 & 6 & 1 \\ 1 & 3 & 2 \\ 2 & 2 & 5 \\ 3 & 0 & 4 \end{pmatrix} \\
&\qquad = \begin{pmatrix} 6 & 10 & 11 & 10 & 10 & 25 & 14 & 30 & 19 \\ 7 & 3 & 10 & 15 & 0 & 20 & 13 & 12 & 20 \\ 10 & 18 & 17 & 10 & 26 & 9 & 12 & 12 & 30 \\ 11 & 6 & 16 & 7 & 12 & 12 & 18 & 0 & 24 \end{pmatrix}.
\end{aligned} \tag{12}
$$

An even more general definition of the p-product was given in [17] and [16]. As mentioned, the p-product includes the common matrix and vector products of linear algebra. It has been proved that those products can be obtained by substituting specific values for p [16]. The properties and some applications of the p-product in image processing can be found in [17]. In the following sections, we discuss the applications of the p-product in the Fourier transform, the Walsh transform, the generalized Walsh transform, and the wavelet transform.

3. The Fourier transform. In this section, we use matrix and the generalized matrix product identities associated with FFTs to develop a novel formulation of the FFT in terms of the p-product. We use the tensor product expressions of the FFT developed by Tolimieri [21] to derive the new algorithms in terms of the p-product formulation.

We first establish notation and state identities that are required in deriving the formulation.

DEFINITION 3.1. *Suppose that $N = rs$. The N-*point strides s permutation matrix $\mathbf{P}(N, s)$ *is defined by*

$$\mathbf{P}(N, s)(\mathbf{x} \otimes \mathbf{y}) = \mathbf{y} \otimes \mathbf{x}, \tag{13}$$

where $\mathbf{x}$ *and* $\mathbf{y}$ *are arbitrary vectors of sizes* r *and* s, *respectively.*

That is to say that the action of $\mathbf{P}(N,s)$ on an arbitrary vector $\mathbf{x}$ of size N is

(14)
$$\begin{aligned}\mathbf{P}(N,s)\mathbf{x} &= \mathbf{P}(N,s)[x_0,x_1,\ldots,x_{N-1}]' \\ &= [x_0,x_s,x_{2s},\ldots,x_{(r-1)s},x_1,x_{+1s},\ldots,x_{(r-1)s+1},x_2,\ldots,x_{s-1},\ldots,x_{rs-1}]'.\end{aligned}$$

DEFINITION 3.2. *For any* $\mathbf{a}\in\mathbb{F}_{l\times m}$, *we define* $\operatorname{col}(\mathbf{a})$, *the* column vector *of* $\mathbf{a}$, *as*

$$\operatorname{col}(\mathbf{a}) = [a_{11},a_{12},\ldots,a_{1m},a_{21},a_{22},\ldots,a_{2m},\ldots,a_{l1},a_{l2},\ldots,a_{lm}]', \tag{15}$$

and the row vector *of* $\mathbf{a}$ *as* $\operatorname{row}(\mathbf{a}) = (\operatorname{col}(\mathbf{a}))'$.

We now state some basic identities concerning the p-product, tensor product, permutation matrices, and Fourier matrices, which were developed in [17] and [21]. We use these identities as building blocks for some of our derivations.

LEMMA 3.1. *Let* $N = rs$ *and* $x\in\mathbb{F}_{r\times s}$. *Then*

$$\operatorname{col}(\mathbf{x}') = \mathbf{P}(N,s)\operatorname{col}(\mathbf{x}). \tag{16}$$

LEMMA 3.2. *Let* $\mathbf{a}\in\mathbb{F}_{l\times m}$ *and* $\mathbf{b}\in\mathbb{F}_{m\times q}$. *Then*

$$\begin{aligned}\operatorname{col}(\mathbf{ab}) &= \mathbf{a}\oplus_m \operatorname{col}(\mathbf{b}),\\ \operatorname{row}(\mathbf{ab}) &= \operatorname{row}(\mathbf{a})\oplus_m \mathbf{b},\\ (\mathbf{ab})' &= \operatorname{row}(\mathbf{a})\oplus_m \operatorname{col}(\mathbf{b}).\end{aligned} \tag{17}$$

LEMMA 3.3. *Let* $\mathbf{a}\in\mathbb{F}_{s\times s}$ *and* $\mathbf{x}\in\mathbb{F}_{rs\times t}$ *with* $t\in\mathbb{Z}^+$. *Then*

$$(\mathbf{a}\otimes\mathbf{I}_r)\mathbf{x} = \mathbf{a}\oplus_s \mathbf{x}, \tag{18}$$

where $\mathbf{I}_r$ *denotes the* $r\times r$ *identity matrix.*

LEMMA 3.4. *Let* $\mathbf{a}\in\mathbb{F}_{s\times s}$, *with*

$$\mathbf{a} = \begin{bmatrix} a_0 & a_1 & \cdots & a_{s-1}\\ a_s & a_{s+1} & \cdots & a_{2s-1}\\ \vdots & \vdots & \vdots & \vdots\\ a_{(s-1)s} & a_{(s-1)s+1} & \cdots & a_{s^2}\end{bmatrix} \tag{19}$$

and $\mathbf{b}\in\mathbb{F}_{r\times r}$, *with*

$$\mathbf{b} = \begin{bmatrix} b_0 & b_1 & \cdots & b_{r-1}\\ b_r & b_{r+1} & \cdots & b_{2r-1}\\ \vdots & \vdots & \vdots & \vdots\\ b_{(r-1)r} & b_{(r-1)r+1} & \cdots & b_{r^2}\end{bmatrix}. \tag{20}$$

Let $N = rs$ *and* $\mathbf{x}$ *be a vector of length* N.

(a) *If* $\mathbf{y}_1 = \mathbf{P}(N,s)(\mathbf{b}\oplus_r \mathbf{x})$ *and* $\mathbf{y}_2 = \operatorname{row}(\mathbf{b})\oplus_r \mathbf{x}$, *then we have*

$$\mathbf{y}_1 = \operatorname{col}(\mathbf{y}_2). \tag{21}$$

(b) *If* $\mathbf{y}\in\mathbb{F}_{s\times r}$, *then we have*

$$\operatorname{row}(\mathbf{a})\oplus_s \mathbf{y} = (\mathbf{a}\oplus_s \operatorname{col}(\mathbf{y}))'. \tag{22}$$

More generally, if $\mathbf{x}$ is an $N \times t$ matrix, then the next lemma can be proved.

LEMMA 3.5. *Let $N = rs, \mathbf{b}$ be the matrix defined in Lemma 3.4, and $\mathbf{x}$ be an $N \times t$ matrix. Suppose*

$$\mathbf{y}_1 = \mathbf{P}(N,s)(\mathbf{b} \oplus_r \mathbf{x}) \quad \text{and} \quad \mathbf{y}_2 = \text{row}(\mathbf{b}) \oplus_r \mathbf{x}. \tag{23}$$

Then $\text{col}(\mathbf{y}_1) = \text{col}(\mathbf{y}_2)$.

LEMMA 3.6. *Let $N = rs$ with r and s being any integers. Then the N-point Fourier transform can be factored as*

$$\mathbf{F}(N) = (\mathbf{F}(s) \otimes \mathbf{I}_r)\mathbf{T}_r(N)\mathbf{P}(N,s)(\mathbf{F}(r) \otimes \mathbf{I}_s), \tag{24}$$

where $\mathbf{P}(N,s)$ is an N-point stride s permutation matrix and $\mathbf{T}_r(N)$ is a diagonal matrix as

$$\mathbf{T}_r(N) = \text{diag}(1,\dots,1;1,w,\dots,w^{r-1};\dots;1,w^{s-1},\dots,w^{(r-1)(s-1)}), \tag{25}$$

with $w = e^{2\pi i/N}$.

We now use the above identities to express the Fourier matrices first in terms of the tensor product and then in terms of the p-product.

Let $N = p_1p_2\cdots p_n$, with p_i being a prime number, then the N-point Fourier transform of any N-vector $\mathbf{x}$ is given by the formula

$$y_k = \sum_{j=0}^{N-1} w^{jk}x_j, \qquad 0 \le k < N, w = e^{2\pi i/N}. \tag{26}$$

THEOREM 3.1. *Let $N = p_1p_2\cdots p_n$, with p_i being a prime number. Then the tensor product formula of N-point Fourier transform is*

$$\mathbf{F}(N) = \left(\prod_{i=1}^{n-1}((\mathbf{F}(p_i) \otimes \mathbf{I}_{\delta_i})(\mathbf{T}_{i+1} \otimes \mathbf{I}_{\lambda_{i+1}})(\mathbf{P}_{i+1} \otimes \mathbf{I}_{\lambda_{i+1}}))\right)(\mathbf{F}(p_n) \otimes \mathbf{I}_{\delta_n}),$$
$$\text{with} \quad \delta_i = N/p_i, \lambda_i = \frac{N}{p_1p_2\cdots p_i}, \tag{27}$$

where $\mathbf{T}_{i+1}$ is the $p_1p_2\cdots p_{i+1} \times p_1p_2\cdots p_{i+1}$ diagonal matrix

(28)
$$\begin{aligned}\mathbf{T}_{i+1} &= \mathbf{T}_{p_{i+1}}(p_1p_2\cdots p_{i+1}) \\ &= (\text{diag}(1,\dots,1;1,w,\dots,w^{p_{i+1}-1};\dots;1,w^{p_1p_2\cdots p_i-1},\dots,w^{(p_{i+1}-1)(p_1\cdots p_i-1)}))^{\lambda_i},\end{aligned}$$

and $\mathbf{P}_{i+1}$ is a $p_1p_2\cdots p_{i+1}$-point stride p_{i+1} permutation matrix.

Proof. We prove this by induction. By Lemma 3.6, we know (21) is valid for $n = 2$. Assume that the equation holds for $K = p_1p_2\cdots p_{n-1}$, and prove this for $N = p_1p_2\cdots p_n$. Since $N = p_1p_2\cdots p_n = Kp_n$, applying Lemma 3.6 once more, we obtain

$$\mathbf{F}(N) = (\mathbf{F}(K) \otimes \mathbf{I}_{p_n})\mathbf{T}_{p_n}(N)\mathbf{P}(N,p_n)(\mathbf{F}(p_n) \otimes \mathbf{I}_K). \tag{29}$$

By hypothesis

$$\mathbf{F}(K) = \left(\prod_{i=1}^{n-2}(\mathbf{F}(p_i) \otimes \mathbf{I}_{\delta_i})(\mathbf{T}_{i+1} \otimes \mathbf{I}_{\lambda_{i+1}})(\mathbf{P}_{i+1} \otimes \mathbf{I}_{\lambda_{i+1}})\right)(\mathbf{F}(p_{n-1}) \otimes \mathbf{I}_{\delta_{n-1}}), \tag{30}$$

with $\mathbf{T}_i, \mathbf{P}_i, \lambda_i$, and δ_i defined by (21) and (22), respectively. Combining the two equations, (23) and (24), the result follows. □

Now we use Theorem 3.1 to express the Fourier transform of any N-vector $\mathbf{x}$ in terms of the p-product. The notation $\mathbf{ones}(n)$ denotes a row vector of size n all of whose elements are equal to 1.

THEOREM 3.2. *Let $N = p_1 p_2 \cdots p_n$, with p_i a prime number and $p_i \leq p_{i+1}$. Then the p-product Fourier transform of any N-point vector $\mathbf{x}$ is given by*

$$\mathbf{y} = \frac{1}{\sqrt{N}}\mathbf{F}(N)\mathbf{x} = \frac{1}{\sqrt{N}}(\mathbf{w}_1 \oplus_{p_1} \mathbf{D}_2 * (\mathbf{w}_2 \oplus_{p_2} \mathbf{D}_3 * (\cdots (\mathbf{D}_n * (\mathbf{w}_n \oplus_{p_n} \mathbf{x})))))', \tag{31}$$

where $\mathbf{w}_i = \mathrm{row}(\mathbf{F}(p_i))$,

$$\begin{aligned} &\mathbf{D}_i = [\mathbf{d}_i^0, \mathbf{d}_i^1, \ldots, \mathbf{d}_i^{p_i-1}] \otimes \mathbf{ones}(\lambda_i), \\ &\textit{with } \mathbf{d}_i = ([1, w, w^2, \ldots, w^{\gamma_i-1}]')^{\lambda_i}, \\ &\lambda_i = N/p_1 p_2 \cdots p_i, \gamma_i = p_1 p_2 \cdots p_{i-1}, \end{aligned} \tag{32}$$

is an

$$p_1 p_2 \cdots p_{i-1} \times p_i p_{i+1} \cdots p_n \tag{33}$$

matrix associated with the diagonal elements of $\mathbf{T}_i$ with $w = w^{2\pi i/N}$ and $*$ denotes the Hadamard matrix product, defined componentwise.

Proof. We organize the computation of (21) into stages by setting

$$\begin{aligned} &\mathbf{Y}_n = (\mathbf{T}_n \otimes \mathbf{I}_{\lambda_n})(\mathbf{P}_n \otimes \mathbf{I}_{\lambda_n})(\mathbf{F}(p_n) \otimes \mathbf{I}_{\delta_n}), \\ &\mathbf{Y}_{n-1} = (\mathbf{T}_{n-1} \otimes \mathbf{I}_{\lambda_{n-1}})(\mathbf{P}_{n-1} \otimes \mathbf{I}_{\lambda_{n-1}})(\mathbf{F}(p_{n-1}) \otimes \mathbf{I}_{\delta_{n-1}}), \\ &\vdots \\ &\mathbf{Y}_i = (\mathbf{T}_i \otimes \mathbf{I}_{\lambda_i})(\mathbf{P}_i \otimes \mathbf{I}_{\lambda_i})(\mathbf{F}(p_i) \otimes \mathbf{I}_{\delta_i}), \\ &\vdots \\ &\mathbf{Y}_2 = (\mathbf{T}_2 \otimes \mathbf{I}_{\lambda_2})(\mathbf{P}_2 \otimes \mathbf{I}_{\lambda_2})(\mathbf{F}(p_2) \otimes \mathbf{I}_{\delta_2}), \\ &\mathbf{Y}_1 = \mathbf{F}(p_1) \otimes \mathbf{I}_{\delta_1}. \end{aligned} \tag{34}$$

To prove the theorem, we first verify the formula

$$\mathrm{col}(\mathbf{T}_k \mathbf{P}_k (\mathbf{F}(p_k) \otimes \mathbf{I}_{p_1 p_2 \cdots p_{k-1}})\mathbf{x}) = \mathrm{col}(\mathbf{D}_k * (\mathbf{w}_k \oplus_{p_k} \mathbf{x})), \tag{35}$$

where $\mathbf{D}_k$ is a matrix having the same size as $\mathbf{w}_k \oplus_{p_k} \mathbf{x}$ and associated with the diagonal matrix $\mathbf{T}_k, \mathbf{w}_k = \mathrm{row}(\mathbf{F}(p_k))$, and $\mathbf{x}$ is a $p_1 p_2 \cdots p_k \times t$ matrix with t being a positive integer.

In fact, let $\mathbf{x}$ be a designated matrix. Then by Lemmas 3.3 and 3.5, we have

$$\begin{aligned} \mathrm{col}(\mathbf{P}_k(\mathbf{F}(p_k) \otimes \mathbf{I}_{p_1 \cdots p_{k-1}})\mathbf{x}) &= \mathrm{col}(\mathbf{P}_k(\mathbf{F}(p_k) \oplus_{p_k} \mathbf{x})) \\ &= \mathrm{col}(\mathrm{row}(\mathbf{F}(p_k)) \oplus_{p_k} \mathbf{x}) = \mathrm{col}(\mathbf{w}_k \oplus_{p_k} \mathbf{x}). \end{aligned} \tag{36}$$

Now to $\mathbf{T}_k$, associate a $p_1 \cdots p_{k-1} \times p_k t$ matrix $\mathbf{D}_k$ such that the above formula is true (because of the lengthy specification, the details of finding $\mathbf{D}_k$ have been omitted here).

Now we apply (29) to prove the theorem. Let $\mathbf{x}$ be an N-vector. Then

$$
\begin{aligned}
\mathbf{Y}_n\mathbf{x} &= (\mathbf{T}_n \otimes \mathbf{I}_{\lambda_n})(\mathbf{P}_n \otimes \mathbf{I}_{\lambda_n})(\mathbf{F}(p_n) \otimes \mathbf{I}_{\delta_n})\mathbf{x} \\
&= \mathbf{T}_n\mathbf{P}_n(\mathbf{F}(p_n) \otimes \mathbf{I}_{p_1\cdots p_{n-1}})\mathbf{x} \\
&= \operatorname{col}(\mathbf{D}_n * (\mathbf{w}_n \oplus_{p_n} \mathbf{x})) = \operatorname{col}(\mathbf{y}_n), \\
\mathbf{Y}_{n-1}\operatorname{col}(\mathbf{y}_n) &= (\mathbf{T}_{n-1} \otimes \mathbf{I}_{\lambda_{n-1}})(\mathbf{P}_{n-1} \otimes \mathbf{I}_{\lambda_{n-1}})(\mathbf{F}(p_{n-1}) \otimes \mathbf{I}_{\delta_{n-1}})\operatorname{col}(\mathbf{y}_n) \\
&= (\mathbf{T}_{n-1} \otimes \mathbf{I}_{p_n})(\mathbf{P}_{n-1} \otimes \mathbf{I}_{p_n})(\mathbf{F}(p_{n-1}) \otimes \mathbf{I}_{p_1p_2\cdots p_{n-2}p_n})\operatorname{col}(\mathbf{y}_n) \\
&= ((\mathbf{T}_{n-1}\mathbf{P}_{n-1}(\mathbf{F}(p_{n-1}) \otimes \mathbf{I}_{p_1\cdots p_{n-2}})) \otimes \mathbf{I}_{p_n})\operatorname{col}(\mathbf{y}_n) \\
&= \operatorname{col}(\mathbf{T}_{n-1}\mathbf{P}_{n-1}(\mathbf{F}(p_{n-1}) \otimes \mathbf{I}_{p_1\cdots p_{n-2}})\mathbf{y}_n) \\
&= \operatorname{col}(\mathbf{D}_{n-1} * (\mathbf{w}_{n-1} \oplus_{p_{n-1}} \mathbf{y}_n)) = \operatorname{col}(\mathbf{y}_{n-1}).
\end{aligned} \tag{37}
$$

Similarly, for any integer $1 \le i \le n-2$, we have

(38)

$$
\begin{aligned}
\mathbf{Y}_i\operatorname{col}(\mathbf{y}_{i+1}) &= (\mathbf{T}_i \otimes \mathbf{I}_{\lambda_i})(\mathbf{P}_i \otimes \mathbf{I}_{\lambda_i})(\mathbf{F}(p_i) \otimes \mathbf{I}_{\delta_i})\operatorname{col}(\mathbf{y}_i) \\
&= (\mathbf{T}_i \otimes \mathbf{I}_{p_{i+1}\cdots p_n})(\mathbf{P}_i \otimes \mathbf{I}_{p_{i+1}\cdots p_n})(\mathbf{F}(p_i) \otimes \mathbf{I}_{p_1\cdots p_{i-1}p_{i+1}\cdots p_n})\operatorname{col}(\mathbf{y}_i) \\
&= ((\mathbf{T}_i\mathbf{P}_i(\mathbf{F}(p_i) \otimes \mathbf{I}_{p_1\cdots p_{i-1}})) \otimes \mathbf{I}_{p_{i+1}\cdots p_n})\operatorname{col}(\mathbf{y}_i) \\
&= \operatorname{col}(\mathbf{T}_i\mathbf{P}_i(\mathbf{F}(p_i) \otimes \mathbf{I}_{p_1\cdots p_{i-1}})\mathbf{y}_i) \\
&= \operatorname{col}(\mathbf{D}_i * (\mathbf{w}_i \oplus_{p_i} \mathbf{y}_i)) = \operatorname{col}(\mathbf{y}_i).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\mathbf{y} &= \frac{1}{\sqrt{N}}\mathbf{F}(N)\mathbf{x} = \frac{1}{\sqrt{N}}\mathbf{Y}_1\mathbf{Y}_2\cdots\mathbf{Y}_n\mathbf{x} \\
&= \frac{1}{\sqrt{N}}\operatorname{col}(\mathbf{w}_1 \oplus_{p_1} (\mathbf{D}_2 * (\mathbf{w}_2 \oplus_{p_2} (\cdots (\mathbf{D}_n * (\mathbf{w}_n \oplus_{p_n} \mathbf{x})))))).
\end{aligned} \tag{39}
$$

Since $\mathbf{y}$ is a vector, the result follows. □

Special case. When $p_1 = p_2 = \cdots = p_n = 2$, i.e., $N = 2^n$, the p-product Fourier transform of any N-point vector $\mathbf{x}$ is given by

$$
\mathbf{y} = \frac{1}{\sqrt{N}}\mathbf{F}(N)\mathbf{x} = \frac{1}{\sqrt{N}}(\mathbf{w} \oplus_2 \mathbf{D}_2 * (\mathbf{w} \oplus_2 \mathbf{D}_3 * (\cdots (\mathbf{D}_n * (\mathbf{w} \oplus_2 \mathbf{x})))))', \tag{40}
$$

where $\mathbf{w}$ is the row vector corresponding to $\mathbf{F}(2)$, i.e.,

$$
\mathbf{w} = \operatorname{row}(\mathbf{F}(2)) = [1, 1, 1, -1], \tag{41}
$$

and

$$
\begin{aligned}
&\mathbf{D}_i = [\mathbf{d}_i^0, \mathbf{d}_i] \otimes \mathbf{ones}(2^{n-i}) \quad \text{for } i = 2, \ldots, n, \\
&\text{where } \mathbf{d}_i = ([1, w, w^2, \ldots, w^{(2^{i-1}-1)}]2^{n-i})'
\end{aligned} \tag{42}
$$

is a $2^{i-1} \times 2^{n-i+1}$ matrix associated with the diagonal elements of $\mathbf{T}_i$.

Notice that the number of arithmetical operations required in the p-product Fourier transform is same as that in the regular FFT, both having the order of $N \log N$. However, the additional reordering process required in the regular FFT plays an important role on modern architecture machines. Stone [19] showed that up to $O(N)$

TABLE 1
Time required to compute Fourier transform of N-vectors on the SUN 3 *Workstation using* MATLAB *with Cooley–Tukey* FFT *and the p-product Fourier transform algorithms.*

	$N=2$	$N=4$	$N=8$	$N=16$	$N=32$	$N=64$	$N=128$	$N=256$	$N=512$	$N=1024$
t_p	0.018	0.033	0.047	0.063	0.086	0.122	0.177	0.279	0.491	0.909
t_c	0.032	0.048	0.066	0.090	0.116	0.167	0.230	0.369	0.592	1.110

t_c Time required for using Cooley–Tukey FFT algorithm.
t_p Time required for using p-product Fourier transform algorithm.

time is required to execute the reverse operation for N being a power of two. This means that the p-product Fourier transform algorithm can save time by more than $O(N)$. The larger the number N, the more time that can be saved by using the new algorithm.

Although the p-product FFT exhibits similarities to the Stockham autosort algorithm (e.g., [21]), it has the advantage of (i) concise expression, which is achieved by replacing the permutation matrix as well as the tensor product with the single p-product; and (ii) a standard implementation with the p-product as a basis of operation. For example, if $N = 8$, then $w = e^{\pi i/4}$, and the p-product Fourier transform of an 8-point vector is $\mathbf{x}$ is given by

$$\mathbf{y} = \frac{1}{\sqrt{8}}\mathbf{F}(8)\mathbf{x} = \frac{1}{\sqrt{8}}(\mathbf{w} \oplus_2 \mathbf{D}_2 * (\mathbf{w} \oplus_2 \mathbf{D}_3 * (\mathbf{w} \oplus_2 \mathbf{x})))', \tag{43}$$

where

$$\mathbf{w} = (1,1,1,-1), \quad \mathbf{D}_2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & w^2 & w^2 \end{pmatrix}, \quad \text{and } \mathbf{D}_3 = \begin{pmatrix} 1 & 1 \\ 1 & w \\ 1 & w^2 \\ 1 & w^3 \end{pmatrix}. \tag{44}$$

The signal flowgraph of this 8-point transform is shown in Fig. 1 and is similar to the ordinary FFT except for the order of the input signal.

We have implemented this new algorithm on a sequential machine (SUN 3 Workstation) using MATLAB. As shown in Table 1, the time required for the new algorithm is less than the time required by the Cooley–Tukey method.

Since the Fourier transform is separable, it is easy to induce the two-dimensional case using the p-product language from the theorem we proved in this section. The only thing we need to do is to define a function ψ_n: $\mathbb{F}^m \to \mathbb{F}_{l\times n}$ as

$$\psi_n([a_1, a_2, \ldots, a_m]) = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \\ a_{n+1} & a_{n+2} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{(l-1)n+1} & a_{(l-1)n+2} & \cdots & a_{ln} \end{bmatrix}, \tag{45}$$

with $m = ln$, then apply this function after performing row and column transforms.

THEOREM 3.3. *Let $N = p_1p_2\cdots p_n$ and $M = q_1q_2\cdots q_m$. For a two-dimensional $N \times M$ array $\mathbf{x}$, its Fourier transform, defined by $\mathbf{y}$, is*

$$\mathbf{y} = \mathbf{F}(N)\mathbf{f}\mathbf{F}(M). \tag{46}$$

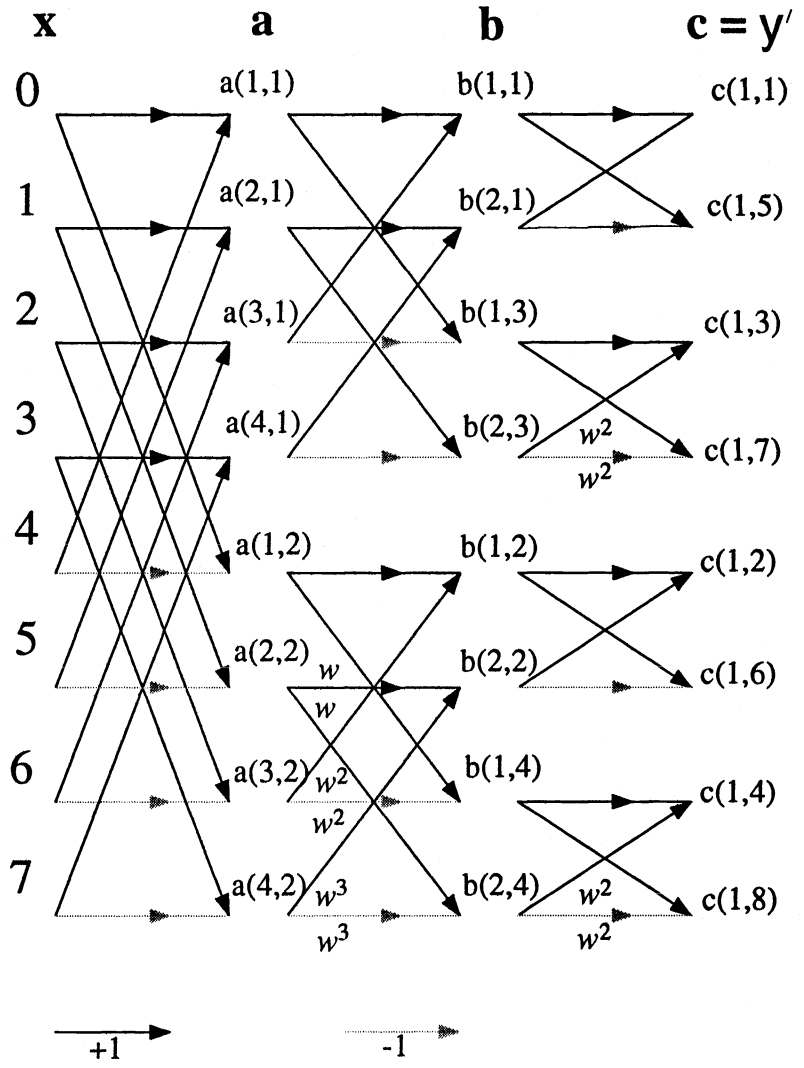


FIG. 1. *Signal flow graph of an 8-point discrete p-product Fourier transform. Multiplying factors* $+1$ *and* -1 *are indicated by solid and dotted branches, respectively.*

Then its p-product factorization is

$$\mathbf{y} = \frac{1}{\sqrt{NM}}\psi_M(\mathbf{w}_1 \oplus_{p_1} \mathbf{D}_2 * (\mathbf{w}_2 \oplus_{p_2} \mathbf{D}_3 * (\cdots(\mathbf{D}_n * (\mathbf{w}_n \oplus_{p_n} \hat{\mathbf{x}}'))))) \tag{47}$$

with

$$\hat{\mathbf{x}} = \mathbf{F}(M)\mathbf{f}' = \psi_N[\mathbf{v}_1 \oplus_{q_1} \mathbf{E}_2 * (\mathbf{v}_2 \oplus_{q_2} \mathbf{E}_3 * (\cdots \mathbf{E}_m * (\mathbf{v}_m \oplus_{p_m} \mathbf{x}'))))], \tag{48}$$

where $\mathbf{w}_i, \mathbf{D}_i$, *and* $\mathbf{v}_i, \mathbf{E}_i$ *are defined as in Theorem* 3.2.

4. The Walsh transform. In this section, we use the similarities and differences among the Fourier transform, the Walsh transform, and the generalized Walsh transform to induce new fast algorithms of the Walsh transform and the generalized Walsh transform from the FFT formula in terms of the p-product presented in §5.

To define the generalized Walsh functions, Chrestenson defined the following Rademacher functions of order α [3].

Let α denote a fixed integer, $\alpha \geq 2$, and put $\omega = e^{2\pi i/\alpha}$.

DEFINITION 4.1. *The* Rademacher functions *of order α are defined by*

$$\phi_0(x) = \omega^k \quad \text{if } k/\alpha \leq x \leq (k+1)/\alpha, \quad k = 0, \ldots, \alpha - 1, \tag{49}$$

and for $n \geq 0$

$$\phi_n(x+1) = \phi_n(x) = \phi_0(\alpha^n x). \tag{50}$$

Thus, under this definition, when $\alpha = 2$, the Rademacher function of index n is a train of rectangular pulses with 2^n cycles in the half open interval $[0, 1)$, taking the values $+1$ or -1. Its period is 1.

Using Rademacher functions we can define the Walsh functions of order α.

DEFINITION 4.2. *The* Walsh functions *of order α are defined by*

$$\psi_0(x) = 1 \tag{51}$$

and if $n = a_1\alpha^{n_1} + \cdots + a_m\alpha^{n_m}$, where $0 < a_j < \alpha$ and $n_1 > n_2 > \cdots > n_m$, then

$$\psi_n(x) = \phi_{n_1}^{a_1}(x) \cdots \phi_{n_m}^{a_m}(x). \tag{52}$$

For convenience we let ψ_α denote the set of Walsh functions of order α. Then ψ_2 is the set of functions defined by Walsh. Different from this, we refer to $\psi_\alpha, \alpha > 2$, as the set of *generalized Walsh functions.* It has been proved that ψ_α is orthonormal and complete in $\mathbf{L}(1, 0)$ [3].

Following the above definition, in the case of $\alpha = 2$, we write the one-dimensional forward Walsh kernel as

$$h(x, u) = \frac{1}{\sqrt{N}} \prod_{i=1}^{n-1} (-1)^{b_i(x) b_{n-1-i}(u)}, \tag{53}$$

where $b_k(z)$ is the kth bit in the binary representation of z. For example, if $n = 3$, and $z = 6$, (110 is binary), we have that $b_0(z) = 0, b_1(z) = 1$, and $b_2(z) = 1$.

By using this kernel, we have the following one-dimensional Walsh transform of a function $\mathbf{f}$

$$g(u) = \frac{1}{\sqrt{N}} \sum_{x=0}^{N-1} f(x) h(x, u) = \frac{1}{\sqrt{N}} \sum_{x=0}^{N-1} f(x) \prod_{i=0}^{n-1} (-1)^{b_i(x) b_{n-1-i}(u)}. \tag{54}$$

Thus $\mathbf{g} = \mathbf{H}_W(N)\mathbf{f}$.

In [18], Shanks proved that (46) can be decomposed as

$$\begin{aligned} g(u) &= g(j_{m-1}, \ldots, j_0) \\ &= \frac{1}{\sqrt{N}} \sum_{k_0=0}^{1} (-1)^{j_{m-1}k_0} \sum_{k_1=0}^{1} (-1)^{j_{m-1}k_1} \cdots \sum_{k_{m-1}=0}^{1} (-1)^{j_0 k_{m-1}} f\left(\sum_{i=0}^{m-1} 2^i k_i \right), \end{aligned} \tag{55}$$

with $u = \sum_{i=0}^{m-1} 2^i j_i$ and the intermediate Walsh transform arrays defined by

$$a_l(k_{m-l-1}, \ldots, k_0; j_{l-1}, \ldots, j_0) = \sum_{k_{m-l+1}=0}^{1} a_{l-1}(k_{m-l}, \ldots, k_0; j_{l-2}, \ldots, j_0)(-1)^{j_{l-1}k_{m-1}}, \tag{56}$$

for $l = 1, 2, \ldots, m$, and

$$a_0(k_{m-1}, \ldots, k_0) = f\left(\sum_{i=0}^{m-1} 2^i k_i\right). \tag{57}$$

For $N = \alpha^m$. The N-point one-dimensional forward generalized Walsh kernel is given by the relation

$$h(x, u) = \frac{1}{\sqrt{N}} \prod_{i=1}^{n-1} w^{b_i(x)b_{n-1-i}(u)}, \tag{58}$$

where $w = e^{2\pi i/\alpha}$ and $b_k(z)$ is the kth bit in the α–nary representation of z, for example, if $\alpha = 3, N = 3^3$, and $z = 6$ (020 is trinary), we have $b_0(z) = 0, b_1(z) = 2$, and $b_2(z) = 0$.

By using this kernel, we have the following one-dimensional generalized Walsh transform of a function $\mathbf{f}$:

$$g(u) = \frac{1}{\sqrt{N}} \sum_{x=0}^{N-1} f(x)h(x, u) = \frac{1}{\sqrt{N}} \sum_{x=0}^{N-1} f(x) \prod_{i=0}^{n-1} w^{b_i(x)b_{n-i-1}(u)}. \tag{59}$$

Thus $\mathbf{g} = \mathbf{H}_W(N)\mathbf{f}$.

Similar to the factorization of the Walsh transform, the N-point generalized Walsh transform can be factored as

$$\begin{aligned} g(u) &= g(j_{m-1}, \ldots, j_0) \\ &= \frac{1}{\sqrt{N}} \sum_{k_0=0}^{\alpha-1} w^{j_{m-1}k_0} \sum_{k_1=0}^{\alpha-1} w^{j_{m-2}k_1} \cdots \sum_{k_{m-1}=0}^{\alpha-1} w^{j_0 k_{m-1}} f\left(\sum_{i=0}^{m-1} \alpha^i k_i\right), \end{aligned} \tag{60}$$

with $u = \sum_{i=0}^{m-1} \alpha^i j_i$. Its similarity to the Cooley–Tukey decompositions of the Fourier transform is obvious. In fact, Shanks [18] and Andrews and Caspari [1] pointed out that the Fourier transform equation differs from the Walsh transform equation and the generalized Walsh transform equation by the introduction of a scalar (or a matrix) multiplication at each stage. In the p-product equation of the Fourier transform, this matrix is $\mathbf{D}_i$ in stage i. Hence, by removing the matrices $\mathbf{D}_i$ in (25), theorems for the p-product Walsh transform and the p-product generalized Walsh transform are induced as follows.

THEOREM 4.1. *Let $N = 2^m$. For a function $\mathbf{f} = [f(0), f(1), \ldots, f(N-1)]'$, its Walsh transform in terms of the p-product is given by*

$$\mathbf{g} = \mathbf{H}_W(N)\mathbf{f} = \frac{1}{\sqrt{N}}[\mathbf{w}_1 \oplus_2 (\mathbf{w}_2 \oplus_2 (\cdots (\mathbf{w}_m \oplus_2 \mathbf{f})))]', \tag{61}$$

where $\mathbf{w}_j = [1, 1, 1, -1]$ *for* $j = 1, 2, \ldots, m$.

We call $\mathbf{w} = [1, 1, 1, -1]$ the *core matrix* of the Walsh transform for the dimensions of powers of 2.

THEOREM 4.2. *Let* $N = \alpha^m$. *For an* N*-length function* $\mathbf{f}$, *its generalized Walsh transform in terms of the p-product is given by*

$$\mathbf{g} = \mathbf{H}_W(N)\mathbf{f} = \frac{1}{\sqrt{N}}[\mathbf{w}_1 \oplus_\alpha (\mathbf{w}_2 \oplus_\alpha (\cdots (\mathbf{w}_m \oplus_\alpha \mathbf{f})))]', \tag{62}$$

where $\mathbf{w}_i$ *is the row vector corresponding to the Fourier matrix* $\mathbf{F}(\alpha)$ *and is called the core matrix of the generalized Walsh transform for dimensions of powers of* α.

Combining the above two theorems, we obtain a more general formula for N being the factors of the prime numbers, i.e., $N = p_1 p_2 \cdots p_n$.

THEOREM 4.3. *Let* $N = p_1 p_2 \cdots p_n$ *with* p_i *being a prime number and* $p_i \leq p_{i+1}$. *Then the p-product generalized Walsh transform of any* N*-point vector* $\mathbf{f}$ *is given by*

$$\mathbf{g} = \mathbf{H}_W(N)\mathbf{f} = \frac{1}{\sqrt{N}}(\mathbf{w}_1 \oplus_{p_1} (\mathbf{w}_2 \oplus_{p_2} (\cdots (\mathbf{w}_n \oplus_{p_n} \mathbf{f}))))', \tag{63}$$

where $\mathbf{w}_i$ *is the row vector corresponding to the Fourier matrix* $\mathbf{F}(p_i)$.

The same as the Fourier transform, both Walsh and generalized Walsh transforms are separable. Therefore, to induce the two-dimensional Walsh transform formulation in terms of the p-product, we apply the function ψ_n defined in §5 after performing row and column transforms.

THEOREM 4.4. *Let* $N = p_1 p_2 \cdots p_n, M = q_1 q_2 \cdots q_m$, *and* $\mathbf{f}$ *be an* $N \times M$ *array. Then its generalized Walsh transform in terms of the p-product is given by*

$$\mathbf{g} = \frac{1}{\sqrt{NM}}\psi_M(\mathbf{w}_1 \oplus_{p_1} (\mathbf{w}_2 \oplus_{p_2} (\cdots (\mathbf{w}_n \oplus_{p_n} \hat{\mathbf{f}}')))), \tag{64}$$

with

$$\hat{\mathbf{f}} = \psi_N(\mathbf{v}_1 \oplus_{q_1} (\mathbf{v}_2 \oplus_{q_2} (\cdots (\mathbf{v}_m \oplus_{q_m} \mathbf{f}')))), \tag{65}$$

where $\mathbf{w}_i$ *and* $\mathbf{v}_i$ *are the row vectors corresponding to* $\mathbf{F}(p_i)$ *and* $\mathbf{F}(q_i)$, *respectively.*

Notice that the p-product Walsh transform (53) deals only with values $+1$ and -1. There is no multiplication involved. Meanwhile, the reordering process required in most Walsh transform algorithms has been eliminated here. This makes the new p-product Walsh transform faster and more efficient. Due to the reordering process, researchers have had difficulties implementing fast versions of the generalized Walsh transform. This roadblock has now been eliminated. The p-product generalized Walsh transform algorithm provides an implementation on the order of $N \log N$ without the reordering process. This is much faster than the traditional method that required N^2 operations. Finally, this section also demonstrates that with similar implementation techniques, we can use the p-product to establish simplified algorithms for efficient matrix operations for a class of generalized tensor matrices. More precisely, any orthogonal or nonorthogonal transforms obtainable from a series of matrix tensor products can be expressed simply in terms of the p-product and efficiently implemented on digital computers. Thus, in addition to the transforms discussed here, other implementable transforms are the Hadamard transform, the generalized transform, and a variety of other similar transforms.

5. The wavelet transform. In 1988, Daubechies defined the notion of the "multiplier 2" compactly supported discrete wavelet transform and obtained conditions for smoothness and polynomial representation by multiplier 2 wavelet series [10]. In particular, she defined a scaling function $\phi(x)$ as a compactly supported solution of

$$\phi(x) = \sum_{k=0}^{2g-1} a_k \phi(2x - k), \tag{66}$$

where $a_0, a_1, \ldots, a_{2g-1}$ are the scaling coefficients. Associated with this scaling function of the wavelet system there is another set of coefficients, $b_k = (-1)^k a_{2g-1-k}$, that defines the wavelet function as

$$\psi(x) = \sum_{k=0}^{2g-1} b_k \phi(2x - k). \tag{67}$$

Using these definitions, Heller et al. [5] introduced wavelet matrices as generalizations of the $2 \times 2g$ matrix of the form

$$\begin{pmatrix} a_0 & \cdots & a_{2g-1} \\ b_0 & \cdots & b_{2g-1} \end{pmatrix}, \tag{68}$$

where the a's and b's are defined as above. It is not difficult to ascertain that this satisfies the wavelet scaling property and that

$$\sum_k a_k = 2 \quad \text{and} \quad \sum_k b_k = 0. \tag{69}$$

To generalize this concept one may define

$$\mathbf{a} = \begin{pmatrix} a_0^0 & \cdots & a_{2g-1}^0 \\ a_0^1 & \cdots & a_{2g-1}^1 \end{pmatrix}, \tag{70}$$

where $a_i^0 = a_i$ and $a_i^1 = b_i$. The general $m \times mg$ matrix is then of the form

$$\begin{pmatrix} a_0^0 & a_1^0 & \cdots & a_{mg-1}^0 \\ a_0^1 & a_1^1 & \cdots & a_{mg-1}^1 \\ \vdots & \vdots & & \vdots \\ a_0^{m-1} & a_1^{m-1} & \cdots & a_{mg-1}^{m-1} \end{pmatrix}, \tag{71}$$

with the wavelet scaling conditions

$$\sum_k a_k^s = m\delta^{r,0} \quad \text{and} \quad \sum_k \bar{a}_{k+ml'}^{r'} a_{k+ml}^r = m\delta^{r',r}\delta_{l',l}, \tag{72}$$

where m is the *rank* of the matrix, and g denotes the *genus* of the wavelet matrix, i.e., the number of $m \times m$ blocks in the matrix. The vector a^0 is called the *scaling vector* and for $0 < s < m$, $\mathbf{a}^s$ is called a *wavelet vector.*

Note that the wavelet matrices of rank m correspond to a wavelet system with multiplier m, replacing the multiplier 2 used by both Daubechies [10] and Mallat [13].

Analogous to the case $m = 2$, for the above definition of the wavelet matrix $\mathbf{a}$, a scaling function ψ^0 and $m - 1$ fundamental wavelets are defined by the system of equations

$$\psi^r(x) = \sum_k a_k^r \psi^0(mx - k), \tag{73}$$

where $0 \leq r < m$. The scaling function is to be thought of as a low-pass function while the fundamental wavelets are high-pass functions. If we define a set of auxiliary functions by the formula

$$\psi_{jk}^r := m^{j/2} \psi^r(m^j x - k), \tag{74}$$

where $j, k \in \mathbb{Z}$, then Resnikoff [15] has shown that the set

$$\{\psi_{jk}^r(x) : 0 \leq r < m, j, k \in \mathbb{Z}\} \tag{75}$$

is an orthonormal basis for $\mathbf{L}^2(\mathbb{R})$. The support of $\psi_{jk}^r(x)$ has length equal to the length of the support of the scaling function $\psi^0(x)$ divided by m^j, where the quantity j is called the *scale* of the wavelet function. Thus, any function f in $\mathbf{L}^2(\mathbb{R})$ can be represented as a wavelet series with a wavelet coefficient matrix $\mathbf{a}$.

A wavelet matrix for which $g = 1$, i.e., a square wavelet matrix, is said to a *Haar* matrix. A number of classical examples of specific matrices that have different origins in mathematics and signal processing can all be seen to be Haar wavelet matrices of specific types. These include the finite Fourier transform matrices, the discrete cosine transform matrix, Hadamard and Walsh matrices, Rademacher matrices, and Chebyshev matrices.

Let $f(x) = \sum_n f_n(x)$, where f_n is a sequence of functions $x \to f_n(x)$ defined on some infinite set (e.g., $\mathbb{Z}$ or $\mathbb{R}$). The function $f(x)$ will have a meaning that is prescribed by the type of convergence that is assumed. Let us suppose that for each x, only finitely many of the numbers $f_n(x)$ are nonzero and assume that $a_k^s = 0$ unless $0 \leq k < mg$. The following theorem proved by Heller et al. in [5] exhibits a locally finite compact wavelet matrix series for an arbitrary discrete function f.

THEOREM 5.1. *Let f: $\mathbb{Z} \to \mathbb{C}$ be an arbitrary function defined on the integers and let* $\mathbf{a}$ *be a compact wavelet matrix of* rank m *and* genus g *defined by* (63). *Then f has a unique wavelet matrix expansion*

$$f(n) = \sum_{l \in \mathbb{Z}} c_l a_{n-ml} + \sum_{k \in \mathbb{Z}} \sum_{s=1}^{m-1} c_k^s a_{n-mk}^s, \tag{76}$$

where

$$c_l = \frac{1}{m} \sum_n f(n) \bar{a}_{n-ml} \tag{77}$$

and

$$c_k^s = \frac{1}{m} \sum_n f(n) \bar{a}_{n-mk}^s. \tag{78}$$

The wavelet matrix expansion is locally finite, i.e., for given n only finitely many terms of the series are different from zero.

Remark. Using the language of signal processing, the first term in (65) is the "low-pass" part of the expansion and the second term is the "high-pass" part of the expansion.

Now, if we denote $c_l^0 = c_l$ and $\bar{a}_{n-mk}^0 = \bar{a}_{n-mk}$, then (66) and (67) can be combined into the single formula

$$c_k^s = \frac{1}{m} \sum_n f(n) \bar{a}_{n-mk}^s, \qquad 0 \le s < m. \tag{79}$$

Using the above assumption on $\mathbf{a}$, we have

$$c_k^s = \frac{1}{m} \sum_{n=mk} f(n) \bar{a}_{n-mk}^s. \tag{80}$$

Since theoretically k ranges from $-\infty$ to $+\infty$, in what follows, we analyze (69) for the two cases $k \ge 0$ and $k < 0$.

For $k \ge 0$, we have

$$c_k^s = \frac{1}{m} \sum_{n=mk}^{mg+mk-1} f(n) \bar{a}_{n-mk}^s = \frac{1}{m} \sum_{n=0}^{mg-1} f(n+km) \bar{a}_n^s. \tag{81}$$

Decomposing the summations in (70) into g smaller ones of length m, we obtain

$$\begin{aligned} c_k^s &= \frac{1}{m} \sum_{n=0}^{mg-1} f(n+km) \bar{a}_n^s \\ &= \frac{1}{m} \left[\sum_{p=0}^{m-1} f(p+km) \bar{a}_p^s + \sum_{p=0}^{m-1} f(m+p+km) \bar{a}_{m+p}^s \right. \\ &\qquad \left. + \cdots + \sum_{p=0}^{m-1} f(m(g-1)+p+km) \bar{a}_{m(g-1)+p}^s \right]. \end{aligned} \tag{82}$$

Since the length of f is N and $f(n) = 0$ for $n > N$, the length of the vector c_+^s is N/m. That is, c_+^s is of the form $c_+^s = [c_0^s, c_1^s, \ldots, c_{N/m-1}^s]$. For $0 \le j < g$, let

$$\mathbf{h}_j = [f(jm), \ldots, f(jm+N-1)] \oplus_m \begin{bmatrix} \bar{a}_{mj}^s \\ \vdots \\ \bar{a}_{m(j+1)-1}^s \end{bmatrix}. \tag{83}$$

By definition of the p-product, the size of $\mathbf{h}_j$ is $1 \times N/m$. It is not difficult to prove that the first summation of c_k^s in (71) corresponds to $h_0(k)$, the second one to $h_1(k)$, and the jth one to $h_{j-1}(k)$.

Let

$$\bar{\mathbf{w}}_0^s = \begin{bmatrix} \bar{a}_0^s \\ \vdots \\ \bar{a}_{m-1}^s \end{bmatrix}, \ldots, \quad \bar{\mathbf{w}}_j^s = \begin{bmatrix} \bar{a}_{mj}^s \\ \vdots \\ \bar{a}_{m(j+1)-1}^s \end{bmatrix}, \ldots, \quad \bar{\mathbf{w}}_{g-1}^s = \begin{bmatrix} \bar{a}_{m(g-1)}^s \\ \vdots \\ \bar{a}_{mg-1}^s \end{bmatrix}, \tag{84}$$

and let

$$f_k = [f(mk), f(mk+1), \ldots, f(N-1), 0, \ldots, 0] \text{ be of size } 1 \times N. \tag{85}$$

Then

$$\mathbf{h}_j = f_j \oplus_m \bar{\mathbf{w}}_j^s. \tag{86}$$

Combining (71) and (75), we have

$$c_+^s = \frac{1}{m}\sum_{i=1}^{g-1} \mathbf{h}_j = \frac{1}{m}[f \oplus_m \bar{\mathbf{w}}_0^s + \cdots + f_{g-1} \oplus_m \bar{\mathbf{w}}_{g-1}^s], \qquad 0 \le s < m. \tag{87}$$

On the other hand, in the case of $k < 0$, (69) becomes

$$c_{-k}^s = \frac{1}{m}\sum_{n=-mk}^{mg-mk-1} f(n)\bar{a}_{n+mk}^s = \frac{1}{m}\sum_{n=0}^{mg-1} f(n-mk)\bar{a}_n^s. \tag{88}$$

Since $f(n) = 0$ unless $0 \le n < N$, we have

$$c_{-k}^s = \frac{1}{m}\sum_{n=mk}^{mg-1} f(n-mk)\bar{a}_n^s. \tag{89}$$

Thus the size of $c_-^s = [c_{-g+1}^s, \ldots, c_{-2}^s, c_{-1}^s]$ is $1 \times (g-1)$ and only the first $m(g-1)$ terms of f are used. Therefore, we denote

$$e = [f(0), f(1), \ldots, f(m(g-1)-1)] \tag{90}$$

and

$$e_{-i} = [0, 0, \ldots, 0, f(0), \ldots, f(m(g-i-1)-1)] \tag{91}$$

having the same size as e for $1 \le i < g-2$. Then by following the same procedure as above, we can decompose (77) as

$$\begin{aligned} c_-^s &= [c_{-g+1}^s, \ldots, c_{-2}^s, c_{-1}^s] \\ &= \frac{1}{m}[e \oplus_m \bar{\mathbf{w}}_{g-1}^s + e_{-1} \oplus_m \bar{\mathbf{w}}_{g-2}^s + \cdots + e_{-g+2} \oplus_m \bar{\mathbf{w}}_1^s] \end{aligned} \tag{92}$$

for $0 \le s < m$. Combining (76) and (81), the wavelet transform of f under the wavelet matrix $\mathbf{a}$ is

$$c = [c^0, c^1, \ldots, c^{m-1}]', \tag{93}$$

where

$$\begin{aligned} c^s &= [c_-^s; c_+^s] \\ &= \frac{1}{m}[c \oplus_m \bar{\mathbf{w}}_{g-1}^s + \cdots + e_{-g+2} \oplus_m \bar{\mathbf{w}}_1^s; f \oplus_m \bar{\mathbf{w}}_0^s + \cdots + f_{g-1} \oplus_m \bar{\mathbf{w}}_{g-1}^s] \end{aligned} \tag{94}$$

for $0 \le s < m$.

Before defining the inverse wavelet transform of f, we prove the following lemma.

LEMMA 5.1. *Let* $\bar{\mathbf{w}}_i^s$ *be defined as* (73). *Then*

$$\sum_{0\le s<m}\sum_{0\le i,j<g}(\bar{\mathbf{w}}_i^s \otimes (\mathbf{w}_j^s)') = \begin{cases} m\mathbf{I}_m & \text{if } i=j, \\ 0 & \text{if } i\ne j. \end{cases} \tag{95}$$

Proof. It holds that

$$\begin{aligned} &\sum_{0\le s<m}\sum_{0\le i,j<g}(\bar{\mathbf{w}}_i^s \otimes (\mathbf{w}_j^s)') \\ &= \sum_{0\le s<m}\sum_{i,j}\left(\begin{bmatrix} \bar{a}_{mi}^s \\ \vdots \\ \bar{a}_{m(i+1)-1} \end{bmatrix} \otimes [a_{mj}^s, a_{mj+1}^s, \dots, a_{m(j+1)-1}]\right) \\ &= \sum_s\sum_{i,j}\begin{bmatrix} \bar{a}_{mi}^s a_{mj}^s & \bar{a}_{mi}^s a_{mj+1}^s & \cdots & \bar{a}_{mi}^s a_{m(j+1)-1}^s \\ \vdots & \vdots & & \vdots \\ \bar{a}_{m(i+1)-1}^s a_{mj}^s & \bar{a}_{m(i+1)-1}^s a_{mj+1}^s & \cdots & \bar{a}_{m(i+1)-1}^s a_{m(j+1)-1}^s \end{bmatrix}. \end{aligned} \tag{96}$$

By the wavelet scaling conditions given on (64), we have

$$\delta(n',n) = \frac{1}{m}\sum_{0\le s<m}\sum_{l\in\mathbb{Z}}\bar{a}_{n-ml}^s a_{n'-ml}^s \tag{97}$$

and the results follow. □

LEMMA 5.2. *Let* $\mathbf{a}\in\mathbb{F}_{l\times n}, \mathbf{b}\in\mathbb{F}_{l\times n}$, *and* $\mathbf{c}\in\mathbb{F}_{n\times q}$. *Then*

$$[\mathbf{a}|\mathbf{b}]\oplus_n \mathbf{c} = [\mathbf{ac}|\mathbf{bc}]. \tag{98}$$

In particular, suppose $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k \in \mathbb{F}_{l\times n}$, *then*

$$[\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k]\oplus_n \mathbf{c} = [\mathbf{a}_1\mathbf{c}, \mathbf{a}_2\mathbf{c}, \dots, \mathbf{a}_k\mathbf{c}]. \tag{99}$$

Now we use above lemmas to prove the following theorem.

THEOREM 5.2. *Let* $f:\mathbb{Z}\to\mathbb{C}$ *be an arbitrary function defined on the integers and let* $\mathbf{a}$ *be a compact wavelet matrix of* rank m *and* genus g *defined by* (63). *Let* $e = [f(0), f(1), \dots, f(m(g-1)-1)]$. *Then the wavelet transform of* f *is given by*

$$\mathbf{c} = [\mathbf{c}^0, \mathbf{c}^1, \dots, \mathbf{c}^{m-1}]', \tag{100}$$

where

$$\begin{aligned} \mathbf{c}^s &= [\mathbf{c}_-^s; \mathbf{c}_+^s] \\ &= \frac{1}{m}[e\oplus_m \bar{\mathbf{w}}_{g-1}^s + \cdots + e_{-g+2}\oplus_m \bar{\mathbf{w}}_1^s; f\oplus_m \bar{\mathbf{w}}_0^s + \cdots + f_{g-1}\oplus_m \bar{\mathbf{w}}_{g-1}^s] \end{aligned} \tag{101}$$

for $0\le s<m$, *and its inverse transform is given by*

$$\begin{aligned} f &= \sum_{0\le s<m}\mathbf{d}^s, \\ \text{where } \mathbf{d}^s &= \mathbf{c}_0^s\otimes(\mathbf{w}_0^s)' + \mathbf{c}_1^s\otimes(\mathbf{w}_1^s)' + \cdots + \mathbf{c}_{g-1}^s\otimes(\mathbf{w}_{g-1}^s)', \\ \mathbf{c}_i^s &= [c_{-i}^s, c_{-i+1}^s, \dots, c_{-i+N/m-1}^s], \end{aligned} \tag{102}$$

and $\bar{\mathbf{w}}_i^s, f_i$, *and* e_{-i} *are as defined by* (73), (74), *and* (80), *respectively.*

Proof. In the above discussion, we have shown (90). Now, we use this equation and Lemmas 5.1 and 5.2 to prove its inverse formulation, i.e., (91).

Write

$$f_i = [V_i^0, V_i^1, \ldots, V_i^{N/m-1}], \tag{103}$$

with $V_i^j = [f_i(mj), \ldots, f_i(m(j+1)-1)]$ and

$$e_{-i} = [V_{-i}^0, V_{-i}^1, \ldots, V_{-i}^{g-2}]. \tag{104}$$

Then from the notation of f_i, we have

$$\begin{aligned} V_i^j &= [f_i(mj), \ldots, f_i(m(j+1)-1)] \\ &= [f(m(j+i)), \ldots, f(m(j+i+1)-1)] \\ &= V_0^{j+i}. \end{aligned} \tag{105}$$

Also note that

$$\begin{aligned} \mathbf{c}^s &= [\mathbf{c}_-^s; \mathbf{c}_+^s] \\ &= \frac{1}{m}[e \oplus_m \bar{\mathbf{w}}_{g-1}^s + \cdots + e_{-g+2} \oplus_m \bar{\mathbf{w}}_1^s; f \oplus_m \bar{\mathbf{w}}_0^s + \cdots + f_{g-1} \oplus_m \bar{\mathbf{w}}_{g-1}^s]. \end{aligned} \tag{106}$$

Using Lemma 5.2 and substituting f_i and e_{-i} by V_i^j, (95) becomes

(107)

$$\begin{aligned} \mathbf{c}^s &= \frac{1}{m}\left[\sum_{i=0}^{g-2} V_{-i}^0 \bar{\mathbf{w}}_{g-1-i}^s, \ldots, \sum_{i=0}^{g-2} V_{-i}^{g-2} \bar{\mathbf{w}}_{g-1-i}^s; \sum_{k=0}^{g-1} V_k^0 \bar{\mathbf{w}}_k^s, \ldots, \sum_{k=0}^{g-1} V_k^{N/m-1} \bar{\mathbf{w}}_k^s\right] \\ &= \frac{1}{m}[c_{-g+1}^s, c_{-g+2}^s, \ldots, c_0^s, c_1^s, \ldots, c_{N/m-1}^s]. \end{aligned}$$

By the definition of $\mathbf{c}_i^s$, we have

$$\begin{aligned} \mathbf{c}_0^s &= \frac{1}{m}[c_0^s, c_1^s, \ldots, c_{N/m-1}^s], \\ \mathbf{c}_1^s &= \frac{1}{m}[c_{-1}^s, c_0^s, \ldots, c_{N/m-2}^s], \\ &\vdots \\ \mathbf{c}_i^s &= \frac{1}{m}[c_{-i}^s, c_{-i+1}^s, \ldots, c_{N/m-i+1}^s], \\ &\vdots \\ \mathbf{c}_{g-1}^s &= \frac{1}{m}[c_{-g+1}^s, c_{-g+2}^s, \ldots, c_{N/m-g}^s]. \end{aligned} \tag{108}$$

Substituting these into $\mathbf{d}^s$, we obtain

(109)

$$\begin{aligned}\mathbf{d}^s &= \mathbf{c}_0^s \otimes (\mathbf{w}_0^s)' + \mathbf{c}_1^s \otimes (\mathbf{w}_1^s)' + \cdots + \mathbf{c}_{g-1}^s \otimes (\mathbf{w}_{g-1}^s)' \\ &= \frac{1}{m}\left[\sum_{k=0}^{g-1} V_k^0(\bar{\mathbf{w}}_k^s \otimes (\mathbf{w}_0^s)'), \ldots, \sum_{k=0}^{g-1} V_k^{N/m-1}(\bar{\mathbf{w}}_k^s \otimes (\mathbf{w}_0^s)')\right] \\ &\quad + \cdots + \frac{1}{m}\left[\sum_{i=0}^{g-2} V_{-i}^0(\bar{\mathbf{w}}_{g-1-i}^s \otimes (\mathbf{w}_{g-1}^s)'), \ldots, \sum_{k=0}^{g-1} V_k^{N/m-g}(\bar{\mathbf{w}}_k^s \otimes (\mathbf{w}_{g-1}^s)')\right] \\ &= \frac{1}{m}\left[\sum_{j=0}^{g-1}\sum_{i=0}^{g-1} V_{i-j}^0(\bar{\mathbf{w}}_i^s \otimes (\mathbf{w}_j^s)'),\right. \\ &\qquad \left.\sum_{j=0}^{g-1}\sum_{i=0}^{g-1} V_{i-j}^1(\bar{\mathbf{w}}_i^s \otimes (\mathbf{w}_j^s)'), \ldots, \sum_{j=0}^{g-1}\sum_{i=0}^{g-1} V_j^{N/m-i-1}(\bar{\mathbf{w}}_j^s \otimes (\mathbf{w}_i^s)')\right]\end{aligned}$$

and

(110)

$$\begin{aligned}\sum_{s=0}^{m-1} \mathbf{d}^s &= \frac{1}{m}\sum_{s=0}^{m-1}\left[\sum_{i,j} V_{i-j}^0(\bar{\mathbf{w}}_i^s \otimes (\mathbf{w}_j^s)'),\right. \\ &\qquad \left.\sum_{i,j} V_{i-j}^1(\bar{\mathbf{w}}_i^s \otimes (\mathbf{w}_j^s)'), \ldots, \sum_{i,j} V_{i-j}^{N/m-1}(\bar{\mathbf{w}}_i^s \otimes (\mathbf{w}_j^s)')\right] \\ &= \frac{1}{m}\left[\sum_{s}\sum_{i,j} V_{i-j}^0(\bar{\mathbf{w}}_i^s \otimes (\mathbf{w}_j^s)'),\right. \\ &\qquad \left.\sum_{s}\sum_{i,j} V_{i-j}^1(\bar{\mathbf{w}}_i^s \otimes (\mathbf{w}_j^s)'), \ldots, \sum_{s}\sum_{i,j} V_{i-j}^{N/m-1}(\bar{\mathbf{w}}_i^s \otimes (\mathbf{w}_j^s)')\right].\end{aligned}$$

Then by Lemma 5.1, we have

$$\sum_s \mathbf{d}^s = \frac{1}{m}[mV_0^0, mV_0^1, \ldots, mV_0^{N/m-1}] = f. \qquad \square \tag{111}$$

Note that the expressions of the wavelet transform of f and its inverse are simple and, with the exception of the p-product, the only operations required are shifts, which can be easily accomplished on any digital computer. In (76), there are g p-products and each p-product requires N operations, while in (81), there are $g-1$ p-products and each p-product requires $m(g-1)$ operations. Hence, the total number of operations required for computing $\mathbf{c}^s$ is $Ng + m(g-1)^2$. However, since each p-product is independent of the other, after shifting f to form $f_1, \ldots, f_{s-1}$, and shifting e to form $e_{-1}, e_{-2}, \ldots, e_{-g+2}$, all p-product computations can be executed simultaneously in parallel as shown in Fig. 2. Therefore, the computing time of the p-product wavelet transform depends only on the length of the function f, which is g times less than the existing methods presented in the literature [2], [4], [10]. On the other hand, notice that the size of $\mathbf{c}_-^s$ is $1 \times (g-1)$ and only affects the left boundary of the transform $\mathbf{c}^s$. If g and m are very small, which happens in most applications, or if the boundary

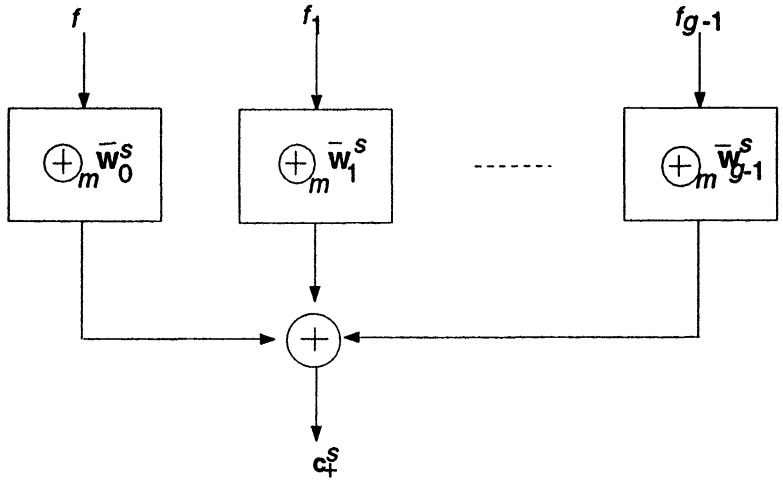


FIG. 2. *Parallel computation of* c_+^s *from* (76).

effect is not so important, which is the case on big images, we can ignore $\mathbf{c}_-^s$ and set it to zero. Then fewer processes will be needed and the formula will become simpler [23].

6. Conclusions. We have introduced a new matrix product. Using this product, we presented a general approach to fast transform computation. Specific examples included a new formulation of the FFT, the fast Walsh transform, and the fast generalized Walsh transform. In contrast to other fast transform algorithms, these p-product formulations do not require any reordering process, which is extremely important on supercomputers where the data flow is usually the major time-consuming part of the computation. A new formulation of the wavelet transform has also been presented here. This formulation is given in terms of the matrix p-product and provides not only a concise expression for wavelet transforms, but also a novel algorithm which is g times faster than regular methods for a wavelet matrix $\mathbf{a}$ of rank m and genus g. The applications presented here demonstrate that this new matrix product will be of great use in traditional transform theory as well as in the development of new transforms.

REFERENCES

[1] H. C. ANDREWS AND K. L. CASPARI, *A generalized technique for spectral analysis,* IEEE Transactions on Computers, C-19 (1970), pp. 16–25.

[2] G. BEYLKIN, R. COIFMAN, I. DAUBECHIES, S. MALLAT, Y. MEYER, L. ROPHAEL, AND B. RUSKAI, EDS., *Wavelets and Their Applications,* Jones and Barllett, Cambridge, MA, 1992.

[3] H. E. CHRESTENSON, *A class of generalized Walsh functions,* Pacific J. Math., 5 (1995), pp. 17–34.

[4] C. K. CHUI, ED., *Introduction to Wavelets,* Academic Press, Boston, 1992.

[5] ———, *Wavelets—A Tutorial in Theory and Applications,* Academic Press, Boston, 1992.

[6] J. W. COOLEY, P. A. LEWIS, AND P. D. WELCH, *The fast Fourier transform and its applications,* IEEE Trans. Educ., E-12 (1969), pp. 27–34.

[7] J. W. COOLEY AND J. W. TUKEY, *An algorithm for the machine calculation of complex Fourier series,* Math. Comput., 19 (1965), pp. 297–301.

[8] R. CUNINGHAME-GREEN, *Minimax Algebra: Lecture Notes in Economics and Mathematical Systems* 166, Springer-Verlag, New York, 1979.

[9] Z. CVETANOVIC, *Performance analysis of* FFT *algorithm on a shared-memory parallel architecture,* IBM J. Res. Development, 31 (1987), pp. 435–451.

[10] I. DAUBECHIES, *Orthonormal bases of wavelets,* Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[11] R. KRONLAND-MARTINET, J. MORLET, AND A. GROSSMANN, *Analysis of sound patterns through wavelet transform,* Internat. J. Pattern Recognition Artificial Intelligence, 1988.

[12] S. G. MALLAT, *A compact multiresolution representation: the wavelet model,* in Proc. IEEE Workshop Computer Vision, Miami, FL, December 1987.

[13] ———, *A theory of multiresolution signal decomposition: the wavelet representation,* IEEE Pattern Analysis and Machine Intelligence, 11 (1989), pp. 674–693.

[14] R. E. A. C. PALEY, *A remarkable series of orthogonal functions,* Proc. London Math. Soc., 34 (1932), pp. 241–279.

[15] H. L. RESNIKOFF, *Wavelets and adaptive signal processing,* Optical Engrg., 31 (1992), pp. 1229–1234.

[16] G. X. RITTER, *Heterogeneous matrix products,* in *Image Algebra and Morphological Image Processing* II, Vol. 1568, Proc. SPIE, San Diego, CA, July 1991, pp. 92–100.

[17] G. X. RITTER AND H. ZHU, *The generalized matrix product and its applications,* J. Math. Imaging Vision, 1 (1992), pp. 201–213.

[18] J. L. SHANKS, *Computation of the fast Walsh-Fourier transform,* IEEE Trans. Computers, C-18 (1969), pp. 457–459.

[19] H. S. STONE, *High-Performance Computer Architecture,* Addison-Wesley, Menlo Park, CA, 1987.

[20] F. THEILHEIMER, *A matrix version of the fast Fourier transform,* IEEE Trans. Audio Electroacoustics, AU-17 (1969), pp. 158–161.

[21] R. TOLIMIERI, M. AN, AND C. LU, *Algorithms for Discrete Fourier Transform and Convolution,* Springer-Verlag, New York, 1989.

[22] J. L. WALSH, *A closed set of normal orthogonal functions,* Amer. J. Math., 45 (1923), pp. 5–24.

[23] H. ZHU AND G. X. RITTER, *The generalized matrix product and the wavelet transform,* J. Math. Imaging Vision, 3 (1993), pp. 95–104.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 602–627, April 1995

018

OBLIQUE PROJECTION METHODS FOR LARGE SCALE MODEL REDUCTION*

IMAD M. JAIMOUKHA† AND EBRAHIM M. KASENALLY†

Abstract. The aim of this paper is to consider approximating a linear transfer function $F(s)$ of McMillan degree N, by one of McMillan degree m in which $N \gg m$ and where N is large. Krylov subspace methods are employed to construct bases to parts of the controllability and observability subspaces associated with the state space realisation of $F(s)$. Low rank approximate grammians are computed via the solutions to low dimensional Lyapunov equations and computable expressions for the approximation errors incurred are derived. We show that the low rank approximate grammians are the exact grammians to a perturbed linear system in which the perturbation is restricted to the transition matrix, and furthermore, this perturbation has at most rank = 2. This paper demonstrates that this perturbed linear system is equivalent to a low dimensional linear system with state dimension no greater than m. Finally, exact low dimensional expressions for the $\mathcal{L}^\infty$ norm of the errors are derived. The model reduction of discrete time linear systems is considered via the use of the same Krylov schemes. Finally, the behaviour of these algorithms is illustrated on two large scale examples.

Key words. Lanczos, Arnoldi, iterative methods, model reduction, Krylov subspace methods, Lyapunov matrix equation, large scale systems

AMS subject classifications. 65F10, 65F15, 93A15, 93B05, 93B07, 93B20

1. Introduction.

The need for model reduction arises in many areas of engineering, where high order mathematical models are used to describe complex dynamical behaviour. These occur whenever models are described by partial differential equations that culminate in large linear finite element or finite difference models. For practical reasons, it is desirable to replace these high order models by low order approximations. For example, in control system applications, high order models may result in high order controllers and the subsequent implementation of these controllers is cumbersome and expensive. Consider a stable linear state-space model of the form

$$\dot{x}(t) = Ax(t) + bu(t), \tag{1}$$

$$y(t) = cx(t) + du(t), \tag{2}$$

in which $x(t)$ is the state vector of dimension N, and $u(t)$ and $y(t)$ are scalar functions representing the input and the output of the system, respectively. The matrix A and vectors b, c, and d are real with their dimensions fixed by those of $x(t)$, $u(t)$, and $y(t)$. The associated transfer function is given by $F(s) = c(sI - A)^{-1}b + d$. The task of any model reduction algorithm is to find an approximate stable model

$$\dot{x}_m(t) = A_m x_m(t) + b_m u(t), \tag{3}$$

$$y_m(t) = c_m x_m(t) + d_m u(t), \tag{4}$$

in which $x_m(t) \in \mathbb{R}^m$, with $m \ll N$ and the low order transfer function is given by $F_m(s) = c_m(sI - A_m)^{-1}b_m + d_m$. Well-established model reduction methods such as optimal Hankel norm [6] and balanced truncation [13] begin by solving the linear matrix equations

$$AP + PA' + bb' = 0, \tag{5}$$

$$A'Q + QA + c'c = 0, \tag{6}$$

* Received by the editors June 23, 1993; accepted for publication (in revised form) by P. Van Dooren, March 25, 1994.

† Interdisciplinary Research Centre for Process Systems Engineering, Imperial College, Exhibition Road, London SW7-2BY, England (jaimouka@ps.ic.ac.uk and kasenall@ps.ic.ac.uk).

which admit unique symmetric solutions if and only if $\lambda_i(A) + \bar{\lambda}_j(A) \neq 0$ for all i, j, and where λ_i denotes the ith eigenvalue and the overbar represents the complex conjugate. The requisite for P and Q stems from the easily computable $\mathcal{L}^\infty$ error bound [6]

$$\|F(s) - F_m(s)\|_\infty \leq 2 \sum_{m+1}^{N} \sigma_i(F(s)), \tag{7}$$

where the σ_i's are the Hankel singular values of $F(s)$ defined as $\sigma_i = \lambda_i^{1/2}(PQ)$ and arranged in decreasing order of magnitude and where the norm $\|F(s)\|_\infty$ is defined as $\|F(s)\|_\infty = \sup_{\omega \in \mathbb{R}} \{\sigma_{\max}\{F(j\omega)\}\}$ in which $\sigma_{\max}(\cdot)$ denotes the largest singular value. Furthermore, P and Q are used by both methods to form the balancing transformation the effect of which is to yield $P = Q = \Sigma$ in the new coordinate system where Σ is a diagonal matrix of the Hankel singular values. Safonov and Chiang [18] developed a numerically robust variant of Moore's balanced truncation algorithm that does not require the formation of the balancing transformations; however, this variant does not obviate the need to compute P and Q. The motivation for using Krylov subspace methods in this paper is to enable low order approximate models to be computed while effecting all the computations in the low dimension. This technique was used successfully in [1] to perform model reduction in the light of a control system design for a fusion reactor. One of the aims of this paper is to justify this approach and give computable error expressions.

Related to this work are [4], [5], which use Krylov subspace methods to obtain bases for the controllability and observability spaces. Furthermore, in [5] Boley and Golub presented a means of computing a minimal realisation of a linear dynamical system from the coefficients generated in the course of the Lanczos process. The Lanczos process was also exploited by Parlett in [14] to obtain minimal realisations. In that paper, the rank of the Hankel matrix was used to determine the order of the minimal realisation. Furthermore, it was demonstrated that a minimal realisation could be constructed from the data generated by the Lanczos process. A similar approach was adopted in [7] in which the minimal realisation and its order were found to be related to the different types of breakdowns encountered in the Lanczos process. Here, too, the onset of breakdown was also given in terms of properties of the Hankel matrix. Recently, the presentations in [3], [19] reviewed the use of projection methods for large scale control problems. Both papers suggest the use of Krylov subspace methods as an effective tool for the model reduction of large scale linear dynamical systems; however, no algorithms were provided.

A drawback associated with the methods above is that the computed minimal realisation may still have a high dimension. This deficiency is remedied by obtaining approximate reduced order models with low state dimension. A key issue in the development of our model reduction schemes is the efficient computation of low rank approximate solutions to the controllability and observability Lyapunov equations. This paper exploits the approach developed in our previous work [10]–[12] and that of Saad [17], both of which employ classical Krylov subspace techniques. In [17], Saad considers the low rank approximate solutions to (5) by imposing a Galerkin condition on the residual error, and [11] extends his work to the general case via the use of block schemes and gives a computable expression for the associated residual error norm. Furthermore, we addressed the problem of computing a low rank approximate solution to (5) which meets an optimality condition. The generalized minimal residual

(GMRES) method presented in [11] minimises the Frobenius norm of the residual error for which an exact computable expression is also derived. The focus of this paper is to consider the solution to the coupled Lyapunov equations (5) and (6). Two Lyapunov equation solvers are presented for which we derive computable residual error and a priori and a posteriori backward error expressions. Two model reduction algorithms are introduced for which computable $\mathcal{L}^\infty$ error expressions are provided. For ease of presentation, the developments are carried out for single-input, single-output systems, and the findings are then extended to multivariable problems and discrete-time dynamical systems.

The following summarises the contributions in this paper. Section 2 describes the type of approximations employed and justifies the use of Krylov subspace methods to solve large Lyapunov equations. Approximate solutions to (5) and (6) are obtained together with exact expressions for the errors incurred. Furthermore, it is demonstrated that the two approximate solutions are the exact solutions to a pair of perturbed Lyapunov equations. In §3, we show that the latter are the controllability and observability Lyapunov equations of a perturbed linear system; furthermore, this perturbed linear system is equivalent to a linear system of state dimension no greater than m. Computable $\mathcal{L}^\infty$ error expressions are provided that enable one to gauge the progress of the iterative method for increasing m. Section 4 employs the solution techniques of §§2 and 3 to obtain model reduction schemes for discrete time systems. Two illustrative examples in §5 show how the Lyapunov equation solvers and model reduction algorithms behave in practice, and, finally, the conclusions are found in §6.

2. Krylov subspace techniques. In practice, solutions to large Lyapunov equations (5) and (6) frequently admit good low rank approximations. In addition, one is generally interested in computing only the dominant eigenspace of the exact solution P^*, rather than P^* itself, since the dominant eigenspace of P^* is known to be associated with the dominant modes of the system described by (1) and (2) [1]. In what follows, what can be said of (5) can also be said of (6). Thus we will limit our discussion to (5) and invoke (6) only when necessary.

Ideally, we want to compute a rank m approximation P_m where $m \ll N$ such that $\|P^* - P_m\|_F$ is minimised. Throughout this paper we make use of the Frobenius norm defined as $\|Z\|_F = \sqrt{\text{trace}(ZZ')}$ in which Z' denotes the conjugate transpose of Z.

Consider the Schur decomposition of P^* given by $P^* = U\Sigma U'$ in which $U \in \mathbb{R}^{N\times N}$ is an orthogonal matrix and $\Sigma = \text{diag}\{\sigma_1, \sigma_2, \ldots, \sigma_N\}$ is a matrix of eigenvalues ordered such that $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_N \geq 0$. Then the optimal rank m Frobenius norm approximation of P^* is given by

$$P_m := U \begin{bmatrix} \Sigma_m & 0 \\ 0 & 0 \end{bmatrix} U' = U_m \Sigma_m U_m', \tag{8}$$

where $\Sigma_m = \text{diag}\{\sigma_1, \sigma_2, \ldots, \sigma_m\}$ (i.e., the first m diagonal elements of Σ) and $U_m \in \mathbb{R}^{N\times m}$ is a matrix of the first m columns of U. In [10], [11], [12], [17] a low rank approximation is computed by choosing an orthogonal matrix $V_m \in \mathbb{R}^{N\times m}$ and calculating the exact solution X_m to the reduced order Lyapunov equation

$$(V_m' A V_m) X_m + X_m (V_m' A' V_m) + V_m' b b' V_m = 0. \tag{9}$$

The estimate of P^* is then given by $P_m = V_m X_m V_m'$. Compared with the optimal approximation given in (8), it is apparent that to compute a good estimate of P^*, V_m

must be an accurate approximation of U_m or, in other words, the m most dominant eigenvectors of P^*. Unfortunately, minimising $\|P^* - P_m\|_F$ is intractable when P^* is unknown.

We therefore turn our attention to the problem of selecting V_m and X_m. Let $\mathcal{K}_m(A, b)$ be the m-dimensional Krylov space defined as

$$\mathcal{K}_m(A, b) = \text{span}\left\{[b \; Ab \; A^2b \; \cdots \; A^{m-1}b]\right\}, \tag{10}$$

then, following [17], we select V_m to be an orthogonal basis of $\mathcal{K}_m(A, b)$. Throughout the remainder of this paper, we exploit approximations to the solution P^* which have the form

$$P_m = V_m X_m V_m', \tag{11}$$

where $X_m \in \mathbb{R}^{m\times m}$ is an arbitrary symmetric matrix. The key point here is that even though $P_m \in \mathbb{R}^{N\times N}$, it may be efficiently stored as the product of smaller matrices $V_m \in \mathbb{R}^{N\times m}$ and $X_m \in \mathbb{R}^{m\times m}$. We observe that P_m is symmetric for symmetric X_m and $\text{rank}(P_m) = \text{rank}(X_m) \leq m$. Similarly, associated with (6), one seeks an approximate solution of the form $Q_m := W_m Y_m W_m'$ where W_m is selected to be an orthogonal basis of m-dimensional Krylov space defined as

$$\mathcal{L}_m(A', c') = \text{span}\left\{[c' \; (A')c' \; (A')^2c' \; \cdots \; (A')^{m-1}c']\right\}. \tag{12}$$

The remainder of this section is devoted to the appropriate selection of symmetric matrices X_m and Y_m. We begin by defining the residual error functions associated with a particular choice of X_m and Y_m as

$$R_m(X_m) := A(V_m X_m V_m') + (V_m X_m V_m')A' + bb', \tag{13}$$

$$S_m(Y_m) := A'(W_m Y_m W_m') + (W_m Y_m W_m')A + c'c. \tag{14}$$

The solution techniques presented in this paper are based on seeking a symmetric X_m and Y_m so as to give $R_m(X_m)$ and $S_m(Y_m)$ desirable properties. Sections 2.1 and 2.2 address the problem of constructing different X_m and Y_m such that $R_m(X_m)$ and $S_m(Y_m)$ have an orthogonality property with respect to the Krylov spaces defined in (10) and (12). Computable expressions for $\|R_m(X_m)\|_F$ and $\|S_m(Y_m)\|_F$ are provided. We demonstrate that the two approximate grammians are the exact solutions to a set of perturbed Lyapunov equations, thus providing an alternative means of gauging the progress of the iterative process for increasing m.

2.1. The Arnoldi process and Lyapunov equations. Next we use the well-established Arnoldi algorithm [20] to calculate the orthonormal bases V_m and W_m for the Krylov subspaces $\mathcal{K}_m$ and $\mathcal{L}_m$, respectively. The basic outline of the Arnoldi process is given next in terms of (A, b).

ARNOLDI PROCESS

- Initialise: Compute $\beta = \|b\|_2$ and set $v_1 := b/\beta$.
- Iterate: Do $j = 1, \ldots, m$
  - Compute j coefficients h_{ij} so that $\hat{v} := Av_j - \sum_{i=1}^{j} v_i h_{ij}$ is orthogonal to all the previous v_i's.
  - Compute $h_{j+1,j} := \|\hat{v}\|_2$. If $h_{j+1,j} = 0$ stop, else $v_{j+1} := \hat{v}/h_{j+1,j}$.
  - End Do.

By construction, the Arnoldi process produces the matrix $V_m = [v_1, v_2, \ldots, v_m]$, which forms an orthogonal basis for the Krylov subspace $\mathcal{K}_m(A, b)$. The process also yields an $m \times m$ upper Hessenberg matrix H_m, which satisfies the relation

$$AV_m = V_m H_m + v_{m+1} h_{m+1,m} e'_m, \tag{15}$$

in which e_m denotes the last column of the m-dimensional identity matrix. We observe that (15) is a combination of the first two steps of the iterative loop. From (15) it is easy to verify that $H_m = V'_m A V_m$ since $[V_m \ v_{m+1}]$ is part of an orthogonal matrix. Applying the Arnoldi process to (A', c') yields an orthogonal basis $W_m := [w_1, w_2, \ldots, w_m]$ for the Krylov space $\mathcal{L}_m(A', c')$ in which $w_1 := c/\delta$ and $\delta := \|c'\|_2$. Furthermore, G_m is a lower Hessenberg matrix which satisfies

$$W'_m A = G_m W'_m + e_m g_{m,m+1} w'_{m+1}. \tag{16}$$

For convenience, we define

$$\hat{H}_m := (W'_m V_m)^{-1} W'_m A V_m = H_m + (W'_m V_m)^{-1} W'_m v_{m+1} h_{m+1,m} e'_m, \tag{17}$$

$$\hat{G}_m := W'_m A V_m (W'_m V_m)^{-1} = G_m + e_m g_{m,m+1} w'_{m+1} V_m (W'_m V_m)^{-1} \tag{18}$$

for nonsingular $(W'_m V_m)$. Observe that $\hat{H}_m$ and $\hat{G}_m$ are upper Hessenberg and lower Hessenberg matrices, respectively. The Arnoldi process generates an upper Hessenberg H_m and $(W'_m V_m)^{-1} W'_m v_{m+1} h_{m+1,m} e'_m$ has nonzero elements only in its last column; thus $\hat{H}_m$ has the same structure as H_m. Similarly, $\hat{G}_m$ is lower Hessenberg and $e_m g_{m,m+1} w'_{m+1} V_m (W'_m V_m)^{-1}$ has nonzero elements only in its last row. For implementation details of the Arnoldi process and its breakdown-free variants, we refer the reader to [4], [11].

Assuming that the approximate solution to (5) has the form $P_m := V_m X_m V'_m$ for some arbitrary symmetric $X_m \in \mathbb{R}^{m \times m}$, the residual error function associated with any solution is given by (13). Substituting (15) and (17) into (13) and assuming that $(W'_m V_m)$ is nonsingular allows the error function in (13) to be written as

$$\begin{aligned} R_m(X_m) = [V_m \quad & (I - V_m (W'_m V_m)^{-1} W'_m) v_{m+1}] \\ & \times \begin{bmatrix} \hat{H}_m X_m + X_m \hat{H}'_m + e_1 \beta^2 e'_1 & X_m e_m h_{m+1,m} \\ h_{m+1,m} e'_m X_m & 0 \end{bmatrix} \\ & \times \begin{bmatrix} V'_m \\ v'_{m+1} (I - W_m (V'_m W_m)^{-1} V'_m) \end{bmatrix}. \end{aligned} \tag{19}$$

Similarly, associated with (6), the residual error function for any given approximate solution of the form $Q_m = W_m Y_m W'_m$ is defined in (14). Substituting (16) and (18) into (14) allows the error function to be written as

$$\begin{aligned} S_m(Y_m) = [W_m \quad & (I - W_m (V'_m W_m)^{-1} V'_m) w_{m+1}] \\ & \times \begin{bmatrix} \hat{G}'_m Y_m + Y_m \hat{G}_m + e_1 \delta^2 e'_1 & Y_m e_m g_{m,m+1} \\ g_{m,m+1} e'_m Y_m & 0 \end{bmatrix} \\ & \times \begin{bmatrix} W'_m \\ w'_{m+1} (I - V_m (W'_m V_m)^{-1} W'_m) \end{bmatrix} \end{aligned} \tag{20}$$

for nonsingular $(W'_m V_m)$. The orthogonality conditions imposed in [10]–[12] sought to determine low rank approximate solutions $P_m := V_m X_m V'_m$ such that

$V_m' R_m(X_m) V_m = 0$. In contrast, the Arnoldi–Lyapunov solver considered in this section seeks symmetric matrices X_m and Y_m such that the residual errors $R_m(X_m)$ and $S_m(Y_m)$ satisfy orthogonality properties with respect to the Krylov subspaces $\mathcal{L}_m(A', c')$ and $\mathcal{K}_m(A, b)$, respectively. We now state the problem we wish to address.

PROBLEM 2.1. Find the approximate solutions $P_m := V_m X_m V_m'$ and $Q_m := W_m Y_m W_m'$ to (5) and (6), respectively, that satisfy the Galerkin type conditions $W_m' R_m(X_m) W_m = 0$ and $V_m' S_m(Y_m) V_m = 0$.

The following theorem gives the solution to Problem 2.1.

THEOREM 2.1. *Suppose that m steps of the Arnoldi process have been taken, that $(W_m' V_m)$ is nonsingular, and that the residual errors associated with* (5) *and* (6) *are defined by* (19) *and* (20). *Furthermore, suppose that $\lambda_i(\hat{H}_m) + \bar{\lambda}_j(\hat{H}_m) \neq 0$ for all i, j and $\lambda_i(\hat{G}_m) + \bar{\lambda}_j(\hat{G}_m) \neq 0$ for all i, j then,*

(a) $W_m' R_m(X_m) W_m = 0$ *if and only if $X_m = X_m^A$, where X_m^A satisfies*

$$\hat{H}_m X_m^A + X_m^A \hat{H}_m' + e_1 \beta^2 e_1' = 0, \tag{21}$$

where e_1 is the first column of the $m \times m$ identity matrix. Under these conditions,

$$\begin{aligned} \|R_m^A\|_F &:= \|R_m(X_m^A)\|_F \\ &= \sqrt{2h_{m+1,m} v_{m+1}' (I - W_m (V_m' W_m)^{-1} V_m') R_m(X_m^A) V_m X_m^A e_m} \end{aligned} \tag{22}$$

$$= \left\| \begin{bmatrix} H_m X_m^A + X_m^A H_m' + e_1 \beta^2 e_1' & X_m^A e_m h_{m+1,m} \\ h_{m+1,m} e_m' X_m^A & 0 \end{bmatrix} \right\|_F. \tag{23}$$

(b) $V_m' S_m(Y_m) V_m = 0$ *if and only if $Y_m = Y_m^A$, where Y_m^A satisfies*

$$\hat{G}_m' Y_m^A + Y_m^A \hat{G}_m + e_1 \delta^2 e_1' = 0. \tag{24}$$

Under these conditions,

$$\begin{aligned} \|S_m^A\|_F &:= \|S_m(Y_m^A)\|_F \\ &= \sqrt{2g_{m,m+1} w_{m+1}' (I - V_m (W_m' V_m)^{-1} W_m') S_m(Y_m^A) W_m Y_m^A e_m} \end{aligned} \tag{25}$$

$$= \left\| \begin{bmatrix} G_m' Y_m^A + Y_m^A G_m + e_1 \delta^2 e_1' & Y_m^A e_m g_{m,m+1} \\ g_{m,m+1} e_m' Y_m^A & 0 \end{bmatrix} \right\|_F. \tag{26}$$

Proof. Pre and postmultiplying (19) by W_m' and W_m, respectively, gives

$$\begin{aligned} &W_m' R_m(X_m) W_m \\ &= \begin{bmatrix} W_m' V_m & 0 \end{bmatrix} \begin{bmatrix} \hat{H}_m X_m + X_m \hat{H}_m' + e_1 \beta^2 e_1' & X_m e_m h_{m+1,m} \\ h_{m+1,m} e_m' X_m & 0 \end{bmatrix} \begin{bmatrix} V_m' W_m \\ 0 \end{bmatrix} \\ &= (W_m' V_m)(\hat{H}_m X_m + X_m \hat{H}_m' + e_1 \beta^2 e_1')(V_m' W_m). \end{aligned}$$

The result follows immediately since $(W_m' V_m)$ is assumed to be nonsingular. Substituting (21) into (19) gives, following some calculation,

$$\|R_m^A\|_F^2 := \|R_m(X_m^A)\|_F^2 = \operatorname{trace}\{R_m^A R_m^{A\prime}\} \tag{27}$$

$$= 2h_{m+1,m} v_{m+1}' (I - W_m (V_m' W_m)^{-1} V_m') R_m(X_m^A) V_m X_m^A e_m, \tag{28}$$

which establishes (22). Substituting $X_m := X_m^A$ into (13) allows us to factorise $R_m(X_m^A)$ as

$$R_m(X_m^A) = [V_m \;\; v_{m+1}] \begin{bmatrix} H_m X_m^A + X_m^A H_m' + e_1 \beta^2 e_1' & X_m^A e_m h_{m+1,m} \\ h_{m+1,m} e_m' X_m^A & 0 \end{bmatrix} \begin{bmatrix} V_m' \\ v_{m+1}' \end{bmatrix},$$

from which (23) follows immediately since $[V_m \;\; v_{m+1}]$ is part of an orthogonal matrix and completes the proof of part (a).

The proof to part (b) is identical to that of part (a) except that it uses (14), (20), and (24). □

An implication of the above result is that as m is increased, the residuals are confined to progressively smaller and smaller subspaces of $\mathbb{R}^{N\times N}$. This however, does not imply that the Arnoldi process will produce a sequence of nonincreasing residual error norms. The residual error norms in (22) and (25) provide a useful stopping criterion in a practical implementation of the algorithm as they allow one to economically evaluate the error norms and gauge the quality of the low rank approximations. The key points here are, first, that (21) and (24) are Lyapunov equations of dimension m that can be solved accurately using the Bartels–Stewart algorithm [2]; second, P_m and Q_m may be efficiently stored as the product of low order matrices; and, third, the residual error norm does not require the formation of the approximate solutions at each step. Instead (22), (23), (25), and (26) may be computed via low dimensional matrix products.

Remark 2.1. Computing the residual error norms $\|R_m^A\|_F$ and $\|S_m^A\|_F$ at the end of each Arnoldi process iteration using (22) and (25) may be limited to approximately $6Nm$ floating point operations if the computations are restricted to matrix vector products. Alternatively, approximately $O(m^3)$ floating point operations are needed to evaluate either (23) or (26). Thus as m increases, one would switch from evaluating (23) and (26) to computing (22) and (25) as it became more economical to do so.

The following result gives one perturbation, in exact arithmetic, of the data in (5) (and similarly for (6)) for which the low rank solutions given in Theorem 2.1 are the exact solutions. This result is reminiscent of the backward error analysis of [9].

COROLLARY 2.2. *Suppose that m steps of the Arnoldi process have been taken and that* $P_m := V_m X_m^A V_m'$ *and* $Q_m := W_m Y_m^A W_m'$ *are the low rank approximate solutions to* (5) *and* (6), *respectively, and furthermore that* X_m^A *and* Y_m^A *satisfy* (21) *and* (24), *respectively. Then*

$$(A - \Delta_1)P_m + P_m(A - \Delta_1)' + bb' = 0, \tag{29}$$

$$(A - \Delta_2)'Q_m + Q_m(A - \Delta_2) + c'c = 0, \tag{30}$$

where

$$\Delta_1 = (I - V_m(W_m'V_m)^{-1}W_m')v_{m+1}h_{m+1,m}v_m', \tag{31}$$

$$\Delta_2 = w_m g_{m,m+1} w_{m+1}'(I - V_m(W_m'V_m)^{-1}W_m'), \tag{32}$$

and

$$\|\Delta_1\|_F^2 = h_{m+1,m}^2\{1 + \|(W_m'V_m)^{-1}W_m'v_{m+1}\|_2^2\}$$

and

$$\|\Delta_2\|_F^2 = g_{m,m+1}^2\{1 + \|(V_m'W_m)^{-1}V_m'w_{m+1}\|_2^2\}.$$

Proof. Substituting $X_m := X_m^A$ into (19) gives

$$
\begin{aligned}
&A(V_m X_m^A V_m') + (V_m X_m^A V_m')A' + bb' \\
&= (I - V_m(W_m'V_m)^{-1}W_m')v_{m+1}h_{m+1,m}e_m' X_m^A V_m' \\
&\quad + V_m X_m^A e_m h_{m+1,m} v_{m+1}'(I - W_m(V_m'W_m)^{-1}V_m').
\end{aligned} \tag{33}
$$

Equation (29) follows by rearranging (33) and noting that $e_m' = v_m' V_m$. The expression for $\|\Delta_1\|_F$ follows from the fact that v_m and V_m are parts of an orthogonal matrix.

Similarly, (30) and the expression for $\|\Delta_2\|_F$ follow by substituting $Y_m := Y_m^A$ into (20) and using the facts that w_m and W_m are parts of an orthogonal matrix. Finally, observe that Δ_1 and Δ_2 are at most rank-1 perturbations and that $\|\Delta_1\|_F$ and $\|\Delta_2\|_F$ may be evaluated without the need to form X_m^A or Y_m^A. □

Remark 2.2. Observe that $\Delta_3 := \Delta_1 + \Delta_2$ is also a perturbation on the data in A such that

$$(A - \Delta_3)P_m + P_m(A - \Delta_3)' + bb' = 0, \tag{34}$$

$$(A - \Delta_3)'Q_m + Q_m(A - \Delta_3) + c'c = 0. \tag{35}$$

Furthermore, Δ_3 is at most a rank-2 perturbation, which may be factorised as

$$
\begin{aligned}
\Delta_3 = [w_m \;\; (I - V_m(W_m'V_m)^{-1}W_m')v_{m+1}] \\
\times \begin{bmatrix} 0 & g_{m,m+1} \\ h_{m+1,m} & 0 \end{bmatrix} \begin{bmatrix} v_m' \\ w_{m+1}'(I - V_m(W_m'V_m)^{-1}W_m') \end{bmatrix}.
\end{aligned} \tag{36}
$$

Finally, a direct calculation will verify that $W_m'\Delta_1 V_m = W_m'\Delta_2 V_m = W_m'\Delta_3 V_m = 0$.

From (31) and (32) one observes that Δ_1 and Δ_2 depend only on the data generated in the course of the Arnoldi process and, as such, one can envisage an iterative scheme in which the evolutions of $||\Delta_1||_F$ and $||\Delta_2||_F$ were monitored for increasing m. This scheme would require $4Nm$ operations above each Arnoldi step and $2N(m+1)$ storage locations for V_{m+1} and W_{m+1} and a further $3m^2$ for H_m, G_m and $W_m'V_m$, respectively, the latter being updatable through matrix vector products. In contrast, an implementation that monitors the evolution of the residual error requires $25m^3 + 2min\{Nm,\ m^3\}$ floating point operations and a storage of approximately $2N(m+1)+7m^2$ locations making it computationally cheaper but with more storage needs than a backward error checking scheme.

The following procedures summarise the Arnoldi–Lyapunov solvers proposed in this section; the first monitors the residual error while the second checks the backward error evolution.

ARNOLDI–LYAPUNOV SOLVER (RESIDUAL ERROR)

- Start: Specify a tolerance $\epsilon > 0$, set an integer parameter m.
- Perform m steps of the Arnoldi process to compute $\hat{H}_m$, $h_{m+1,m}$, V_m, v_{m+1}, and β.
- Perform m steps of the Arnoldi process to compute $\hat{G}_m$, $g_{m,m+1}$, W_m, w_{m+1}, and δ.
- Compute the symmetric matrices X_m^A and Y_m^A, which uniquely satisfy the low dimensional Lyapunov equations $\hat{H}_m X_m + X_m \hat{H}_m' + e_1\beta^2 e_1' = 0$ and $\hat{G}_m' Y_m + Y_m \hat{G}_m + e_1\delta^2 e_1' = 0$, respectively.

- Evaluate $\|R_m^A\|_F$ and $\|S_m^A\|_F$ using (22) and (25) or (23) and (26), respectively. If either $\|R_m^A\|_F > \epsilon$ or $\|S_m^A\|_F > \epsilon$ increase m and continue the Arnoldi process, otherwise, form the approximate solutions: $P_m := V_m X_m^A V_m'$ and $Q_m := W_m Y_m^A W_m'$.

ARNOLDI–LYAPUNOV SOLVER (BACKWARD ERROR)

- Start: Specify a tolerance $\epsilon > 0$, set an integer parameter m.
- Perform m steps of the Arnoldi process to compute H_m, $h_{m+1,m}$, V_m, v_{m+1}, and β.
- Perform m steps of the Arnoldi process to compute G_m, $g_{m,m+1}$, W_m, w_{m+1}, and δ.
- Evaluate $\|\Delta_1\|_F$ and $\|\Delta_2\|_F$ using Corollary 2.2, if either $\|\Delta_1\|_F > \epsilon$ or $\|\Delta_2\|_F > \epsilon$ increase m and continue the Arnoldi process, otherwise, compute the symmetric matrices X_m^A and Y_m^A which uniquely satisfy the low dimensional Lyapunov equations $\hat{H}_m X_m + X_m \hat{H}_m' + e_1 \beta^2 e_1' = 0$ and $\hat{G}_m' Y_m + Y_m \hat{G}_m + e_1 \delta^2 e_1' = 0$, respectively.
- Form the approximate solutions: $P_m := V_m X_m^A V_m'$ and $Q_m := W_m Y_m^A W_m'$.

As pointed out earlier, $\|\Delta_1\|_F$ and $\|\Delta_2\|_F$ depend entirely on the data generated in the course of the Arnoldi process, thus they are a priori perturbation bounds that may be conservative. A posteriori backward error bounds are discussed in §2.3.

2.2. The Lanczos process and Lyapunov equations. The nonsymmetric Lanczos process is an alternative algorithm used to simultaneously construct bases for $\mathcal{K}_m(A, b)$ and $\mathcal{L}_m(A', c')$. The process requires A, its transpose, and two starting vectors b and c' to construct bases for parts of the controllability and observability subspaces that meet a biorthogonality condition (i.e., $W_m' V_m = I$). The aim of this section is not to consider the different aspects of the Lanczos process, but rather to show how this algorithm may be used to efficiently solve large Lyapunov equations. For an analysis of the Lanczos process and some of its breakdown-free variants, we refer the reader to [5], [14]–[16] and the references therein. A simple version of the nonsymmetric Lanczos process is given here.

LANCZOS PROCESS

1. Start: Set $\beta_1 := \sqrt{|cb|}$ and $\delta_1 := \beta_1 \cdot \text{sign}[cb]$ and define $v_1 := b/\delta_1$ $w_1 := c'/\beta_1$.
2. Iterate: For $j = 1, 2, \ldots,$ do:
   - $\alpha_j := (Av_j, w_j)$,
   - $\hat{v}_{j+1} := Av_j - \alpha_j v_j - \beta_j v_{j-1}$ (when $j = 1$, take $\beta_1 v_0 = 0$),
   - $\hat{w}_{j+1} := A^T w_j - \alpha_j w_j - \delta_j w_{j-1}$ (when $j = 1$, take $\delta_1 w_0 = 0$),
   - $\beta_{j+1} := \sqrt{|(\hat{v}_{j+1}, \hat{w}_{j+1})|}$, $\delta_{j+1} := \beta_{j+1} \cdot \text{sign}[(\hat{v}_{j+1}, \hat{w}_{j+1})]$,
   - $v_{j+1} := \hat{v}_{j+1}/\delta_{j+1}$, $w_{j+1} := \hat{w}_{j+1}/\beta_{j+1}$.

If we denote $V_m = [v_1, v_2, \ldots, v_m]$ and similarly $W_m = [w_1, w_2, \ldots, w_m]$, we then have $W_m' V_m = I$ where $I \in \mathbb{R}^{m \times m}$ is the identity matrix. Let us denote by T_m the tridiagonal matrix $T_m \equiv \text{Tridiag}[\delta_{i+1}, \alpha_i, \beta_{i+1}]$. Then it is easy to verify that

$$AV_m = V_m T_m + \delta_{m+1} v_{m+1} e_m', \tag{37}$$

$$A'W_m = W_m T_m' + \beta_{m+1} w_{m+1} e_m', \tag{38}$$

and $W_m' A V_m = T_m$, where the vector e_m is the mth column of the identity matrix in $\mathbb{R}^{m \times m}$. Observe that there are infinitely different ways of selecting the scalars δ_{j+1} and β_{j+1} as long as they satisfy $(\hat{v}_{j+1}, \hat{w}_{j+1}) = \delta_{j+1}\beta_{j+1}$. In our application, we select $v_1 := b/\sqrt{|cb|}$ and $w_1 := c'/\sqrt{|cb|} \cdot \text{sign}[cb]$. Then setting $\delta_1 = \sqrt{|cb|}$, we have $b = V_m e_1 \delta_1$, where e_1 is the first column of the identity matrix in $\mathbb{R}^{m \times m}$. Similarly,

setting $\beta_1 = \sqrt{|cb|} \cdot \text{sign}[cb]$ yields $c' = W_m e_1 \beta_1$. Assuming that the approximate solution to (5) has the form $P_m := V_m X_m V_m'$, substituting (37) into (13) gives

$$(39)\quad R_m(X_m) = [V_m \; v_{m+1}] \begin{bmatrix} T_m X_m + X_m T_m' + e_1 \delta_1^2 e_1' & X_m e_m \delta_{m+1} \\ \delta_{m+1} e_m' X_m & 0 \end{bmatrix} \begin{bmatrix} V_m' \\ v_{m+1}' \end{bmatrix}.$$

Similarly, suppose that the approximate solution to (6) has the form $Q_m := W_m Y_m W_m'$. Then on substituting into (14), we get

$$(40)\quad S_m(Y_m) = [W_m \; w_{m+1}] \begin{bmatrix} T_m' Y_m + Y_m T_m + e_1 \beta_1^2 e_1' & Y_m e_m \beta_{m+1} \\ \beta_{m+1} e_m' Y_m & 0 \end{bmatrix} \begin{bmatrix} W_m' \\ w_{m+1}' \end{bmatrix}.$$

The Lanczos–Lyapunov solver proposed here seeks symmetric X_m and Y_m that solve Problem 2.1. The following theorem gives the solution to Problem 2.1 in the context of the Lanczos process.

THEOREM 2.3. *Suppose that m steps of the Lanczos process have been taken and that the residual errors are defined by* (39) *and* (40). *Then if* $\lambda_i(T_m) + \bar{\lambda}_j(T_m) \neq 0$ *for all* i, j

(a) $W_m' R_m(X_m) W_m = 0$ *if and only if* $X_m = X_m^L$ *where* X_m^L *satisfies*

$$(41)\qquad T_m X_m^L + X_m^L T_m' + e_1 \delta_1^2 e_1' = 0.$$

If these conditions are met then the residual error norm is given by

$$(42)\qquad \|R_m^L\|_F := \|R_m(X_m^L)\|_F = \sqrt{2\delta_{m+1} v_{m+1}' R_m(X_m^L) V_m X_m^L e_m}.$$

(b) $V_m' S_m(Y_m) V_m = 0$ *if and only if* $Y_m = Y_m^L$ *where* X_m^L *satisfies*

$$(43)\qquad T_m' Y_m^L + Y_m^L T_m + e_1 \beta_1^2 e_1' = 0.$$

If these conditions are met then the residual error norm is given by

$$(44)\qquad \|S_m^L\|_F := \|S_m(Y_m^L)\|_F = \sqrt{2\beta_{m+1} w_{m+1}' S_m(L_m^L) W_m Y_m^L e_m}.$$

Proof. The proof is essentially the same as that of Theorem 2.1 except that it uses (39) and (40). □

Each low dimensional Lyapunov equation may be solved at a cost of $12.5m^3$ floating point operations and $2.5m^2$ storage locations. Observe that the fourth step of the Lanczos process reveals that δ_{j+1} and β_{j+1} differ at most by a sign. A consequence of this is that T_m is almost symmetric in the sense that its superdiagonal entries are equal to its subdiagonal elements up to an occasional sign. It is therefore natural to ask whether this structure may be exploited in the solution to the low dimensional Lyapunov equations (41) and (43). Thus, we seek a sign matrix J such that $T_m' = JT_mJ$, pre and postmultiplying (41) by J shows that $Y_m^L := JX_m^L J$. Hence a saving of $12.5m^3$ operations may be incurred since one would only have to solve (41) and form the matrix J. A simple way to generate the sign matrix J is to employ the following scheme.

```
J(1,1) := -1
        For i := 2 to m
                J(i,i) := 1 · sign[J(i-1,i-1)T_m(i,i-1)]
        End.
```

The following corollary gives perturbations on the data in (5) and (6) such that the low rank approximate solutions given in Theorem 2.3 are exact.

COROLLARY 2.4. *Suppose that m steps of the Lanczos process have been taken and that* $P_m := V_m X_m^L V_m'$ *and* $Q_m := W_m Y_m^L W_m'$ *are the low rank approximate solutions to* (5) *and* (6), *respectively, and furthermore that* X_m^L *and* Y_m^L *satisfy* (41) *and* (43), *respectively. Then*

$$(A - \Delta_1)P_m + P_m(A - \Delta_1)' + bb' = 0, \tag{45}$$

$$(A - \Delta_2)'Q_m + Q_m(A - \Delta_2) + c'c = 0, \tag{46}$$

where $\Delta_1 = v_{m+1}\delta_{m+1}w_m'$ *and* $\Delta_2 = v_m\beta_{m+1}w_{m+1}'$. *Furthermore, it holds that* $\|\Delta_1\|_F = \|\hat{v}_{m+1}\|_2\|w_m\|_2$ *and* $\|\Delta_2\|_F = \|v_m\|_2\|\hat{w}_{m+1}\|_2$.

Proof. The proof is essentially the same as that of Corollary 2.2. □

It is interesting to note that $\|\Delta_1\|_F$ and $\|\Delta_2\|_F$ may be evaluated directly from the data generated in the course of the last two steps of the Lanczos process. Compared to a residual error checking scheme, an implementation based on monitoring the backward error evolution would require less computations and storage, i.e., $\|\Delta_1\|_F$ requires $2N$ operations while evaluating $\|R_m^L\|_F$ consumes in excess of $3Nm$ flops. Observe that $\Delta_3 := \Delta_1 + \Delta_2$ is a perturbation on the data in A such that $(A - \Delta_3)P_m + P_m(A - \Delta_3)' + bb' = 0$ and $(A - \Delta_3)'Q_m + Q_m(A - \Delta_3) + c'c = 0$. Furthermore, Δ_3 is at most a rank-2 perturbation that may be factorised as

$$\Delta_3 = [v_m \;\; v_{m+1}] \begin{bmatrix} 0 & \beta_{m+1} \\ \delta_{m+1} & 0 \end{bmatrix} \begin{bmatrix} w_m' \\ w_{m+1}' \end{bmatrix}. \tag{47}$$

Finally, a direct calculation will verify that $W_m'\Delta_1 V_m = W_m'\Delta_2 V_m = W_m'\Delta_3 V_m = 0$.

In contrast to the Arnoldi method, the Lanczos process yields matrices V_m and W_m that are no longer orthogonal. Consequently, manipulations with either V_m or W_m may suffer from numerical difficulties because of possible poor conditioning. There is frequently a loss of biorthogonality that may be checked by rebi-orthogonalising the newly computed v_{j+1} and w_{j+1} against V_m and W_m [20]. Finally, we observe that the Lyapunov equation solvers stemming from the Lanczos process enjoy lower complexity and storage requirements than Arnoldi solvers presented in §2.1.

2.3. A posteriori backward error. Even though one might elect to monitor the evolutions of Δ_1, Δ_2 or Δ_3 for increasing m, there is no guarantee that their norms are nonincreasing. In fact our computational experience has revealed that $\|\Delta_i\|_F$ for $i = 1, 2, 3$ may behave erratically for increasing m. The conservative nature of a priori expression for $\|\Delta_i\|_F$ for $i = 1, 2, 3$ leads us to seek alternatives that enable us to assess the quality of the results obtained when exploiting the techniques proposed in this paper. It is interesting to consider whether, for given P_m and Q_m, there exist other, possibly smaller Δ_1 and Δ_2 for which (29) and (30) hold. Thus the aim of this section is to derive a posteriori expressions Δ_4, Δ_5 and Δ_6 for Δ_1, Δ_2 and Δ_3, respectively.

COROLLARY 2.5. *Suppose that m steps of the Arnoldi process have been taken and that* X_m^A *and* Y_m^A *are the solutions to the m-dimensional Lyapunov equations* (21) *and* (24), *respectively. Suppose that*

$$X_m^A = [U_r \; U_\perp] \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_r' \\ U_\perp' \end{bmatrix}$$

and

$$Y_m^A = [U_s \ \tilde{U}_\perp] \begin{bmatrix} \Sigma_s & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_s' \\ \tilde{U}_\perp' \end{bmatrix}$$

are Schur decompositions in which $\Sigma_r \in \mathbb{R}^{r\times r}$ *and* $\Sigma_s \in \mathbb{R}^{s\times s}$ *are nonsingular and* $U_r \in \mathbb{R}^{m\times r}$ *and* $U_s \in \mathbb{R}^{m\times s}$. *Then*

$$(A - \Delta_4)P_m + P_m(A - \Delta_4)' + bb' = 0, \tag{48}$$

$$(A - \Delta_5)'Q_m + Q_m(A - \Delta_5) + c'c = 0, \tag{49}$$

where

$$\Delta_4 = (I - V_m(W_m'V_m)^{-1}W_m')v_{m+1}h_{m+1,m}e_m'U_rU_r'V_m', \tag{50}$$

$$\Delta_5 = W_mU_sU_s'e_mg_{m,m+1}w_{m+1}'(I - V_m(W_m'V_m)^{-1}W_m'). \tag{51}$$

Furthermore, the norms are given by $\|\Delta_4\|_F^2 = h_{m+1,m}^2\{1+\|(W_m'V_m)^{-1}W_m'v_{m+1}\|_2^2\}\cdot \|e_m'U_r\|_2^2$ *and* $\|\Delta_5\|_F^2 = g_{m,m+1}^2\{1 + \|(V_m'W_m)^{-1}V_m'w_{m+1}\|_2^2\}\|e_m'U_s\|_2^2$.

Proof. This proof is essentially the same as Corollary 2.2 except that it uses $X_m := X_m^A := U_r\Sigma_rU_r'$ and $Y_m := Y_m^A := U_s\Sigma_sU_s'$ and the facts that U_r and U_s are parts of orthogonal matrices. A more detailed proof may be found in [12]. □

Remark 2.3. It is clear from (31) and (50) that $\|\Delta_4\|_F \le \|\Delta_1\|_F$ since $\|e_m'U_r\|_2 \le 1$. Similarly, $\|\Delta_5\|_F \le \|\Delta_2\|_F$ since $\|e_m'U_s\|_2 \le 1$, in the event that X_m^A and Y_m^A have rank equal to m $\|\Delta_1\|_F = \|\Delta_4\|_F$ and $\|\Delta_2\|_F = \|\Delta_5\|_F$, respectively. Finally, observe that Δ_4 and Δ_5 are rank-1 perturbations and that $\Delta_4 + \Delta_5 =: \Delta_6$ is a perturbation that simultaneously satisfies (34) and (35).

A posteriori perturbation bounds may be derived for the nonsymmetric Lanczos process that we now state without proof.

COROLLARY 2.6. *Suppose that* m *steps of the Lanczos process have been successfully completed and that* X_m^L *and* Y_m^L *are the solutions to the* m*-dimensional Lyapunov equations* (41) *and* (43), *respectively. Suppose that*

$$X_m^A = [U_r \ U_\perp] \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_r' \\ U_\perp' \end{bmatrix}$$

and

$$Y_m^A = [U_s \ \tilde{U}_\perp] \begin{bmatrix} \Sigma_s & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_s' \\ \tilde{U}_\perp' \end{bmatrix}$$

are Schur decompositions in which $\Sigma_r \in \mathbb{R}^{r\times r}$ *and* $\Sigma_s \in \mathbb{R}^{s\times s}$ *are nonsingular and* $U_r \in \mathbb{R}^{m\times r}$ *and* $U_s \in \mathbb{R}^{m\times s}$. *Then*

$$(A - \Delta_4)P_m + P_m(A - \Delta_4)' + bb' = 0, \tag{52}$$

$$(A - \Delta_5)'Q_m + Q_m(A - \Delta_5) + c'c = 0, \tag{53}$$

where

$$\Delta_4 = v_{m+1}\delta_{m+1}e_m'U_rU_r'W_m', \tag{54}$$

$$\Delta_5 = V_mU_sU_s'e_m\beta_{m+1}w_{m+1}', \tag{55}$$

and the norms are given by $\|\Delta_4\|_F = \|\hat{v}_{m+1}\|_2 \cdot \|W_m U_r U_r' e_m\|_2$ *and* $\|\Delta_5\|_F^2 = \|\hat{w}_{m+1}\|_2 \cdot \|V_m U_s U_s' e_m\|_2$.

Observe that as with the Arnoldi process, if X_m^L and Y_m^L have rank equal to m, then Δ_4 and Δ_5 defined in (52) and (53) degenerate into $\Delta_1 = v_{m+1}\delta_{m+1}w_m'$ and $\Delta_2 = v_m\beta_{m+1}w_{m+1}'$, respectively. Furthermore, $\Delta_6 := \Delta_4 + \Delta_5$ is a rank-2 perturbation of the matrix A such that $(A - \Delta_6)P_m + P_m(A - \Delta_6)' + bb' = 0$ and $(A - \Delta_6)'Q_m + Q_m(A - \Delta_6) + c'c = 0$.

3. Model reduction using Krylov subspace methods. The aim of this section is to consider the Krylov subspace techniques described above to provide computationally efficient model reduction schemes for large scale systems. Denoting a transfer function by $F(s)$

$$F(s) = d + c(sI - A)^{-1}b \stackrel{s}{=} \left[\begin{array}{c|c} A & b \\ \hline c & d \end{array}\right]; \quad A \in \mathbb{R}^{N\times N}, \quad b, c' \in \mathbb{R}^N, \quad \text{and} \quad d \in \mathbb{R}. \tag{56}$$

The task is to determine a reduced order model $F_m(s)$, where

$$\begin{aligned} F_m(s) &= d_m + c_m(sI - A_m)^{-1}b_m \stackrel{s}{=} \left[\begin{array}{c|c} A_m & b_m \\ \hline c_m & d_m \end{array}\right]; \\ A_m &\in \mathbb{R}^{m\times m}, \quad b_m, c_m' \in \mathbb{R}^m \quad \text{and} \quad d_m \in \mathbb{R}, \end{aligned} \tag{57}$$

that approximates the high dimensional model $F(s)$. Associated with the linear system in (56), we define the controllability and observability Lyapunov equations

$$AP + PA' + bb' = 0, \tag{58}$$

$$A'Q + QA + c'c = 0, \tag{59}$$

respectively. The low rank approximate solutions to (58) and (59) may be computed via low dimensional calculations as demonstrated in Theorem 2.1. It is natural to question whether the approximate grammians P_m and Q_m are the exact controllability and observability grammians of a perturbed linear system. From Remark 2.2, it is apparent that $P_m = V_m X_m^A V_m'$ and $Q_m = W_m Y_m^A W_m'$ are, respectively, the controllability and observability grammians of $F_\Delta(s)$, where

$$F_\Delta(s) = d + c(sI - A + \Delta_3)^{-1}b \stackrel{s}{=} \left[\begin{array}{c|c} A - \Delta_3 & b \\ \hline c & d \end{array}\right], \tag{60}$$

where X_m^A and Y_m^A satisfy (21) and (24), respectively, and Δ_3 is defined in (36). Alternatively, one may employ the relations derived for the Lanczos process in §2.2 to conclude that $P_m = V_m X_m^L V_m'$ and $Q_m = W_m Y_m^L W_m'$ are, respectively, the controllability and observability grammians of $F_\Delta(s)$ in which X_m^L and Y_m^L satisfy (41) and (43), respectively, and Δ_3 is defined in (47). The main results of this section show that (60) is equivalent to a low dimensional linear system and gives computable expression for an $\mathcal{L}^\infty$ error between the high dimensional and low order approximate model.

From (56), $F(s) = d + cf_b(s) = d + f_c(s)b$, where $f_b(s) = (sI - A)^{-1}b$ and $f_c(s) = c(sI - A)^{-1}$; for later reference, $f_b(s)$ and $f_c(s)$ may be considered as the solutions to the coupled linear systems

$$(sI - A)f_b(s) = b, \tag{61}$$

$$f_c(s)(sI - A) = c, \tag{62}$$

respectively. The focus of what follows is to approximate $F(s)$ by obtaining approximate solutions to the linear systems (61) and (62). The approximate solutions $f_{b,m}(s)$ and $f_{c,m}(s)$ to the linear systems (61) and (62) are constructed to satisfy the following two conditions: (i) $f_{b,m}(s) \in \mathcal{K}_m(A,b)$, i.e., $f_{b,m}(s) = V_m F_b(s)$, such that $\mathcal{L}_m(A',c') \perp \{(sI-A)f_{b,m}(s)-b\}$; (ii) $f'_{c,m}(s) \in \mathcal{L}_m(A',c')$, i.e., $f_{c,m}(s) = F_c(s)W'_m$, such that $\{f_{c,m}(s)(sI-A)-c\} \perp \mathcal{K}_m(A,b)$. Since $f_{b,m}(s)$ and $f_{c,m}(s)$ are approximate solutions to (61) and (62), and $F(s) = d + cf_b(s) = d + f_c(s)b$, we consider $F_{m,1}(s) = d + cf_{b,m}(s)$ and $F_{m,2}(s) = d + f_{c,m}(s)b$ as low order approximations to $F(s)$. The problem we wish to solve may be stated as follows.

PROBLEM 3.1. Find approximate solutions $f_{b,m}(s) = V_m F_b(s)$ and $f_{c,m}(s) = F_c(s)W'_m$ to (61) and (62), respectively, which satisfy the Galerkin type conditions

$$W'_m\{(sI-A)V_mF_b(s)-b\} = 0 \quad \forall s, \tag{63}$$

$$\{F_c(s)W'_m(sI-A)-c\}V_m = 0 \quad \forall s. \tag{64}$$

The following is referred to as a basis change in the state space realisation of $F(s)$

$$F(s) \stackrel{s}{=} \left[\begin{array}{c|c} A & b \\ \hline c & d \end{array}\right] \stackrel{T}{\rightarrow} F(s) \stackrel{s}{=} \left[\begin{array}{c|c} TAT^{-1} & Tb \\ \hline cT^{-1} & d \end{array}\right], \tag{65}$$

where T is nonsingular. The next lemma is needed in the proof of the main result.

LEMMA 3.1. *Suppose that m steps of the Arnoldi process have been completed and that Δ_3 is given by* (36), *then,* $-W'_m(A-\Delta_3) + W'_mAV_m(W'_mV_m)^{-1}W'_m = 0$.

Proof. It holds that

$$\begin{aligned} &W'_m\Delta_3 - W'_mA(I - V_m(W'_mV_m)^{-1}W'_m) \\ &\quad = W'_mw_mg_{m,m+1}w'_{m+1}(I - V_m(W'_mV_m)^{-1}W'_m) \\ &\qquad -e_mg_{m,m+1}w'_{m+1}(I - V_m(W'_mV_m)^{-1}W'_m) \\ &\quad = 0, \end{aligned}$$

since $W'_mw_m = e_m$. □

The following theorem gives the solution to Problem 3.1.

THEOREM 3.2. *Suppose that m steps of the Arnoldi process have been taken and that W'_mV_m is nonsingular, then the following are true.*

(a) *The Galerkin conditions in* (63) *and* (64) *are satisfied if and only if $F_b(s) = (sI-\hat{H}_m)^{-1}e_1\beta$ and $F_c(s) = \delta e'_1(sI-\hat{G}_m)^{-1}$. Under these conditions, the residual error norms are*

$$\|b-(sI-A)V_mF_b(s)\|_\infty = h_{m+1,m}\left\|\begin{bmatrix} (W'_mV_m)^{-1}W'_mv_{m+1} \\ 1 \end{bmatrix}\right\|_2 \|e'_mF_b(s)\|_\infty, \tag{66}$$

$$\|c-F_c(s)W'_m(sI-A)\|_\infty = g_{m,m+1}\left\|\begin{bmatrix} (V'_mW_m)^{-1}V'_mw_{m+1} \\ 1 \end{bmatrix}\right\|_2 \|F_c(s)e_m\|_\infty. \tag{67}$$

(b) *$F_\Delta(s)$, $F_{m,1}(s)$, and $F_{m,2}(s)$ are different realisations of the same transfer function, namely,*

$$F_\Delta(s) \equiv F_{m,1}(s), \tag{68}$$

$$F_\Delta(s) \equiv F_{m,2}(s), \tag{69}$$

where

$$F_\Delta(s) \stackrel{s}{=} \left[\begin{array}{c|c} A-\Delta_3 & b \\ \hline c & d \end{array}\right], \tag{70}$$

$$F_{m,1}(s) = d + cV_mF_b(s) \stackrel{s}{=} \left[\begin{array}{c|c} (W_m'V_m)^{-1}W_m'AV_m & (W_m'V_m)^{-1}W_m'b \\ \hline cV_m & d \end{array}\right] = \left[\begin{array}{c|c} \hat{H}_m & e_1\beta \\ \hline c_m & d \end{array}\right], \tag{71}$$

$$F_{m,2}(s) = d + F_c(s)W_m'b \stackrel{s}{=} \left[\begin{array}{c|c} W_m'AV_m(W_m'V_m)^{-1} & W_m'b \\ \hline cV_m(W_m'V_m)^{-1} & d \end{array}\right] = \left[\begin{array}{c|c} \hat{G}_m & b_m \\ \hline \delta e_1' & d \end{array}\right]. \tag{72}$$

(c) X_m^A *and* $(V_m'W_m)Y_m^A(W_m'V_m)$ *are the controllability and observability grammians of* $F_{m,1}(s)$.

(d) $(W_m'V_m)X_m^A(V_m'W_m)$ *and* Y_m^A *are the controllability and observability grammians of* $F_{m,2}(s)$.

Proof. The residue associated with the approximate solution to (61) is $(sI - A)V_mF_b(s) - b$, premultiplying by W_m' and substituting (17) leads to

$$\begin{aligned} W_m'\{(sI-A)V_mF_b - b\} &= W_m'V_m\{(sI-(W_m'V_m)^{-1}W_m'AV_m)F_b(s) \\ &\quad -(W_m'V_m)^{-1}W_m'b\} \\ &= (W_m'V_m)^{-1}\left\{(sI-\hat{H}_m)F_b(s) - e_1\beta\right\}. \end{aligned}$$

The result follows immediately since $(W_m'V_m)$ is assumed to be nonsingular. Similarly, postmultiplying the residue associated with the approximate solution to (62) and substituting (18) leads to $F_c(s) = \delta e_1'(sI - \hat{G}_m)^{-1}$. For the $\mathcal{L}^\infty$ error bounds, we have

$$\begin{aligned} &\left\|b - (sI-A)V_m(sI-\hat{H}_m)^{-1}e_1\beta\right\|_\infty \\ &\quad = \left\|V_me_1\beta - (V_ms - V_mH_m - v_{m+1}h_{m+1,m}e_m')(sI-\hat{H}_m)^{-1}e_1\beta\right\|_\infty \\ &\quad = h_{m+1,m}\left\|(I - V_m(W_m'V_m)^{-1}W_m')v_{m+1}\right\|_2\|e_m'F_b(s)\|_\infty, \end{aligned}$$

from which follows the relation in (66); similarly, for (67), which completes the proof of part (a).

We establish that $F_\Delta(s) \equiv F_{m,1}(s)$ by taking the difference between the two transfer function models

$$F_\Delta(s) - F_{m,1}(s) \stackrel{s}{=} \left[\begin{array}{cc|c} A-\Delta_3 & 0 & b \\ 0 & (W_m'V_m)^{-1}W_m'AV_m & (W_m'V_m)^{-1}W_m'b \\ \hline c & -cV_m & 0 \end{array}\right]. \tag{73}$$

Consider the following basis transformation T where

$$T = \left[\begin{array}{cc} I & 0 \\ -(W_m'V_m)^{-1}W_m' & I \end{array}\right]; \quad \text{and} \quad T^{-1} = \left[\begin{array}{cc} I & 0 \\ (W_m'V_m)^{-1}W_m' & I \end{array}\right],$$

which yields

$$F_\Delta(s) - F_{m,1}(s) \stackrel{s}{=} \left[\begin{array}{cc|c} A-\Delta_3 & 0 & b \\ (W_m'V_m)^{-1}W_m'(AV_m(W_m'V_m)^{-1}W_m' - A + \Delta_3) & (W_m'V_m)^{-1}W_m'AV_m & 0 \\ \hline 0 & -cV_m & 0 \end{array}\right], \tag{74}$$

since $c' = W_m e_1 \delta$. By Lemma 3.1, the (2,1) block of (74) is zero, from which we conclude that $F_\Delta(s) - F_{m,1}(s) = 0$ since the realization in (74) has N unobservable and m uncontrollable modes thus establishing the first equivalence. $F_\Delta(s) \equiv F_{m,2}(s)$ follows by applying the basis transformation $T = (W_m' V_m)$ to $F_{m,2}(s)$ to give $F_{m,2}(s) \xrightarrow{T} F_{m,1}(s)$; the second equivalence is immediate, thus completing the proof of part (b).

Using $(W_m' V_m)^{-1} W_m' b = e_1 \beta$, it follows that X_m^A is the controllability grammian of (71). For the observability grammian, we observe that

$$(W_m' V_m)^{-1} W_m' A V_m = (W_m' V_m)^{-1} (G_m W_m' + g_{m,m+1} e_m w_{m+1}') V_m, \tag{75}$$

$$c V_m = \delta e_1' W_m' V_m. \tag{76}$$

Substituting (75) and (76) into (71) gives the observability Lyapunov equation

$$\begin{aligned} &Z (W_m' V_m)^{-1} (G_m W_m' + g_{m,m+1} e_m w_{m+1}') V_m \\ &+ V_m' (W_m G_m' + w_{m+1} e_m' g_{m,m+1}) (V_m' W_m)^{-1} Z + V_m' W_m e_1 \delta^2 e_1' W_m' V_m = 0, \end{aligned} \tag{77}$$

for the realization in (71); the unknown grammian is Z. Pre and postmultiplying (77) by $(V_m' W_m)^{-1}$ and $(W_m' V_m)^{-1}$ yields

$$\begin{aligned} &(V_m' W_m)^{-1} Z (W_m' V_m)^{-1} (G_m + g_{m,m+1} e_m w_{m+1}' (W_m' V_m)^{-1}) \\ &+ (G_m' + (V_m' W_m)^{-1} w_{m+1} e_m' g_{m,m+1}) (V_m' W_m)^{-1} Z (W_m' V_m)^{-1} + e_1 \delta^2 e_1' = 0, \end{aligned} \tag{78}$$

from which we deduce that $Z = (V_m' W_m) Y_m^A (W_m' V_m)$ is the observability grammian of (71) and completes the proof of part (c).

The proof to part (d) is identical to part (c) except that it uses (72). □

Remark 3.1. We note that the Galerkin type conditions of (63) and (64) are analogous to parts (a) and (b) of Theorem 2.1, respectively; similarly, the residual error norm expressions in (66) and (67) are analogous to (22) and (25), respectively. In a practical implementation, the equalities in (66) and (67) enable us to economically monitor the progress of the iterative process at each step. An open issue is that of obtaining computable expressions for the $\mathcal{L}^\infty$ error $\|F(s) - F_{m,i}(s)\|_\infty$.

Remark 3.2. An implication of the orthogonality property reported in Remark 2.2 is that the perturbation is confined to a progressively smaller subspace of $\mathbb{R}^{N \times N}$ for increasing m. Thus Δ_3 may be employed as a stopping condition for an iterative model reduction algorithm in which $\|\Delta_3\|_F$ reports on the size of the perturbation that is effected on A to obtain a reduced order model of state dimension no greater that m.

Remark 3.3. It is interesting to observe that the realisations of $F_{m,1}(s)$ and $F_{m,2}(s)$ given in Theorem 3.2 depend only on the data generated in the course of the two Arnoldi processes, and that despite the relations between the low dimensional realisations and the low order Lyapunov equations in (21) and (24), one can construct $F_{m,1}(s)$ and $F_{m,2}(s)$ without having to form X_m^A and Y_m^A.

It is clear from Corollary 2.2 that Δ_3 defined in (36) is not the only perturbation that leads to the reduced order models of Theorem 3.2. The effect of Δ_3 is to perturb A in such a way that the nonminimal modes in the perturbed system are simultaneously uncontrollable and unobservable, while, to obtain a reduced order model, it is sufficient to perturb A so as to obtain either $N - m$ uncontrollable or unobservable modes.

COROLLARY 3.3. *Suppose that m steps of the Arnoldi processes have been completed and that* Δ_1 *and* Δ_2 *are defined by* (31) *and* (32) *respectively, then,*

$$F_{m,3}(s) \overset{s}{:=} \left[\begin{array}{c|c} A - \Delta_1 & b \\ \hline c & d \end{array}\right] \equiv \left[\begin{array}{c|c} (W_m' V_m)^{-1} W_m' A V_m & (W_m' V_m)^{-1} W_m' b \\ \hline c V_m & d \end{array}\right]$$

$$= \left[\begin{array}{c|c} \hat{H}_m & e_1\beta \\ \hline c_m & d \end{array}\right] \stackrel{s}{=} F_{m,1}(s),$$

$$F_{m,4}(s) := \left[\begin{array}{c|c} A-\Delta_2 & b \\ \hline c & d \end{array}\right] \equiv \left[\begin{array}{c|c} W_m'AV_m(W_m'V_m)^{-1} & W_m'b \\ \hline cV_m(W_m'V_m)^{-1} & d \end{array}\right]$$

$$= \left[\begin{array}{c|c} \hat{G}_m & b_m \\ \hline \delta e_1' & d \end{array}\right] \stackrel{s}{=} F_{m,2}(s).$$

Proof. We establish that $F_{m,3}(s) \equiv F_{m,1}(s)$ by taking the difference between the two transfer function models

$$F_{m,3}(s) - F_{m,1}(s) \stackrel{s}{=} \left[\begin{array}{cc|c} A-\Delta_1 & 0 & b \\ 0 & (W_m'V_m)^{-1}W_m'AV_m & (W_m'V_m)^{-1}W_m'b \\ \hline c & -cV_m & 0 \end{array}\right]. \tag{79}$$

Consider the following basis transformation T where

$$T = \begin{bmatrix} I & -V_m \\ 0 & I \end{bmatrix}, \quad \text{and} \quad T^{-1} = \begin{bmatrix} I & V_m \\ 0 & I \end{bmatrix},$$

which yields

(80)

$$F_{m,3}(s) - F_{m,1}(s) \stackrel{s}{=} \left[\begin{array}{cc|c} A-\Delta_1 & (A-\Delta_1)V_m - V_m(W_m'V_m)^{-1}W_m'AV_m & 0 \\ 0 & (W_m'V_m)^{-1}W_m'AV_m & (W_m'V_m)^{-1}W_m'b \\ \hline c & 0 & 0 \end{array}\right].$$

Since $b = V_m e_1 \beta$, a routine calculation shows that the (1,2) block of (81) is zero, from which we conclude that $F_{m,3}(s) - F_{m,1}(s) = 0$ since the realisation in (81) has m unobservable and N uncontrollable modes thus establishing the first equivalence. $F_{m,4}(s) \equiv F_{m,2}(s)$ follows in an analogous way, but by applying the transformation

$$T = \begin{bmatrix} I & 0 \\ -W_m' & I \end{bmatrix}$$

to the difference $F_{m,4}(s) - F_{m,2}(s)$, thus completing the proof. □

It is apparent from Theorem 3.2 and Corollary 3.3 that

$$\left[\begin{array}{c|c} A-\Delta_3 & b \\ \hline c & d \end{array}\right] \equiv \left[\begin{array}{c|c} A-\Delta_1 & b \\ \hline c & d \end{array}\right] \quad \text{and} \quad \left[\begin{array}{c|c} A-\Delta_3 & b \\ \hline c & d \end{array}\right] \equiv \left[\begin{array}{c|c} A-\Delta_2 & b \\ \hline c & d \end{array}\right], \tag{81}$$

which may be demonstrated by, respectively, applying the basis changes

$$T = \begin{bmatrix} I & 0 \\ V_m(W_m'V_m)^{-1}W_m' & I \end{bmatrix} \quad \text{and} \quad T = \begin{bmatrix} I & V_m(W_m'V_m)^{-1}W_m' \\ 0 & I \end{bmatrix}$$

to the differences between each transfer function in (81). The perturbation Δ_1 to the transition matrix of $F(s)$ yields $N-m$ uncontrollable modes, while the perturbation Δ_2 gives rise to $N-m$ unobservable modes. It is interesting to observe that despite the fact that Δ_1, Δ_2 and Δ_3 have different Frobenius norms, each perturbed linear system is a different realisation of the same transfer function.

The following theorem states the findings above in the context of the Lanczos process.

THEOREM 3.4. *Suppose that m steps of the Lanczos process have been taken and that Δ_1, Δ_2 and Δ_3 are defined in Corollary 2.4 and (47), then, the following are true.*

(a) *The Galerkin conditions in (63) and (64) are satisfied if and only if $F_b(s) = (sI - T_m)^{-1}e_1\delta_1$ and $F_c(s) = \beta_1 e_1'(sI - T_m)^{-1}$; furthermore, the residual error norms are*

$$\|b - (sI - A)V_m F_b(s)\|_\infty = \|\hat{v}_{m+1}\|_2 \cdot \|e_m' F_b(s)\|_\infty, \tag{82}$$

$$\|c - F_c(s)W_m'(sI - A)\|_\infty = \|\hat{w}_{m+1}\|_2 \cdot \|F_c(s)e_m\|_\infty. \tag{83}$$

(b) *$F_\Delta(s)$ and $F_m(s)$ are different realisations of the same transfer function, namely, $F_\Delta(s) \equiv F_m(s)$ where*

(84)

$$F_\Delta(s) \stackrel{s}{=} \left[\begin{array}{c|c} A - \Delta_3 & b \\ \hline c & d \end{array}\right] \quad \text{and} \quad F_m(s) \stackrel{s}{=} \left[\begin{array}{c|c} W_m'AV_m & e_1\delta_1 \\ \hline \beta_1 e_1' & d \end{array}\right] = \left[\begin{array}{c|c} T_m & e_1\delta_1 \\ \hline \beta_1 e_1' & d \end{array}\right].$$

(c) *X_m^L and Y_m^L are the controllability and observability grammians of $F_m(s)$.*

(d) $$\left[\begin{array}{c|c} A - \Delta_1 & b \\ \hline c & d \end{array}\right] \equiv \left[\begin{array}{c|c} A - \Delta_2 & b \\ \hline c & d \end{array}\right] \equiv \left[\begin{array}{c|c} A - \Delta_3 & b \\ \hline c & d \end{array}\right] \equiv \left[\begin{array}{c|c} T_m & e_1\delta_1 \\ \hline \beta_1 e_1' & d \end{array}\right].$$

Proof. The proof is essentially the same as that of Theorem 3.2 and Corollary 3.3 except that it uses the fact that $W_m'(A - \Delta_3) - W_m'AV_mW_m' = 0$. □

The appeal of the Lanczos process in model reduction stems from the simplicity of the formulas derived in Theorem 3.4. The reduced order model and the backward error perturbations are expressed in terms of the data generated in the course of the Lanczos process. Thus despite the relations with Lyapunov equations, the reduced order model may be constructed without the need to form X_m^L and Y_m^L. The Arnoldi–Lanczos schemes suggested in this paper are not always guaranteed to yield stable reduced order models. A possible remedy to this problem is to employ an implicitly restarted Lanczos process to compute stable reduced order models. This method was proposed in [8].

The following procedure summarises the findings of this section.

KRYLOV SUBSPACE MODEL REDUCTION ALGORITHM

- Start: Specify tolerances $\gamma > 0$ and $\epsilon > 0$, set an integer parameter m.
  Arnoldi
- Perform m steps of the Arnoldi process with (A, b) to produce $\hat{H}_m$, $h_{m+1,m}$, V_m, v_{m+1}, and β.
- Perform m steps of the Arnoldi process with (A', c') to produce $\hat{G}_m$, $g_{m,m+1}$, W_m, w_{m+1}, and δ.
- Form the reduced order model from either (71) or (72).
- Test the $\mathcal{L}^\infty$ errors in (66) and (67), if either (66) $> \epsilon$ or (67) $> \epsilon$ increase m and continue the Arnoldi process.
  Lanczos
- Perform m steps of the Lanczos process with A, b, c' to produce T_m, V_m, v_{m+1}, W_m, w_{m+1}, δ_1, δ_{m+1}, β_1, and β_{m+1}.
- Form the reduced order model from (84).

- Test the $\mathcal{L}^\infty$ errors in (82) and (83), if either (82) $> \epsilon$ or (83) $> \epsilon$ increase m and continue the Lanczos process.

Remark 3.4. The Arnoldi- and Lanczos-based model reduction algorithms presented above are mathematically equivalent in the sense that they produce the same low order transfer function. This equivalence can be established by using (15), (16), (37), and (38) and the fact that V_m (similarly, W_m) computed from the Arnoldi and the Lanczos methods are different bases for the same space.

In practice, it is desirable to perform model reduction of multivariable linear systems in which $F(s) = D + C(sI - A)^{-1}B$ where $B \in \mathbb{R}^{N\times p}$ and $C \in \mathbb{R}^{q\times N}$. To this end, one may employ block Krylov schemes to compute the bases V_m and W_m of $\mathcal{K}_m(A, B)$ and $\mathcal{L}_m(A', C')$, respectively. The reduced order models obtained using such schemes are given by

$$F_{m,1}(s) \stackrel{s}{=} \left[\begin{array}{c|c} (W_m'V_m)^{-1}W_m'AV_m & (W_m'V_m)^{-1}W_m'B \\ \hline CV_m & D \end{array}\right] \quad \text{and}$$

$$F_{m,2}(s) \stackrel{s}{=} \left[\begin{array}{c|c} W_m'AV_m(W_m'V_m)^{-1} & W_m'B \\ \hline CV_m(W_m'V_m)^{-1} & D \end{array}\right]$$

for the block Arnoldi process, assuming that $(W_m'V_m)$ is nonsingular, and by

$$F_m \stackrel{s}{=} \left[\begin{array}{c|c} W_m'AV_m & W_m'B \\ \hline CV_m & D \end{array}\right]$$

for the block Lanczos process. We refer the reader to [5], [11], [15] for the implementation details of block Krylov methods. In the absence of breakdown, one may derive a nonsymmetric block Lanczos algorithm based on a generalisation of the vector scheme suggested in [5]. The behaviour of such an algorithm remains an open area of research in the presence of breakdowns.

Suppose that m steps of the Arnoldi process have been taken and that either $h_{m+1,m} = 0$ or $g_{m,m+1} = 0$. Three possible scenarios may then arise. The first, $h_{m+1,m} = 0$, implies that v_{m+1} may not be computed and the process yields the exact solution; $P^* = P_m = V_m X_m^A V_m'$; $Q_m = W_m Y_m^A W_m'$ is a low rank approximate solution to (6). The matrix V_m forms an orthogonal basis for the controllable space, which implies that the reduced order models given by Theorem 3.2 are equivalent to the high order model $F(s)$ in (56). The second, $g_{m,m+1} = 0$, implies that w_{m+1} may not be computed and the process yields the exact solution; $Q^* = Q_m = W_m Y_m^A W_m'$; $P_m = V_m X_m^A V_m'$ is a low rank approximate solution to (5). In this case, W_m forms an orthogonal basis for the observability space and the reduced order models of Theorem 3.2 are equivalent to $F(s)$. The third, $h_{m+1,m} = 0$ and $g_{m,m+1} = 0$, implies that v_{m+1} and w_{m+1} may not be formed and the process yields exact grammians; here, $F_{m,i}(s) \equiv F(s)$ for $i = 1, 2$. A similar type of breakdown is experienced in the Lanczos process if either $\hat{v}_{m+1} = 0$ or $\hat{w}_{m+1} = 0$. The key observation associated with these types of breakdowns is that the reduced order models are minimal realisations of the high order $F(s)$; this connection was first established in [14]. The Lanczos process might suffer from a breakdown in which $\hat{v}_{m+1} \neq 0$ and $\hat{w}_{m+1} \neq 0$ and yet $\hat{v}_{m+1}' w_{m+1} = 0$; similarly, for the Arnoldi scheme, $(W_m'V_m)$ might be singular. The implications of such breakdown on the model reduction algorithms is not well understood and is the focus of ongoing research.

Observe that if the solutions to the m-dimensional Lyapunov equations have rank $< m$, $F_{m,i}(s)$ for $i = 1, 2$ are not minimal realisations. By exploiting X_m and Y_m one

readily computes the balancing transformations and the balanced $F_{m,i}(s)$'s may then be truncated to yield minimal realisations [13]. An alternative, and numerically sound approach, is to utilise the hybrid scheme proposed in [1]. This procedure combines the Krylov subspace methods to the balancing procedure via orthogonal transformations without computing the potentially ill-conditioned balancing transformation. For the implementation details, we refer the reader to [1].

4. Model reduction of large scale discrete time systems. The focus of this section is to extend the findings above to obtain model reduction schemes suited to large scale discrete time systems. The extensions are fairly straightforward and are stated without proofs.

The need for discrete time model reduction schemes arises when high order models described by difference equations are approximated by those of lower dimension. High order models arise in such areas as the modeling of digital circuits, communication networks, and model identification. Consider the discrete time linear dynamical system

$$x_{k+1} = Ax_k + bu_k \qquad A \in \mathbb{R}^{N\times N},\ b \in \mathbb{R}^N, \tag{85}$$

$$y_k = cx_k + du_k \qquad c' \in \mathbb{R}^N,\ d \in \mathbb{R}. \tag{86}$$

Associated with this linear system are the controllability and observability Lyapunov equations defined as

$$APA' - P + bb' = 0, \tag{87}$$

$$A'QA - Q + c'c = 0, \tag{88}$$

respectively. Low rank solutions $P_m = V_m X_m V_m'$ and $Q_m = W_m Y_m W_m'$ are sought such that the residuals $AP_mA' - P_m + bb'$ and $A'Q_mA - Q_m + c'c$ satisfy Galerkin-type conditions on the Krylov spaces $\mathcal{L}_m(A', c')$ and $\mathcal{K}_m(A, b)$, respectively. Assuming that $W_m'V_m$ is nonsingular allows the residual error to be expressed as

$$\begin{aligned} R_m(X_m) = [V_m \quad & (I - V_m(W_m'V_m)^{-1}W_m')v_{m+1}] \\ & \times \begin{bmatrix} \hat{H}_m X_m \hat{H}_m' - X_m + e_1\beta^2 e_1' & X_m e_m h_{m+1,m} \\ h_{m+1,m} e_m' X_m & h_{m+1,m}^2 e_m' X_m e_m \end{bmatrix} \\ & \times \begin{bmatrix} V_m' \\ v_{m+1}'(I - W_m(V_m'W_m)^{-1}V_m') \end{bmatrix}. \end{aligned} \tag{89}$$

Similarly, associated with (88), the residual error function for any given approximate solution of the form $Q_m = W_m Y_m W_m'$ may be written as

$$\begin{aligned} S_m(Y_m) = [W_m \quad & (I - W_m(V_m'W_m)^{-1}V_m')w_{m+1}] \\ & \times \begin{bmatrix} \hat{G}_m' Y_m \hat{G}_m - Y_m + e_1\delta^2 e_1' & Y_m e_m g_{m,m+1} \\ g_{m,m+1} e_m' Y_m & g_{m,m+1}^2 e_m' Y_m e_m \end{bmatrix} \\ & \times \begin{bmatrix} W_m' \\ w_{m+1}'(I - V_m(W_m'V_m)^{-1}W_m') \end{bmatrix}, \end{aligned} \tag{90}$$

where V_m, v_{m+1}, $h_{m+1,m}$ and W_m, w_{m+1}, $g_{m,m+1}$ are defined from the data generated in the course of the two Arnoldi processes associated with $\mathcal{K}_m(A, b)$ and $\mathcal{L}_m(A', c')$ respectively. Furthermore, $\hat{H}_m$ and $\hat{G}_m$ are defined in (17) and (18), respectively. The following theorem solves Problem 2.1 in the context of discrete time systems.

THEOREM 4.1. *Suppose that m steps of the Arnoldi process have been taken and that the residual errors associated with* (87) *and* (88) *are defined by* (89) *and* (90). *Furthermore, suppose that $|\lambda_i(\hat{H}_m)(\bar{\lambda}_j(\hat{H}_m))^{-1}| \neq 1$ for all i, j and $|\lambda_i(\hat{G}_m)(\bar{\lambda}_j(\hat{G}_m))^{-1}| \neq 1$ for all i, j then,*

(a) $W_m' R_m(X_m) W_m = 0$ *if and only if $X_m = X_m^A$, where X_m^A satisfies*

$$\hat{H}_m X_m^A \hat{H}_m' - X_m^A + e_1 \beta^2 e_1' = 0. \tag{91}$$

Under these conditions,

$$\|R_m^A\|_F := \|R_m(X_m^A)\|_F = \left\| \begin{bmatrix} H_m X_m^A H_m' - X_m^A + e_1\beta^2 e_1' & X_m^A e_m h_{m+1,m} \\ h_{m+1,m} e_m' X_m^A & h_{m+1,m}^2 e_m' X_m e_m \end{bmatrix} \right\|_F.$$

(b) $V_m' S_m(Y_m) V_m = 0$ *if and only if $Y_m = Y_m^A$, where Y_m^A satisfies*

$$\hat{G}_m' Y_m^A \hat{G}_m - Y_m^A + e_1 \delta^2 e_1' = 0. \tag{92}$$

Under these conditions,

$$\|S_m^A\|_F := \|S_m(Y_m^A)\|_F = \left\| \begin{bmatrix} G_m' Y_m^A G_m - Y_m^A + e_1\delta^2 e_1' & Y_m^A e_m g_{m,m+1} \\ g_{m,m+1} e_m' Y_m^A & g_{m,m+1}^2 e_m' Y_m e_m \end{bmatrix} \right\|_F.$$

(c) *There exist* rank-1 *perturbations $\Delta_1 = (I - V_m(W_m'V_m)^{-1}W_m')v_{m+1}h_{m+1,m}v_m'$ and $\Delta_2 = w_m\, g_{m,m+1} w_{m+1}'(I - V_m(W_m'V_m)^{-1}W_m')$ such that*

$$(A - \Delta_1)P_m(A - \Delta_1)' - P_m + bb' = 0,$$
$$(A - \Delta_2)'Q_m(A - \Delta_2) - Q_m + c'c = 0,$$

and

$$\|\Delta_1\|_F^2 = h_{m+1,m}^2\{1 + \|(W_m'V_m)^{-1}W_m'v_{m+1}\|_2^2\}$$

and

$$\|\Delta_2\|_F^2 = g_{m,m+1}^2\{1 + \|(V_m'W_m)^{-1}V_m'w_{m+1}\|_2^2\}.$$

Remark 4.1. Observe that $\Delta_3 := \Delta_1 + \Delta_2$ is a perturbation on the data in A such that

$$(A - \Delta_3)P_m(A - \Delta_3)' - P_m + bb' = 0,$$
$$(A - \Delta_3)'Q_m(A - \Delta_3) - Q_m + c'c = 0.$$

Furthermore, Δ_3 is at most a rank-2 perturbation which may be factorised as

$$\Delta_3 = [w_m \;\; (I - V_m(W_m'V_m)^{-1}W_m')v_{m+1}] \begin{bmatrix} 0 & g_{m,m+1} \\ h_{m+1,m} & 0 \end{bmatrix} \times \begin{bmatrix} v_m' \\ w_{m+1}'(I - V_m(W_m'V_m)^{-1}W_m') \end{bmatrix}.$$

Finally, a direct calculation will verify that $W_m'\Delta_1 V_m = W_m'\Delta_2 V_m = W_m'\Delta_3 V_m = 0$.

Note that despite the form of the Lyapunov equations in (87) and (88), the results of Theorem 4.1 are the same as those obtained in §2.1. Theorem 3.2 may be restated in its exact form as presented in §3. The main differences are that V_m and W_m are derived from the two Arnoldi processes associated with $\mathcal{K}_m(A, b)$ and $\mathcal{L}_m(A', c')$, respectively, and where A, b, and c' originate from the discrete time model defined in (85) and (86); furthermore, X_m^A and Y_m^A are the solutions to the low dimensional discrete time Lyapunov equations (91) and (92), respectively. One may apply the Lanczos process to the discrete time model reduction problem in a similar fashion as presented in §3 to obtain a variant of Theorem 3.4 the derivation of which we omit.

5. Numerical experiments. The purpose of this section is to illustrate with the help of two examples the behaviour of the error formulas presented in §§2 and 3. The tests reported here were performed on a Sparc-10 Sun workstation using PRO-MATLAB 4.0 which carries out operations to a unit round-off of 2.22×10^{-16}.

Example 1. The aim of the first example is to illustrate the behaviour of the backward error formulas associated with the Lyapunov equation solvers of §2. This example has been derived from the discretisation of a partial differential equation of the form

$$
\begin{aligned}
(93) \qquad -\Delta u(x,y,t) + \alpha \frac{\partial^2 u(x,y,t)}{\partial x \partial y} + 2\beta e^{-xy}\left(\frac{\partial u(x,y,t)}{\partial x} + \frac{\partial u(x,y,t)}{\partial y}\right) \\
- 2\gamma u(x,y,t) = \frac{\partial u(x,y,t)}{\partial t}, \\
x, y \in \Omega, \\
u(x,y,0) = u_0 \; x, y \in \Omega, \\
u(x,y,t) = \sigma(x,y), \; t \geq 0, \; x, y \in \partial\Omega,
\end{aligned}
$$

where Ω denotes the set of points in the unit square $(0,1) \times (1,0)$ and $\partial\Omega$ denotes the boundary. Setting $\alpha := 3$, $\beta := 0.5$, and $\gamma := 20$, the discretisation is carried out using 102 points in both the x and y directions leading to a linear system of the form

$$
(94) \qquad F(s) = \left\{ \begin{array}{l} \dot{w}(t) = Aw(t) + bg(t), \\ f(t) = cw(t), \end{array} \right.
$$

where A is nonsymmetric sparse matrix of dimension $N = 100^2$ with $nz = 88804$ nonzero elements (i.e., with a density of approximately 0.09%). The vector e is the vector of ones and $b := e * 0.1$, finally, c is a random vector in $\mathbb{R}^{1 \times N}$. Tests with other choices of b and c showed similar results.

Figures 1 and 2 report on the evolutions of $\|\Delta_i\|_F$, $i = 1, 2, 4, 5$ for the Arnoldi and Lanczos–Lyapunov equation solvers. The tests show that $\|\Delta_{4,5}\|_F$ (the lower traces) are significantly smaller than $\|\Delta_{1,2}\|_F$ (the upper traces) when either scheme is employed. Figures 1 and 2 confirm the fact that the a priori backward error bounds are conservative and do not provide an accurate means of gauging the progress of the iterative process. However, the a posteriori bounds indicate that accurate approximate solutions may be obtained for small m. The a posteriori perturbations Δ_4 and Δ_5 may not be the smallest perturbations for which one can compute an mth order approximate model; the problem of determining such perturbations remains open.

Example 2. The aim of this experiment is to test the effectiveness of the model reduction schemes proposed in §3. The problem is set up with $A \in \mathbb{R}^{N \times N}$ where $N = 100$ and the top left-hand 2×2 block of A is set to

$$
\begin{bmatrix} -1 & 100 \\ -100 & -1 \end{bmatrix},
$$

while the remaining nonzero elements of A are uniformly distributed in $[0, -1]$ and are all located on the leading diagonal. Consequently, all the system poles are real except for two that are $-1 \pm 100j$. The first five elements of b and c are uniformly distributed in $[0, 1]$ while the 95 remaining elements are uniformly distributed in $[0, 1/25]$. The infinity norm of $F(s)$ is given by the radius of the smallest circle centered at the origin to enclose the Nyquist plot of Fig. 3; alternatively, it may be computed from $\max |F(j\omega)|$

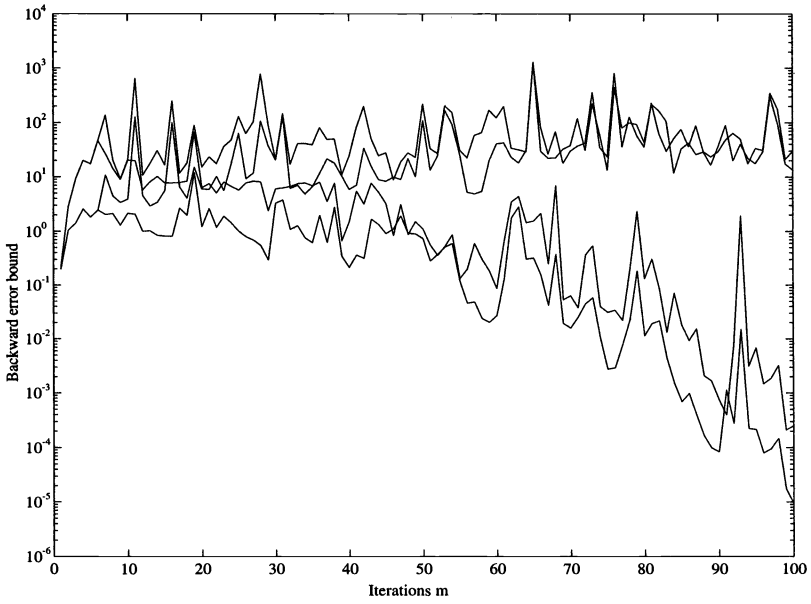


FIG. 1. *Evolutions of* $\|\Delta_{1,2}\|_F$ *(upper traces) and* $\|\Delta_{4,5}\|_F$ *(lower traces) using the Arnoldi-based method in Example* 1.

for all $\omega \in \mathbb{R}$. In this case, $\|F(j\omega)\|_\infty = 3.0716$. Table 1 shows the evolution of the $\mathcal{L}^\infty$ error expression of (66) and (67) denoted here by Err1 and Err2, respectively, for the Arnoldi model reduction algorithm. The table indicates that Err1 and Err2 fall in

TABLE 1
Residual error norms associated with the Arnoldi model reduction scheme in Example 2.

m	Err1	Err2	m	Err1	Err2	m	Err1	Err2
1	1.2998e+1	1.3398e+1	2	1.2912e+0	1.2052e+1	3	5.2256e−1	1.7498e−1
4	8.2001e−1	3.4535e−1	5	4.5508e−1	3.2630e−1	6	2.0198e−1	1.9200e−1
7	1.2076e−1	1.0679e−1	8	7.4857e−2	6.8587e−2	9	5.6592e−2	6.2173e−2
10	4.8653e−2	5.6193e−2	11	5.7017e−2	5.9666e−2	12	5.9939e−2	5.5473e−2
13	7.2936e−2	6.9952e−2	14	7.1812e−2	7.5504e−2	15	9.5773e−2	8.8814e−2

magnitude as m increases. However, as is well known, Galerkin conditions of the type in (63) and (64) do not guarantee a nonincreasing evolution of Err1 and Err2. To maintain orthogonal and biorthogonal bases for $\mathcal{K}_m(A, b)$ and $\mathcal{L}_m(A', c')$, we resorted to reorthogonalisation and rebiorthogonalisation for the Arnoldi and Lanczos processes, respectively. As predicted by Proposition 3.4, the sequence of Lanczos errors, (82) and (83), were the same as Table 1. Figure 3 compares the frequency responses of $F(s)$ and three low order approximate realisations obtained from the Arnoldi model reduction scheme. The low order approximate models have four, six, and eight states, respectively. Each frequency response is performed over $[10^{-3}\text{rads}/s, 10^{3}\text{rads}/s]$ with the lowest frequency point close to the positive real axis and the highest frequency point almost at the origin. Observe that the four frequency responses are indistinguishable over high frequencies with a progressively smaller error over low frequencies

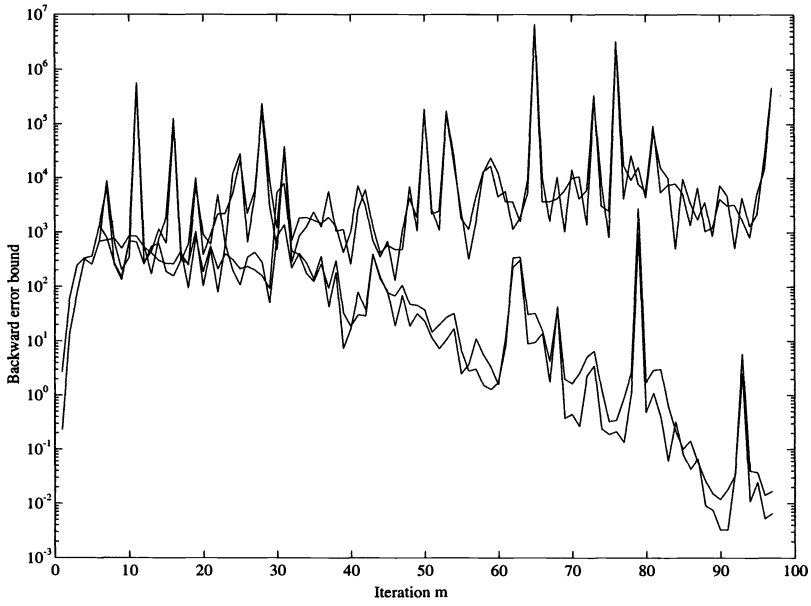


FIG. 2. *Evolutions of* $\|\Delta_{1,2}\|_F$ *(upper traces) and* $\|\Delta_{4,5}\|_F$ *(lower traces) using the Lanczos-based method in Example* 1.

as the state dimension is increased. This behaviour is reflected in the $\mathcal{L}^\infty$ error being significantly smaller than $\|F(s)\|_\infty$ for increasing state dimension as shown in Table 2. Table 2 also lists the $\mathcal{L}^\infty$ forward error associated with the balanced truncation algorithm [13].

The tabulated results indicate that the balanced truncation algorithm enjoys a rapid forward error decay for increasing state dimension while the convergence rate of the Krylov based method is much slower. This superior convergence behaviour is due in part to the large volume of computation associated with solving (5) and (6).

TABLE 2
Forward error $\|F(s) - F_m(s)\|_\infty$ *using the Arnoldi–Lanczos and balanced truncation model reduction schemes.*

m	Krylov	bal/trun	m	Krylov	bal/trun	m	Krylov	bal/trun
1	2.8417e+0	7.4438e−1	5	6.6195e−1	1.4232e−2	9	1.2821e−1	3.9387e−6
2	3.0724e+0	1.4023e+0	6	2.3708e−1	1.5040e−3	10	1.1953e−1	6.6290e−7
3	2.2053e+0	7.4556e−1	7	1.6710e−1	2.1184e−4	11	1.1108e−1	7.8445e−8
4	1.5677e+0	1.1623e−1	8	1.4313e−1	2.2361e−5	12	1.0036e−1	5.7788e−9

6. Conclusions. The aim of this paper has been to present and test several model reduction algorithms suitable for computing low dimensional approximate models of large scale continuous and discrete time systems. By exploiting Krylov subspace methods, approximate solutions to the controllability and observability Lyapunov equations have been found for which the residual errors satisfy Galerkin type conditions. Furthermore, low dimensional expressions have been derived for the residual error norms and a priori and a posteriori backward perturbation norms.

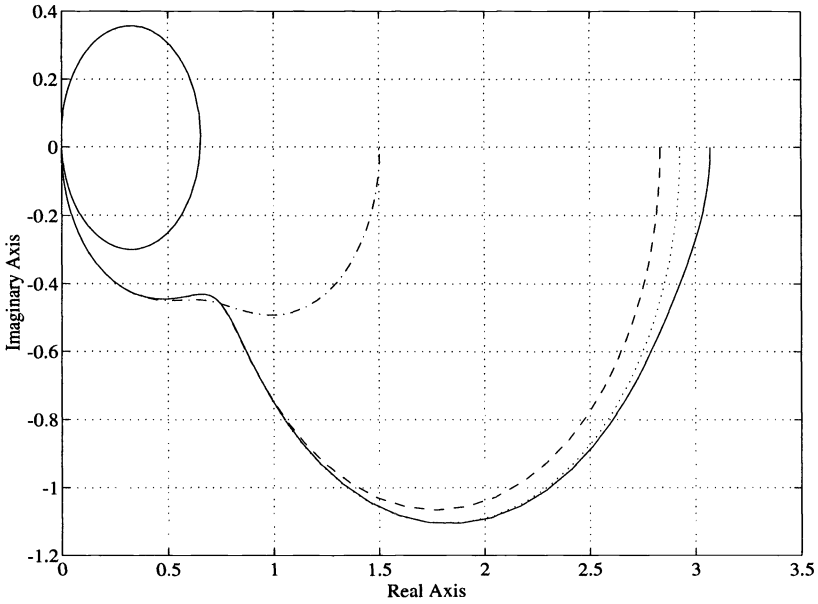


FIG. 3. *Frequency responses for the four models of Arnoldi based approximations, Original* —; *four-state* -.-.-.-; *six-state* - - - -; *and eight-state* .....

We show that the low rank approximate grammians are the exact grammians to a perturbed linear system in which the perturbation has at most rank = 2. We demonstrate that this perturbed linear system is equivalent to a low dimensional linear system with state dimension no greater than m. Furthermore, exact low dimensional expressions for an associated $\mathcal{L}^\infty$ error are presented.

Acknowledgment. The authors wish to thank Professor Paul Van Dooren for helpful discussions and advice on the subject matter of this paper.

REFERENCES

[1] M. M. M. AL-HUSARI, B. HENDEL, I. M. JAIMOUKHA, E. M. KASENALLY, D. J. N. LIMEBEER, AND A. PORTONE, *Vertical stabilisation of Tokamak plasmas*, 30th Conf. Decision and Control, Brighton, England, 1991, pp. 1165–1170.

[2] R. H. BARTELS AND W. STEWART, *Solution of the matrix equation* $AX + XB = C$, Comm. ACM, 15 (1972), pp. 820–826.

[3] D. L. BOLEY, *Krylov space methods on state-space control models*, CS tech. report, TR92-18, University of Minnesota, Minneapolis, 1992.

[4] D. L. BOLEY AND G. H. GOLUB, *The Lanczos-Arnoldi algorithm and controllability*, Systems Control Lett., 4 (1984), pp. 317–327.

[5] ———, *The nonsymmetric Lanczos algorithm and controllability*, Systems Control Lett., 16 (1991), pp. 97–105.

[6] K. GLOVER, *All optimal Hankel norm approximations of linear multivariable systems and their* $\mathcal{L}^\infty$ *error bounds*, Internat. J. Control, 39 (1984), pp. 1115–1193.

[7] G. H. GOLUB, B. KÅGSTRÖM AND P. VAN DOOREN, *Direct block tridiagonalisation of single-input single-output systems*, Systems Control Lett., 18 (1992), pp. 109–120.

[8] E. J. GRIMME, D. C. SORENSEN, AND P. VAN DOOREN, *An implicitly restarted non-symmetric Lanczos algorithm for sparse or unstructured matrices*, Appl. Math. Lett., 1994, pp. 75–80.

[9] N. J. HIGHAM, *Perturbation theory and backward error for $AX - XB = C$*, BIT, 33 (1993), pp. 124–136.

[10] I. M. JAIMOUKHA, E. M KASENALLY, AND D. J. N. LIMEBEER, *Numerical solution of large scale Lyapunov equations using Krylov subspace methods*, 31st Conf. Decision and Control, Tucson, AZ, 1992, pp. 1927–1932.

[11] I. M. JAIMOUKHA AND E. M. KASENALLY, *Krylov subspace methods for solving large Lyapunov equations*, SIAM J. Numer. Anal., 31 (1994), pp. 227–251.

[12] E. M. KASENALLY, *Analysis of some Krylov subspace methods for large matrix equations*, IRC-PSE report, No C93-14, Imperial College, University of London, 1992.

[13] B. C. MOORE, *Principal component analysis in linear systems: controllability, observability and model reduction*, IEEE Trans. Auto. Contr., AC-26 (1981), pp. 17–31.

[14] B. PARLETT, *Reduction to tridiagonal form and minimal realisations*, SIAM J. Math. Anal. Appl., 13 (1992), pp. 567–593.

[15] Y. SAAD, *On the rates of convergence of the Lanczos and block Lanczos methods*, SIAM J. Numer. Anal., 17 (1980), pp. 687–706.

[16] ———, *The Lanczos bi-orthogonalisation algorithm and other oblique projection methods for solving large unsymmetric systems*, SIAM J. Numer. Anal., 19 (1982), pp. 485–506.

[17] ———, *Numerical solution of large Lyapunov equations*, Signal Processing, Scattering, Operator Theory and Numerical Methods, M. A. Kaashoek, J. H. Van Schuppen and A. C. M. Ran, eds., Birkhäuser, Boston, 1990, pp. 503–511.

[18] M. G. SAFONOV AND R. Y. CHIANG, *A Schur method for balanced model reduction*, Proc. American Control Conf., Atlanta, GA, 1988.

[19] P. VAN DOOREN, *Numerical linear algebra techniques for large scale matrix problems in systems and control*, 31st Conf. Decision and Control, Tucson, Arizona, 1992, pp. 1933–1938.

[20] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 628–645, April 1995

019

FAST TRANSFORM BASED PRECONDITIONERS FOR TOEPLITZ EQUATIONS*

E. BOMAN† AND I. KOLTRACHT‡

Abstract. We present a new preconditioner for $n \times n$ symmetric, positive definite Toeplitz systems. This preconditioner is an element of the n–dimensional vector space of matrices that are diagonalized by the discrete sine transform. Conditions are given for which the preconditioner is positive definite and for which the preconditioned system has asymptotically clustered eigenvalues. The diagonal form of the preconditioner can be calculated in $O(n \log(n))$ operations if $n = 2^k - 1$. Thus only n additional parameters need be stored. Moreover, complex arithmetic is not needed. To use the preconditioner effectively, we develop a new technique for computing a fast convolution using the discrete sine transform (also requiring only real arithmetic). The results of numerical experimentation with this preconditioner are presented. Our preconditioner is comparable, and in some cases superior, to the standard circulant preconditioner of Tony Chan. Possible generalizations for other fast transforms are also indicated.

Key words. Toeplitz matrix, conjugate gradient algorithm, preconditioner, fast sine transform

AMS subject classifications. 65F10, 65T20, 65Y20, 65F99

0. Introduction.

The preconditioned conjugate gradient algorithm (PCGA) is traditionally used to solve sparse systems of linear equations; see [6] or [13], for example. In recent years, starting with the proposal of Strang [14], this algorithm has also been used for computations with dense structured matrices, in particular for symmetric, positive definite systems of equations

$$Ax = b, \tag{1}$$

where the coefficient array is Toeplitz (see [2] and [3], for example.) Toeplitz matrices are constant along the diagonals $A = (a_{i-j})_{i,j=0}^{n-1}$. It is assumed that A is Toeplitz throughout this paper.

The successful application of the PCGA to (1) relies on the existence of a good preconditioner; that is, a matrix P such that the following properties are satisfied.

Property 1. The spectrum of $P^{-1}A$ is clustered.

Property 2. P is positive definite.

Property 3. The complexity of computing P is comparable to the complexity of computing Ax.

Property 4. The complexity of solving $Pz = b$, for arbitrary b, is comparable to the complexity of computing Ax.

Since Strang's proposal several investigators have been successful in choosing preconditioners for Toeplitz problems from among the set of circulants. Circulant matrices form a subspace of the vector space of Toeplitz matrices and may be characterized as those matrices that are diagonal in the orthogonal basis defined by the columns of the discrete Fourier transform (DFT); see [4]. Thus if P is circulant and F is the DFT then there is a diagonal matrix Λ such that $P = F\Lambda F^*$. Since Λ can

* Received by the editors August 23, 1993; accepted for publication (in revised form) by R. J. Plemmons, March 29, 1994.

† United Technologies Research Center, East Hartford, Connecticut 06119 (boman@math.uconn.edu).

‡ University of Connecticut, U-9, Storrs, Connecticut 06269 (koltrach@uconnvm.edu) This author's work was supported in part by National Science Foundation grants DMS-9007030 and DMS-9306357.

be computed in $O(n\log(n))$ operations (see [4]), this factorization allows one to solve $Pz = b$ for arbitrary $b \in R^n$ in $O(n\log(n))$ operations provided $n = 2^k$ for some integer k. Since the computation of a matrix–vector product where the matrix is Toeplitz can be performed in at most $O(n\log(n))$ flops one arrives at an $O(n\log(n))$ algorithm for solving Toeplitz systems, provided that the spectrum of $P^{-1}A$ is clustered as explained in the next section. This is faster than more traditional $O(n^2)$ algorithms like Levinson's algorithm ([10]), for example.

Strang built a circulant preconditioner for Toeplitz problems in [14] by copying the central diagonals of A around to complete the circulant. That is, if $A = \left(a_{|i-j|}\right)_{i\,j=1}^{n}$ then Strang's preconditioner is given by

$$p_{i\,j} = \begin{cases} a_{|i-j|} & \text{if } |i-j| \le \frac{n}{2}, \\ a_{n-|i-j|} & \text{if } |i-j| > \frac{n}{2}. \end{cases}$$

In [3], Tony Chan built a circulant preconditioner by taking

$$P = \underset{M\,\text{circulant}}{\operatorname{argmin}} \|M - A\|_F,$$

where F denotes the Frobenius norm, and in [15] Tyrtyshnikov built a circulant preconditioner by taking

$$P^{-1} = \underset{M\,\text{circulant}}{\operatorname{argmin}} \|I - MA\|_F.$$

In [1], Raymond Chan showed that as n grows, all of the above preconditioners give similar asymptotic clustering of the spectrum of $P^{-1}A$ when A is a finite section of a singly infinite Toeplitz matrix associated with a Weiner class function.

In this paper we study preconditioners based on the fast sine transform (FST), called S_1-diagonal preconditioners, and compare them with Tony Chan's circulant preconditioner on some common Toeplitz problems. We also indicate generalizations to other fast transforms. S_1-diagonal preconditioners can be implemented quickly and do not require complex arithmetic (as do the circulant preconditioners listed above). They are banded when the Toeplitz matrix is banded and, in this case, perform better than the circulant preconditioner in our numerical experiments.

This paper is organized as follows. Section 1 is a brief outline of the PCGA. Section 2 develops the underlying theory that supports the generation of new preconditioners for Toeplitz equations, based on a given fast transform, and shows how the circulant matrices fit into this broader scheme.

In §3 we show how to apply a symmetric Toeplitz matrix to an arbitrary vector by embedding the matrix into an S_1-diagonal matrix.

Section 4 develops a new preconditioner for banded, symmetric Toeplitz systems. This new preconditioner is based on the discrete sine transform in the same way that circulants are based on the DFT. When A is banded, the spectrum of the product $P^{-1}A$ will not only be clustered, but all eigenvalues are equal to 1 except for a few outliers. The number of the outlying eigenvalues depends only on the bandwidth of A.

Under certain conditions this preconditioner also works for full Toeplitz systems. Asymptotic clustering of the eigenvalues of $P^{-1}A$ when A is full is shown in §4.2.

The results of numerical experimentation with the new preconditioners are presented in §5, and the Appendix indicates how preconditioners based on some other fast transforms may be generated.

1. The preconditioned conjugate gradient algorithm (PCGA). The PCGA for solving (1), as given in [6], follows here.

ALGORITHM 1. Set $x_0 = 0$, and $r_0 = b$. Then for $k = 1$, to convergence repeat the following.

(i) If $r_{k-1} = 0$ set $x = x_{k-1}$ and stop.

(ii) Otherwise,

1. Solve $Pz_{k-1} = r_{k-1}$ (This is the preconditioning step.)
2. Set $\beta_k = z_{k-1}^T r_{k-1} / z_{k-2}^T r_{k-2}, \quad \beta_1 \equiv 0$
3. Set $p_k = z_{k-1} + \beta_k p_{k-1}, \quad p_1 \equiv z_0$
4. Set $\alpha_k = z_{k-1}^T r_{k-1} / p_k^T A p_k$
5. Set $x_k = x_{k-1} + \alpha_k p_k$
6. Set $r_k = r_{k-1} - \alpha_k A p_k$

Applying the PCGA to (1) is equivalent to applying the conjugate gradient algorithm to $\tilde{A}\tilde{x} = \tilde{b}$, where $\tilde{A} = P^{-1/2} A P^{-1/2}$, $\tilde{x} = P^{1/2} x$, and $\tilde{b} = P^{-1/2} b$ (see [6]). The important feature of this algorithm is that if it is performed in exact arithmetic, it will converge to the correct solution in $k \leq n$ iterations where k is the number of distinct eigenvalues of $\tilde{A}$. This is because after the jth iteration of the PCGA,

$$\left(\tilde{A}x_j - \tilde{b}\right)^T \tilde{A}^{-1} \left(\tilde{A}x_j - \tilde{b}\right),$$

has minimal norm over all vectors spanned by the Krylov vectors: $\left\{\tilde{b}, \tilde{A}\tilde{b}, \tilde{A}^2\tilde{b}, \ldots, \tilde{A}^j\tilde{b}\right\}$ (see [12]).

Clearly the choice of the preconditioner is all important. A preconditioner that reduces the number of distinct eigenvalues of $P^{-1/2} A P^{-1/2}$ also reduces the number of iterations required for convergence. We remark that the spectra of $P^{-1/2} A P^{-1/2}$ and $P^{-1}A$ are equal since the two matrices are similar. We will work with the latter matrix hereafter.

In practice it is rarely possible to actually reduce the number of distinct eigenvalues. However, Jennings [8] shows that if the eigenvalues of $P^{-1}A$ are "clustered" — if the eigenvalues occur in $q < n$ clusters — the convergence characteristics of the PCGA are almost as good as if there were only q distinct eigenvalues.

2. T−diagonal matrices.

DEFINITION 1. *Let T be an arbitrary, nonsingular matrix. A matrix M is T−diagonal if $T^{-1}MT$ is diagonal.*

DEFINITION 2. *For a given T, denote by D_T the n-dimensional vector space of all T−diagonal matrices.*

Much of the recent work on preconditioners for Toeplitz systems is focused on circulant matrices that are F−diagonal, where F denotes the DFT,

$$F = \frac{1}{\sqrt{n}} \left(\exp\left(\frac{jk2\pi i}{n} \right) \right)_{j,k=0}^{n-1}.$$

The key to the use of circulant matrices as preconditioners for Toeplitz systems of equations is the following well-known property (see [4]). If C is circulant and c_1 is the first row of C, then $\sqrt{n} c_1 F$ is a row-vector whose elements are the eigenvalues of C.

This is a special case of the next proposition. It shows that if T and $M \in D_T$ are given, it is not necessary to compute $T^{-1}MT$ to produce the diagonal form (eigenvalues) of M.

DEFINITION 3. *Let* $x = (x_1, \ldots, x_n)$. *Then* $\Delta(x) \equiv \operatorname{diag}\{x_1, \ldots, x_n\}$.

Thus if $x_i \neq 0$, $i = 1, \ldots, n$, then $[\Delta(x)]^{-1} \equiv \operatorname{diag}\left\{\frac{1}{x_1}, \ldots, \frac{1}{x_n}\right\}$.

Some care must be taken here. We will need to refer to Δ^{-1}, which is the inverse of Δ as an operator *not* the inverse of $\operatorname{diag}\{x_1, \ldots, x_n\}$, thus $\Delta^{-1}(\operatorname{diag}\{x_1, \ldots, x_n\}) = (x_1, \ldots, x_n)$.

PROPOSITION 2.1. *Let* T *be a nonsingular matrix that has at least one row* τ_k, *which is everywhere nonzero. Let* μ_k *be the corresponding row of* M. *Then*

$$T^{-1}MT = \Delta\left(\mu_k T\left[\Delta(\tau_k)\right]^{-1}\right).$$

Proof. Since $M \in D_T$ there is a diagonal matrix, Λ such that $MT = T\Lambda$. Thus

$$\mu_k T = \tau_k \Lambda = \Delta^{-1}(\Lambda)\,\Delta(\tau_k).$$

Therefore $\Delta\left(\mu_k T\left[\Delta(\tau_k)\right]^{-1}\right) = \Lambda$. □

Observe that the computation of Λ is dominated by the product $\mu_k T$ that generally is $O(n^2)$. If T admits a fast calculation, for example if T is the DFT and $n = 2^k$, then $\mu_k T$ can be computed in $O(n\log(n))$ operations.

If M is circulant and T is the DFT, then take $k = 1$ and observe that $\tau_1 = \frac{1}{\sqrt{n}}(1, \ldots, 1)$ to recover the previously mentioned property of circulant matrices.

2.1. The diagonal space of the discrete sine transform, (DST1). Let S_1 denote the first discrete sine transform (there are at least two; see [16])

$$S_1 = \sqrt{\frac{2}{n+1}}\left(\sin\left(\frac{ij\pi}{n+1}\right)\right)_{i,j=1}^{n}.$$

In this section a basis for D_{S_1} is explicitly displayed.

It is well known that $S_1^2 = I$, $S_1 = S_1^T$ and that S_1 can be applied to a vector in $O(n\log(n))$ flops as long as $n = 2^k - 1$ for some integer k. This will be assumed hereafter. For a detailed exposition of the properties of the DST1, see [16].

Recall that the $n-$dimensional vector space of $n \times n$ circulant matrices is spanned by the set $\{C_p\}_{p=0}^{n-1}$, where C_1 is defined by the following relation. If $x = (x_1, \ldots, x_n)^T$ then $C_1 x \equiv (x_n, x_1, \ldots, x_{n-1})^T$ and $C_p = C_1^p$. Similarly, a basis for D_{S_1} is given by choosing ζ so that

$$\zeta(i,j) = \begin{cases} 1 & \text{if } |i-j| = 1, \\ 0 & \text{otherwise,} \end{cases}$$

(see ζ_1 displayed below). A basis is then given by $\{I, \zeta^1, \ldots, \zeta^{n-1}\}$ just like the circulant basis. Since nothing that follows depends on this basis we simply remark that it exists.

For our purposes we have found it more convenient to use the basis $\{\zeta_p\}_{p=0}^{n-1}$ where

$$\zeta_p(i,j) = \begin{cases} 1 & \text{if } |i-j| = p, \\ -1 & \text{if } i+j = p, \\ -1 & \text{if } i+j = 2(n+1)-p, \\ 0 & \text{otherwise.} \end{cases}$$

To display the structure of the ζ_p's more clearly, we now display the complete basis for the 5×5 case. Of course, ζ_0 is the identity. The other four basis matrices are given below.

$$\zeta_1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \qquad \zeta_2 = \begin{pmatrix} -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 \end{pmatrix},$$

$$\zeta_3 = \begin{pmatrix} 0 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 & 0 \end{pmatrix}, \qquad \zeta_4 = \begin{pmatrix} 0 & 0 & -1 & 0 & 1 \\ 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 \\ 1 & 0 & -1 & 0 & 0 \end{pmatrix}.$$

That this is in fact a basis is the subject of the next lemma. That it is more convenient is taken up in §4. The dimension of ζ_p will always be clear from context.

LEMMA 2.2. *Let $n \in Z^+$ be given. Then $\{\zeta_p\}_{p=0}^{n-1}$ is a basis for D_{S_1}. Moreover if $p > 0$ then the spectrum of ζ_p is*

$$\left\{ 2\cos\left(\frac{p\pi k}{n+1}\right) \right\}_{k=1}^{n}.$$

Proof. Clearly $\{\zeta_p\}_{p=0}^{n-1}$ is a linearly independent set and it is large enough to span D_{S_1}. All that remains is to show that for all $p = 0, \ldots, n-1$, ζ_p is in D_{S_1} and that the spectrum is as given. Denote by s_k the kth column of S_1. Calculating the components of $\zeta_p s_k = (\nu_1, \ldots, \nu_n)^T$ directly will show that every column of S_1 is an eigenvector of ζ_p and, incidentally, gives the spectrum of ζ_p, which will complete the proof.

There are five cases: (i) $i < p$, (ii) $i = p$, (iii) $p < i < n-p+1$, (iv) $i = n-p+1$, and (v) $i > n-p+1$. If $i < p$ then

$$\nu_i = \sqrt{\frac{2}{n+1}} \left[-\sin\left(\frac{(p-i)\pi k}{n+1}\right) + \sin\left(\frac{(p+i)\pi k}{n+1}\right) \right],$$

which by an elementary trigonometric identity is

$$\nu_i = 2\cos\left(\frac{p\pi k}{n+1}\right) \sqrt{\frac{2}{n+1}} \sin\left(\frac{i\pi k}{n+1}\right).$$

The other four cases are similar. □

Remark. D_{S_1} is clearly a subspace of the space of Toeplitz plus Hankel matrices. It may prove useful to precondition such matrices with a preconditioner drawn from D_{S_1}, but we have not done so here.

3. Efficient matrix–vector multiplication. Before proceeding to develop our preconditioner, we first indicate how the diagonal space D_{S_1} may be used to apply a symmetric Toeplitz matrix $A \in R^{n \times n}$ to a vector $x \in R^n$ in $O(n \log(n))$ operations. Step 4 of the PCGA requires this product, which is commonly performed by embedding A into a larger circulant matrix, C, and x into a a larger vector, $\bar{x}$. Applying C

to $\bar{x}$ requires only $O(n\log(n))$ operations and the desired product Ax emerges from the nature of the embeddings. The difficulty of this method is that the efficient application of C to $\bar{x}$ requires the use of the DFT thus requiring complex arithmetic. But if A and x are both real their product will also be real and we would like to avoid the added computational and storage burden of using complex arithmetic to compute real results.

Our method is similar to the method just outlined in that we embed the Toeplitz matrix in an S_1–diagonal matrix. No complex arithmetic is required at any step.

The method is best demonstrated by example. Let

$$A = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 \\ a_1 & a_0 & a_1 & a_2 \\ a_2 & a_1 & a_0 & a_1 \\ a_3 & a_2 & a_1 & a_0 \end{pmatrix} \quad \text{and} \quad x = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

We seek a matrix $\hat{A} \in D_{S_1}$ such that A is a submatrix of $\hat{A}$. Clearly,

$$\hat{A} = \begin{pmatrix} (a_0 - a_2) & (a_1 - a_3) & a_2 & a_3 & 0 & 0 \\ (a_1 - a_3) & \boldsymbol{a_0} & \boldsymbol{a_1} & \boldsymbol{a_2} & \boldsymbol{a_3} & 0 \\ a_2 & \boldsymbol{a_1} & \boldsymbol{a_0} & \boldsymbol{a_1} & \boldsymbol{a_2} & a_3 \\ a_3 & \boldsymbol{a_2} & \boldsymbol{a_1} & \boldsymbol{a_0} & \boldsymbol{a_1} & a_2 \\ 0 & \boldsymbol{a_3} & \boldsymbol{a_2} & \boldsymbol{a_1} & \boldsymbol{a_0} & (a_1 - a_3) \\ 0 & 0 & a_3 & a_2 & (a_1 - a_3) & (a_0 - a_2) \end{pmatrix}$$

is such a matrix since $\hat{A} = \sum_{i=0}^{3} a_i \zeta_i$. The embedding of A is indicated with boldface.

Similarly, embed x in $\hat{x}$

$$\hat{x} = \begin{pmatrix} 0 \\ x_0 \\ x_1 \\ x_2 \\ x_3 \\ 0 \end{pmatrix}, \quad \text{so that} \quad \hat{A}\hat{x} = \begin{pmatrix} * \\ Ax \\ * \end{pmatrix}.$$

Since by construction $\hat{A}$ is in D_{S_1}, there is a diagonal matrix Λ such that $\hat{A} = S_1 \Lambda S_1$. Thus the product $\hat{A}\hat{x}$ can be computed in $O(n\log(n))$ operations once Λ is known. Proposition 2.1 gives an algorithm for computing Λ in $O(n\log(n))$ operations. Thus the entire computation is again $O(n\log(n))$.

In general an $n \times n$ real Toeplitz matrix A can be embedded in the S_1–diagonal matrix,

$$\hat{A} = \begin{pmatrix} A_1 - H_1 & A_2 - H_2 & 0 \\ (A_2 - H_2)^T & A & J(A_2 - H_2)^T J \\ 0 & J(A_2 - H_2)J & J(A_1 - H_1)J \end{pmatrix},$$

where

$$J \equiv \begin{pmatrix} 0 & \dots & 1 \\ \vdots & \iddots & \vdots \\ 1 & \dots & 0 \end{pmatrix}$$

and A_1, H_1, A_2, and H_2 are defined as follows. Take $k = (n-3)/2$ if n is odd and $k = (n-4)/2$ if n is even. Then

$$A_1 = \begin{pmatrix} a_0 & \cdots & a_k \\ \vdots & \ddots & \vdots \\ a_k & \cdots & a_0 \end{pmatrix}, \qquad H_1 = \begin{pmatrix} a_2 & \cdots & a_{k+2} \\ \vdots & \iddots & \vdots \\ a_{k+2} & \cdots & a_{n-1} \end{pmatrix},$$

$$A_2 = \begin{pmatrix} a_{k+1} & \cdots & a_{n-1} & 0 & \cdots & 0 \\ \vdots & \ddots & & \ddots & \ddots & \vdots \\ a_1 & \cdots & a_{k+1} & \cdots & a_{n-1} & 0 \end{pmatrix},$$

and

$$H_2 = \begin{pmatrix} a_{k+3} & \cdots & a_{n-1} & 0 & \cdots & 0 \\ \vdots & \iddots & \iddots & & \iddots & \vdots \\ a_{n-1} & 0 & & \iddots & & 0 \\ 0 & \cdots & 0 & \cdots & \cdots & 0 \end{pmatrix},$$

Finally, it is not necessary to compute all of $\hat{A}$. By Proposition 2.1 all that is required to calculate the diagonal form (eigenvalues) of $\hat{A}$ is its first row.

4. S_1–diagonal preconditioners.

4.1. Banded matrices. The goal of this section is to construct an S_1–diagonal preconditioner, P_b, which approximates the banded Toeplitz matrix,

$$A = \begin{pmatrix} a_0 & a_1 & \dots & a_b & & & 0 \\ a_1 & a_0 & \dots & & a_b & & \\ \vdots & & \cdot & & & \ddots & \\ a_b & & & \cdot & & a_b & \\ & a_b & & & \cdot & & a_b \\ & & \ddots & & & \cdot & \vdots \\ & & & a_b & \dots & a_0 & a_1 \\ 0 & & & a_b & \dots & a_1 & a_0 \end{pmatrix} \in \mathcal{R}^{n\times n}$$

with a banded matrix from D_{S_1} and to give conditions on A for which it meets the criteria for a good preconditioner as set forth in the first section.

Recall that $\zeta_p = T_p + H_p$, where T_p is Toeplitz and H_p is Hankel. Thus any element of D_{S_1} will be Toeplitz plus Hankel. We wish to choose P_b in D_{S_1} such that its Toeplitz part is A. In view of Lemma 2.2 it is clear how the preconditioner is to be constructed. Clearly,

$$P_b(n) \equiv P_b = \sum_{i=0}^{b} a_i \zeta_i \tag{2}$$

is the desired preconditioner. This is why the basis $\{\zeta_p\}_{p=0}^{n-1}$ is convenient, the Toeplitz

part of the preconditioner is exactly A. An example is instructive. Let

$$A = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 & 0 & 0 & 0 & 0 & 0 \\ a_1 & a_0 & a_1 & a_2 & a_3 & 0 & 0 & 0 & 0 \\ a_2 & a_1 & a_0 & a_1 & a_2 & a_3 & 0 & 0 & 0 \\ a_3 & a_2 & a_1 & a_0 & a_1 & a_2 & a_3 & 0 & 0 \\ 0 & a_3 & a_2 & a_1 & a_0 & a_1 & a_2 & a_3 & 0 \\ 0 & 0 & a_3 & a_2 & a_1 & a_0 & a_1 & a_2 & a_3 \\ 0 & 0 & 0 & a_3 & a_2 & a_1 & a_0 & a_1 & a_2 \\ 0 & 0 & 0 & 0 & a_3 & a_2 & a_1 & a_0 & a_1 \\ 0 & 0 & 0 & 0 & 0 & a_3 & a_2 & a_1 & a_0 \end{pmatrix}.$$

Then $b = 3$ and P_3 is given by

$$\begin{aligned} P_3 &= a_0 I + a_1\zeta_1 + a_2\zeta_2 + a_3\zeta_3 \\ &= A - \begin{pmatrix} a_2 & a_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ a_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_3 & a_2 \end{pmatrix}. \end{aligned}$$

In general, as in the above example, it is clear that $A = P_b + H$ where, in block form, H is

$$H = \begin{pmatrix} G & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \hat{G} \end{pmatrix}, \tag{3}$$

and G and $\hat{G}$ are given by

$$G = \begin{pmatrix} a_2 & a_3 & \dots & a_b \\ a_3 & & \iddots & 0 \\ \vdots & \iddots & \iddots & \vdots \\ a_b & 0 & \dots & 0 \end{pmatrix}, \qquad \hat{G} = \begin{pmatrix} 0 & \dots & 0 & a_b \\ \vdots & \iddots & \iddots & \vdots \\ 0 & \iddots & & a_3 \\ a_b & \dots & a_3 & a_2 \end{pmatrix}.$$

Note that the blocks in (3) are not necessarily the same size. In particular, G and $\hat{G}$ are $(b-1) \times (b-1)$ and the central block of zeros is $(n - 2(b-1)) \times (n - 2(b-1))$.

First observe that P_n satisfies Property 1. Since H has at most $2(b-1)$ nonempty columns $\operatorname{rank}(H) \leq 2(b-1)$. Thus if P_b is nonsingular, which we address below, then $P_b^{-1}A = I + P_b^{-1}H$ and all but $2(b-1)$ of the eigenvalues are equal to 1. Thus $P^{-1}A$ has at most $2b-1$ distinct eigenvalues. Therefore P_b satisfies Property 1 if $b \ll n$, which is true for instance when n is increasing while b remains fixed. An identical argument and conclusion are possible for Strang's circulant preconditioner. Indeed, our preconditioner is the S_1–diagonal analogue of Strang's preconditioner.

Regarding Property 4, notice that the complexity of computing Ax is $O(bn)$, while solving $P_n z_{k-1} = r_{k-1}$ is $O(n\log(n))$. (Recall that this is the preconditioning step of

the PCGA.) Although asymptotically solving $P_n z_{k-1} = r_{k-1}$ is slower than computing Ax, this does not appear to be a drawback for the problems we consider in the framework of this paper. In fact, in all of our numerical work we have $bn \geq n\log(n)$. Note also that when A is nonbanded (see §4.2), Ax is computed in $O(n\log(n))$ operations so that in this case Property 4 is satisfied. Moreover, as seen in §3 Ax can be computed using only real arithmetic.

Property 3 is satisfied in the banded case by computing the diagonal form of P_b via (2). Lemma 2.2 gives the spectrum of each of the ζ_p's so that $\zeta_p = S_1 \Lambda_p S_1$ for some known Λ_p. Thus the diagonal form of the preconditioner is given by $S_1 P_b S_1 = \sum_{i=0}^{b} a_i \Lambda_i$. This formula requires $O(bn)$ operations to compute.

The following method is $O(n\log(n))$ regardless of the bandwidth of A. Let Z be the shift operator, i.e., $(x_1, \ldots, x_n) Z \equiv (x_2, \ldots, x_n, 0)$. In view of Proposition 2.1 only the first row of P_b is required to calculate the spectrum of P_b. Denoting by ρ and τ the first rows of P_b and A, ρ is computed as follows. From $P_b = \sum_{i=0}^{b} a_i \zeta_i$,

$$\rho = a_0 \begin{pmatrix} 1 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}^T + a_1 \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}^T + a_2 \begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}^T + \cdots + a_b \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -1 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}^T = \tau\left(I - Z^2\right).$$

Thus ρ may be computed in $O(n)$ operations via the formula, $\rho = \tau\left(I - Z^2\right)$. Applying Proposition 2.1 proves the following lemma.

LEMMA 4.1. *Let $\Lambda = S_1 P_b S_1$ be the diagonal form of P_b and let σ denote the first row of S_1. Then*

$$\Lambda = \Delta\left(\tau\left(I - Z^2\right) S_1 \left[\Delta(\sigma)\right]^{-1}\right). \tag{4}$$

Proof. From Proposition 2.1 if P_b is an element of D_{S_1}, then

$$S_1 P_b S_1 = \Delta\left(\rho S_1 \left[\Delta(\sigma)\right]^{-1}\right).$$

(Recall that $S_1 = S_1^{-1}$.) Observing that $\rho = \tau\left(I - Z^2\right)$ completes the proof. □

It is now clear that the preconditioner $P_b(n)$ given in (2) satisfies Properties 1, 3, and 4 given in the Introduction.

Only Property 2 remains to be shown. The next theorem gives conditions on A for which P_b is positive definite.

THEOREM 4.2. *If A is an $n \times n$ section of an infinite, positive definite, Toeplitz matrix, $A(\infty)$ with bandwidth $2b+1 < n$ then P_b, is also positive definite.*

Proof. Since $A(\infty)$ is positive definite, $A(\infty) = U(\infty)U(\infty)^T$ where $U(\infty)$ is an upper triangular Toeplitz matrix (see e.g., [5, proof of Thm. 1.1]),

$$U(\infty) = \begin{pmatrix} c_0 & c_1 & \ldots & c_b & 0 & \cdots & \cdots \\ 0 & c_0 & \ldots & c_{b-1} & c_b & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

and hence

$$A = UU^T, \tag{5}$$

where

$$U = \begin{pmatrix} c_0 & c_1 & \cdots & c_b & 0 & \cdots & \cdots & 0 \\ 0 & c_0 & \cdots & c_{b-1} & c_b & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & c_0 & c_1 & \cdots & c_b \end{pmatrix} \in R^{n\times(n+b)}.$$

Remark. All of the zeros of the polynomial $c_0 z^b + \cdots + c_b$ are in the open unit disk (see [5]).

For our purposes it is more convenient to write U in the following block form:

$$U = \begin{pmatrix} \xi & U_1 & 0 \\ 0 & U_2 & 0 \\ 0 & U_3 & \beta \end{pmatrix},$$

where

$$\xi = \begin{pmatrix} c_0 & \cdots & c_{b-2} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & c_0 \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} c_b & \cdots & 0 \\ \vdots & \ddots & \vdots \\ c_2 & \cdots & c_b \end{pmatrix}.$$

Let $A[a:b, c:d]$ be the submatrix of A consisting of the intersection of rows a through b and columns c through d.

As before,

$$P_b = A - H \text{ and } H = \begin{pmatrix} G & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \hat{G} \end{pmatrix}.$$

Note that since $2b < n+1$,

$$A(b+1:2b-1, 1:b-1) = \begin{pmatrix} a_b & a_{b-1} & \cdots & a_2 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & a_{b-1} \\ 0 & \cdots & 0 & a_b \end{pmatrix} = GJ.$$

Recall from §2.1 that J is the anti-identity. From the block representation above it is clear that $\xi\beta^T = A(b+1:2b-1, 1:b-1)$ so that $G = \xi\beta^T J$. Similarly $\hat{G} = \beta\xi^T J$.

Observe next that $P_b = \hat{U}\hat{L}$ where

$$\hat{U} = \begin{pmatrix} I & U_1 & 0 \\ 0 & U_2 & 0 \\ 0 & U_3 & I \end{pmatrix} \quad \text{and} \quad \hat{L} = \begin{pmatrix} (\xi\xi^T - \xi\beta^T J) & 0 & 0 \\ U_1^T & U_2^T & U_3^T \\ 0 & 0 & (\beta\beta^T - \beta\xi^T J) \end{pmatrix},$$

as may be seen by mulitplying $\hat{U}$ and $\hat{L}$ directly and comparing the result with $UU^T = A$.

With these definitions in place the theorem can now be proved. Since A and H are both symmetric and $P_b = A - H$, we need only show that all of the eigenvalues of P_b are positive.

Suppose λ is an eigenvalue of P_b with corresponding nonzero eigenvector $x = (x_1, \ldots, x_n)^T$. Then at least one of $y \equiv x \pm Jx$ is also a nonzero eigenvector of P_b with the same eigenvalue. (This follows from the identity $JP_b = P_b J$.) Therefore, we may asssume without loss of generality that y is an eigenvector corresponding to λ such that $Jy = \pm y$ and $\|y\| = 1$.

Partition y as

$$y = \begin{pmatrix} y_1 \\ \hat{y} \\ y_2 \end{pmatrix},$$

where y_1 and y_2 are of length $b-1$. Thus $y_1 = \pm Jy_2$. Moreover ξy_1 and βy_2 are defined.

Next observe that

$$y^T \hat{U} = \left(y_1^T, z^T, y_2^T\right),$$

$$\hat{L}y = \begin{pmatrix} (\xi\xi^T - \xi\beta^T J)y_1 \\ z \\ (\beta\beta^T - \beta\xi^T J)y_2 \end{pmatrix},$$

and that

$$y^T U = \left(y_1^T \xi, z^T, y_2^T \beta\right),$$

$$U^T y = \begin{pmatrix} \xi^T y_1 \\ z \\ \beta^T y_2 \end{pmatrix}. \tag{6}$$

Thus

$$\begin{aligned}
\lambda &= y^T P_b y \\
&= \begin{pmatrix} y_1^T, & \hat{y}^T, & y_2^T \end{pmatrix} \hat{U}\hat{L} \begin{pmatrix} y_1 \\ \hat{y} \\ y_2 \end{pmatrix} \\
&= \begin{pmatrix} y_1^T, & z^T, & y_2^T \end{pmatrix} \begin{pmatrix} (\xi\xi^T - \xi\beta^T J)y_1 \\ z \\ (\beta\beta^T - \beta\xi^T J)y_2 \end{pmatrix} \\
&= y_1^T(\xi\xi^T - \xi\beta^T J)y_1 + y_2^T(\beta\beta^T - \beta\xi^T J)y_2 + z^T z \\
&= y_1^T\xi\xi^T y_1 - y_1^T\xi\beta^T J y_1 + y_2^T\beta\beta^T y_2 - y_2^T\beta\xi^T J y_2 + z^T z \\
&= y_1^T\xi\xi^T y_1 - y_1^T\xi\beta^T J y_1 + y_1^T J\beta\beta^T J y_1 - y_1^T J\beta\xi^T y_1 + z^T z \\
&= y_1^T(\xi - J\beta)(\xi^T - \beta^T J)y_1 + z^T z \\
&= \left\|(\xi^T - \beta^T J)y_1\right\|_2^2 + \|z\|_2^2 \\
&\geq 0.
\end{aligned}$$

We now show that this last inequality is actually strict since $z \neq 0$. The proof is by contradiction. Suppose $z = 0$.

Defining P and Q as

$$P = \begin{pmatrix} c_0 & & 0 \\ \vdots & \ddots & \\ c_b & \cdots & c_0 \end{pmatrix}, \qquad Q = \begin{pmatrix} c_b & & 0 \\ \vdots & \ddots & \\ c_0 & \cdots & c_b \end{pmatrix},$$

and repartitioning y as

$$y = \begin{pmatrix} \omega_1 \\ v \\ \omega_2 \end{pmatrix},$$

where $P\omega_1$ is defined and $\omega_1 = \pm J\omega_2$, we can write

$$Ay = UU^T y = \begin{pmatrix} P^T P\omega_1 \\ 0 \\ QQ^T\omega_2 \end{pmatrix}. \tag{7}$$

Observe that neither c_0 nor c_b are equal to zero since $c_0 c_b = a_b$ and by assumption $a_b \neq 0$. Therefore both $P^T P$ and QQ^T are nonsingular. Moreover since $JA = AJ$ it follows that $P^T P\omega_1 = \pm JQQ^T\omega_2$. In the case of a negative sign, we have $\omega_1^T P^T P\omega_1 = -\omega_2^T QQ^T\omega_2$. The left-hand side of this equality is nonnegative and the right-hand side is nonpositive. Thus $\omega_1 = \omega_2 = 0$, which implies by (7) that $y = 0$ as well. This contradicts the assumption that y is a nonzero eigenvector of A.

In the case of a positive sign, write

$$JP^T PJ\omega_2 = QQ^T\omega_2.$$

Since $JP^T PJ = JP^T JJPJ = PP^T$, we have

$$\begin{pmatrix} c_0 & & 0 \\ \vdots & \ddots & \\ c_b & \cdots & c_0 \end{pmatrix}\begin{pmatrix} c_0 & \cdots & c_b \\ & \ddots & \vdots \\ 0 & & c_0 \end{pmatrix}\omega_2 = \begin{pmatrix} c_b & & 0 \\ \vdots & \ddots & \\ c_0 & \cdots & c_b \end{pmatrix}\begin{pmatrix} c_b & \cdots & c_0 \\ & \ddots & \vdots \\ 0 & & c_b \end{pmatrix}\omega_2. \tag{8}$$

Since $z = 0$ we can rewrite (6) in more detail as

$$\begin{pmatrix} c_0 & & & & 0 \\ c_1 & \cdot & & & \\ \vdots & \cdot & \cdot & & \\ c_b & \cdot & \cdot & \cdot & \\ & \cdot & \cdot & \cdot & c_0 \\ & & \cdot & \cdot & c_1 \\ & & & \cdot & \vdots \\ 0 & & & & c_b \end{pmatrix}\begin{pmatrix} \omega_1 \\ v \\ \omega_2 \end{pmatrix} = \begin{pmatrix} \phi_0 \\ \vdots \\ \phi_{b-2} \\ 0 \\ \vdots \\ 0 \\ \psi_0 \\ \vdots \\ \psi_{b-2} \end{pmatrix}.$$

Inspecting row $b+1$ counting from below, we see that

$$(c_b, \ldots, c_0)\omega_2 = 0. \tag{9}$$

In combination with (8), this implies that

$$\begin{pmatrix} c_0 & & 0 \\ \vdots & \ddots & \\ c_b & \cdots & c_0 \end{pmatrix} \begin{pmatrix} c_0 & \cdots & c_b \\ & \ddots & \vdots \\ 0 & & c_0 \end{pmatrix} \omega_2$$
$$= \begin{pmatrix} 0 & & & 0 \\ 0 & c_b & & \\ \vdots & \vdots & \ddots & \\ 0 & c_1 & \cdots & c_b \end{pmatrix} \begin{pmatrix} 0 & \cdots & 0 \\ 0 & c_b & \cdots & c_1 \\ & & \ddots & \vdots \\ 0 & & & c_b \end{pmatrix} \omega_2,$$

which in turn is equal to

$$\begin{pmatrix} 0 & & & 0 \\ c_b & \ddots & & \\ \vdots & \ddots & \ddots & \\ c_1 & \cdots & c_b & 0 \end{pmatrix} \begin{pmatrix} 0 & c_b & \cdots & c_1 \\ & \ddots & \ddots & \vdots \\ & & \ddots & c_b \\ 0 & & & 0 \end{pmatrix} \omega_2.$$

Thus we have derived the identity

$$\left[\begin{pmatrix} c_0 & & & 0 \\ \vdots & \ddots & & \\ \vdots & & \ddots & \\ c_b & \cdots & \cdots & c_0 \end{pmatrix} \begin{pmatrix} c_0 & \cdots & \cdots & c_b \\ & \ddots & & \vdots \\ & & \ddots & \vdots \\ 0 & & & c_0 \end{pmatrix} \right.$$
$$\left. - \begin{pmatrix} 0 & & & 0 \\ c_b & \ddots & & \\ \vdots & \ddots & \ddots & \\ c_1 & \cdots & c_b & 0 \end{pmatrix} \begin{pmatrix} 0 & c_b & \cdots & c_1 \\ & \ddots & \ddots & \vdots \\ & & \ddots & c_b \\ 0 & & & 0 \end{pmatrix} \right] \omega_2 = 0.$$

It is known (see, for example, Theorem 3 in [9]) that if all zeros of the polynomial $c_0 z^b + \cdots + c_b$ are in the open unit disk (as they are in this case; see the remark at the beginning of the proof), then the vector $(c_0, \ldots, c_b)^T$ is the first column of the inverse of some positive definite Toeplitz matrix, say Φ. Then by the Gohberg–Semencul formula ([5, p. 86])

$$c_0 \Phi^{-1} = \begin{pmatrix} c_0 & & & 0 \\ \vdots & \ddots & & \\ \vdots & & \ddots & \\ c_b & \cdots & \cdots & c_0 \end{pmatrix} \begin{pmatrix} c_0 & \cdots & \cdots & c_b \\ & \ddots & & \vdots \\ & & \ddots & \vdots \\ 0 & & & c_0 \end{pmatrix}$$
$$- \begin{pmatrix} 0 & & & 0 \\ c_b & \ddots & & \\ \vdots & & \ddots & \\ c_1 & \cdots & c_b & 0 \end{pmatrix} \begin{pmatrix} 0 & c_b & \cdots & c_1 \\ & \ddots & \ddots & \vdots \\ & & \ddots & c_b \\ 0 & & & 0 \end{pmatrix}.$$

Since Φ^{-1} is positive definite whenever Φ is we conclude that $\omega_2 = 0$. This contradiction completes the proof. □

Next we show that the inverses of P_b are in fact uniformly bounded or, equivalently, that the minimal eigenvalues of P_b are bounded away from zero independent of n. Let $\lambda_{\min}(n)$ be the minimal eigenvalue of P_b with corresponding eigenvector

$$y(n) = \begin{pmatrix} y_1(n) \\ \hat{y}(n) \\ \pm J y_1(n) \end{pmatrix} = \begin{pmatrix} \omega_1(n) \\ v(n) \\ \pm J\omega_1(n) \end{pmatrix}$$

partitioned as before so that $\xi y_1(n)$ and $P\omega_1(n)$ are defined and normalized so that $\|y(n)\| = 1$. The proof is similar to the last part of the proof of Theorem 4.2, and we use the same notation except that we use n to indicate the order of the involved matrix.

COROLLARY 4.3. *For $n > 2b+1$, there is a constant $\kappa > 0$ such that $\lambda_{\min}(n) > \kappa$.*

Proof. If $\liminf \lambda_{\min}(n) > 0$ there is nothing to prove so without loss of generality assume $\lim_{n\to\infty} \lambda_{\min}(n) = 0$ and hence $\lim_{n\to\infty} \|z(n)\| = 0$. This implies that (9) becomes $\lim_{n\to\infty}(c_b, \ldots, c_0)\omega_2(n) = 0$. It follows as in the proof of Theorem 4.2 that

$$(JPP^TJ \pm QQ^T)\omega_2(n) \to 0,$$

since P and Q are defined independent of n. In the case of a positive sign the matrix is positive definite and fixed. Hence as $n \to \infty$, $\omega_2(n) \to 0$.

In the case of a negative sign, as n increases without bound, it follows that

$$\left[\begin{pmatrix} c_0 & & & 0 \\ \vdots & \ddots & & \\ \vdots & & \ddots & \\ c_b & \cdots & \cdots & c_0 \end{pmatrix} \begin{pmatrix} c_0 & \cdots & \cdots & c_b \\ & \ddots & & \vdots \\ & & \ddots & \vdots \\ 0 & & & c_0 \end{pmatrix} - \begin{pmatrix} 0 & & & 0 \\ c_b & \ddots & & \\ \vdots & \ddots & \ddots & \\ c_1 & \cdots & c_b & 0 \end{pmatrix} \begin{pmatrix} 0 & c_b & \cdots & c_1 \\ & \ddots & \ddots & \vdots \\ & & \ddots & c_b \\ 0 & & & 0 \end{pmatrix}\right] \omega_2(n) \to 0,$$

as $n \to \infty$ so that once again $\omega_2(n) \to 0$.

In both cases it follows from (7) that $A_n y(n) \to 0$. Since A is positive definite it follows that $\|A^{-1}(n)\| > \|A^{-1}\|$ for all n. Therefore $y(n) \to 0$ as $n \to \infty$. This contradiction completes the proof. □

4.2. Nonbanded matrices. If $A = (a_{|i-j|})_{i,j=1}^n$ is not banded, then it is natural to use

$$P_n(n) = P_n = \sum_{i=0}^{n} a_i \zeta_i \tag{10}$$

as a preconditioner instead of (2). Note that for fixed n, the cost of computing the spectrum of $P_b(n)$ via (4) is fixed regardless of the value of b. In particular the spectrum of $P_n(n)$ may be computed at the same cost. In general, $P_n(n)$ may not satisfy the

first two properties. (This is also true of Strang's circulant preconditioner [14].) If however, the Toeplitz matrix $A = A(n)$ is a finite $n \times n$ section of a singly infinite positive definite Toeplitz matrix $A(\infty) = \left(a_{|i-j|}\right)_{i\,j=1}^{\infty}$ such that $\sum_{i=-\infty}^{\infty} |a_i| < \infty$ (in other words $A(\infty)$ is generated by the Wiener class function $f(\theta) = \sum_{k=-\infty}^{\infty} a_k e^{ik\theta}$), then for n sufficiently large these two properities will be satisfied as the following standard argument shows.

Since $\sum_{i=-\infty}^{\infty} |a_i| < \infty$, then given $\varepsilon > 0$, there is an integer $b > 0$ such that $\sum_{|i|>b} |a_i| < \varepsilon$. If we write $A(\infty) = A_b(\infty) + A_\varepsilon(\infty)$, where $A_b(\infty)$ has bandwidth $2b+1$, then clearly

$$\|A_\varepsilon(\infty)\| \leq \sum_{|i|>b} |a_i| < 2\varepsilon.$$

If ε is chosen such that

$$\Delta \equiv \inf \frac{x^T A(\infty) x}{x^T x} \geq 2\varepsilon > 0,$$

then clearly $A_b(\infty)$ is also positive definite, and hence for n large enough $P_b(n)$ is positive definite too, by Theorem 4.2. By Corollary 4.3 we can further constrain ε such that $\lambda_{\min}(P_b(n)) > 2\varepsilon$ for all sufficiently large n.

Writing $P(n) = P_b(n) + P_\varepsilon(n)$ we see that for such ε, $P(n)$ is positive definite whenever $P_b(n)$ is positive definite. Indeed, it is clear that $\|\zeta_p\| \leq 2$ for any p and hence

$$\|P_\varepsilon(n)\| \leq 2 \sum_{i>b} |a_i| = \sum_{|i|>b} |a_i| < 2\varepsilon.$$

Since $A(n) - P(n) = [A_b(n) - P_b(n)] + [A_\varepsilon(n) - P_\varepsilon(n)]$ we get

$$I - A^{-1}(n)P(n) = A^{-1}(n)\left[A_b(n) - P_b(n)\right] + A^{-1}(n)\left[A_\varepsilon(n) - P_\varepsilon(n)\right].$$

Therefore $\left\|A^{-1}(n)\left[A_\varepsilon(n) - P_\varepsilon(n)\right]\right\| \leq 4\varepsilon/\Delta$. Defining $\tilde{\varepsilon}$ to be $4\varepsilon/\Delta$ we see that the the interval $(1-\tilde{\varepsilon}, 1+\tilde{\varepsilon})$ contains the spectrum of $A^{-1}(n)P(n)$ except possibly for $2b+1$ outliers. Therefore the interval $\left(1-\tilde{\varepsilon}+O(\tilde{\varepsilon}^2), 1+\tilde{\varepsilon}+O(\tilde{\varepsilon}^2)\right)$ contains spectrum of $P^{-1}(n)A(n)$ except possibly for $2b+1$ outliers and hence is asymptotically clustered.

We remark that if the problem is poorly conditioned in the sense that $1/\Delta = \left\|A^{-1}(\infty)\right\|_2$ is large, then a large b may be needed to obtain a satisfactory clustering of the eigenvalues of $P^{-1}(n)A(n)$.

We also remark that if A is positive definite itself but not necessarily a section of an infinite matrix, then a positive definite S_1-diagonal preconditioner can be built as follows. Let $D = \Delta(S_1 A S_1)$ and let $P = S_1 D S_1$. Then clearly P is positive definite. It is also the nearest element of D_{S_1} to A in Frobenius norm. This P is the analogue of the optimal circulant preconditioner of T. Chan [3]. We remark that the determination of P via the computation of $S_1 A S_1$ is prohibitively expensive as it requires $O(n^2 \log(n))$ flops. A fast $O(n \log(n))$ method for computing P will be suggested elsewhere.

5. Numerical results. To test the S_1-diagonal preconditioner, we have implemented the PCGA on the Connection Machine 200 at United Technologies Research Center in East Hartford, Connecticut. This computer was configured with 16,384 bit serial processors, 512 floating point processors, and a Vax 6320 front end.

The PCGA was used to solve $Ax = b$ for each of the following matrices and $b = (1, \ldots, 1)^T$.

Matrix 1. $A = \left[\frac{1}{(|i-j|+1)^{1.1}}\right]_{i,j=1}^{n}$.

Matrix 2. $A = \left[\frac{1}{|i-j|+1}\right]_{i,j=1}^{n}$.

Matrix 3. Bandwidth = 41, (see text).

Matrix 4. Bandwidth = 201, (see text).

Matrix 5. $A = \left[\frac{\cos(j)}{j+1}\right]_{i,j=1}^{n}$.

Matrix 6. $A = \left[\frac{1}{(|i-j|+1)^2}\right]_{i,j=1}^{n}$.

Matrix 7. $A = \left[\frac{1}{2^{|i-j|}}\right]_{i,j=1}^{n}$.

Except for Matrix 3 and Matrix 4, most of these matrices have appeared in the literature previously (see [14], [1], and [3]). These were generated in the following manner. It is well known that for $\rho_i \in R$, $|\rho_i| < 1$, $i = 1, \ldots, k$, $k < n - 1$ the coefficients of $\prod_{i=1}^{k} (1 - z\rho_i)\left(1 - z^{-1}\rho_i\right)$ define for any n an $n \times n$ positive definite, symmetric, Toeplitz matrix with bandwidth $2k + 1$.

Table 1 shows the convergence results for Matrices 1–4. N is the problem size; I, C, and S, represent, respectively, no preconditioning, Tony Chan's circulant preconditioner, and the S_1–diagonal preconditioner. The body of the table gives the iteration count for each matrix and each preconditioner.

TABLE 1

	Matrix 1			Matrix 2			Matrix 3			Matrix 4		
N	I	C	S	I	C	S	I	C	S	I	C	S
255	19	5	5	21	5	5	255	47	9	147	10	7
511	20	5	5	22	5	5	511	37	8	169	9	7
1023	21	5	5	23	5	5	837	29	9	190	9	7
2047	22	5	5	24	6	5	860	21	9	204	9	7
4095	22	6	5	25	6	5	874	17	9	206	9	7
8191	22	6	5	25	6	6	876	16	10	204	9	7

The cases $k = 20$ and $k = 100$, $\rho_i = -d + i\Delta$, $i = 1, \ldots, k$ where $\Delta = \frac{2d}{k}$ and $d = 0.75$ were then used to generate matrices 3 and 4, respectively.

The stopping criterion $\|r_i\|_\infty < 10^{-7}$ was used in all cases, except that the algorithm was terminated if the number of iterations ever exceeded the order of the matrix.

Table 2 shows the convergence results for Matrices 5–7. The format is the same as in Table 1.

TABLE 2

	Matrix 5			Matrix 6			Matrix 7		
N	I	C	S	I	C	S	I	C	S
1023	21	7	7	11	4	4	16	3	3
2047	23	7	7	11	4	4	16	3	3
4095	23	7	7	10	4	4	15	3	3
8191	24	7	7	10	4	4	15	3	3
16383	25	7	7	10	4	4	14	3	3
32767	25	7	6	9	4	4	14	3	3

Since no efficient implementation of the FST was available it was performed via the fast Fourier transform (FFT) as in [11]. The FFT from the Connection Machine Scientific Software Library was used. If an efficient FST were available the algorithm could be faster overall since the FST can be faster than the FFT; see [7]. Moreover, when the coefficient matrix of the problem is real, using the FST eliminates the need for complex arithmetic. Thus storage can also be reduced.

Note that for the nonbanded problems the S_1-diagonal preconditioner is competitive with the circulant preconditioner, while for the banded problems it is clearly superior.

6. Appendix. Other diagonal spaces. This section displays bases for the diagonal spaces of some common fast transforms. All proofs are very similar to the proof of Lemma 2.2 and are therefore omitted.

Denote by S_2, C_1, and C_2 the second discrete sine transform and the two discrete cosine transforms defined in [16]. That is,

$$S_2 \equiv \left(\sin\left(\frac{i(2j-1)\pi}{2n}\right)\right)_{i,j=1}^{n},$$

$$C_2 \equiv \left(\cos\left(\frac{i(2j+1)\pi}{2n}\right)\right)_{i,j=0}^{n-1}, \quad \text{where } k_i = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } i = 1, \\ 1 & \text{otherwise.} \end{cases}$$

$$C_1 \equiv (c_{ij})_{i,j=0}^{n}, \text{ where } c_{ij} \equiv \begin{cases} \frac{1}{2} & \text{if } j = 0, \\ \frac{(-1)^i}{2} & \text{if } j = n, \\ \cos\left(\frac{i(2j+1)\pi}{2n}\right) & \text{otherwise.} \end{cases}$$

LEMMA 6.1. *Let $n \in Z^+$ be given. Then $\{\eta_p\}_{p=0}^{n-1}$ is a basis for D_{S_2}, where η_0 is the identity and*

$$\eta_p(i,j) = \begin{cases} 1 & \text{if } |i-j| = p \text{ and } i < n, \\ -1 & \text{if } i+j = p+1, \\ 1 & \text{if } 2n-(i+j) = p \text{ and } i < n, \\ 2 & \text{if } i = n \text{ and } j = n-p, \\ 0 & \text{otherwise.} \end{cases}$$

The spectrum of η_p, $p > 0$ is $\left\{ 2\cos\left(\frac{(2k-1)p\pi}{2n}\right)\right\}_{k=1.}^{n}$

LEMMA 6.2. *Let $n \in Z^+$ be given. Then $\{\xi_p\}_{p=1}^{n}$ is a basis for D_{C_1} where ξ_0 is the identity and*

$$\xi_p(i,j) = \begin{cases} 2 & \text{if } i = 1 \text{ and } j = p+1, \\ 2 & \text{if } i = n \text{ and } j = n-p, \\ 1 & \text{if } |i-j| = p,\ i \neq 1,\ i \neq n, \\ 1 & \text{if } i+j = p+2,\ j \neq p+1, \\ 1 & \text{if } i+j = 2n-p,\ j \neq n-p, \\ 0 & \text{otherwise.} \end{cases}$$

The spectrum of ξ_p $p > 0$ is $\left\{2\cos\left(\frac{kp\pi}{n-1}\right)\right\}_{k=0.}^{n-1}$

LEMMA 6.3. *Let $n \in Z^+$ be given. Then $\{\xi_p\}_{p=1}^n$ is a basis for $D_{C_2^T}$ where χ_0 is the identity and*

$$\chi_p(i,j) = \begin{cases} 1 & \text{if } |i-j| = p, \\ 1 & \text{if } i+j = p+1, \\ 1 & \text{if } i+j = 2n-p+1, \\ 0 & \text{otherwise.} \end{cases}$$

The spectrum of χ_p is $\left\{ 2\cos\left(\frac{kp\pi}{n}\right) \right\}_{k=1}^n$.

Of course, bases abound for each of the diagonal spaces above but for constructing preconditioners for Toeplitz matrices the bases given have certain advantages. Principally, if we take $P_b = \sum_{i=0}^{n-1} a_i\beta_i$ where $\{\beta_i\}$ is one of $\{\eta_i\}$, $\{\xi_i\}$, or $\{\chi_i\}$ then

$$A - P_b = \begin{pmatrix} G & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \hat{G} \end{pmatrix},$$

as in (3). Thus most of the eigenvalues of $P^{-1}A$ are equal to one and the number of outlying eigenvalues will depend linearly on b.

Also as with the S_1-diagonal preconditioner of §4 all of these transforms are real so complex arithmetic can be avoided by using them rather than the DFT.

Acknowledgments. We would like to thank the referees for many helpful comments and suggestions.

REFERENCES

[1] R. CHAN, *The spectrum of a family of circulant preconditioned Toeplitz systems*, SIAM J. Numer. Anal., 26 (1989), pp. 503–506.

[2] R. CHAN, X. JIN, AND M.-C. YEUNG, *The circulant operator in the Banach algebra of matrices*, Linear Algebra Appl., 149 (1991), pp. 41–53.

[3] T. F. CHAN, *An optimal circulant preconditioner for Toeplitz systems*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 766–771.

[4] P. DAVIS, *Circulant Matrices*, John Wiley and Sons, New York, 1979.

[5] I. GOHBERG AND I. FEĬLDMAN, *Convolution Equations and Projection Methods for their Solution*, American Mathematical Society, Providence, RI, 1974. Translated from Russian by F. M. Goldware.

[6] G. GOLUB AND C. VAN LOAN, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1983.

[7] A. K. JAIN, *Fundamentals of Digital Image Processing*, System Sciences Series, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[8] A. JENNINGS, *Influence of eigenvalue spectrum on the convergence rate of the conjugate gradient method*, J. Instit. Math. Appl., 20 (1977), pp. 61–72.

[9] I. KOLTRACHT AND M. NEUMANN, *On the inverse M-matrix problem for real symmetric positive–definite Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 310–320.

[10] N. LEVINSON, *The Weiner rms error criterion in filter design and prediction*, J. Math. Phys., 25 (1947), pp. 261–278.

[11] W. H. PRESS, B. P. FLANNERY, S. A. TEUKOLSKY, AND W. T. VETTERLING, *Numerical Recipes: The Art of Scientific Computing (Fortran Version)*, Cambridge University Press, Cambridge, MA, 1989.

[12] J. STOER AND R. BULIRSCH, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980. Translated from German by R. Bartels, W. Gautschi, and C. Witzgall.

[13] G. STRANG, *Introduction to Applied Mathematics*, Wellesly–Cambridge Press, Wellesly, MA, 1986.

[14] ———, *A proposal for Toeplitz matrix calculations*, Stud. Appl. Math., 74 (1986), pp. 171–176.

[15] E. E. TYRTYSHNIKOV, *Optimal and super-optimal circulant preconditioners*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 459–473.

[16] C. VAN LOAN, *Computational Frameworks for the Fast Fourier Transform*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 646–654, April 1995

020

THE EUCLIDIAN DISTANCE MATRIX COMPLETION PROBLEM*

MIHÁLY BAKONYI† AND CHARLES R. JOHNSON‡

Abstract. Motivated by the molecular conformation problem, completions of partial Euclidian distance matrices are studied. It is proved that any partial distance matrix with a chordal graph can be completed to a distance matrix. Given any nonchordal graph G, it is shown that there is a partial distance matrix A with graph G such that A does not admit any distance matrix completions. Finally, the connection between distance matrix completions and positive semidefinite completions is outlined.

Key words. Euclidian distance matrix, partial matrix, completion, positive (semi)definite matrix, circumhypersphere

AMS subject classifications. 15A47, 05C50

1. Introduction.

Let $||\quad||$ denote Euclidian length on $\mathbf{R}^k$. For two points A and B, we use $d(A,B)$ for $||A-B||$. The matrix $D=(d_{ij})_{i,j=1}^n$ is a (*Euclidian*) *distance matrix* if there exist $P_1,\ldots,P_n \in \mathbf{R}^k$ such that $d_{ij}=d(P_i,P_j)^2$. A great deal is known about distance matrices (e.g., [2], [7], [9]). For example in [9], a symmetric matrix $D=(d_{ij})_{i,j=1}^n$, with $d_{ii}=0$, $i=1,\ldots,n$, is a distance matrix if and only if D is negative semidefinite on the orthogonal complement of the vector $e=(1,1,\ldots,1)^T$. This is equivalent to the statement that the bordered matrix

$$\begin{pmatrix} 0 & e^T \\ e & D \end{pmatrix} \tag{1}$$

has only one positive eigenvalue or to the fact that the Schur complement of the upper left $2-by-2$ principal submatrix

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

in (1) is negative semidefinite. Furthermore, the rank of this Schur complement is the minimum dimension k in which the points $P_1,\ldots,P_n$ may lie. In this case we say that D is a *distance matrix* in $\mathbf{R}^k$. In §1, we use these characterizations to recover a result concerning circumhyperspheres [7].

We call an $n-by-n$ array $A=(a_{ij})_{i,j=1}^n$ a *partial distance matrix* in $\mathbf{R}^k$ if

(i) every entry a_{ij} of A is either "specified" or "unspecified" (free to be chosen);

(ii) a_{ii} is specified as 0, $i=1,\ldots,n$, and a_{ji} is specified (and equal to a_{ij}) if and only a_{ij} is specified; and

(iii) every fully specified principal submatrix of A is itself a distance matrix in $\mathbf{R}^k$.

A *completion* of a partial distance matrix is a choice of values for each of the unspecified entries, resulting in a conventional matrix. The *distance matrix completion*

* Received by the editors June 1, 1993; accepted for publication (in revised form) by John Maybee, April 7, 1994.

† Department of Mathematics, Georgia State University, Atlanta, Georgia 30303 (matmmb@gsusgi2.gsu.edu). The work of this author was supported in part by Georgia State University.

‡ Department of Mathematics, The College of William and Mary, Williamsburg, Virginia 23187-8795 (crjohnso@cs.wm.edu). The work of this author was supported in part by National Science Foundation grant DMS 92-00899 and Office of Naval Research contract N0014-90-J-1739.

problem then asks which partial distance matrices have distance matrix completions. It is clear that assumption (iii) above is necessary for this. This "inheritance property" is important in other previously studied completion problems, such as those involving positive definiteness, inertia, rank, and contractions.

Our principal result here is that if the undirected graph of the specified entries of a partial distance matrix is chordal, then it necessarily has a distance matrix completion. This is based upon an analysis of the case of partial distance matrices with one pair of symmetrically placed unspecified entries, the "one variable problem," together with one-step-at-a-time technology for chordal graphs developed in [8]. For any nonchordal graph there exists partial distance matrices without distance matrix completions.

Though it should also be of interest in the classical subject of distance geometry, we were motivated by the molecular mapping, or "conformation," problem. This is the problem of deducing the possible shapes of a molecule from partial (or inaccurate) information about interatomic distances. For many compounds, not all of the interatomic distances may be measured accurately, but the shape (essentially determined by the distance matrix) is crucial to understanding how the molecule functions.

For terminology and results concerning graph theory we essentially follow [6]. An *undirected graph* is a pair $G = (V, E)$ in which V, the *vertex set*, is a finite set (usually $V = \{1, \ldots, n\}$), and the *edge set* E is a symmetric binary relation on V. The *adjacency set* of a vertex v is denoted by $\mathrm{Adj}(v)$, i.e., $w \in \mathrm{Adj}(v)$ if $\{v, w\} \in E$. Given a subset $S \subseteq V$, define the *subgraph induced* by S by $G_S = (S, E_S)$, in which $E_S = \{\{x, y\} \in E | x \in S \text{ and } y \in S\}$. A *complete graph* is one with the property that every pair of distinct vertices is adjacent. A subset $K \subseteq V$ is a *clique* if the induced graph on K is complete. The complement $\bar{G} = (V, \bar{E})$ of a graph $G = (V, E)$ is defined by $\bar{E} = \{\{i, j\} | i \neq j, \text{ and } \{i, j\} \notin E\}$.

A *path* $[v_1, \ldots, v_k]$ is a sequence of vertices such that $\{v_j, v_{j+1}\} \in E$ for $j = 1, \ldots, k-1$. A *cycle* of length $k > 2$ is a path $[v_1, \ldots, v_k, v_1]$ in which $v_1, \ldots, v_k$ are distinct. A graph G is called *chordal* if every cycle of length greater than three possesses a chord, i.e., an edge joining two nonconsecutive vertices of the cycle. A subset $S \subset V$ is called a $u - v$ *vertex separator* for the nonadjacent vertices u and v if the removal of S from the graph separates u and v into distinct connected components. If no proper subset of S contains a $u - v$ separator, then S is a *minimal* $u - v$ *separator.* It is known ([6, Thm. 4.1]) that an undirected graph is chordal if and only if every minimal vertex separator is a clique.

In §4 we are concerned with the connections between positive semidefinite completions and distance matrix completions.

A partial matrix $A = (a_{ij})_{i,j}^n$ is called (*combinatorially*) *symmetric* if

(i) a_{ii} is specified, $i = 1, \ldots, n$, and

(ii) a_{ij} is specified if and only if a_{ji} is also specified.

All partial distance matrices are symmetric. With a symmetric partial matrix $A = (a_{ij})_{i,j=1}^n$ we associate the undirected graph $G = (V, E)$ with $V = \{1, 2, \ldots, n\}$ and $E = \{\{i, j\} | a_{ij} \textit{ is specified}\}$.

A symmetric partial matrix A is called *partial positive semidefinite* if all fully specified principal submatrices of A are positive semidefinite. In [8] it has been proved that any partial positive semidefinite matrix, the graph of whose specified entries is chordal, can be completed to a positive semidefinite matrix. We translate this result into a distance problem among points on a hypersphere. We also treat by this approach the problem of the existence of a positive semidefinite completion of a

partial positive semidefinite matrix having a nonchordal graph.

Throughout the paper, for a matrix $A = (a_{ij})_{i,j=1}^n$ and an index set $\alpha \subset \{1, \ldots, n\}$, $A(\alpha)$ denotes the principal submatrix of A whose rows and columns correspond to the index set α. The notation A^- represents the (unique Moore–Penrose) generalized inverse of the Hermitian matrix A.

2. Circumhyperspheres. Using the approach presented in the introduction, one easily recovers the results of [7] concerning the existence of a circumhypersphere for a set of points.

THEOREM 2.1. *Let D be a distance matrix corresponding to the points $P_1, \ldots, P_n$ in $\mathbf{R}^n$. Then, there is a circumhypersphere for $P_1, \ldots, P_n$ if and only if $e^T D^- e \neq 0$. This has radius given by $r^2 = (2e^T D^- e)^{-1}$.*

Proof. Using results on generalized Schur complements (see e.g. [3]), one obtains that the number $i_+(M)$ of positive eigenvalues of a partitioned matrix $M = \left(\begin{smallmatrix} A & B \\ B^* & C \end{smallmatrix}\right)$ is given by

$$i_+(M) = i_+(C) + i_+(A - BC^-B^*) + \operatorname{rank}(B|\ker C), \tag{2}$$

in which $B|\ker C$ means the restriction of B to the null space of C. Let D be a distance matrix corresponding to the points $P_1, \ldots, P_n$ in $\mathbf{R}^n$. By the observation in the introduction, the existence of $O \in \mathbf{R}^n$ such that $d(O, P_i) = r$ for $i = 1, \ldots, n$ is equivalent to the condition that the bordered matrix

$$\hat{D} = \begin{pmatrix} 0 & 1 & e^T \\ 1 & 0 & r^2 e^T \\ e & r^2 e & D \end{pmatrix}$$

has exactly one positive eigenvalue. Let $a = -e^T D^- e$, and then, by (2) we have that

$$i_+(\hat{D}) = i_+(D) + i_+\left(\begin{pmatrix} -a & 1 - r^2 a \\ 1 - r^2 a & -r^4 a \end{pmatrix}\right) + \operatorname{rank}\left(\begin{pmatrix} e^T \\ r^2 e^T \end{pmatrix} | \ker D\right).$$

Since

$$\begin{pmatrix} 0 & e^T \\ e & D \end{pmatrix}$$

has exactly one positive eigenvalue, we have that $i_+(D) = 1$ and $e^T h = 0$ for any $h \in \ker D$. Thus

$$\operatorname{rank}\left(\begin{pmatrix} e^T \\ r^2 e^T \end{pmatrix} | \ker D\right) = 0,$$

and $i_+(\hat{D}) = 1$ if and only if

$$\begin{pmatrix} -a & 1 - r^2 a \\ 1 - r^2 a & -r^4 a \end{pmatrix} \tag{3}$$

is negative semidefinite. This can be realized if and only if $a > 0$. We always have $a \geq 0$ since

$$i_+\left(\begin{pmatrix} 0 & e^T \\ e & D \end{pmatrix}\right) = 1.$$

The smallest r that makes (3) negative semidefinite is given by $r = 1/\sqrt{2a}$, and this completes the proof. □

3. The main results. The following result is a consequence of Corollary 3.2 in [5]. For the sake of completeness we present a proof.

LEMMA 3.1. *Let*

$$R = \begin{pmatrix} a & B & x \\ B^T & C & D \\ x & D^T & f \end{pmatrix}$$

be a real partial positive semidefinite matrix, with x an unknown scalar and

$$\operatorname{rank} \begin{pmatrix} a & B \\ B^T & C \end{pmatrix} = p$$

and

$$\operatorname{rank} \begin{pmatrix} C & D \\ D^T & f \end{pmatrix} = q,$$

with (necessarily) $|p - q| \leq 1$. Then there is real positive semidefinite completion F of R such that $\operatorname{rank} F = \max\{p, q\}$. *Moreover, this completion is unique if and only if* $\operatorname{rank} C = p$ *or* $\operatorname{rank} C = q$.

Proof. Let U be an orthogonal matrix that diagonalizes C, namely, $U^T C U = Y$, in which Y is a positive semidefinite diagonal matrix. Let $\hat{U} = 1 \oplus U \oplus 1$ and

$$\hat{R} = \hat{U}^T R \hat{U} = \begin{pmatrix} a & \hat{B} & x \\ \hat{B}^T & Y & \hat{D} \\ x & \hat{D}^T & f \end{pmatrix}$$

in which $\hat{B} = BU$ and $\hat{D} = U^T D$. Since $\hat{U}$ is orthogonal, the set of numbers that make R positive semidefinite coincides with that making $\hat{R}$ positive semidefinite and $\operatorname{rank} R = \operatorname{rank} \hat{R}$. Since

$$\begin{pmatrix} a & \hat{B} \\ \hat{B}^T & Y \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} Y & \hat{D} \\ \hat{D}^T & f \end{pmatrix}$$

are positive semidefinite, all entries (in the rows and columns corresponding to diagonal entries of Y that are 0) equal zero. We may eliminate those rows and columns and then assume that Y is invertible. Then, $\hat{R}$ is positive semidefinite if and only if the Schur complement S of Y in $\hat{R}$ is positive semidefinite. But,

$$S = \begin{pmatrix} a - \hat{B}Y^{-1}\hat{B}^T & x - \hat{B}Y^{-1}\hat{D} \\ x - \hat{D}^T Y^{-1}\hat{B}^T & f - \hat{D}^T Y^{-1}\hat{D} \end{pmatrix}$$

and $\operatorname{rank} \hat{R} = \operatorname{rank} C + \operatorname{rank} S$. If $\operatorname{rank} C = p$ or $\operatorname{rank} C = q$ then $a - \hat{B}Y^{-1}\hat{B}^T = 0$, respectively $f - \hat{D}^T Y^{-1}\hat{D}^T = 0$, and, thus, we must choose $x = \hat{B}Y^{-1}\hat{D}$. If $\operatorname{rank} C < p = q$, the problem has two solutions given by

$$|x - \hat{B}Y^{-1}\hat{D}|^2 = (a - \hat{B}Y^{-1}\hat{B}^T)(f - \hat{D}^T Y^{-1}\hat{D})$$

that realize the completion of S to a rank-1 negative semidefinite matrix. □

LEMMA 3.2. *The partial distance matrix*

$$R = \begin{pmatrix} 0 & D_{12} & x \\ D_{12}^T & D_{22} & D_{23} \\ x & D_{23}^T & 0 \end{pmatrix}$$

admits at least one completion to a distance matrix F. Moreover, if

$$\begin{pmatrix} 0 & D_{12} \\ D_{12}^T & D_{22} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} D_{22} & D_{23} \\ D_{23}^T & 0 \end{pmatrix}$$

are distance matrices in $\mathbf{R}^p$, respectively, $\mathbf{R}^q$, then x can be chosen so that F is a distance matrix in $\mathbf{R}^s$, $s = \max\{p, q\}$.

Proof. Without loss of generality, we may assume that R is at least $3 - by - 3$, since, otherwise, we may complete with any positive number. Thus, R has at least one fully specified row and column. Interchange the first two rows and the first two columns of R, and then we must complete the partial distance matrix

$$\tilde{R} = \begin{pmatrix} 0 & d_{12} & \tilde{D}_{13} & d_{14} \\ d_{12} & 0 & \tilde{D}_{23} & x \\ \tilde{D}_{13}^T & \tilde{D}_{23}^T & \tilde{D}_{33} & \tilde{D}_{34} \\ d_{14} & x & \tilde{D}_{34}^T & 0 \end{pmatrix}$$

to a distance matrix in $\mathbf{R}^s$. By the remark in the introduction, this latter problem is equivalent to finding completions of the partial matrix

$$\tilde{D} = \begin{pmatrix} 0 & 1 & 1 & e^T & 1 \\ 1 & 0 & d_{12} & \tilde{D}_{13} & d_{14} \\ 1 & d_{12} & 0 & \tilde{D}_{23} & x \\ e & \tilde{D}_{13}^T & \tilde{D}_{23}^T & \tilde{D}_{33} & \tilde{D}_{34} \\ 1 & d_{14} & x & \tilde{D}_{34}^T & 0 \end{pmatrix}$$

to a matrix in which the Schur complement of the upper left $2 - by - 2$ principal submatrix $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ is negative semidefinite and has rank s. This latter Schur complement is of the form

$$S = \begin{pmatrix} a & B & x - d_{12} - d_{14} \\ B^T & C & D \\ x - d_{12} - d_{14} & D^T & f \end{pmatrix},$$

in which

$$\begin{pmatrix} a & B \\ B^T & C \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} C & D \\ D^T & f \end{pmatrix}$$

are negative semidefinite and have ranks less than or equal to s. Then, any negative semidefinite completion of S of rank s given by Lemma 3.1 provides a solution to our distance completion problem. □

Remark 1. From the proof of Lemma 3.2 and the uniqueness part of Lemma 3.1, we obtain that the partial distance matrix in $\mathbf{R}^k$,

$$R = \begin{pmatrix} 0 & D_{12} & x \\ D_{12}^T & D_{22} & D_{23} \\ x & D_{23}^T & 0 \end{pmatrix}$$

admits a unique completion to a distance matrix in $\mathbf{R}^k$ if and only if

$$\operatorname{rank}\left(\begin{pmatrix} 0 & e^T \\ e & D_{22} \end{pmatrix}\right) = k + 2.$$

Our main result is the following theorem.

THEOREM 3.3. *Every partial distance matrix in* $\mathbf{R}^k$, *the graph of whose specified entries is chordal, admits a completion to a distance matrix in* $\mathbf{R}^k$.

Proof. Let R be an $n-by-n$ partial distance matrix in $\mathbf{R}^k$ and assume that the graph $G=(V,E)$ of R is chordal. Then from [8], there exists a sequence of chordal graphs $G=G_0,G_1,\ldots,G_t=K_n$ (the complete graph on n vertices), such that each G_j is obtained by adding exactly one new edge $\{u_j,v_j\}$ to G_{j-1}. Moreover, each G_j, $j=1,\ldots,t$, has only one maximal clique V_j that is not a clique in G_{j-1}.

Consider first the partial submatrix $R(V_1)$, with one pair of unknowns, symmetrically placed on the (u_1,v_1) and (v_1,u_1) positions. Then, by Lemma 3.2, we can specify these entries and obtain a partial distance matrix in $\mathbf{R}^k$ having G_1 as the graph of its specified entries. Then we complete the partial submatrix corresponding to the index set V_2. We continue this one-entry-at-a-time completion procedure until we complete R to a distance matrix in $\mathbf{R}^k$. □

Example. Given any nonchordal graph $G=(V,E)$, $V=\{1,\ldots,n\}$, we show that there exists a partial distance matrix $R=(r_{ij})_{i,j=1}^n$ such that R has no completion to a distance matrix. Assume that the vertices $1,2,\ldots,k\geq 4$ form a chordless cycle in G. Define the partial distance matrix R by

$$r_{ij}=\begin{cases}0 & \text{if } \{i,j\}\in E \quad \text{and} \quad k+1\leq i,j\leq n,\\ 0 & \text{if } |i-j|=1,\ 1\leq i\leq j\leq k,\\ 1 & \text{for any other } \{i,j\}\in E.\end{cases}$$

Then any fully specified principal submatrix of R is either

$$0,\begin{pmatrix}0 & e^T\\ e & 0\end{pmatrix},\begin{pmatrix}0 & 0 & e^T\\ 0 & 0 & e^T\\ e & e & 0\end{pmatrix},\text{ or }\begin{pmatrix}0 & 1 & e^T\\ 1 & 0 & e^T\\ e & e & 0\end{pmatrix},$$

each of them being a distance matrix. Thus R is a partial distance matrix, but R does not admit a completion to a distance matrix. Indeed, the upper left $k-by-k$ principal submatrix cannot be completed to a distance matrix since otherwise there exist points $P_1,\ldots,P_k$ such that $d(P_i,P_{i+1})=0$ for $i=1,\ldots,k-1$ and $d(P_1,P_k)=1$, a contradiction.

THEOREM 3.4. *Let* R *be a partial distance matrix in* $\mathbf{R}^k$, *the graph* $G=(V,E)$ *of whose specified entries is chordal, and let* $\mathcal{S}$ *be the collection of all minimal vertex separators of* G. *Then* R *admits a unique completion to a distance matrix in* $\mathbf{R}^k$ *if and only if*

$$\begin{pmatrix}0 & e^T\\ e & R(S)\end{pmatrix} \text{ has rank } k+2 \text{ for any } S\in\mathcal{S}. \tag{4}$$

Proof. We prove the result by induction on the cardinality m of the complement of the edge set of G. If $m=1$, the result follows from Remark 1. Assume the result is true for any chordal graph whose complement has cardinality less than m. Let $G=(V,E)$ be a chordal graph such that $|\bar{E}|=m$ and let R be a partial distance matrix satisfying (4). Let S be an arbitrary minimal vertex separator of G. By Ex. 12, Chap. IV in [6], there exist vertices u and v belonging to different connected components of G_{V-S} with the property that $S\subset \mathrm{Adj}(u)$ and $S\subset \mathrm{Adj}(v)$.

We first prove that the graph $G'=(V,E\cup\{\{u,v\}\})$ is also chordal. Assume the contrary, which means that there exists a chordless cycle $[u,x_1,\ldots,x_k,v]$, $k\geq 2$, in

G'. By the definition of a minimal vertex separator, at least one $x_l \in S$, $1 \leq l \leq k$. This implies that $\{x_l, u\}, \{x_l, v\} \in E$, a contradiction, showing that G' is chordal.

Then $S \cup \{u, v\}$ is the unique maximal clique in G' that is not a clique in G. As in the proof of Theorem 3.3, consider the principal submatrix $R(S \cup \{u, v\})$ having only one pair of symmetrically placed unspecified entries. Complete $R(S \cup \{u, v\})$ to a distance matrix in $\mathbf{R}^k$ to obtain a partial distance matrix $\tilde{R}$ having G' as the graph of its specified entries.

If

$$\operatorname{rank}\left(\begin{pmatrix} 0 & e^T \\ e & R(S) \end{pmatrix}\right) < k+2,$$

by Remark 1 $R(S \cup \{u, v\})$ has more than one completion to a distance matrix in $\mathbf{R}^k$ and so R admits more than one completion.

If R satisfies (4), then $\tilde{R}$ constructed above is uniquely determined. Since any minimal vertex separator of G' contains a minimal vertex separator of G, $\tilde{R}$ also satisfies condition (4). By the assumption made for $m-1$, $\tilde{R}$ admits a unique completion to a distance matrix in $\mathbf{R}^k$. This implies that R also admits a unique completion to a distance matrix in $\mathbf{R}^k$. □

Let $0 < m < n$ be given integers. Since the graph $G = (V, E)$ with $E = \{\{i, j\} | 0 < |i - j| \leq m\}$ is chordal, Theorem 3.3 and Remark 1 have the following consequence in the "band" case.

COROLLARY 3.5. *Any partial distance matrix $R = (r_{ij})_{i,j=1}^n$ in $\mathbf{R}^k$, with r_{ij} specified if and only if $|i - j| \leq m$, admits a completion to a distance matrix in $\mathbf{R}^k$. Moreover, the completion is unique if and only if all the matrices*

$$\begin{pmatrix} 0 & e^T \\ e & R(l, \ldots, l+m-1) \end{pmatrix}$$

have rank $k+2$ *for any* $l = 1, \ldots, n-m+1$.

4. Connections with positive semidefinite completions.

LEMMA 4.1. *Let $A = (a_{ij})_{i,j=1}^n$ be a symmetric matrix such that $a_{ii} = 1$ for $i = 1, \ldots, n$. Then A is positive semidefinite if and only if there are n points $P_1, \ldots, P_n$ on a hypersphere of radius $\sqrt{2}/2$ in $\mathbf{R}^k$, $k = \operatorname{rank} A$, such that $d(P_i, P_j) = \sqrt{1 - a_{ij}}$ for any $i, j = 1, \ldots, n$.*

Proof. As remarked in the introduction, the existence of the points $P_1, \ldots, P_n$ and O in $\mathbf{R}^k$ such that $d(P_i, P_j) = \sqrt{1 - a_{ij}}$ and $d(O, P_i) = \sqrt{2}/2$ is equivalent to the condition that the Schur complement of the upper left $2 - by - 2$ principal submatrix $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ in the matrix

$$\begin{pmatrix} 0 & 1 & 1 & 1 & \ldots & 1 \\ 1 & 0 & \frac{1}{2} & \frac{1}{2} & \ldots & \frac{1}{2} \\ 1 & \frac{1}{2} & 0 & 1-a_{12} & \ldots & 1-a_{1n} \\ 1 & \frac{1}{2} & 1-a_{12} & 0 & \ldots & 1-a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \frac{1}{2} & 1-a_{1n} & 1-a_{2n} & \ldots & 0 \end{pmatrix}$$

is negative semidefinite and has rank k. A straightforward computation shows that this latter Schur complement equals $-A$. This completes the proof. □

By Lemma 4.1, the result in [8] on positive definite completion of partial positive definite matrices having a chordal graph can be translated into the following.

THEOREM 4.2. *Let $G = (V, E)$ be a chordal graph and $\{K_j\}_{j=1}^m$ the maximal cliques of G. Consider points $P_1, \ldots, P_n$ in $\mathbf{R}^n$ satisfying the following conditions:*

(i) *The distances $d(P_i, P_j)$ are specified if and only if $\{i, j\} \in E$.*

(ii) *Each of the subsets $\{P_l\}_{l \in K_j}$, $j = 1, \ldots, m$, lie on a hypersphere of radius R.*

Then the points $P_1, \ldots, P_n$ can be chosen to lie on a hypersphere of radius R.

Remark 2. The conclusion of Theorem 4.2 is not valid when the graph G is not chordal. Consider G, for example, to be the simple cycle of length 4 and points A, B, C, D such that $d(A, B) = d(B, C) = d(C, D) = 2$ and $d(A, D) = 0$. Then each of the pairs $\{A, B\}$, $\{B, C\}$, $\{C, D\}$, and $\{A, D\}$ lie on a sphere of radius 1, but the smallest radius of a sphere on which $A = D$, B, and C may lie is $2\sqrt{3}/3 > 1$.

As a particular case of Theorem 4.2 we obtain the following corollary, analogous to the result on positive semidefinite completion of banded partial matrices in [4].

THEOREM 4.3. *Let $0 < m < n$ be given integers, and consider points $P_1, \ldots, P_n$ in $\mathbf{R}^n$ satisfying the following conditions:*

(i) *The distances $d(P_i, P_j)$ are specified if and only if $|i - j| \leq m$.*

(ii) *Each of the subsets $\{P_k, \ldots, P_{k+m-1}\}$, $k = 1, \ldots, n - m + 1$, lie on a hypersphere of radius R.*

Then the points $P_1, \ldots, P_n$ can be chosen to lie on a hypersphere of radius R.

We note that an elementary, purely geometric proof of Theorem 4.3 can be provided.

Let us also mention the following positive semidefinite completion problem, which is still unsolved.

(P_1) Given is a partial positive semidefinite matrix $A = (a_{ij})_{i,j=1}^n$ such that the graph of the specified entries of A is not chordal. Determine necessary and sufficient conditions on A in order that A admits at least one positive semidefinite completion.

Without loss of generality we may assume that A has a unit diagonal, since otherwise we may apply a diagonal congruence.

Consider now the following distance problem.

(P_2) Let $G = (V, E)$ be a nonchordal graph and let $\{K_1, \ldots, K_m\}$ be the maximal cliques of G. Consider the points $P_1, \ldots, P_n$ in $\mathbf{R}^n$ satisfying the following conditions.

(C_1) The distances $d(P_i, P_j) = d_{ij}$ are specified if and only if $\{i, j\} \in E$.

(C_2) Each of the subsets $\{P_l\}_{l \in K_j}$, $j = 1, \ldots, m$, lie on a hypersphere of radius R.

Determine the hypersphere of minimum radius (if any) on which the points $P_1, \ldots, P_n$ satisfying (C_1) and (C_2) may lie.

The problems (P_1) and (P_2) are equivalent. Indeed, without loss of generality, we may assume that $R = \sqrt{2}/2$. Consider the partial positive semidefinite matrix A satisfying the conditions of Problem (P_1) and then consider the points $P_1, \ldots, P_n$ in R^n such that $d(P_i, P_j) = \sqrt{1 - a_{ij}}$ for any $(i, j) \in E$. Then, by Lemma 4.1, A admits a positive semidefinite completion if and only if the points $P_1, \ldots, P_n$ may be chosen to lie on a hypersphere of radius $\sqrt{2}/2$.

REFERENCES

[1] W. W. BARRETT, C. R. JOHNSON, AND M. LUNDQUIST, *Determinantal formulae for matrix completions associated with chordal graphs*, Linear Algebra Appl., 121 (1989), pp. 265–289.

[2] L. M. BLUMENTHAL, *Theory and Applications of Distance Geometry*, Clarendon Press, Oxford, 1953.

[3] D. CARLSON, *What are Schur complements anyway?*, Linear Algebra Appl., 74 (1986), pp. 257–275.

[4] H. DYM AND I. GOHBERG, *Extensions of band matrices with band inverses*, Linear Algebra Appl., 36 (1981), pp. 1–24.

[5] R. L. ELLIS AND D. LAY, *Rank-preserving extensions of band matrices*, Linear Multilinear Algebra, 26 (1990), pp. 147–179.

[6] M. C. GOLUMBIC, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.

[7] J. C. GOWER, *Properties of Euclidian and non-Euclidian distance matrices*, Linear Algebra Appl., 67 (1985), pp. 81–97.

[8] R. GRONE, C. R. JOHNSON, E. SA, AND H. WOLKOWICZ, *Positive definite completions of partial Hermitian matrices*, Linear Algebra Appl., 58 (1984), pp. 109–124.

[9] T. L. HAYDEN AND J. WELLS, *Approximation by matrices positive semidefinite on a subspace*, Linear Algebra Appl., 109 (1988), pp. 115–130.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 655–674, April 1995

021

FAST ALGORITHMS FOR CONFLUENT VANDERMONDE LINEAR SYSTEMS AND GENERALIZED TRUMMER'S PROBLEM*

HAO LU†

Abstract. Asymptotically fast algorithms for both dual confluent Vandermonde linear systems and generalized Trummer's problem are presented by using the divide and conquer method. It is shown that dual confluent Vandermonde linear systems can be solved in $O(n \log n \log p)$ operations and generalized Trummer's problem can be done in $O(np \log n \log \frac{n}{p})$ operations if fast polynomial multiplication and division are used. Also a fast algorithm for Hermite evaluation of rational functions is presented.

Key words. confluent Vandermonde linear system, generalized Hilbert matrix, generalized Trummer's problem, Hermite evaluation, Hermite interpolation, divide and conquer

AMS subject classifications. 65F05, 65Y05, 68C25

1. Introduction.

Let $a_0, a_1, \ldots, a_p$ be $p+1$ numbers, $n_0, n_1, \ldots, n_p$ be $p+1$ positive integers and $n = \sum_{i=0}^{p} n_i$. Consider dual confluent Vandermonde linear systems

$$V_c^T \mathbf{x} = \mathbf{c}, \tag{1}$$

where $V_c = (B_0, B_1, \ldots, B_p)$ is a confluent Vandermonde matrix, (see [5], [7], [8], [23]) and B_k is an $n \times n_k$ matrix with (i,j) entry

$$\left.\frac{d^{j-1}(x^{i-1})}{dx^{j-1}}\right|_{x=a_k}.$$

Linear systems (1) and confluent Vandermonde linear systems

$$V_c \mathbf{x} = \mathbf{c} \tag{2}$$

arise in various applications such as construction of quadrature formulae [2], [15], [20], [24] and approximation of linear functionals [3], [29].

In the early part of the 1970s, Björck et al. [6], [5] derived some $O(n^2)$ fast algorithms for (dual) Vandermonde linear systems and (dual) confluent Vandermonde linear systems. Vandermonde-like matrices [9], [17] and confluent Vandermonde-like matrices [18] are generalizations of Vandermonde matrices and confluent Vandermonde matrices, respectively, in which the monomials are replaced by arbitrary polynomials. For the case where the polynomials satisfy a three-term recurrence relation, in 1988 and 1990, Higham [17], [18] generalized the results to (dual) Vandermonde-like linear systems and (dual) confluent Vandermonde-like linear systems. It was shown recently that the operations of solving both Vandermonde linear systems and confluent Vandermonde linear systems can be further reduced if fast polynomial multiplication and division are used. Using the divide and conquer method, Lu [22], [23] presented an $O(n \log^2 n)$ algorithm and an $O(n \log n \log p)$ algorithm for Vandermonde linear systems and confluent Vandermonde linear systems (2), respectively, where throughout the paper, $\log m$ means $\log_2 m$ if $m \geq 2$ and 1 if $0 \leq m < 2$.

* Received by the editors September 30, 1993; accepted for publication (in revised form) by N. J. Higham, March 30, 1994. This work was partially supported by The Netherlands Organization for Scientific Research grant 611-302-025.

† Department of Mathematics, Faculty of Mathematics and Informatics, University of Nijmegen, Toernooiveld, 6525 ED Nijmegen, The Netherlands (na.hlu@na-net.ornl.gov).

Let H_p be a generalized Hilbert matrix with (i,j) entry

$$(H_p)_{ij} = \begin{cases} 1/(t_i - s_j)^p, & i \neq j,\ i,j = 1,2,\ldots,n, \\ 1/(t_i - s_i)^p, & t_i \neq s_i, \\ h_{pi}, & t_i = s_i, \end{cases} \tag{3}$$

where p is a positive integer, t_i, s_i, and h_{pi} are points in the complex plane satisfying $t_i \neq t_j$, $s_i \neq s_j$, $t_i \neq s_j$ for $i \neq j$, $i,j = 1,2,\ldots,n$. For the case $p = 1$, $t_i = s_i = c_i$, and $h_{1i} = 0$, we have a matrix H with (i,j) entry

$$H_{ij} = \begin{cases} \dfrac{1}{c_i - c_j}, & i \neq j, \\ 0, & i = j. \end{cases}$$

Let $\mathbf{b}$ be any n-vector. In 1985, Golub [13] posed Trummer's problem as follows:

> Give an algorithm for computing $H\mathbf{b}$ in less than $O(n^2)$ multiplications. If this is impossible, show that it cannot be done.

Two years later Gerasoulis et al. proposed an $O(n \log^2 n)$ algorithm for Trummer's problem, henceforth the GGS algorithm [12]. Extending the GGS algorithm to include the matrices defined in (3) in the case $t_i \neq s_j$, $i,j = 1,2,\ldots,n$, Gerasoulis showed the existence of a fast algorithm with $O(n \log^2 n)$ time complexity for the multiplications of generalized Hilbert matrices with vectors $H_1\mathbf{b}$, $H_2\mathbf{b}$ [11]. For the general case, I showed that there exists an $O(np \log n \log np)$ algorithm for the multiplication $H_p\mathbf{b}$ if $t_i \neq s_j$, $i,j = 1,2,\ldots,n$ [21].

Consider generalized Trummer's problem, i.e., the multiplications of generalized Hilbert matrices $H_1, H_2, \ldots, H_p$ with vectors

$$H_1\mathbf{b},\ H_2\mathbf{b},\ldots,\ H_p\mathbf{b}. \tag{4}$$

Various applications of the problem can be found in the computation of conformal mappings [30], the numerical evaluation of singular integrals [11], and particle simulations [14], [27].

The purpose of this paper is to construct asymptotically fast algorithms for both dual confluent Vandermonde linear systems (1) and generalized Trummer's problem (4) by using the divide and conquer method and incorporating the fast polynomial arithmetic.

Let $r(x) = p(x)/q(x)$ be a rational function, where $p(x)$ and $q(x)$ are polynomials of degree $\bar{n}$ and $\bar{m}$, respectively. In §2, we present a fast algorithm for Hermite evaluation of $r(x)$, i.e., the computation

$$r^{(k)}(a_i), \qquad k = 0,1,\ldots.n_i - 1, \quad i = 0,1,\ldots,p,$$

where $a_0, a_1, \ldots, a_p$ are $p+1$ points in complex plane and $n_0, n_1, \ldots, n_p$ are $p+1$ positive integers. The algorithm needs at most $O(N \log n + n \log n \log p)$ operations if $N \geq n$ and $O(n \log N \log \frac{Np}{n})$ operations if $N < n$, where $N = \max(\bar{n}, \bar{m})$ and $n = \sum_{i=o}^{p} n_i$. Using the results in §2, I present divide and conquer algorithms for dual confluent Vandermonde linear systems (1) in §3 and for generalized Trummer's problem (4) in §4. It is shown that dual confluent Vandermonde linear systems can

be solved in $O(n \log n \log p)$ operations and generalized Trummer's problem can be done in $O(np \log n \log \frac{n}{p})$ operations if fast polynomial multiplication and division are used. I derive also an $O(np \log np + n \log^2 n)$ algorithm for computing multiplication $H_p\mathbf{b}$ in the case $t_i \neq s_j$, $i, j = 1, 2, \ldots, n$ in §5. Finally, some comments on practical aspects of the implementation of the algorithms for Chebyshev points are made in §6.

Let $A(x)$ and $B(x)$ be two polynomials. For convenience, $\mathrm{quot}(A(x), B(x))$ denotes the quotient of polynomial division $A(x)/B(x)$, i.e., ignoring the remainder $r(x)$: $A(x) = B(x)\mathrm{quot}(A(x), B(x)) + r(x)$, throughout the paper.

2. Hermite evaluation of rational functions. Let $r(x) = p(x)/q(x)$ be a rational function, where $p(x)$ and $q(x)$ are polynomials of degree $\bar{n}$ and $\bar{m}$, respectively. For given $p+1$ numbers $a_0, a_1, \ldots, a_p$ and $p+1$ positive integers $n_0, n_1, \ldots, n_p$, Hermite evaluation of $r(x)$ is to compute

$$r^{(j)}(a_i), \qquad j = 0, 1, \ldots, n_i - 1, \quad i = 0, 1, \ldots, p.$$

In this section, we present a divide and conquer algorithm for the problem.

2.1. Algorithm. To construct an asymptotically fast algorithm for the problem, we expand $r(x)$, $p(x)$, and $q(x)$ in Taylor series at a_i, i.e.,

$$\begin{aligned} r(x) &= \sum_{j=0}^{n_i-1} r_{ij}(x - a_i)^j + O((x - a_i)^{n_i}), \\ p(x) &= \sum_{j=0}^{n_i-1} p_{ij}(x - a_i)^j + O((x - a_i)^{n_i}), \\ q(x) &= \sum_{j=0}^{n_i-1} q_{ij}(x - a_i)^j + O((x - a_i)^{n_i}). \end{aligned}$$

Comparing the coefficients of $(x - a_i)^j$ in $r(x)q(x) = p(x)$ shows that the coefficient vector $\mathbf{r}_i = (r_{i0}, r_{i1}, \ldots, r_{i,n_i-1})^T$ is the solution of the following triangular Toeplitz linear system

$$T_i\mathbf{r}_i = \mathbf{p}_i, \tag{5}$$

where T_i is a triangular Toeplitz matrix of the form

$$T_i = \begin{pmatrix} q_{i0} & & & \\ q_{i1} & q_{i0} & & \\ \vdots & \ddots & \ddots & \\ q_{i,n_i-1} & \cdots & q_{i1} & q_{i0} \end{pmatrix},$$

or simply $\mathrm{triT}(q_{i0}, q_{i1}, \ldots, q_{i,n_i-1})$ for convenience, and $\mathbf{p}_i = (p_{i0}, p_{i1}, \ldots, p_{i,n_i-1})^T$. Hence, we have

$$r^{(j)}(a_i) = j!r_{ij}, \qquad j = 0, 1, \ldots, n_i - 1, \quad i = 0, 1, \ldots, p.$$

To solve triangular Toeplitz linear systems (5) we need to compute the Hermite evaluation of polynomials $p(x)$ and $q(x)$. Let $N = \max(\bar{n}, \bar{m})$. Without loss of

generality, assume $n = \sum_{i=0}^{p} n_i = 2^k$ for some nonnegative integer k. Set

$$b_{m_i+j} = a_i, \qquad m_i = \sum_{t=0}^{i-1} n_t, \qquad j = 1, 2, \ldots, n_i, \quad i = 0, 1, \ldots, p,$$

$$T_{0i}(x) = x - b_i, \qquad i = 1, 2, \ldots, n,$$

$$T_{ji}(x) = T_{j-1,2i-1}(x) T_{j-1,2i}(x), \qquad i = 1, 2, \ldots, 2^{k-j}, \quad j = 1, 2, \ldots, k. \tag{6}$$

With an essential modification of Algorithm 3.2 in [23] for reducing operations we construct an asymptotically fast algorithm for Hermite evaluation of rational functions. Note that it is not hard to construct an algorithm for the problem through a straightforward modification of the algorithm in [23], though the computational complexity needs to be further estimated. The main difference between the two algorithms is how to compute $r^{(i)}(a_l)$, $i = 0, 1, \ldots, n_l - 1$ after finding a proper l. Both algorithms compute $r^{(i)}(a_l)$, $i = 0, 1, \ldots, n_l - 1$ by using polynomials, but the new one avoids using polynomials of high degree. For example, if $n_0 = n_1 = \cdots = n_p = 1$. The following algorithm obtains $r(a_i)$, $i = 0, 1, \ldots, n - 1$ from polynomials of degree zero while the old algorithm evaluates polynomials of high degree for the same purpose. However, the complexity analysis of our new algorithm becomes complicated as we will see.

ALGORITHM HERF.
Given a rational function $r(x) = p(x)/q(x)$, $p+1$ numbers $a_0, a_1, \ldots, a_p$ and $p+1$ positive integers $n_0, n_1, \ldots, n_p$, where $p(x)$ and $q(x)$ are polynomials of degree $\bar{n}$ and $\bar{m}$, respectively, the following algorithm computes Hermite evaluation of $r(x)$, i.e., $r^{(j)}(a_i)$, $j = 0, 1, \ldots, n_i - 1$, $i = 0, 1, \ldots, p$.

Stage I. $b_{m_i+j} = a_i$, $m_0 = 0$, $m_i = \sum_{t=0}^{i-1} n_t$, $j = 1, \ldots, n_i$, $i = 0, \ldots, p$
$T_{0i} = x - b_i$, $i = 1, 2, \ldots, n$, $p_{k1} = p(x)$, $q_{k1} = q(x)$
$S_{k1} = \{0, 1, \ldots, p\}$
For $j = 1 : 1 : k$
  If $N + 1 \geq 2^j$ then
    For $i = 1 : 1 : 2^{k-j}$
      $T_{ji} = T_{j-1,2i-1} T_{j-1,2i}$
    endfor i
  endif
endfor j
if $N \geq n$ then $Q_{k1} = \text{quot}(x^{2^{k+1}-1}, T_{k1})$
  if $\bar{n} \geq n$ then p_{k1}=Div$(p_{k1}, T_{k1}, Q_{k1}, n)$
  endif
  if $\bar{m} \geq n$ then q_{k1}=Div$(q_{k1}, T_{k1}, Q_{k1}, n)$
  endif
endif
Stage II. For $j = k : -1 : 1$
  For $i = 1 : 1 : 2^{k-j}$
    if $S_{ji} = \{l\}$ then Call Value$(p_{ji}, q_{ji}, a_l, n_l)$
      for $m = 0 : 1 : n_l - 1$
        $r^{(m)}(a_l) = i! u_m$
      endfor m

$S_{ji} = \emptyset$
elseif $S_{ji} \neq \emptyset$ then
if $2^{j-1} \leq N < 2^j$ then
for $m = 2i-1, 2i$
$Q_{j-1,m} = \text{quot}(x^{2^j-1}, T_{j-1,m})$
endfor m
elseif $N \geq 2^j$ then $Q_{j-1,2i-1} = \text{quot}(T_{j-1,2i}Q_{ji}, x^{2^j})$
$Q_{j-1,2i} = \text{quot}(T_{j-1,2i-1}Q_{ji}, x^{2^j})$
endif
if $\bar{n} < 2^{j-1}$ then $p_{j-1,2i-1} = p_{j-1,2i} = p_{ji}$
else for $m = 2i-1, 2i$
$p_{j-1,m} = \text{Div}(p_{ji}, T_{j-1,m}, Q_{j-1,m}, 2^{j-1})$
endfor m
endif
if $\bar{m} < 2^{j-1}$ then $q_{j-1,2i-1} = q_{j-1,2i} = q_{ji}$
else for $m = 2i-1, 2i$
$q_{j-1,m} = \text{Div}(q_{ji}, T_{j-1,m}, Q_{j-1,m}, 2^{j-1})$
endfor m
endif
if $b_{(2i-1)2^{j-1}} = b_{(2i-1)2^{j-1}+1} = a_l$ and $a_l \in S_{ji}$ then
call Value$(p_{ji}, q_{ji}, a_l, n_l)$
for $m = 0:1:n_l - 1$
$r^{(m)}(a_l) = i!u_m$
endfor m
$S_{j-1,2i-1}$=$\{t\colon t \in S_{ji}$ and $t < l\}$
$S_{j-1,2i}$=$\{t\colon t \in S_{ji}$ and $t > l\}$
elseif $b_{(2i-1)2^{j-1}} = a_l$ then
$S_{j-1,2i-1}$=$\{t\colon t \in S_{ji}$ and $t \leq l\}$
$S_{j-1,2i}$=$\{t\colon t \in S_{ji}$ and $t > l\}$
endif
endif
endfor i
endfor j

FUNCTION Div$(A(x), B(x), Q(x), n)$. Let $A(x)$ and $B(x)$ be two polynomials of degree $\bar{n}$ and n, respectively, and $Q(x)$ be preprocessing of $B(x)$. Based on Proposition 2.1 (see next page). Function Div computes the remainder $R(x)$ of division $A(x)/B(x)$ and return $R(x)$.

$K(x) = \text{quot}(A(x), x^n)$, $P(x) = \text{quot}(K(x)Q(x), x^{n-1})$
$R(x) = A(x) - P(x)B(x)$
Return $R(x)$
end

ALGORITHM Value$(A(x), B(x), n, a)$. Let n be a positive integer, a be a complex number, $A(x)$ and $B(x)$ be two polynomials. Algorithm Value computes $R^{(m)}(a)/m!$, $m = 0, 1, \ldots, n-1$, where $R(x) = A(x)/B(x)$.

Compute $a_i = \frac{1}{i!}A^{(i)}(a)$, $b_i = \frac{1}{i!}B^{(i)}(a)$, $i = 0, 1, \ldots, n-1$
Solve triangular Toeplitz linear system

$$\mathrm{triT}(b_0, b_1, \ldots, a_{n-1})\mathbf{u} = (a_0, a_1, \ldots, a_{n-1})^T$$

We use the preprocessing of $T_{ij}(x)$ in the algorithm for the sake of further reducing the operations. Given a polynomial $B(x)$ of degree n, by preprocessing $B(x)$, we define the computation of the quotient of x^{2n-1} divided by $B(x)$. Assume $Q_{ji}(x)$ is the quotient of $x^{2^{j+1}-1}$ divided by $T_{ji}(x)$. It is not hard to check from (6) that

$$Q_{j-1,2i-1}(x) = \mathrm{quot}(T_{j-1,2i}Q_{ji}(x), x^{2^j}),$$
$$Q_{j-1,2i}(x) = \mathrm{quot}(T_{j-1,2i-1}(x)Q_{ji}(x), x^{2^j})$$

are the quotients of x^{2^j-1} divided by $T_{j-1,2i}(x)$ and of x^{2^j-1} divided by $T_{j-1,2i}(x)$, respectively.

The correctness of Function Div is based on the following proposition [25] on the preprocessing of polynomials.

PROPOSITION 2.1. *Let*

$$A(x) = \sum_{i=0}^{\tilde{n}} a_i x^i, \qquad B(x) = \sum_{i=0}^{n} b_i x^i, \quad (a_{\tilde{n}} \neq 0,\ b_n \neq 0),$$

$D(x)$ be the result of preprocessing $B(x)$ and $K(x) = \mathrm{quot}(A(x), x^n)$. Then

$$Q(x) = \mathrm{quot}(D(x)K(x), x^{n-1}), \qquad R(x) = A(x) - Q(x)B(x)$$

are the quotient and the remainder of division $A(x)/B(x)$, respectively.

2.2. Correctness and computational complexity. We now prove the correctness and analyze the computational complexity for HERF. To this end, we need the following two propositions. The first one is to estimate the computational cost of Value.

PROPOSITION 2.2. *Let $A(x)$ and $B(x)$ be two polynomials of degree $\tilde{n} - 1$ and $\tilde{m} - 1$, respectively, and $m = \max(\tilde{n}, \tilde{m})$. Algorithm* Value$(A(x), B(x), n, a)$ *needs at most $\tilde{C}\max(n, m)\log(\min(n, m))$ operations if fast polynomial multiplication and division are used, where $\tilde{C}$ is a positive constant independent of n and m.*

Proof. If $m \geq n$, it follows from [23] (see the proof of Proposition 4.2 [23] for details) that the computation, $a_i = A^{(i)}(a)/i!$, $b_i = B^{(i)}(a)/i!$, $i = 0, 1, \ldots, n-1$, needs at most $\tilde{C}_1 m \log n$ operations. Triangular Toeplitz linear systems can actually be solved by fast polynomial division that needs $\tilde{C}_2 n \log n$ operations (see, e.g., [4]), where $\tilde{C}_1$ and $\tilde{C}_2$ are positive constants independent of n and m. Choosing $\tilde{C} = \tilde{C}_1 + \tilde{C}_2$ finishes the proof for the case $m \geq n$.

If $n > m$, a_i's and b_i's can be obtained in $\tilde{C}_1 m \log m$ operations because $a_i = 0$ and $b_i = 0$ if $i \geq m$. In this case, the coefficient matrix of the linear system included in Value is of the form $T = \mathrm{triT}(a_0, a_1, \ldots, a_{m-1}, 0, \ldots, 0)$. Without loss of generality, we assume that $n = tm$ for some positive integer t. To solve triangular Toeplitz linear systems

$$T\mathbf{u} = \mathbf{d}, \tag{7}$$

we partition T into a $t \times t$ block form

$$T = \begin{pmatrix} T_0 & & & \\ T_1 & T_0 & & \\ & \ddots & \ddots & \\ & & T_1 & T_0 \end{pmatrix},$$

where $T_0 = \text{triT}(a_0, a_1, \ldots, a_{m-1})$ and $T_1 = \text{triT}(0, a_{m-1}, \ldots, a_1)^T$, and partition $\mathbf{u}$ and $\mathbf{d}$ consistently with T, i.e.,

$$\mathbf{u} = \begin{pmatrix} \mathbf{u}_0 \\ \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_{t-1} \end{pmatrix}, \qquad \mathbf{d} = \begin{pmatrix} \mathbf{d}_0 \\ \mathbf{d}_1 \\ \vdots \\ \mathbf{d}_{t-1} \end{pmatrix},$$

where $\mathbf{u}_i$ and $\mathbf{d}_i$ are m-vectors. Let $\mathbf{y} = (y_0, y_1, \ldots, y_{m-1})^T$ be the solution of

$$T_0\mathbf{y} = \mathbf{e}_1 = (1, 0, \ldots, 0)^T.$$

One can easily check that $T_0^{-1} = \text{triT}(y_0, y_1, \ldots, y_{m-1})$. Now, we solve triangular Toeplitz linear systems (7) as follows:

Solve $T_0\mathbf{y} = \mathbf{e}_1$, $\mathbf{u}_0 = T_0^{-1}\mathbf{d}_0$
for $i = 1 : 1 : t-1$
  $\mathbf{u}_i = T_0^{-1}(\mathbf{d}_i - T_1\mathbf{u}_{i-1})$,
endfor i.

As mentioned above, $T_0\mathbf{y} = \mathbf{e}_1$ can be solved in $\tilde{C}_2\, m \log m$ operations. Since multiplication of triangular Toeplitz matrices with vectors can be computed by polynomial multiplication, $T_1\mathbf{u}_{i-1}$ and $\mathbf{u}_i$ can be obtained by using multiplication by two polynomials of degree m, respectively. Hence, Value needs at most

$$\tilde{C}_1\, m \log m + \tilde{C}_2\, m \log m + 2\tilde{C}_3 \sum_{i=0}^{t-1} m \log m = \tilde{C}\, n \log m$$

operations, where $\tilde{C}_3$ is a positive constant independent of n and m and $\tilde{C} = \tilde{C}_1 + \tilde{C}_2 + 2\tilde{C}_3$. □

PROPOSITION 2.3. *Let $n = 2^k$ for some nonnegative integer k, $p \le n$ be a nonnegative integer and*

$$C(n,p) \le C\left(\frac{n}{2}, p_1\right) + C\left(\frac{n}{2}, p_2\right) + C_1 \frac{n}{2} \log \frac{n}{2}, \qquad C(n,0) \le C_2\, n \log n, \tag{8}$$

where p_1, p_2 are two nonnegative integers such that $p_1 \le \frac{n}{2}$, $p_2 \le \frac{n}{2}$, and $p_1 + p_2 \le p$, and C_1 and C_2 are two positive constants independent of n and p. Then

$$C(n,p) \le C\, n \log n (\log(p+1) + 1), \tag{9}$$

where $C = \max(C_1, C_2)$.

Proof. We prove the proposition by induction on $n + p$. $C(n,0) \le C_2\, n \log n$ implies that (9) holds for $n + p = 1$.

Under the assumption that $\log 1$ means 1 as mentioned in the intr duction of the paper, $n \log n\, (\log(p+1) + 1) \le 2n \log n \log(p+1)$ for $p \ge 0$. If $p_1 = 0$, using (8) and induction hypothesis shows that

$$\begin{aligned} C(n,p) &\le C\left(\frac{n}{2}, 0\right) + C\left(\frac{n}{2}, p_2\right) + C_1 \frac{n}{2} \log \frac{n}{2} \\ &\le (C_1 + C_2) \frac{n}{2} \log \frac{n}{2} + C \frac{n}{2} \log \frac{n}{2} (\log(p_2 + 1) + 1) \\ &\le Cn \log \frac{n}{2} + C\, n \log \frac{n}{2} \log(p_2 + 1) \\ &\le C\, n \log n\, (\log(p+1) + 1). \end{aligned}$$

If $p_2 = 0$, (9) is derived in a similar way.

Otherwise, both $p_1 \geq 1$ and $p_2 \geq 1$. It is easy to check that $(2(p_1+1)(p_2+1))^{\frac{1}{2}} \leq p+1$. It follows from (8) and induction hypothesis that

$$\begin{aligned} C(n,p) &\leq C\frac{n}{2}\log\frac{n}{2}(\log(p_1+1)+1) + C\frac{n}{2}\log\frac{n}{2}(\log(p_1+1)+1) + C_1\frac{n}{2}\log\frac{n}{2} \\ &\leq C\,n\log\frac{n}{2}\log(2(p_1+1)(p_2+1))^{\frac{1}{2}} + (C+C_1)\frac{n}{2}\log\frac{n}{2} \\ &\leq C\,n\log n(\log(p+1)+1), \end{aligned}$$

which completes the proof. □

THEOREM 2.4. *Given a rational function $r(x) = p(x)/q(x), p+1$ numbers $a_0, a_1, \ldots, a_p$ and $p+1$ positive integers $n_0, n_1, \ldots, n_p$, where $p(x)$ and $q(x)$ are polynomials of degree of $\bar{n}$ and $\bar{m}$, respectively, Algorithm* HERF *computes Hermite evaluation, $r^{(j)}(a_i)$, $j = 0, 1, \ldots, n_i - 1$, $i = 0, 1, \ldots, p$. Furthermore, if fast polynomial multiplication and division are used, the algorithm needs only $O(N\log n + n\log n\log p)$ operations if $N \geq n$ and $O(n\log N\log\frac{Np}{n})$ operations if $N < n$, where $N = \max(\bar{n}, \bar{m})$ and $n = \sum_{i=0}^{p} n_i$.*

Proof. We prove the theorem for the case $N = n-1$ first. The correctness is essentially the same as that of Algorithm 3.2 in [23].

It follows from [23] that Stage I needs at most $O(n\log n\log p)$ operations (see [23, Prop. 4.1]). Denote by $K(n,p)$ the number of operations needed by Stage II. If $p = 0$, $S_{k1} = \{0\}$. The algorithm calls only Value$(p_{k1}(x), q_{k1}(x), n, a_0)$. Proposition 2.2 shows that $K(n,0) \leq \tilde{C}\,n\log n$.

Otherwise, Stage II divides Hermite evaluation of a rational function, denoted by HERF(n,p) temporarily, into two subproblems HERF$(\frac{n}{2}, p_1)$ and HERF$(\frac{n}{2}, p_2)$.

Note that function Div performs actually polynomial multiplication twice. Hence, for fixed i and j, Stage II performs multiplication of polynomials of degree at most 2^{k-i} finite times independent of i and j, polynomial division at most twice (this case occurs only if $j = k$) and the Algorithm Value at most once. Incorporating the fast polynomial arithmetic and using Proposition 2.2 show that there exists a positive constant $\bar{C}$ independent of n and p such that

$$K(n,p) \leq \bar{C}\frac{n}{2}\log\frac{n}{2} + K\left(\frac{n}{2}, p_1\right) + K\left(\frac{n}{2}, p_2\right),$$

where $p_1 + p_2 \leq p$. Proposition 2.3 shows that $K(n,p) \leq C\,n\log n(\log(p+1)+1)$, where $C = \max(\tilde{C}, \bar{C})$. Hence, the overall computational cost is $O(n\log n\log p)$ if $N = n-1$.

If $N \geq n$, we need only to add the computation $Q_{k1} = \text{quot}(x^{2^{k+1}}, T_{k1})$, $p_{k1} = \text{Div}(p_{k1}, T_{k1}, Q_{k1}, n)$ if $\bar{n} \geq n$ and $q_{k1} = \text{Div}(q_{k1}, T_{k1}, Q_{k1}, n)$ if $\bar{m} \geq n$. After the computation, we have

$$r^{(m)}(a_i) = \left(\frac{p_{k1}(x)}{q_{k1}(x)}\right)^{(m)}\Bigg|_{x=a_i}, \qquad m = 0, 1, \ldots, n_i - 1, \quad i = 0, 1, \ldots, p.$$

Since now $\max(\deg(p_{k1}(x)), \deg(q_{k1}(x))) \leq n-1$, the correctness of the algorithm for $N \geq n$ follows immediately. On the other hand, Q_{k1} can be computed by fast polynomial division and new $p_{k1}(x)$ and $q_{k1}(x)$ can be obtained by polynomial multiplication at most four times as we mentioned above. The operations needed are $O(n\log n + N\log n)$. Thus, the overall cost is $O(N\log n + n\log n\log p)$ operations.

If $N < n$, assume that $2^m > N+1 \geq 2^{m-1}$ for some positive integer m. Recalling our algorithm, one finds that except the calculation

$$\left(\frac{p(x)}{q(x)}\right)^{(t)}\Bigg|_{x=a_{l_i}}, \qquad t = 0, 1, \ldots, n_{l_i} - 1 \tag{10}$$

if $b_{i2^{m-1}} = b_{i2^{m-1}+1} = a_{l_i}$ the computation of the algorithm is to compute Hermite evaluation of the form

$$\left(\frac{p(x)}{q(x)}\right)^{(t)}\Bigg|_{x=a_{j_i}}, \qquad t = 0, \ldots, n_{j_i} - 1, \quad j_i = q_{i_1}, \ldots, q_{i_2}, \quad i = 1, \ldots, M, \tag{11}$$

where $q_{i_2} > q_{i_1}$ and $M \leq 2^{k+1-m}$,

$$q_{i_1} = \begin{cases} 0 & \text{if } i = 1, \\ l_{i-1} + 1 & \text{if } i > 1, \end{cases}$$

$$q_{i_2} = \begin{cases} l_i & \text{if } b_{i2^{m-1}} = a_{l_i} \text{ and } b_{i2^{m-1}+1} \neq a_{l_i}, \\ l_i - 1 & \text{if } b_{i2^{m-1}} = b_{i2^{m-1}+1} = a_{l_i}. \end{cases}$$

Thus, the correctness of the algorithm becomes clear.

Let $\tilde{k}$ be the number of all i which satisfy $b_{i2^{m-1}} = b_{i2^{m-1}+1}$. Apparently,

$$\tilde{k} \leq 2^{k+1-m}, \qquad 2^{k-m} \leq \tilde{k} + M < 2^{k-m+2},$$

$$\tilde{k} + \sum_{i=1}^{M} (q_{i_2} - q_{i_1} + 1) = p + 1.$$

It follows from the previous proof of the theorem that for fixed i computing (11) needs

$$\begin{aligned} &\bar{C}_1 N(m-1) + \bar{C}_2 2^{m-1}\,(m-1)\log(q_{i_2} - q_{i_1} + 1) \\ &\leq \bar{C}_1 N \log N + \bar{C}_2 2^{m-1}\,\log N \log(q_{i_2} - q_{i_1} + 1) \end{aligned}$$

operations, where $\bar{C}_1$ and $\bar{C}_2$ are positive constants. Proposition 2.2 shows that operations for computing (10) are $\tilde{C}\,\max(N, n_{l_i})\log(\min(N, n_{l_i})) \leq \tilde{C}\,\max(N, n_{l_i})\log N$. Hence, the overall operations are bounded by

$$\begin{aligned} &\sum_{i=1}^{\tilde{k}} \tilde{C}\,\max(N, n_{l_i})\log N + \log N \sum_{i=1}^{M} (\bar{C}_1 N + \bar{C}_2 2^{m-1}\log(q_{i_2} - q_{i_1} + 1)) \\ &\leq \tilde{C} \sum_{i=1}^{\tilde{k}} (N + n_{l_i})\log N + \bar{C}_1 MN \log N + \bar{C}_2\, 2^{m-1} \log N \sum_{q_{i_2} - q_{i_1} = 1} 1 \\ &\quad + \bar{C}_2\, 2^{m-1} \log N \sum_{q_{i_2} - q_{i_1} > 1} \log(q_{i_2} - q_{i_1} + 1) \\ &\leq \tilde{C}\tilde{k} N \log N + \tilde{C} \sum_{i=1}^{\tilde{k}} n_{l_i} \log N + \bar{C}_1 MN \log N + \bar{C}_2 2^{m-1} M \log N \end{aligned}$$

$$+ \bar{C}_2(M+\tilde{k})2^{m-1}\log N \log \left(\prod_{q_{i_2}-q_{i_1}>1} (q_{i_2}-q_{i_1}+1) \right)^{\frac{1}{M+\tilde{k}}}$$

$$\leq (3\tilde{C}+2\bar{C}_1+\bar{C}_2)\, n\log N + 2\bar{C}_2\, n\log N \log \frac{\sum_{i=1}^{M}(q_{i_2}-q_{i_1}+1)+\tilde{k}}{M+\tilde{k}}$$

$$\leq (3\tilde{C}+2\bar{C}_1+\bar{C}_2)\, n\log N + 2\bar{C}_2\, n\log N \log \frac{2Np}{n}$$

$$= \bar{C}\, n\log N + 2\bar{C}_2\, n\log N \log \frac{Np}{n},$$

where $\bar{C} = 3\tilde{C} + 2\bar{C}_1 + 3\bar{C}_2$. □

The Algorithm HERF can easily be adjusted for Hermite evaluation of polynomials with a straightforward modification. When HERF is used for Hermite evaluation of polynomials $p^{(j)}(a_i)$, $j = 0, 1, \ldots, n_i - 1$, $i = 0, 1, \ldots, p$, we denote it by $\mathrm{HEP}(p(x), (a_i, n_i), i = 0, 1, \ldots, p)$.

3. Solution of confluent Vandermonde linear systems (dual). Let $V_c = (B_0, B_1, \ldots, B_p)$ be a confluent Vandermonde matrix, where B_k is an $n \times n_k$ matrix with (i, j) entry

$$\left.\frac{d^{j-1}(x^{i-1})}{dx^{j-1}}\right|_{x=a_k}$$

and $n = \sum_{i=0}^{p} n_i$, $a_i \neq a_j$, $i \neq j$, $i, j = 0, 1, \ldots, p$. Consider dual confluent Vandermonde linear systems

$$V^T \mathbf{x} = \mathbf{c}. \tag{12}$$

It is shown that confluent Vandermonde matrix V_c is nonsingular if and only if $a_i \neq a_j$, $i \neq j$, $i, j = 0, 1, \ldots, p$ (see [23, Prop. 2.2]). Hence, (12) has a unique solution. Note that $p(x) = x_1 + x_2 x + \cdots + x_n x^{n-1}$ satisfies interpolating condition

$$p^{(j)}(a_i) = c_{m_i+j+1}, \qquad j = 0, 1, \ldots, n_i - 1, \quad i = 0, 1, \ldots, p, \tag{13}$$

where $m_0 = 0$, $m_i = \sum_{j=0}^{i-1} n_j$. Thus, $p(x)$ is a Hermite interpolating polynomial of degree at most $n-1$ for the data $\{a_i, c_{m_i+j+1}\}$, $i = 0, 1, \ldots, p$, $j = 0, 1, \ldots, n_i - 1$. Our purpose is to determine its coefficients of x^i. Let

$$l(x) = (x-a_0)^{n_0}(x-a_1)^{n_1}\cdots(x-a_p)^{n_p},$$

$$l_i(x) = l(x)/(x-a_i)^{n_i}, \qquad i = 0, 1, \ldots, p.$$

As the first step, we represent $p(x)$ as

$$p(x) = \sum_{i=0}^{p} f_i(x) l_i(x),$$

where $f_i(x)$ is a polynomial of degree at most $n_i - 1$. Equation (13) shows that

$$\left.(f_i(x)l_i(x))^{(j)}\right|_{x=a_i} = c_{m_i+j+1}, \qquad j = 0, 1, \ldots, n_i - 1, \quad i = 0, 1, \ldots, p. \tag{14}$$

To obtain the representation of $p(x)$ in terms of the basis $\{x^i\}_{i=0}^{n-1}$, we need to compute the coefficients of $f_i(x)$. Expand $f_i(x)$, $l_i(x)$ in Taylor series at a_i,

$$f_i(x) = p_{i0} + p_{i1}(x-a_i) + \cdots + p_{i,n_i-1}(x-a_i)^{n_i-1},$$

$$l_i(x) = l_{i0} + l_{i1}(x-a_i) + \cdots + l_{i,n_i-1}(x-a_i)^{n_i-1} + O((x-a_i)^{n_i}).$$

Equation (14) implies that $\mathbf{u}_i = (p_{i0}, p_{i1}, \ldots, p_{i,n_i-1})^T$ is the solution of the following triangular Toeplitz linear systems

$$T_i \mathbf{u}_i = \mathbf{d}_i,$$

where T_i is a triangular Toeplitz matrix of the form $T_i = \mathrm{triT}(l_{i0}, l_{i1}, \ldots, l_{i,n_i-1})$ and $\mathbf{d}_i$ is an n_i-vector given by $\mathbf{d}_i = (c_{m_i+1}, c_{m_i+2}, \frac{1}{2!}c_{m_i+3}, \ldots, \frac{1}{(n_i-1)!}c_{n_i})^T$. Furthermore, relation $l_i(x)(x-a_i)^{n_i} = l(x)$ implies that

$$l_{ij} = \frac{l_i^{(j)}(a_i)}{j!} = \frac{l^{(n_i+j)}(a_i)}{(n_i+j)!}, \qquad j = 0, 1, \ldots, n_i - 1.$$

Let $g_0(x), g_1(x), \ldots, g_q(x)$ be $q(\leq p)$ polynomials over the complex number field. For convenience, notation $p(g_i(x), q, (c_j, n_j), p)$ stands for a polynomial of the form $\sum_{i=0}^{q} g_i(x) l_i(x)$. Let $\mathbf{v} = (v_0, v_1, \ldots, v_{m-1})^T$ be an m-vector. Notation $\mathbf{v}(x)$ denotes the polynomial $\sum_{i=0}^{m-1} v_i x^i$. Using this notation, we now present the following algorithm for solving the dual confluent Vandermonde linear systems (12), or for representing a Hermite interpolating polynomial $p(x)$ of degree at most $n-1$ such that

$$p^{(k)}(a_i) = c_{m_i+k+1}, \qquad k = 0, 1, \ldots, n_i - 1, \quad i = 0, 1, \ldots, p$$

in terms of the basis $\{x^i\}$.

ALGORITHM SCV. Let $V_c = (B_0, B_1, \ldots, B_p)$ be a confluent Vandermonde matrix, where B_k is an $n \times n_k$ matrix with (i,j) entry

$$\left.\frac{d^{j-1}(x^{i-1})}{dx^{j-1}}\right|_{x=a_k},$$

where $n = \sum_{i=0}^{p} n_i$, $a_i \neq a_j$, $i \neq j$, $i, j = 0, 1, \ldots, p$. The following algorithm solves the dual confluent Vandermonde linear systems (12). Without loss of generality, we assume here $n = 2^k$ for some positive integer k.

Stage I. Call HEP $(l(x), (a_i, n_i), i = 0, 1, \ldots, p)$
  For $i = 0 : 1 : p$
    For $j = 0 : 1 : n_i - 1$
      $l_{ij} = \frac{l^{(n_i+j)}(a_i)}{(n_i+j)!}$
    endfor j
    $T_i = \mathrm{triT}(l_{i0}, l_{i1}, \ldots, l_{i,n_i-1})$
    $\mathbf{d}_i = (c_{m_i+1}, c_{m_i+2}, \ldots, \frac{c_{n_i}}{(n_i-1)!})^T$
    Solve triangular Toeplitz linear system $T_i \mathbf{u}_i = \mathbf{d}_i$
    $f_i(x)$=Convert$(\mathbf{u}_i(x-a_i), n_i - 1)$
  endfor i

Stage II. $S = \{0, 1, \ldots, p\}$, $p(x) = (f_i(x), p, (a_j, n_j), p)$
  Call CHIP$(p(x), S)$

ALGORITHM CHIP$(p(x), S)$. Let $p(x) = p(f_i(x), q, (c_j, n_j), p)$. The Algorithm CHIP$(p(x), S)$ converts $p(x)$ into the form $\sum h_i x^i$.

For $j = k : -1 : 1$
  For $i = 1 : 1 : 2^{k-j}$
    if $b_{(2i-1)2^{j-1}} = a_l$ and $l \in S$ then
      if $b_{(2i-1)2^{j-1}+1} = a_l$ then
        $R_l(x) = f_l(x)T_{ij}(x)/(x - a_l)^{n_l}$
        $S_1 = \{t \in S,\ t < l\}$
        if $S_1 \neq \emptyset$ then $n_l = 2^{j-1} - \sum_{t=0}^{l-1} n_t$, $q_1 = l - 1$,
          $p_1 = l$, $p_1(x) = (f_i(x), q_1, (a_j, n_j), p_1)$
          Call CHIP$(p_1(x), S_1)$
        else $p_1(x) = 0$
        endif
        $S_2 = \{t,\ t \geq 0,\ t + l + 1 \in S\}$
        if $S_2 \neq \emptyset$ then $n_{p-l} = 2^{j-1} - \sum_{t=l+1}^{p} n_t$
          for $m = 0 : 1 : q - l - 1$
            $f_i(x) = f_{l+1+i}(x)$
          endfor m
          for $m = 0 : 1 : p - l - 1$
            $a_i = a_{l+1+i}$, $n_i = n_{l+1+i}$
          endfor m
          $q_2 = q - l - 1$, $p_2 = p - l$
          $p_2(x) = (f_i(x), q_2, (a_j, n_j), p_2)$
          Call CHIP$(p_2(x), S_2)$
        else $p_2(x) = 0$
        endif
        $p(x) = p_1(x)T_{1-1,2j} + p_2(x)T_{i-1,2j-1} + R_l(x)$
      else $S_1 = \{t,\ t \leq l,\ t \in S\}$, $q_1 = l$, $p_1 = l$
        $p_1(x) = (f_i(x), q_1, (a_j, n_j), p_1)$
        Call CHIP$(p_1(x), S_1)$
        $S_2 = \{t,\ t \geq 0,\ t + l + 1 \in S\}$
        if $S_2 \neq \emptyset$
          for $m = 0 : 1 : q - l - 1$
            $f_i(x) = f_{l+1+i}(x)$
          endfor m
          for $m = 0 : 1 : p - l - 1$
            $a_i = a_{l+1+i}$, $n_i = n_{l+1+i}$
          endfor m
          $q_2 = q - l - 1$, $p_2 = p - l - 1$
          $p_2(x) = (f_i(x), q_2, (a_j, n_j), p_2)$
          Call CHIP$(p_2(x), S_2)$
        else $p_2(x) = 0$
        endif
        $p(x) = p_1(x)T_{1-1,2j} + p_2(x)T_{i-1,2j-1}$
      endif
    elseif $b_{(2i-1)2^{j-1}} = a_l$ and $S \neq \emptyset$
      $n_l = 2^{j-1} - \sum_{t=0}^{l-1} n_t$, $p = l$
      $p(x) = (f_i(x), q, (a_j, n_j), p(x))$

```
            Call CHIP(p(x), S)
            p(x) = T_{j-1,2i}(x)p(x)
        endif
    endfor i
endfor j
```

FUNCTION Convert$(q(x-a), m)$. Given a polynomial $q(x)$ of degree m, function Convert converts $q(x-a)$ into the form $\sum h_i x^i$.

```
Compute q_i = q^(i)(-a)/i!,  i = 0, 1, ..., m
Return q_0 + q_1 x + ··· + q_m x^m
end
```

Note that $0 \in S_i$ if $S_i \neq \emptyset$ and $m \notin S_i$ if $l \notin S_i$ and $m > l$ at any step of the algorithm. Since SCV can be used to determine Hermite interpolating polynomials, we denote SCV occasionally by HIP$(p(x), (a_i, n_i), i = 0, 1, \ldots, p)$ in light of a_i, n_i and p.

THEOREM 3.1. *Algorithm* SCV *solves dual confluent Vandermonde linear system* $V_c^T \mathbf{x} = \mathbf{b}$ *in* $O(n \log n \log p)$ *operations if fast polynomial multiplication and division are used.*

Proof. As stated above, Stage I of Algorithm SCV computes correctly the coefficients of polynomial $f_i(x)$, $i = 0, 1, \ldots, p$, of degree at most $n_i - 1$ such that $p(x) = f_0(x)l_0(x) + f_1(x)l_1(x) + \cdots + f_p(x)l_p(x)$ satisfies the condition (13). We now prove that Stage II converts correctly $p(x)$ into the form $\sum h_i x^i$ by induction on p.

If $p = 0$, we have $T_{k0}(x) = (x - a_0)^n$ and $n_0 = n$. Performing the algorithm shows that $S_1 = S_2 = \emptyset$. Hence $p(x) = f_0 \frac{T_{k0}(x)}{(x-a)^n} = f_0(x)$, which shows the correctness of the algorithm.

If $p > 0$ and $b_{2^{k-1}} = b_{2^{k-1}+1} = a_l$, write $p(x)$ as

$$p(x) = \left(\sum_{i=0}^{l-1} f_i(x)\tilde{l}_i(x) \right) T_{k-1,2}(x) + \left(\sum_{i=l+1}^{p} f_i(x)\tilde{l}_i(x) \right) T_{k-1,1}(x) + f_l(x) \frac{T_{k0}(x)}{(x-a_l)^{n_l}},$$

where $\tilde{l}_i(x) = l_i(x)/T_{k-1,2}(x)$ if $i < l$ and $\tilde{l}_i(x) = l_i(x)/T_{k-1,1}(x)$ if $i > l$. Performing the algorithm for $j = k$ shows that $p_1(x) = \sum_{i=0}^{l-1} f_i(x)\tilde{l}_i(x)$ and $p_2(x) = \sum_{i=l+1}^{p} f_i(x)\tilde{l}_i(x)$. Hence, induction shows that the algorithm represents

$$p(x) = p_1(x)T_{k-1,2}(x) + p_2(x)T_{k-1,1}(x) + R_l(x) \tag{15}$$

correctly in term of the basis $\{x^i\}$. If $p > 0$ and $b_{2^{k-1}} \neq b_{2^{k-1}+1}$, the correctness of the algorithm follows from a similar way.

Note that when $j \leq k - 1$, the situation $b_{(2i-1)2^{j-1}} = a_l$, $S \neq \emptyset$, and $l \notin S$ may occur. Consider CHIP$(\tilde{p}(x), S)$, for example, where $\tilde{p}(x) = (f_i(x), \tilde{q}, (a_j, n_j), \tilde{p})$, and $S_1 = \{0, 1, \ldots, \tilde{q}\}$. In this case $l > \tilde{q}$, it is readily seen that $\tilde{p}(x) = \bar{p}(x)T_{j-1,2i}(x)$, where $\bar{p}(x) = (f_i, \tilde{q}, (a_j, \tilde{n}_j), l)$, $\tilde{n}_i = n_i$ if $i < l$ and $\tilde{n}_l = 2^{j-1} - \sum_{i=0}^{l-1} n_i$. Therefore, the "elseif" of the algorithm treats the case correctly.

It follows from the proof of Theorem 2.4, Stage I needs at most $O(n \log n \log p)$ operations. We now prove that Stage II needs also $O(n \log n \log p)$ operations. To this end, denote by $T(n, q)$ the number of operations needed by CHIP$(p(x), S)$ after all $T_{ji}(x)$ are converted into the form $\sum h_i x^i$, where $p(x) = (f_i(x), q, (a_i, n_i), p)$ and

$S = \{0, 1, \ldots, q\}$. If $q = 0$, it is readily seen that $T(n, q) = 0$ according to the analysis of the correctness.

If $q > 0$, $b_{2^{k-1}} = b_{2^{k-1}+1} = a_l$, and $l \in S$, the algorithm divides the problem CHIP$(p(x), S)$ into two subproblems CHIP$(p_1(x), S_1)$ and CHIP$(p_2(x), S_2)$ after computing $R_l(x)$ and then converts $p(x)$ into the form $\sum h_i x^i$ via (15). Since $(x - a_l)^{n_l} = x^n - n_l a_l x^{n_l - 1} + \cdots + (-1)^{n_l} a_l^{n_l}$, it needs $O(n_l) \leq O(n)$ operations to write $(x - a_l)^{n_l}$ into the form $\sum h_i x^i$. Fast polynomial multiplication and division are used to convert $R_l(x)$ and $p(x)$ into the form $\sum h_i x^i$ in $C_1 \frac{n}{2} \log \frac{n}{2}$ operations, where C_1 is a positive constant independent of n and q. Hence,

$$T(n, q) \leq T\left(\frac{n}{2}, q_1\right) + T\left(\frac{n}{2}, q_2\right) + C_1 \frac{n}{2} \log \frac{n}{2},$$

where $q_1 = l - 1$ and $q_2 = q - l - 1$.

If $b_{2^{k-1}} \neq b_{2^{k-1}+1}$ and $l \in S$, we have similarly

$$T(n, q) \leq T\left(\frac{n}{2}, q_1\right) + T\left(\frac{n}{2}, q_2\right) + C_2 \frac{n}{2} \log \frac{n}{2},$$

where $q_1 = l$ and $q_2 = q - l - 1$ and C_2 is a positive constant independent of n and q.

Note that if the case $b_{2^{k-1}} = a_l$, $l \notin S$ and $S \neq \emptyset$ occurs, we can estimate $T(n, q)$ as follows:

$$T(n, q) = T\left(\frac{n}{2}, q\right) + C_3 \frac{n}{2} \log \frac{n}{2},$$

where C_3 is another positive constant independent of n and q.

Hence, $T(n, 0) = 0$ and

$$T(n, p) \leq T(\frac{n}{2}, q_1) + T(\frac{n}{2}, q_2) + C\frac{n}{2} \log \frac{n}{2}, \tag{16}$$

where $C = \max(C_1, C_2, C_3)$ and $q_1 + q_2 \leq q$. Proposition 2.3 and (16) show that $T(n, p) \leq C\, n \log n (\log(p + 1) + 1)$. □

4. Generalized Trummer's problem I. Let H_p be generalized Hilbert matrices of order n defined by

$$(H_p)_{ij} = \begin{cases} 1/(t_i - s_j)^p, & i \neq j,\ i, j = 1, 2, \ldots, n, \\ 1/(t_i - s_i)^p, & i = j,\ t_i \neq s_i, \\ h_{pi}, & i = j,\ t_i = s_i, \end{cases} \tag{17}$$

where p is a positive integer, t_i, s_i, and h_{pi} are points in the complex plane, and $t_i \neq t_j$, $t_i \neq s_j$, $s_i \neq s_j$ for $i \neq j$. Given an n-vector $\mathbf{b} = (b_1, b_2, \ldots, b_n)^T$, in this section, we consider the generalized Trummer's problem, i.e., the computation of multiplications $H_1\mathbf{b}$, $H_2\mathbf{b}$,...,$H_p\mathbf{b}$. Denote

$$w(x) = (x - s_1)(x - s_2) \cdots (x - s_n),$$

$$w_i(x) = \frac{w(x)}{(x - s_i)}, \qquad i = 1, 2, \ldots, n,$$

$$g(x) = b_1 w_1(x) + b_2 w_2(x) + \cdots + b_n w_n(x),$$

$$r(x) = \frac{g(x)}{w(x)}, \tag{18}$$

$$r_i(x) = \begin{cases} r(x), & x_i \neq s_i, \\ r(x) - \dfrac{b_i w_i(x)}{w(x)}, & x_i = s_i, \end{cases} \tag{19}$$

$$g_i(x) = \begin{cases} g(x), & x_i \neq s_i, \\ g(x) - b_i w_i(x), & x_i = s_i, \end{cases} \tag{20}$$

$$\mathbf{y}_i = H_i \mathbf{b} = (y_{i1}, y_{i2}, \ldots, y_{in})^T.$$

Equations (18), (19), and (20) imply that $w(x)r_i(x) = g_i(x)$. A simple computation shows that

$$y_{mi} = \begin{cases} \displaystyle\sum_{i=1}^{n} \frac{b_i}{(t_i - s_j)^m}, & t_i \neq s_i, \\ \displaystyle\sum_{j=1, j\neq i}^{n} \frac{b_j}{(t_i - s_j)^m} + h_{mi} b_i, & t_i = s_i, \end{cases} \tag{21}$$

$$= \begin{cases} \dfrac{(-1)^{m-1}}{(m-1)!} r_i^{(m-1)}(x)\Big|_{x=t_i}, & t_i \neq s_i, \\ \dfrac{(-1)^{m-1}}{(m-1)!} r_i^{(m-1)}(x)\Big|_{x=t_i} + h_{mi} b_i, & t_i = s_i. \end{cases}$$

Therefore, the generalized Trummer's problem can be computed by Hermite evaluation of rational functions $r_1(x), r_2(x), \ldots, r_n(x)$. Furthermore, expanding $r_i(x)$, $l(x)$, and $g_i(x)$ in Taylor series at t_i shows that

$$r_i(x) = r_{i0} + r_{i1}(x - t_i) + \cdots + r_{i,p-1}(x - t_i)^{p-1} + O((x - t_i)^p),$$
$$w(x) = w_{i0} + w_{i1}(x - t_i) + \cdots + w_{ip}(x - t_i)^p + O((x - t_i)^{p+1}),$$
$$g_i(x) = g_{i0} + g_{i1}(x - t_i) + \cdots + g_{ip}(x - t_i)^p + O((x - t_i)^{p+1}).$$

Since $s_i \neq s_j$ for $i \neq j$, $w(x_i) = w_{i0} \neq 0$ if $t_i \neq s_i$ and $w'(t_i) = w_{i1} \neq 0$ if $t_i = s_i$. Equaling to the coefficients of x^i in $r_i(x)w(x) = g_i(x)$ shows that $\mathbf{r}_i = (r_{i0}, r_{i1}, \ldots, r_{i,p-1})^T$ is the solution of the following triangular Toeplitz linear system

$$\mathrm{triT}(w_{i0}, w_{i1}, \ldots, w_{i,p-1})\mathbf{r}_i = \mathbf{d}_{i1}$$

if $t_i \neq s_i$, or

$$\mathrm{triT}(w_{i1}, w_{i2}, \ldots, w_{ip})\mathbf{r}_i = \mathbf{d}_{i2}$$

if $t_i = s_i$, where $\mathbf{d}_{i1} = (g_{i0}, g_{i1}, \ldots, g_{i,p-1})^T$ and $\mathbf{d}_{i2} = (g_{i1}, g_{i2}, \ldots, g_{ip})^T$.

On the other hand, relation $(x - s_i)w_i(x) = w(x)$ implies that

$$w_i^{(m+1)}(x)(x - s_i) + (m+1)w_i^{(m)}(x) = w^{(m+1)}(x).$$

Hence, if $t_i = s_i$, the mth derivative of $w_i(x)$ at point t_i can be calculated by

$$w_i^{(m)}(t_i) = \frac{1}{m+1} w^{(m+1)}(t_i),$$

which together with (20) shows that the mth derivative of $g_i(x)$ at point t_i is given by

$$g_i^{(m)}(t_i) = \begin{cases} g^{(m)}(t_i) & t_i \neq s_i. \\ g^{(m)}(t_i) - \dfrac{b_i}{m+1} w^{(m+1)}(t_i) & t_i = s_i. \end{cases}$$

Therefore, generalized Trummer's problem $H_1\mathbf{b}$, $H_2\mathbf{b}$, ..., $H_p\mathbf{b}$ can be computed as follows.

ALGORITHM G-TRUMMER I. Given p generalized Hilbert matrices H_1, H_2, ..., H_p of order n defined by (17) and an n-vector $\mathbf{b} = (b_1, b_2, \ldots, b_n)^T$, Algorithm G-Trummer I computes generalized Trummer's problem, $H_1\mathbf{b}$, $H_2\mathbf{b}$, ..., $H_p\mathbf{b}$. Without loss of generality, we assume here $n = 2^k$ for some positive integer k.

Stage I. Set $T_i = x - s_i, \qquad i = 1, 2, \ldots, n$
For $j = 1 : 1 : k$
For $i = 1 : 1 : 2^{k-j}$
$T_{j,i}(x) = T_{j-1,2i-1}(x) T_{j-1,2i}(x)$
endif i
endif j
Stage II. $g(x) = \sum_{i=1}^{n} b_i w_i(x)$, $S = \{0, 1, \ldots, n-1\}$
Call CHIP$(g(x), S)$, $w(x) = T_{k1}(x)$
Call HEP$(w(x), (t_i, p+2), i = 1, \ldots, n)$
Call HEP$(g(x), (t_i, p+1), i = 1, \ldots, n)$
For $i = 1 : 1 : n$
if $t_i \neq s_i$ then $T_i = \text{triT}\big(w(t_i), w'(t_i), \ldots, \frac{w^{(p-1)}(t_i)}{(p-1)!}\big)$
$\mathbf{g}_i = \big(g(t_i), g'(t_i), \ldots, \frac{g^{(p-1)}(t_i)}{(p-1)!}\big)^T$
else $T_i = \text{triT}\big(w'(t_i), \frac{w^{(2)}(t_i)}{2!}, \ldots, \frac{w^{(p)}(t_i)}{p!}\big)^T$
for $m = 1 : 1 : p$
$g_{im} = \frac{g^{(m)}(t_i)}{m!} - \frac{b_i w^{(m+1)}(t_i)}{(m+1)!}$
endfor m
$\mathbf{g}_i = (g_{i1}, g_{i2}, \ldots, g_{ip})^T$
endif
Solve triangular Toeplitz linear system $T_i \mathbf{x}_i = \mathbf{g}_i$
For $j = 0 : 1 : p-1$
if $t_i \neq s_i$ then $y_{j+1,i} = (-1)^j x_{ij}$
else $y_{j+1,i} = (-1)^j x_{ij} + h_{j+1,i}\, b_i$
endif
endfor j
endfor i

THEOREM 4.1. *Given a set of Hilbert matrices* $\{H_1, H_2, \ldots, H_p\}$ *defined by* (17) *and a complex vector* $\mathbf{b} = (b_1, b_2, \ldots, b_n)^T$, *Algorithm* G-Trummer I *computes generalized Trummer's problem* $H_1\mathbf{b}, H_2\mathbf{b}, \ldots, H_p\mathbf{b}$ *in* $O(np \log n \log \frac{n}{p})$ *operations if fast polynomial multiplication and division are used.*

Proof. The correctness of the algorithm follows from our previous discussion of the section and its computational complexity follows from Theorems 2.4 and 3.1 immediately. □

5. Generalized Trummer's problem II. In this section, we consider how to compute the multiplication $H_p\mathbf{b}$ fast, where H_p is a generalized Hilbert matrix given by (17), if $t_i \neq s_i$, $i = 1, 2, \ldots, n$. Denote $q(x) = (x - t_1)(x - t_2)\cdots(x - t_n)$, $p(x) = (q(x))^p$ and

$$\mathbf{y} = H_p\mathbf{b} = (y_1, y_2, \ldots, y_n)^T.$$

In the case $t_i \neq s_i$, $i = 1, 2, \ldots, n$, (21) shows that the components of $\mathbf{y}$ are given by

$$y_i = \frac{(-1)^{p-1}}{(p-1)!} r^{(p-1)}(t_i), \qquad i = 1, 2, \ldots, n.$$

Since $w(x)$ and $p(x)$ are prime, there exist unique polynomials $u(x)$ of degree at most $np - 1$ and $v(x)$ of degree at most $n - 1$ such that

$$u(x)w(x) + v(x)p(x) = 1. \tag{22}$$

To compute the $(p-1)$th derivative of $r(x)$ at points t_i, $i = 1, 2, \ldots, n$, we formulate $r(x)$ from (19) and (22) that

$$r(x) = u(x)g(x) + \frac{g(x)v(x)p(x)}{w(x)},$$

which implies that

$$r^{(p-1)}(t_i) = (u(x)g(x))^{(p-1)}\Big|_{x=t_i}, \qquad i = 1, 2, \ldots, n.$$

Putting $x = s_i$ in (22) shows that $v(x)$ is a interpolating polynomial such that

$$v(s_i) = \frac{1}{p(s_i)}, \qquad i = 1, 2, \ldots, n,$$

and furthermore, we obtain again from (22) that

$$u(x) = \frac{1 - v(x)p(x)}{w(x)}.$$

Summarizing our discussion, we have the following algorithm for the matrix-vector product $H_p\mathbf{b}$ in the case $t_i \neq s_i$, $i = 1, 2, \ldots, n$.

ALGORITHM G-TRUMMER II. Given a generalized Hilbert matrix H_p defined by (17) satisfying, in addition, $t_i \neq s_i$, $i = 1, 2, \ldots, n$ and an n-vector $\mathbf{b} = (b_1, b_2, \ldots, b_n)^T$, the following algorithm computes the matrix-vector product $H_p\mathbf{b}$.

Stage I. Set $T_{0i}(x) = x - s_i$, $R_{0i} = x - t_i$, $i = 1, 2, \ldots, n$

For $j = 1:1:k$
  For $i = 1:1:2^{k-j}$
    $T_{ji}(x) = T_{j-1,2i-1}(x)\, T_{j-1,2i}(x)$
    $R_{ji}(x) = R_{j-1,2i-1}(x)\, R_{j-1,2i}(x)$
  endfor i
endfor j

Stage II. $w(x) = T_{k1}(x)$, $q(x) = R_{k1}(x)$
Call HEP$(q(x), (s_i, 1),\ i = 1, 2, \ldots, n)$
For $i = 1:1:n$
  $p(s_i) = (q(s_i))^p$
endfor i
Call HIP$(v(x), (s_i, p(s_i)^{-1}, 1),\ i = 1, 2, \ldots, n)$
$S = \{0, 1, \ldots, n-1\}$, $g(x) = \sum_{i=1}^{n} b_i w_i(x)$
Call CHIP$(g(x), S)$
$p(x) = (q(x))^p$, $u(x) = (1 - v(x)p(x))/w(x)$
$\bar{u}(x) = (u(x)g(x))^{(p-1)}$, $\tilde{u}(x) = \bar{u}(x) (\mathrm{mod}\, q(x))$
Call HEP$(\tilde{u},\ (t_i, 1)\quad i = 1, 2, \ldots, n)$
For $i = 1:1:n$
  $y_i = \frac{(-1)^{p-1}}{(p-1)!}\, \tilde{u}(t_i)$
endfor i

To analyze the computational complexity, we need the following proposition on computing the power of polynomials.

PROPOSITION 5.1. *Given are the coefficients of polynomial $f(x)$ of degree n and a positive integer p. Then the coefficients of $(f(x))^p$ can be determined in $O(np \log np)$ operations if fast polynomial multiplication is used. In particular, the pth power of any complex number can be computed in $O(\log p)$ operations.*

Proof. Without loss of generality, assume that $p = 2^t$ for some positive integer t, the conclusion follows immediately from $(f(x))^p = (f(x))^{p/2}(f(x))^{p/2}$, fast polynomial multiplication and induction. □

THEOREM 5.2. *Given a Hilbert matrix H_p of order n defined by* (17) *satisfying $t_i \neq s_i$, $i = 1, 2, \ldots, n$ and a complex n-vector $\mathbf{b}$, Algorithm* G-Trummer II *computes the matrix-vector product $H_p\mathbf{b}$ in $O(np \log np + n \log^2 n)$ operations if fast polynomial multiplication and division are used.*

Proof. The correctness of the algorithm follows immediately from our discussion. As for computational complexity, Stage I needs $O(n \log^2 n)$ operations (see, e.g., [23, Prop. 4.1]).

Theorems 2.4 and 3.1 show that performing HEP, HIP, and CHIP needs $O(n \log^2 n)$ operations. Proposition 5.1 shows that computation $(q(s_i))^p$, $i = 1, 2, \ldots, n$ and $p(x) = (q(x))^p$ need $O(n \log p)$ and $O(np \log np)$ operations, respectively. If fast polynomial multiplication and division are used, $u(x) = \frac{1 - v(x)u(x)}{w(x)}$, $\bar{u}(x) = (u(x)g(x))^{(p-1)}$, and $\tilde{u}(x) = \bar{u}(x) (\mathrm{mod}\, q(x))$ need $O(np \log n)$ operations. Hence, the overall operations is bounded by $O(np \log np + n \log^2 n)$. □

Note that the G-Trummer II needs at most $O(n \log^2 n)$ operations if $p \leq O(\log n)$ and $O(np \log np)$ if $p > O(\log n)$.

6. Conclusions. If $p \ll n$ for generalized Hilbert matrices and $n_i \ll n$ for confluent Vandermonde matrices, it is not necessary to use a fast solver for triangular Toeplitz linear systems in our algorithms. In this case G-Trummer I reduces to the Gerasoulis algorithm essentially if, in addition, $p = 1, 2$ and $t_i \neq s_i$, $i = 1, 2, \ldots, n$.

In general, it is difficult to analyze the stability of all algorithms in the paper, but it becomes possible for some practical cases. If we choose t_i and s_i of a generalized Hilbert matrix to be Chebyshev points, the computational complexity of G-Trummer I and G-Trummer II can be further reduced and the algorithms can hopefully be implemented in a stable way, though further work is needed. For example, if $p = 1$, $t_i = \cos((2i-1)\pi/2n)$ and $s_i = \cos(i\pi/n+1)$, following [11], we can easily present an $O(n \log n)$ stable implementation of G-Trummer I through a straightforward modification. In this case, the algorithm G-Trummer I shares the same complexity and the same stability with the Gerasoulis algorithm. It is shown in [11] that the performance of the Gerasoulis algorithm with $t_i = \cos((2i-1)\pi/2n)$ and $s_i = \cos(i\pi/n+1)$ is much faster and more stable than the $O(n^2)$ algorithm of common matrix multiplication because it requires the application of the fast Fourier transform (FFT) twice.

Acknowledgments. I am grateful to Dr. Nicholas J. Higham for valuable comments and suggestions on the manuscript.

REFERENCES

[1] A. Aho, J. E. Hopcroft, and J. D. Ullman, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.

[2] C. T. H. Baker and M. S. Derakhshan, *Fast generation of quadrature rules with some special properties*, in Numerical Integration: Recent Developments, Software and Applications, P. Keast and G. Fairweather, eds., D. Reidel Publishing Company, Dordrecht, Holland, 1987, pp. 53–60.

[3] C. Ballester and V. Pereyra, *On the construction of discrete approximations to linear differential expressions*, Math. Comp., 21 (1967), pp. 297–302.

[4] D. Bini and V. Pan, *Polynomial division and its computational complexity*, J. Complexity, 2 (1986), pp. 179–203.

[5] Å. Björck and T. Elfving, *Algorithms for confluent Vandermonde systems*, Numer. Math., 21 (1973), pp. 130–137.

[6] Å. Björck and V. Pereyra, *Solution of Vandermonde systems of equations*, Math. Comp., 24 (1970), pp. 893–903.

[7] G. Galimberti and V. Pereyra, *Solving confluent Vandermonde systems of Hermite type*, Numer. Math., 18 (1971), pp. 44–60.

[8] W. Gautschi, *On inverse of Vandermonde and confluent Vandermonde matrices*, Numer. Math., 4 (1962), pp. 117–123.

[9] ———, *The condition of Vandermonde-like matrices involving orthogonal polynomials*, Linear Algebra Appl., 52/53 (1983), pp. 293–300.

[10] W. Gautschi and J. Wimp, *Computing the Hilbert transform of a Jacobi weight function*, BIT, 27 (1987), pp. 203–215.

[11] A. Gerasoulis, *A fast algorithm for the multiplication of generalized Hilbert matrices with vectors*, Math. Comp., 50 (1988), pp 179–188.

[12] A. Gerasoulis, M. D. Grigoriadis, and L. Sun, *A fast algorithm for Trummer's problem*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. s135–s138.

[13] G. H. Golub, *Trummer's problem*, SIGACT News, ACM Special Interest Group on Automata and Computability Theory, 17 (1985), p. 17.2-12.

[14] L. Greengard and V. Rokhlin, *A fast algorithm for particle simulations*, J. Comput. Phys., 73 (1987), pp. 325–348.

[15] S.-Å. Gustafson, *Control and estimation of computational errors in the evaluation of interpolation formulae and quadrature rules*, Math. Comp., 24 (1970), pp. 847–854.

[16] N.J. Higham, *Error analysis of the Björck-Pereyra algorithm for solving Vandermonde systems*, Numer. Math., 50 (1987), pp. 613–632.

[17] ———, *Fast solution of Vandermonde-like systems involving orthogonal polynomials*, IMA J. Numer. Anal., 8 (1988), pp. 473–486.

[18] ———, *Stability analysis of algorithm for solving confluent Vandermonde-like systems*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 23–41.

[19] ———, *Iterative refinement enhances the stability of QR factorization methods for solving linear equations*, BIT, 31 (1991), pp. 447–468.

[20] J. KAUTSKY AND S. ELHAY, *Calculation of weights of interpolatory quadratures*, Numer. Math., 40 (1982), pp. 407–422.

[21] H. LU, *On computational complexity of the multiplication of generalized Hilbert matrices with vectors and solution of certain generalized Hilbert linear systems*, Chinese Sci. Bull., 13 (1989), pp. 963–966. (In Chinese.) English translation, Chinese Sci. Bull., 35 (1990), pp. 974–978.

[22] ———, *Computational complexity of Vandermonde linear systems*, Chinese Sci. Bull., 9 (1990), pp. 654–656. (In Chinese.)

[23] ———, *Fast solution of confluent Vandermonde linear systems*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1277–1289.

[24] J.N. LYNESS, *Some quadrature rules for finite trigonometric and related integrals*, in Numerical Integration: Recent Developments, Software and Applications, P. Keast and G. Fairweather, eds., D. Reidel Publishing Company, Dordrecht, Holland, 1987, pp. 17–33.

[25] R. MOENCK AND A. BORODIN, *Fast modular transforms via division*, Proc. 13th Annual Symp. on Switching and Automata Theory, 1972, pp. 90–96.

[26] L. REICHEL AND G. OPFER, *Chebyshev–Vandermonde systems*, Math. Comp., 57 (1991), pp. 703–721.

[27] V. ROKHLIN, *Rapid solution of integral equations of classical potential theory*, J. Comput. Phys., 60 (1985), pp. 187–207.

[28] W.P. TANG AND G.H. GOLUB, *The block decomposition of a Vandermonde and its applications*, BIT, 21 (1981), pp. 505–517.

[29] J.F. TRAUB, *Associated polynomials and uniform methods for the solution of linear problems*, SIAM Rev.,8 (1966), pp. 277–301.

[30] M. TRUMMER, *An efficient implementation of a conformal mapping method using the Szegö kernel*, SIAM J. Numer. Anal., 23 (1986), pp. 853–872.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 675–687, April 1995

022

BACKWARD ERROR ANALYSIS FOR THE CONSTRAINED AND WEIGHTED LINEAR LEAST SQUARES PROBLEM WHEN USING THE WEIGHTED QR FACTORIZATION*

MÅRTEN GULLIKSSON†

Abstract. Backward errors are derived for the solution of the constrained and weighted linear least squares problem when using the weighted QR factorization with column pivoting, [*SIAM J. Matrix Anal. Appl.*, 13 (1992), pp. 1298–1313]. On the way to achieving this goal we attain a detailed description of the errors produced when using our M-invariant reflections.

Key words. rounding errors, least squares, QR factorization, weights

AMS subject classifications. 65F25, 65F35, 65G05

1. Introduction.

We analyze the algorithm presented in [9] for solving the constrained and weighted linear least squares problem

$$\min_{x \in \mathbf{R}^n} (b_2 - A_2 x)^T W_2 (b_2 - A_2 x) \quad \text{s.t.} \quad A_1 x = b_1, \tag{1}$$

where $A_1 \in \mathbf{R}^{p \times n}, A_2 \in \mathbf{R}^{q \times n}, b_1 \in \mathbf{R}^p, b_2 \in \mathbf{R}^q$, and W_2 is a diagonal positive definite weight matrix. It is assumed that $p + q = m \geq n > p$. The algorithm was only proved to be backward stable for linear least squares with linear constraints. The main goal is to make a backward error analysis of the weighted QR factorization and of the algorithm based on the weighted QR factorization to solve (1).

To our knowledge direct algorithms for solving systems of linear equations or linear least squares (possibly weighted) are all based on orthogonal or Gauss transformations. Our new concept of M-invariant reflections has given rise to a different approach in the error analysis. The error analysis is based on two properties of the M-invariant reflection Q. The first property is that the 2-norm of a vector, x, is preserved in a certain weighted norm, i.e., $\|N^{1/2}Qx\| = \|N^{1/2}x\|$, where N is a positive semidefinite diagonal matrix. The second property is that vectors in this weighted norm, up to a factor two in norm, are not amplified by the reflection. The norm invariance is a generalization of the norm invariance of orthogonal transformations and the growth of vectors is a generalization of the growth in Gauss transformations. Our results concerning the error analysis are new but similar to the results in [15] for unconstrained weighted least squares problems. In [1], least squares with linear constraints and weighted least squares with *one* weight is analyzed but it is difficult to compare their results with ours. Other work on error analysis for the least squares problem can be found in [2]–[5], [11], and [12].

The only other algorithm for the constrained and weighted least squares problem proved to be backward stable is Paige's algorithm; see [13] and [14]. Paige uses the dual formulation of (1). Paige's approach has many attractive properties for general weight matrices, but when W_2 is diagonal it requires slightly more work than our algorithm and is definitely more complicated.

We use the perturbation analysis developed in [16] because it covers our class of weighted and constrained least squares problems (as well as the rank deficient case).

* Received by the editors March 23, 1993; accepted for publication (in revised form) by N. J. Higham, April 7, 1994.

† Institute for Information Processing, University of Umeå, S-901 87 Umeå, Sweden (marten@cs.umu.se).

The outline of the article is as follows. First, in §2 we introduce some special notation. Section 3 introduces the system equations, which is the problem formulation we use in the analysis. In §4 we present appropriate assumptions for the existence and uniqueness of the weighted QR factorization. The next three sections then describe and analyze the use of M-invariant reflections to solve (1). Backward errors for the weighted QR factorization and backward errors for the solution to (1) are given in §§8 and 9, respectively. In §10 we also use the perturbation theory to get explicit forward error bounds on the solution. Section 11 summarizes the results and describes some possible future work.

2. Some special notation. We use some notations that need to be explained. The variable ϵ, with possible sub or superindices, of order $\mathbf{u}$, where $\mathbf{u}$ is the round-off unit of the floating point system, always has an explicitly stated upper limit. The greek letter δ is used in connection with the perturbation analysis to denote small quantities. Round-off errors, backward and forward, are often, but not always, denoted by Δ followed by a matrix or a vector, e.g., ΔA and Δb. The operator for element-by-element multiplication of two matrices or vectors is denoted $\odot$. The 2-norm of a vector x and a matrix X is denoted by $\|x\|$ and $\|X\|$. Numerically calculated quantities, where the arithmetic precision is finite, are furnished with a hat to distinguish them from their exact counterparts. If $n\mathbf{u} < 1$ we follow [12] and define $\gamma_n = n\mathbf{u}/(1 - n\mathbf{u})$, which is used frequently in the analysis. The floating point model used in the analysis is the same as the one in [12].

3. The system equations. If we define $M_2 = W_2^{-1}$ then an equivalent formulation of (1) is

$$\begin{bmatrix} 0 & 0 & A_1 \\ 0 & M_2 & A_2 \\ A_1^T & A_2^T & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ x \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ 0 \end{bmatrix}, \tag{2}$$

where λ_1 is the vector of Lagrange multipliers and $M_2\lambda_2$ is the residual. We call the system of equations in (2) the *system equations* and the matrix in (2) the *system matrix*.

It is easily seen that problem (1) has a unique solution if and only if $\operatorname{rank}(A_1) = p$ and $\operatorname{rank}(A) = n$ and we assume that these two conditions are fulfilled.

Define the $m \times n$ matrix A and the vector b of length m as

$$A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}, \qquad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}.$$

If we take

$$M = \begin{bmatrix} 0 & 0 \\ 0 & M_2 \end{bmatrix},$$

we can write (2) in a more general way as

$$\begin{bmatrix} M & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}, \tag{3}$$

where $M\lambda$ is the residual. We make the assumption that $M = \operatorname{diag}(\mu_1, \ldots, \mu_m)$, where $\mu_1 \leq \cdots \leq \mu_m$.

It is convenient to write the system of equations in (3) as

$$Ax \overset{M}{\simeq} b. \tag{4}$$

For the ordinary least squares problem we have $M = I_m$ and for an unweighted constrained least squares problem we have $M_2 = I_{m-p}$.

4. The weighted QR factorization. Any matrix Q satisfying the condition

$$QMQ^T = M, \tag{5}$$

will be called *M-invariant.*

The weighted QR factorization of $A \in \mathbf{R}^{m\times n}$ is defined as

$$A\Pi = Q \begin{bmatrix} R \\ 0 \end{bmatrix}, \qquad QMQ^T = M, \tag{6}$$

where $Q \in \mathbf{R}^{m\times m}$ and $R \in \mathbf{R}^{n\times n}$ is an upper triangular matrix and Π is a permutation matrix. From [9] we have the following theorem.

THEOREM 4.1. *Given $M = \operatorname{diag}(\mu_1, \ldots, \mu_m)$ where $\mu_1 = \cdots = \mu_p = 0$, $0 < \mu_{p+1} \le \cdots \le \mu_m \le 1$ and a matrix $A \in \mathbf{R}^{m\times n}$ partitioned as*

$$A = \begin{matrix} \\ p \\ m-p \end{matrix} \begin{matrix} \begin{matrix} p & n-p \end{matrix} \\ \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \end{matrix}$$

then a factorization

$$A = Q \begin{bmatrix} R \\ 0 \end{bmatrix}, \qquad QMQ^T = M$$

with R upper triangular and nonsingular and Q nonsingular exists if and only if A has linearly independent columns and A_{11} is nonsingular. There is a unique nonsingular R if and only if M is nonsingular.

The permutation matrix in (6) is, as we shall see, necessary for getting a numerically stable algorithm. If Theorem 4.1 is stated with column permutations, Π, the condition of nonsingularity on A_{11} for R being nonsingular can be replaced by the condition that $[A_{11}, A_{12}]$ has linearly independent rows.

Note that if the system matrix in (3) is nonsingular then the weighted QR factorization with column permutations exists and R and Q are nonsingular (R and Q need not be unique).

Assuming that there are no constraints and defining $W = M^{-1}$ then the relation in (5), corresponding to the problem formulation (1), is $Q^TWQ = W$. However, this relation is not as useful as the one in (5) because very large or infinite weights in W are better represented as very small elements or zero elements in M.

From (3) we get with (6)

$$\begin{bmatrix} M & \begin{bmatrix} R \\ 0 \end{bmatrix} \\ [R^T, 0] & 0 \end{bmatrix} \begin{bmatrix} Q^T\lambda \\ \Pi^T x \end{bmatrix} = \begin{bmatrix} Q^{-1}b \\ 0 \end{bmatrix}. \tag{7}$$

If we assume that the system matrix in (3) is nonsingular and partition

$$M = \operatorname{diag}(M_n, M_{m-n}),$$

where $M_n = \operatorname{diag}(\mu_1, \ldots, \mu_n)$ and $M_{m-n} = \operatorname{diag}(\mu_{n+1}, \ldots, \mu_m)$, we get

$$x = \Pi \left[R^{-1}, 0\right] Q^{-1} b \tag{8}$$

and

$$\lambda = Q^{-T} \begin{bmatrix} 0 & \\ & M_{m-n}^{-1} \end{bmatrix} Q^{-1} b. \tag{9}$$

5. Calculating the weighted QR factorization using M-invariant reflections. To keep things as simple as possible we consider the first transformation where, after pivoting, the first column in A, d_1, is made parallel to e_1. For the sake of simplicity we assume that the sign of the first element in $d_1, d_1^{(1)}$ is greater than or equal to zero. We want Q_1 to fulfill

$$Q_1 d_1 = -\alpha_1 e_1, \quad \alpha_1 \neq 0, \quad Q_1^2 = I.$$

It is fairly easy to see that Q_1 can be written as

$$Q_1 = I - \frac{(d_1 + \alpha_1 e_1)(N_1 d_1 + \alpha_1 e_1)^T}{\alpha_1 (\alpha_1 + d_1^{(1)})},$$

where $\alpha_1 = \|N_1^{1/2} d_1\|$ and $N_1 = \operatorname{diag}(1, \mu_1/\mu_2, \ldots, \mu_1/\mu_m)$; see [9] for further details.

We now turn to the general algorithm and are forced to introduce some additional notation to describe the details for solving problem (4) with our weighted QR factorization. If we look at iteration k we define in exact arithmetic

$$A_k = \begin{bmatrix} A_{11}^{(k)} & A_{12}^{(k)} \\ 0 & A_{22}^{(k)} \end{bmatrix} = Q_{k-1} \ldots Q_1 A \Pi_1^T \ldots \Pi_{k-1}^T, \qquad A_1 = A,$$

where $A_{11}^{(k)}$ is upper triangular and Π_k is a permutation matrix. In a similar way we define

$$b_k = \begin{bmatrix} b_1^{(k)} \\ b_2^{(k)} \end{bmatrix} = Q_{k-1} b_{k-1}, \qquad b_1 = b.$$

The first column in $A_{22}^{(k)}$ *after* the column pivoting is denoted d_k and a general column in $A_{22}^{(k)}$ is denoted by c_k.

We assume that the first element in d_k, $d_1^{(k)}$ is greater than or equal to zero.

The M-invariant reflection at iteration k is

$$Q_k = \begin{bmatrix} I_{k-1} & 0 \\ 0 & \tilde{Q}_k \end{bmatrix}, \tag{10}$$

where

$$\tilde{Q}_k = I - \frac{(d_k + \alpha_k e_1)(N_k d_k + \alpha_k e_1)^T}{\alpha_k (\alpha_k + d_1^{(k)})},$$

$\alpha_k = \|N_k^{1/2} d_k\|$ and $N_k = \operatorname{diag}(1, \mu_{k+1}/\mu_k, \ldots, \mu_m/\mu_k)$. It is easily seen that $\tilde{Q}_k d_k = -\alpha_k e_1$ and $\tilde{Q}_k^2 = I_{m-k+1}$.

The column pivoting is made so that

$$\|N_k^{1/2} d_k\| = \max_j \|N_j^{1/2} c_j\|, \tag{11}$$

where the maximum in (11) is taken over all columns, c_k, in $A_{22}^{(k)}$.

Row scaling can be made to minimize the norm of the M-invariant transformation matrices. This is done by rescaling $A_{22}^{(k)}$ with a diagonal matrix D so that the rows of $A_{22}^{(k)}$ have approximately the same size. The matrix M is transformed to $D^{1/2}MD^{1/2}$ and then sorted by permuting rows in $A_{22}^{(k)}$, i.e., the new M-matrix looks like $\bar{M} = PD^{1/2}MD^{1/2}P^T$ where P is a permutation matrix and we put a bar above the scaled quantities. One way of scaling the rows of $A_{22}^{(k)}$ would be to multiply with $D^{1/2} = \operatorname{diag}(t_i)$, where

$$t_i = \begin{cases} \|d_k\|/|d_i^{(k)}|, & d_i^{(k)} \neq 0, \\ 1, & d_i^{(k)} = 0, \end{cases}$$

because then we get $D^{1/2}d_k = \operatorname{sign}(d_k)\|d_k\|$ and $\bar{\alpha}_k = \|\bar{N}_k^{1/2} D^{1/2} d_k\| \geq \|d_k\|$, which with Lemma 7.1 implies that $\|Q_k\| \leq 8m$. Note however that D may contain large elements and we have transformed the problem such that the large elments in Q have been put in the right-hand side $\bar{b} = D^{1/2}b$. It is this transformation that motivates why row pivoting is not to be recommended if the diagonal in M has been sorted.

Finally note that for any $x \in \mathbf{R}^{m-k+1}, k = 1, \ldots, n$

$$\|N_k^{1/2} \tilde{Q}_k x\| = \|N_k^{1/2} x\|, \tag{12}$$

which is one of the important properties of the M-invariant matrices, Q_k, that makes the error analysis go through.

6. Detailed results of rounding errors for M-invariant reflections. The following lemma is a short description of the rounding errors produced when applying a weighted Householder reflection. For the proof we refer to [8]. Remember that calculated quantities have a hat.

LEMMA 6.1. *With finite precision arithmetic and assuming that the column pivoting are known a priori, we have*

$$\hat{A}_{k+1} = fl(\hat{Q}_k \hat{A}_k) = (Q_k + \Delta Q_k)\hat{A}_k,$$

where

$$\Delta Q_k = \begin{bmatrix} 0 & 0 \\ 0 & \Delta \tilde{Q}_k \end{bmatrix}.$$

Assume that $\hat{c}_k$ is a column in $\hat{A}_{22}^{(k)}$. Then with $\beta_k = 1 + |\hat{d}_1^{(k)}|/\alpha_k \geq 1$ we have

$$\beta_k \Delta \tilde{Q}_k \hat{c}_k = \hat{d}_k \odot \eta_k + \xi_k \alpha_k e_1 + \hat{c}_k \odot \delta_k,$$

where $|\eta_i^{(k)}| \leq \gamma_{10(m+2)}$, $|\xi_k| \leq \gamma_{6m+14}$ and $|\delta_i^{(k)}| \leq \mathbf{u}$.

Note that this lemma enables us to write $\hat{A}_{n+1} = (Q_n + \Delta Q_n) \ldots (Q_1 + \Delta Q_1)A$ and a similar relation for $\hat{b}_{n+1}$ that will be used frequently.

7. Element growth. We have already seen in [9] that there is a possibility of exponential growth of elements just as for Gauss transformations with partial pivoting. The lemma that follows describes how an M-invariant reflection can enlarge a vector.

LEMMA 7.1. *Assume that* $d, x \in \mathbf{R}^m, d_1 \geq 0$. *Then for any* M*-invariant reflection*

$$Q = I_m - \frac{(d+\alpha e_1)(Nd+\alpha e_1)^T}{\alpha(\alpha+d_1)}, \tag{13}$$

where $N = \operatorname{diag}(1, \mu_1/\mu_2, \ldots, \mu_1/\mu_m)$, $\alpha = \|N^{1/2}d\|$ *and* $\zeta = (\|d\|^2 + 2d_1\alpha + \alpha^2)/2$ *we have*

$$\|Qx\|^2 = \|x\|^2 + 2\frac{(Nd+\alpha e_1)^T x}{\alpha(\alpha+d_1)}\left(\zeta\frac{(Nd+\alpha e_1)^T x}{\alpha(\alpha+d_1)} - (d+\alpha e_1)^T x\right). \tag{14}$$

Furthermore,

$$\|Q\| = \eta + \sqrt{\eta^2 - 1}, \qquad \eta = \frac{\|d+\alpha e_1\|\,\|Nd+\alpha e_1\|}{\alpha(\alpha+d_1)}, \tag{15}$$

and if $\|N^{1/2}x\| \leq \alpha$, *we have*

$$\|Qx\| \leq \|x\| + \|d\|. \tag{16}$$

Proof. For any M-invariant reflection $Q \in \mathbf{R}^{m\times m}$ we have a projection P such that $Q = I - 2P$. It is fairly easy to see that for any $x \in \mathbf{R}^m$ we have

$$\|Qx\|^2 = \|x\|^2 + 4\|Px\|\left(\|Px\| - \frac{(Px)^T x}{\|Px\|}\right). \tag{17}$$

In our case when Q is defined in (13), we can write

$$P = \frac{1}{2}\frac{(d+\alpha e_1)(Nd+\alpha e_1)^T}{\alpha(\alpha+d_1)}$$

if we assume that $d_1 \geq 0$. We assume without loss of generality that $(Nd+\alpha e_1)^T x \geq 0$ and from (17) we get (14). Our aim is to get an upper bound for

$$\phi = 2\frac{(Nd+\alpha e_1)^T x}{\alpha(\alpha+d_1)}\left(\zeta\frac{(Nd+\alpha e_1)^T x}{\alpha(\alpha+d_1)} - (d+\alpha e_1)^T x\right) \tag{18}$$

when $\|N^{1/2}x\| \leq \alpha$ to get the result in (16). It is obvious that the maximum is attained when $N = \operatorname{diag}(I_p, 0_q), p+q = m$ and therefore it is sufficient to consider $N = \operatorname{diag}(1,1,0,\ldots,0)$. Moreover, we may assume that $d_2 = 0$ because the factor $(Nd+\alpha e_1)^T x/[\alpha(\alpha+d_1)]$ is larger for smaller d_2 when $\|N^{1/2}x\| \leq \alpha$. We thus have $d = [d_1, 0, d_3, 0, \ldots, 0]^T$, $\alpha = d_1$, and consequently $x = [x_1, 0, x_3, 0, \ldots, 0]^T$ where $|x_1| \leq d_1$. Putting these x and d into the definition of ϕ we have

$$\phi = 2x_1/d_1\left(\frac{\|d\|^2 + 3d_1^2}{2}\frac{x_1}{d_1} - (2d_1x_1 + d_3x_3)\right)$$

and thus

$$\phi \leq \|d\|^2 - d_1^2 - 2d_3x_3 \leq \|d\|^2 + |d_1|\,|x_1| + 2|d_3|\,|x_3| \leq \|d\|^2 + 2\|x\|\,\|d\|,$$

which implies that $\|Qx\| \leq \|x\| + \|d\|$. For verification of (15) we refer to [9]. □

The next lemma shows that, in contrast to Householder and Gauss transformations, M-invariant reflections will not preserve the norm of the errors when they are reflected back to the original matrix A.

LEMMA 7.2. *There exists an M-invariant reflection*

$$Q = I_m - \frac{(d + \alpha e_1)(Nd + \alpha e_1)}{\alpha(\alpha + |d_1|)},$$

where $N = \operatorname{diag}(1, \mu_1/\mu_2, \ldots, \mu_1/\mu_m)$, $\alpha = \|N^{1/2}d\|$, and a vector $x = \left[0, x_2^T\right]^T \in \mathbf{R}^m$ with $\|N^{1/2}x\| \leq \alpha$ such that $\sup_{x,d} \|Qx\| = \|x\| + \|d\|$.

Proof. If we choose $N = \operatorname{diag}(1, 1, 0, \ldots, 0)$, $d = [0, d_2, d_3, 0, \ldots, 0]$, and $x = [0, d_2, -d_3, 0, \ldots, 0]$ we get $\alpha = |d_2|, d^T N x = d_2^2$. From (18) we have $\phi = 2d_2^2 + d_3^2 - 2d_2^2 + 2d_3^2 = 3d_3^2$. When $\|d_2\|$ approaches zero we get the maximal growth, i.e., $\|Qd\| = 2\|d\|$. Note that the condition number $\|Q\|\,\|Q^{-1}\|$ tends to infinity as $\|d_2\|$ tends to zero. □

8. Backward error for the weighted QR factorization. The following theorem gives a bound on the normwise backward error for the weighted QR factorization.

THEOREM 8.1. *Let $\hat{R}$ be the computed upper triangular matrix in the weighted QR factorization using column pivoting and $Q = Q_1 \cdots Q_n$ the product of the exact M-invariant reflections defined in* (10). *Then*

$$Q \begin{bmatrix} \hat{R} \\ 0 \end{bmatrix} \Pi = A + G, \qquad \|G\| \leq n^3 \rho \|A\| \gamma_{16m+35}(1 + \gamma_{16m+35})^n,$$

where the growth factor ρ defined in (21) *satisfies $1 \leq \rho \leq 2^n$.*

Proof. There are two main properties of the M-invariant reflections that enable us to prove the theorem. The structure in $\Delta\widetilde{Q}_k\widehat{c}_k$ enables us to use Lemma 7.1 and together with the norm invariant property in (12) we can limit the backward error as we now show.

We have from Lemma 6.1 that $\hat{A}_{n+1} = (Q_n + \Delta Q_n) \cdots (Q_1 + \Delta Q_1)A$, and thus

$$\hat{A}_{n+1} = \begin{bmatrix} \hat{R} \\ 0 \end{bmatrix} = \sum_{i=1}^{n} Q_n \ldots Q_{i+1} \Delta Q_i Q_{i-1} \ldots Q_1 A + Q^{-1}K, \tag{19}$$

where K is of second order. Using that $Q_k \hat{A}_k = \hat{A}_{k+1} - \Delta Q_k \hat{A}_k$ in (19) we get

$$Q \begin{bmatrix} \hat{R} \\ 0 \end{bmatrix} = A + \sum_{i=1}^{n} Q_n \ldots Q_i \Delta Q_i \hat{A}_i + E_2 = A + E_1 + E_2, \tag{20}$$

where we have put the new second order terms in E_2.

Consider a column in $\hat{A}_i$, $\hat{c}_j^{(i)}$, then we get from Lemma 7.1, equation (16), that

$$\|Q_n \ldots Q_i \Delta Q_i \hat{c}_j^{(i)}\| \leq \gamma_{10(m+2)} \sum_{l=1}^{i} \|\hat{d}_l\| + \gamma_{6m+14} \left(\alpha_i + \sum_{l=1}^{i-1} \|\hat{d}_l\| \right) + \mathbf{u} \left(\|\hat{c}_j^{(i)}\| + \sum_{l=1}^{i-1} \|\hat{d}_l\| \right).$$

Again from Lemma 7.1 we get $\|\hat{d}_l\| \leq 2^l \max_j \|a_j\|$, $\|\hat{c}_j^{(i)}\| \leq 2^i \max_j \|a_j\|$, and $\|N^{1/2}\hat{d}_i\| \leq \|\hat{d}_i\| \leq 2^i \max_j \|a_j\|$. If we define the growth factor, ρ, as

$$\rho = \frac{\max_{i,j} \|\hat{c}_j^{(i)}\|}{\max_i \|a_i\|} \leq 2^n, \tag{21}$$

we get $\|Q_1 \dots Q_i \Delta Q_i \hat{c}_j^{(i)}\| \leq n\rho\|A\|\gamma_{16m+35}$ and thus $\|E_1\| \leq n^2\rho\|A\|\gamma_{16m+35}$. The second order error matrix E_2 can be bounded using the same arguments as the ones used for E_1. By using the binomial expansion pattern easily obtained by limiting E_2 term by term, we get

$$\|E_2\| \leq n\rho\|A\| \sum_{i=2}^{n} (n-i+1) \begin{pmatrix} n \\ i \end{pmatrix} \gamma_{16m+35}^i.$$

Finally we have $\|E\| \leq \|E_1\| + \|E_2\| \leq n^2\rho\|A\|((1+\gamma_{16m+35})^n - 1)$, which completes the proof. □

It is probably possible, but not certain, that Theorem 8.1 can be attained from the analysis in [15] but then with additional analysis and another technique than we have used.

9. Backward errors for the solution. For the error in the solution we begin by presenting some results of error analysis concerning the important special case of the unweighted and constrained linear least squares problem

$$\min_{x \in \mathbf{R}^n} \|b_2 - A_2 x\| \quad \text{s.t.} \quad A_1 x = b_1. \tag{22}$$

If $A_1 \in \mathbf{R}^{p \times n}$ and $q = m - p$ then the matrix M for problem (22) is $M = \operatorname{diag}(0_p, I_q)$. For the proof of the following theorem we refer to [7].

THEOREM 9.1. *If $\hat{x}$ is the computed solution we get by solving problem* (22) *with the weighted QR factorization using column pivoting, then*

$$(A + F)\hat{x} \overset{M}{\simeq} b + f.$$

For the error matrix $F^T = \left[F_1^T, F_2^T\right]$, we have

$$\begin{bmatrix} \|F_1\|_F \\ \|F_2\|_F \end{bmatrix} \leq \begin{bmatrix} c_1 p^2 \|A_1\|_F \mathbf{u} \\ c_2(q(n-p) + \sqrt{n}\, p(n+p^2)\rho)\, \|A_2\|_F \mathbf{u} \end{bmatrix} + O(\mathbf{u}^2),$$

where $1 \leq \rho \leq 2^p$ is a growth factor. For $f^T = \left[f_1^T, f_2^T\right]$, we have

$$\|f_1\| = 0, \ \|f_2\| \leq c_f(p + (n-p)(m-n+q))\, \|b_2\|\mathbf{u} + O(\mathbf{u}^2).$$

The three constants c_1, c_2, and c_f are of modest size and independent of the problem size.

It is important to add that the perturbed problem for the calculated solution has the same λ (for a well-conditioned problem the relative normwise error will be small). The upper limits bounding the error matrix F and error vector f in the theorem can certainly be made tighter but we are only interested in the dominating factors in the bounds. Two interesting special cases are when $p = 0$ and $p = n - 1$, i.e., we have

no constraints and we have maximal number of constraints. When $p = 0$ we have the following upper bounds on the error

$$\|F\|_F \leq c_F mn \|A_1\| \mathbf{u} + O(\mathbf{u}^2)$$

and

$$\|f\| \leq c_f n(2m - n)\|b_2\| \mathbf{u} + O(\mathbf{u}^2),$$

which are nearly as good as the ones given in [10]. When $p = n - 1$ we have

$$\begin{bmatrix} \|F_1\|_F \\ \|F_2\|_F \end{bmatrix} \leq \begin{bmatrix} c_{12} n^2 \|A_1\|_F \mathbf{u} \\ c_{22}(m - n + 1 + \sqrt{n}\, n^2(n+1)\rho)\|A_2\|_F \mathbf{u} \end{bmatrix} + O(\mathbf{u}^2) \tag{23}$$

and

$$\|f\| \leq c_3(2m - n)\|b_2\| \mathbf{u} + O(\mathbf{u}^2).$$

The case when $p = n - 1$ should be compared with the existing error bounds for solving a system of linear equations $Ax = b$ where $A \in \mathbf{R}^{n \times n}$ and $b \in \mathbf{R}^n$ with Gauss transformations. The computed solution of this system, $\hat{x}$, can be shown (see [6, p. 67]) to satisfy the perturbed system of equations

$$(A + \Delta A)\hat{x} = b, \ \|\Delta A\|_\infty \leq c_4 n^3 \rho_{\text{Gauss}} \|A\|_\infty \mathbf{u} + O(\mathbf{u}^2). \tag{24}$$

If row or column pivoting is performed during the elimination process then the growth factor ρ_{Gauss} satisfies $1 \leq \rho_{\text{Gauss}} \leq 2^n$. We see that if $m \approx n$ our bound (23) is almost as good as the bound stated in (24). Moreover, it is a well-known fact that it is very seldom that the growth factor for Gaussian elimination with partial pivoting is large. We may thus conclude that *if* we use column pivoting *our* growth factor will be of modest size.

The next result is for problems where $\lambda = 0$. An example of such a problem is zero residual problems where M is invertible.

THEOREM 9.2. *If* $\gamma_{36m+122} < 1$ *and* $\lambda = 0$ *then the approximate solution* $\hat{x}$ *of* (4) *computed with the weighted* QR *factorization using column pivoting satisfies*

$$(A + F)\hat{x} \overset{M}{\simeq} b, \qquad \|F\| \leq c_1 n^3 \rho \|A\| \gamma_{c_2 m + c_3}(1 + \gamma_{c_2 m + c_3})^n,$$

where $c_i, i = 1, 2, 3$ *are positive integers of modest size independent of the problem size and* ρ *defined in* (21).

Proof. The computed solution $\hat{x}$ of (4) is gotten by solving $\hat{R}x = \hat{b}_1$ where $\hat{b}_1$ is the first n elments of $\hat{b}_{n+1} = \text{fl}(\hat{Q}_n \ldots \hat{Q}_1 b)$. Standard results on error analysis give that

$$(\hat{R} + \Delta\hat{R})\hat{x} = \hat{b}_1, \qquad |\Delta\hat{R}| \leq \gamma_{n+1}|\hat{R}|.$$

The assumption that $\lambda = 0$ makes it possible to reflect *all* the errors made in the transformation of b back to A. This is easily seen if we look at the system equations

$$\begin{bmatrix} M_n & 0 & \hat{R} + \Delta\hat{R} \\ 0 & M_{m-n} & 0 \\ (\hat{R} + \Delta\hat{R})^T & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \hat{x} \end{bmatrix} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ 0 \end{bmatrix}.$$

From Lemma 6.1 we get $(Q_1 + \Delta P_1)\dots(Q_n + \Delta P_n)[\hat{R}^T + \Delta\hat{R}^T, 0]^T\hat{x} = b$ or using the result in Theorem 8.1 that $(A + G + F)\hat{x} = b$. Invoking Lemma 6.1 gives $\|F\| \leq c_1 n^3 \rho \|A\| \gamma_{c_2 m + c_3}(1 + \gamma_{c_2 m + c_3})^n$ where $c_i, i = 1, 2, 3$ are small integers independent of m and n. □

Even if the theorem is a special case we notice that the weight matrix is intact as well as the symmetry *and* that the new λ for the perturbed problem, $\tilde{\lambda} = 0$.

The next theorem is similar to the one given above but now we attack the general problem (4).

THEOREM 9.3. *If $\sigma = max_{1 \leq k \leq n} \sigma_k$, where σ_k is defined in* (25), *then the approximate solution $\hat{x}$ of* (4) *computed with the weighted QR factorization using column pivoting satisfies*

$$(A + F)\hat{x} \overset{M}{\simeq} b + f,$$

where

$$\|F\| \leq c_1 n^3 \rho \|A\| \gamma_{c_2 m + c_3}(1 + \gamma_{c_2 m + c_3})^n,$$

$$\|f\| \leq c_4 n(n\rho\sigma\|A\| + \|b\|)\gamma_{c_5 m + c_6}(1 + \gamma_{c_5 m + c_6})^n,$$

$c_i, i = 1, \dots, 6$ are positive integers of modest size independent of the problem size, and ρ is defined in (21).

Proof. The difference from the last theorem is that we get a possible magnification σ of the right-hand side. This is due to the fact that if we do not reflect the error in the right-hand side back to $\hat{R} + \Delta\hat{R}$, Lemmas 7.1 and 6.1 only give us

$$\|Q_1 \dots Q_i \Delta Q_i \hat{b}_k\| \leq \gamma_{22m+49}\sigma_k \sum_{i=1}^{k} \|\hat{d}_i\| + \|\hat{b}_2^{(k)}\|\mathbf{u},$$

where

$$\sigma_k = \frac{\|N_k^{1/2}\hat{b}_2^{(k)}\|}{\|N_k^{1/2}\hat{d}_k\|}. \tag{25}$$

Using Lemmas 7.1 and 6.1 again on $\|\hat{b}_2^{(k)}\|$ and ignoring second order terms, we get $\|\hat{b}_2^{(k)}\| \leq \|b\| + \sum_{i=1}^{k-1} \sigma_i \|\hat{d}_i\|$ and can conclude that

$$\|Q_1 \dots Q_i \Delta Q_i \hat{b}_k\| \leq \gamma_{22m+49}\sigma \sum_{i=1}^{k} \|\hat{d}_i\| + \|b\|\mathbf{u}.$$

The rest of the proof consists of collecting the higher order terms. □

The theorem is perhaps not as good as one would hope the algorithm would give, but it is impossible to get rid of the factor σ in the right-hand side without altering the original problem. However, the result in Theorem 9.3 is at least as good as the result attained in [15] for weighted, unconstrained least squares but the factor σ is given a slightly different form. Indeed, it may be possible to use the analysis in [15] to prove Theorem 9.1 but with additional analysis and with a technique other than we have used.

It is possible to scale the matrix A and the solution x to keep σ small and for this scaled problem the backward error will be small. Imagining b as an extra column in A this scaling means that b_k is never exchanged with any other column in A_k.

We stress the importance of sorting the weights in our algorithm. In fact, it is easy to show that solving (1) with the weighted QR factorization without sorting the weights is an unstable algorithm. This corresponds to the well-known fact that the algorithm in [15] is unstable without row pivoting.

One alternative to Theorem 9.3 is to expand the problem class in some way. First we consider the case when the system matrix is perturbed in an unsymmetric way. The proof of the theorem is similar to the proof of Theorem 9.2 and therefore it is omitted; see [8] for the details.

THEOREM 9.4. *The approximate solution $\hat{x}$ of* (4) *computed with the weighted QR factorization using column pivoting satisfies*

$$\begin{bmatrix} M & A+E_1 \\ (A+E_2)^T & 0 \end{bmatrix} \begin{bmatrix} \tilde{\lambda} \\ \hat{x} \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix},$$

where

$$\|E_j\| \leq c_1 n^3 \rho \|A\| \gamma_{c_2 m + c_3} (1 + \gamma_{c_2 m + c_3})^n \quad j = 1, 2,$$

and $c_i, i = 1, 2, 3$ are positive integers of modest size independent of the problem size m and n and ρ is defined in (21).

The theorem might seem a bit suspect because we have managed to get rid of σ in Theorem 9.3 without any scaling. This is unfortunately not the case because when reflecting the errors from the right-hand side to the matrix we must transform

$$\lambda = Q_1^T \dots Q_n^T \begin{bmatrix} 0 & 0 \\ 0 & M_{m-n}^{-1} \end{bmatrix} Q_n \dots Q_1 b$$

to

$$\tilde{\lambda} = (Q_1 + \Delta T_1)^{-T} \dots (Q_n + \Delta T_n)^{-T} \begin{bmatrix} 0 & 0 \\ 0 & M_{m-n}^{-1} \end{bmatrix} (Q_1 + \Delta Z_1) \dots (Q_n + \Delta Z_n) b$$

for some small error matrices ΔZ_i and ΔT_i. It is not evident that λ is of the same size as $\tilde{\lambda}$.

The final theorem is interesting because here we have put the errors emanating from the transformation of b in A *and* M. In other words we have an ordinary backward error result *if we consider the class of constrained and weighted linear least squares problems.*

THEOREM 9.5. *The approximate solution $\hat{x}$ of* (4) *computed with the weighted QR factorization using column pivoting satisfies*

$$\begin{bmatrix} M + \Delta M & A + E \\ (A+E)^T & 0 \end{bmatrix} \begin{bmatrix} \tilde{\lambda} \\ \hat{x} \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix},$$

where

$$\|E\| \leq c_1 n^3 \rho \|A\| \gamma_{c_2 m + c_3} (1 + \gamma_{c_2 m + c_3})^n$$

and $\Delta M = (\Delta M)^T$ has the componentwise bound

$$|\delta \mu_{ij}| \leq c_4 n \rho \mu_i \gamma_{c_5 m^2 + c_6} (1 + \gamma_{c_5 m^2 + c_6})^{2n} \quad i = 1, \dots, m, \qquad j = i, \dots, m.$$

The integers $c_i \geq 0, i = 1, \ldots, 6$ *are of modest size and independent of the problem size and* ρ *is defined in* (21).

Proof. Using the same technique as in the proof of Theorem 9.4 we multiply from the left with $(Q_1 + \Delta P_1) \ldots (Q_n + \Delta P_n)$ but from the right with the transpose $(Q_n + \Delta P_n)^T \ldots (Q_1 + \Delta P_1)^T$. The symmetry is retained but M is transformed to

$$\bar{M} = (Q_1 + \Delta P_1) \ldots (Q_n + \Delta P_n) M (Q_n + \Delta P_n)^T \ldots (Q_1 + \Delta P_1)^T \tag{26}$$

or $\bar{M} = M + \Delta M$, where the symmetry of M and $\bar{M}$ implies the symmetry of ΔM. We now assume, without loss of generality, that M is invertible. Consider the first order terms $G_i = Q_1 \ldots Q_{i-1} \Delta P_i M Q_{i-1}^T \ldots Q_1^T$ in (26), where ΔP_i has the same size and structure as ΔQ_i in Lemma 6.1. We are now able to use the same arguments as in the proof of Theorem 8.1 to show that the lower $(m - i + 1) \times (m - i + 1)$ block of $M^{-1/2} G_i M^{-1/2}$ only contains elements smaller than $\bar{c}_1 n \rho \mu_i \gamma_{\bar{c}_2 m^2 + \bar{c}_3}$ for some constants $\bar{c}_1, \bar{c}_2$ and $\bar{c}_3$ (the rest of $M^{-1/2} G_i M^{-1/2}$ is zero). The rest of the proof consists of collecting the higher order terms, which gives the exponential factor in the theorem. □

10. Perturbation bounds and explicit bounds on the calculated solution. Following [16] the inverse of the system matrix

$$\begin{bmatrix} M & A \\ A^T & 0 \end{bmatrix}^{-1} = \begin{bmatrix} H & B^T \\ B & -BMB^T \end{bmatrix},$$

where B is a generalized inverse of A, i.e., $BA = I_n$, and H satisfies $HA = 0$. If we define $\kappa = \|A\| \, \|B\|$, $\kappa_x = \|A\| \, \|BM^{1/2}\| / \|M^{1/2}\|$, and $\kappa_\lambda = \|M\| \, \|H\|$ we have from [16] (see also [17]) that

$$\frac{\|\delta\lambda\|}{\|\lambda\|} \leq \kappa \epsilon_A + \kappa_\lambda \left(\epsilon_M + \frac{\|A\| \, \|x\|}{\|M\| \, \|\lambda\|} \epsilon_A + \frac{\|b\|}{\|M\| \, \|\lambda\|} \epsilon_b \right) + \mathcal{O}(\epsilon^2)$$

and

$$\frac{\|\delta x\|}{\|x\|} \leq \kappa \left(\epsilon_A + + \frac{\|b\|}{\|A\| \, \|x\|} \epsilon_b + \frac{\|M\| \, \|\lambda\|}{\|A\| \, \|x\|} \epsilon_M \right) + \kappa_x^2 \frac{\|M\| \, \|\lambda\|}{\|A\| \, \|x\|} \epsilon_A + \mathcal{O}(\epsilon^2),$$

where $\epsilon_M = \|\delta M\| / \|M\|$, $\epsilon_A = \|\delta A\| / \|A\|$, $\epsilon_b = \|\delta b\| / \|b\|$, and $\epsilon = \max\{\epsilon_M, \epsilon_A, \epsilon_b\}$. The bounds in Theorem 9.3 or 9.5 can now be applied to get normwise relative errors in x and λ if we have an approximate solution $\hat{x}$ *and* an approximation of λ.

The factor σ in Theorem 9.3 is a nuisance because it seems as if the relative error in x will become large even for well-conditioned problems, i.e., the method is potentially unstable. For a general weight matrix we have not shown that the relative error will not be amplified by a large σ but for the constrained and unweighted least squares problem and zero residual problems we have come further. For constrained least squares, it is easily seen that $\sigma = \max \|b_{21}^{(k)}\| / \|d_1^{(k)}\|$, where $d_1^{(k)}$ and $b_{21}^{(k)}$ are the parts of d_k and $b_2^{(k)}$ corresponding to the constraints. Furthermore, we have from the constraints that $b_1 = A_1 x$ and with the column pivoting we have $\|b_1\| \leq n \|d_1^{(k)}\| \, \|x\|$, i.e., $\|x\| \geq \|b_1\| / (n \|d_1^{(k)}\|)$. Both b_1 and A_1 are transformed by orthogonal transformations and thus we have $\|x\| \geq \|b_{21}^{(k)}\| / (n \|d_1^{(k)}\|)$ and consequently $\sigma / \|x\| \leq n$. Zero residual problems are treated similarly but to get the inequality $\|x\| \geq \|N_k^{1/2} b_2^{(k)}\| / \|N_k^{1/2} d_k\|$ we must use that $b = Ax$ together with the property in (12).

11. Conclusions and further work. We have proved that the algorithm computing the weighted QR factorization is backward stable and that the algorithm for the constrained and weighted linear least squares problem is backward stable when the right-hand side b is properly scaled.

Future work could consist of an investigation of how the backward error bounds depend on the right-hand side. Givens rotations can be generalized to M-invariant Givens rotations and it remains to prove their stability and the backward stability when using them for the constrained and weighted linear least squares problem. Finally, it is quite possible to use M-invariant transformations in a weighted Gram–Schmidt algorithm and the error analysis used in this article is well suited for analyzing this algorithm.

REFERENCES

[1] J. L. BARLOW AND S. L. HANDY, *The direct solution of weighted and equality constrained least squares problems*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 704–716.

[2] Å. BJÖRCK, *Iterative refinement of linear least squares solutions* I, BIT, 7 (1967), pp. 257–278.

[3] ———, *Component-wise perturbation analysis and error bounds for linear least square solutions*, BIT, 31 (1990), pp. 238–244.

[4] ———, *Error analysis of the modified Gram–Schmidt method for solving underdetermined linear systems*, Tech. Report LiTH-MAT-R-1990-06, Department of Mathematics, Linköping University, 1990.

[5] ———, *Error analysis of least squares algorithms*, in Numerical Linear Algebra, Digital Signal Processing and Parallel Algorithms, G. H. Golub and P. V. Dooren, eds., NATO ASI Series, Berlin, 1991, Springer-Verlag, pp. 41–73.

[6] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computation*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[7] M. E. GULLIKSSON, *Error analysis of an algorithm for constrained linear least squares*, Tech. Report UMINF 91-01, Inst. of Info. Proc., Univ. of Umeå, S-901 87 Umeå, Sweden, 1991.

[8] ———, *Backward error analysis for the constrained and weighted linear least squares problem when using the modified QR decomposition*, Tech. Report UMINF 93.01, Inst. of Info. Proc., Univ. of Umeå, S-901 87 Umeå, Sweden, 1993.

[9] M. E. GULLIKSSON AND P.-Å. WEDIN, *Modifying the QR decomposition to weighted and constrained linear least squares*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1298–1313.

[10] R. J. HANSON AND C. L. LAWSON, *Solving least squares problems*, Prentice Hall, Englewood Cliffs, NJ, 1974.

[11] N. J. HIGHAM, *Computing error bounds for regression problems*, in Statistical Analysis of Measurement Error Models and Applications, Contemporary Mathematics 112, P. J. Brown and W. A. Fuller, eds., Amer. Math. Soc., Providence, RI, 1990, pp. 195–208.

[12] ———, *Iterative refinement enhances the stability of QR factorization methods for solving linear equations*, BIT, 31 (1991), pp. 447–468.

[13] C. C. PAIGE, *Computer solution and perturbation analysis of generalized linear least squares problems*, Math. Comput., 33 (1964), pp. 171–183.

[14] ———, *Fast numerically stable computaions for generalized linear least squares problems*, SIAM J. Numer. Anal. Appl., 16 (1979), pp. 165–171.

[15] M. J. D. POWELL AND J. K. REID, *On applying Householder's method to linear least squares problems*, in Proc. IFIP Congress 68, A. J. M. Morell, ed., Amsterdam, 1969, North Holland, pp. 122–126.

[16] P.-Å. WEDIN, *Perturbation theory and condition numbers for generalized and constrained linear least squares problems*, Tech. Report UMINF 125.85, Inst. of Info. Proc., Univ. of Umeå, Umeå, Sweden, 1985.

[17] ———, *Perturbation results and condition numbers for outer inverses and especially for projections*, Tech. Report UMINF 124.85, Inst. of Info. Proc. , Univ. of Umeå, Umeå, Sweden, 1985.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 688–696, April 1995

023

APPROXIMATIONS TO SOLUTIONS TO SYSTEMS OF LINEAR INEQUALITIES *

OSMAN GÜLER†, ALAN J. HOFFMAN‡, AND URIEL G. ROTHBLUM§

Abstract. In this paper we consider a result of Hoffman [*J. Res. Nat. Bur. Stand.*, 49 (1952) pp. 263–265] about approximate solutions to systems of linear inequalities. We obtain a new representation for a corresponding Lipschitz bound via singular values. We also provide geometric representations of these bounds via extreme points. The latter have been developed independently by Bergthaller and Singer [*Linear Algebra Appl.*, 169 (1992), pp. 111–129] and Li [*Linear Algebra Appl.*, 187 (1993), pp. 15–40], but, our proofs are simpler. We obtain a particularly simple proof of Hoffman's existence result which relies only on linear programming duality.

Key words. linear systems, approximations, singular values

AMS subject classifications. 15A18, 15A39, P0C05

1. Introduction.

For a vector $u \in R^k$, let u^+ be the vector in R^k with $(u^+)_i = \max\{u_i, 0\}$ for each $i = 1, \ldots, k$. The following result was first established by Hoffman [8]; for an alternative proof see Robinson [21].

THEOREM 1.1 (Hoffman [8]). *Let $A \in R^{m \times n}$ and let $\| \quad \|_\alpha$ and $\| \quad \|_\beta$ be norms on R^n and on R^m, respectively. Then there exists a scalar $K_{\alpha\beta}(A)$, such that for each $b \in R^m$ for which the set $\{x \in R^n : Ax \leq b\} \neq \emptyset$ and for each $x' \in R^n$,*

$$\min_{Ax \leq b} \|x - x'\|_\alpha \leq K_{\alpha\beta}(A) \|(Ax' - b)^+\|_\beta. \tag{1.1}$$

In particular, the minimum on the left-hand side of (1.1) *is attained.*

We refer to a coefficient $K_{\alpha\beta}(A)$ that satisfies the conclusion of Theorem 1.1 as a *Lipschitz bound* of A. In this paper we obtain a particularly simple proof of Hoffman's existence result that relies only on linear programming duality. As such it extends from the real field to arbitrary ordered fields. The new proof simplifies that of Mangasarian and Shiau [18] who also used linear programming duality. We also provide geometric representations of the bounds via extreme points. The latter generalizes results of Mangasarian [15] and Mangasarian and Shiau [18] who considered the case where $\| \quad \|_\alpha$ is the l_∞ norm; they overlap recent results of Bergthaller and Singer [3] and Li [11], which were derived independently. Finally, we use singular values of submatrices to express relaxations of the Lipschitz bounds for the case where both $\| \quad \|_\alpha$ and $\| \quad \|_\beta$ are the l_2 norms.

Theorem 1.1 has many applications in optimization, perturbation theory, and sensitivity analysis. A complete list of references is beyond the scope of this paper and we mention only a few examples. Agmon [1], Motzkin and Schoenberg [19], and

* Received by the editors September 24, 1992; accepted for publication (in revised form) by R. W. Cottle, April 15, 1994.

† Department of Mathematics and Statistics, University of Maryland at Baltimore County, Baltimore, Maryland 21228 (guler@math.umb).

‡ IBM Thomas J. Watson Research Center, Post Office Box 218, Yorktown Heights, New York 10598 (hoffman@yktvm.bitnet).

§ Faculty of Industrial Engineering and Management, Technion-Israel Institute of Technology, Haifa 32000, Israel (ierur10@technion.technion.ac.il). The research of this author was supported in part by Office of Naval Research grant N00014-92-J1142 and Air Force grants 89-0512 and 90-0008.

Goffin [5] use Theorem 1.1 to prove the linear convergence rate of relaxation algorithms. Polyak and Treti'akov [20] use Theorem 1.1 to prove the finite convergence of the proximal minimization algorithm for solving linear programs, and Güler [7] uses it to establish a finite convergence of an accelerated version of the Polyak–Treti'akov algorithm. Robinson [21] applies it to study the solution set of perturbed linear programming problems. Theorem 1.1 has been applied recently to obtain results in convex programming. Mangasarian and Shiau [17], [18] use it to obtain error bounds for monotone linear complementarity problems. Iusem and De Pierro [10] use a version of Theorem 1.1 to prove the (asymptotic) linear convergence of Hildreth's "row action" algorithm for solving large scale quadratic programs arising from image reconstruction problems in computerized tomography. Luo and Tseng [13] use a weaker version of Theorem 1.1 to develop convergence measure for solving a class of convex programs, which is used to prove the (asymptotic) linear convergence rate of some popular algorithms, including the gradient projection method and the matrix-splitting algorithms for linear complementarity problems. Ye et al. [23] used Theorem 1.1 to determine convergence rates of a particular interior point method.

There have been a number of recent generalizations of Theorem 1.1 and studies of the bound $K_{\alpha\beta}(A)$ that appears in (1.1). Robinson [22] extended this bound to a system of convex inequalities which defines a bounded feasible region with a nonempty interior, and Mangasarian [16] extended it further to a system of differentiable convex inequalities which satisfy Slater's condition and an asymptotic constraint qualification instead of the boundedness assumption on the feasible region. Auslender and Crouzeix [2] relaxed the differentiability assumption and replaced it, among others, by the assumption that the underlying functions are proper and closed and that the interior of their effective domain contains the feasible set. Hu and Wang [9] generalized Theorem 1.1 to infinite systems of linear inequalities. Cook et al. [4, Thm. 1] extended Theorem 1.1 to integer problems. Bergthaller and Singer [3] and Li [11] sharpened these extensions and obtained results that are related to ours and that are valid in infinite dimensional spaces. Luo and Tseng [14] studied uniform boundedness of the constant $K_{\alpha\beta}(A)$ as the underlying matrix A and the right-hand side b of the corresponding inequality systems are perturbed.

Some of the above references consider a variant of Theorem 1.1 where both linear equalities and inequalities are present in the left-hand side of (1.1). But, the standard representation of linear equalities through pairs of linear inequalities reduces this apparently more general situation to the more restricted one that is considered here. When linear equalities are not present, those previous results relate to those of the present paper.

2. Proof via linear programming duality. We denote the l_∞ and l_1 norms in R^n and $R^{1\times n}$ by $\|\ \ \|_\infty$ and $\|\ \ \|_1$, respectively, e.g., for a vector a in R^n, $\|a\|_\infty = \max\{|a_i|\colon i = 1, \ldots, n\}$ and $\|x\|_1 = \sum_{i=1}^n |a_i|$. We next use linear programming duality to obtain an explicit representation of the coefficient $K_{\alpha\beta}(A)$ for the special instance of Theorem 1.1 with the l_∞ norms, thereby obtaining a new elementary proof of Theorem 1.1 for that special case. The general case of Theorem 1.1 is then derived from the equivalence of all norms in R^n.

THEOREM 2.1. *Let* $A \in R^{m\times n}$,

$$\sigma(A) \equiv \{\lambda \in R^m\colon \lambda \geq 0, \|\lambda^T A\|_1 \leq 1\} \tag{2.1}$$

and

$$K(A) \equiv \max\{\|\lambda\|_1\colon \lambda \textit{ is an extreme point of } \sigma(A)\}. \tag{2.2}$$

Then for each $b \in R^m$ for which the set $\{x \in R^n : Ax \leq b\} \neq \emptyset$ and for each $x' \in R^n$,

$$\min_{Ax \leq b} \|x - x'\|_\infty \leq K(A)\|(Ax' - b)^+\|_\infty. \tag{2.3}$$

In particular, the maximum on the right-hand side of (2.2) *and the minimum on the left-hand side of* (2.3) *are attained.*

Proof. Let $b \in R^m$ and $x' \in R^n$ be given where $\{x \in R^n : Ax \leq b\} \neq \emptyset$. Standard arguments show that the minimum of the left-hand side of (2.3) is attained. Thus, we have that

$$\min_{Ax \leq b} \|x - x'\|_\infty = \min_{A(x-x') \leq b - Ax'} \|x - x'\|_\infty = \min_{Az \geq Ax' - b} \|z\|_\infty = \min_{Az \geq a} \|z\|_\infty, \tag{2.4}$$

where $a \equiv Ax' - b$. In particular, $\{z \in R^n : Az \geq a\} \neq \emptyset$ and all minima in (2.4) are finite and are attained.

For $i = 1, \ldots, n$, let e^i be the ith unit vector in R^n, let $F \equiv \{e^i\colon\ i = 1, \ldots, n\} \cup \{-e^i\colon\ i = 1, \ldots, n\}$, and let B be the $2^n \times n$ matrix whose rows are all the possible arrangements of $+1$'s and -1's. Then the convex hull of F, denoted $\operatorname{conv}(F)$, is the set $\{u \in R^n : Bu \leq \mathbf{1}\}$, where $\mathbf{1}$ denotes the vector all of whose coordinates are 1. It follows that for each vector $z \in R^n$,

$$\|z\|_\infty = \max\{z^T f\colon f \in F) = \max\{z^T f\colon f \in \operatorname{conv}(F)\} = \max\{z^T f\colon Bf \leq \mathbf{1}\}.$$

Also, $\operatorname{conv}(F)$ is the l_1 unit ball, i.e.,

$$\|u\|_1 \leq 1 \quad \text{if and only if } Bu \leq \mathbf{1}.$$

These facts and the duality theorem for linear programming imply that

$$\begin{aligned} \min_{Az \geq a} \|z\|_\infty &= \min_{Az \geq a} \max_{Bf \leq \mathbf{1}} z^T f = \min_{Az \geq a} \min_{\substack{y^T B = z^T \\ y \geq 0}} y^T \mathbf{1} = \min_{\substack{Az \geq a \\ B^T y - z = 0 \\ y \geq 0}} y^T \mathbf{1} \\ &= \max_{\substack{\mu^T B^T \leq \mathbf{1}^T \\ \lambda^T A - \mu^T = 0 \\ \lambda \geq 0}} \lambda^T a = \max_{\substack{\lambda^T A B^T \leq \mathbf{1}^T \\ \lambda \geq 0}} \lambda^T a \\ &= \max_{\substack{\|A^T \lambda\|_1 \leq 1 \\ \lambda \geq 0}} \lambda^T a = \max_{\lambda \in \sigma(A)} \lambda^T a. \end{aligned} \tag{2.5}$$

We note that equality of the optimal primal and dual objective values of linear programs must be justified, e.g., by verification that either one is finite. Indeed, we observed in (2.4) that the first term of (2.5) is finite. We conclude that the min and max at the extreme ends of (2.5) are attained and are finite. But, the last term of (2.5) is the maximal value of a linear functional on a pointed polyhedral set (i.e., a polyhedral set containing no lines) and therefore equals the maximal value of this linear functional over the extreme points of this set. For each such extreme point λ,

$$\lambda^T a \leq \lambda^T (a)^+ \leq \|\lambda\|_1 \|(a)^+\|_\infty. \tag{2.6}$$

As $\sigma(A)$ is a polyhedral set, it has finitely many extreme points. Thus, the right-hand side of (2.2) is well defined and (2.3) follows immediately from (2.4)–(2.6) and the definition of a. □

Our proof of Theorem 2.1 simplifies the proof of Mangasarian and Shiau [18, Thm. 2.2] who consider a more general situation where the l_∞ norm on the right-hand side of (2.3) is replaced by an arbitrary norm and the l_1 norm in (2.2) is replaced by the dual of that norm. In §3, their generalization of Theorem 2.1 is further extended by replacing the l_∞ norm on the left-hand side of (2.3) by another arbitrary norm and the l_1 norm in (2.1) by the dual of that second norm.

The following known corollary of Theorem 2.1 shows that the correspondence where $b \in R^m$ is mapped into $\{x \in R^n : Ax \leq b\}$ is uniformly lower semicontinuous; see Mangasarian and Shiau [18, Thm. 2.2]. A proof is included for the sake of completeness.

COROLLARY 2.2. *Let $A \in R^{m\times n}$ and let $K(A)$ be defined as in Theorem 2.1. Let $x' \in R^n$ and $b' \in R^m$ satisfy $Ax' \leq b'$, and let $b \in R^m$ have $\{x \in R^n : Ax \leq b\} \neq \emptyset$. Then there exists a vector $x \in R^n$ satisfying $Ax \leq b$ with*

$$\|x' - x\|_\infty \leq K(A)\|b' - b\|_\infty. \tag{2.7}$$

Proof. Theorem 2.1 implies the existence of a vector $x \in R^n$ satisfying $Ax \leq b$ and

$$\|x' - x\|_\infty \leq K(A)\|(Ax' - b)^+\|_\infty. \tag{2.8}$$

Let $s \equiv b' - Ax' \in R^m$. As $s \geq 0$, we have that

$$\begin{aligned} \|b' - b\|_\infty &= \|(Ax' + s) - b\|_\infty = \|(Ax' - b) + s\|_\infty = \max_i\{|(Ax' - b)_i + s_i|\} \\ &\geq \max_i\{[(Ax' - b)_i]^+\} = \|(Ax' - b)^+\|_\infty, \end{aligned} \tag{2.9}$$

and (2.7) follows directly from (2.8) and (2.9). □

We derived Corollary 2.2 from Theorem 2.1. We next observe that we also have the reverse implication. Specifically, suppose that $A \in R^{m\times n}$ and $K(A) \in R$ satisfy the conclusion of Corollary 2.2 and that $b \in R^m$, $x' \in R^m$, and $\{x \in R^n : Ax \leq b\} \neq \emptyset$. Let $b' \equiv b \vee Ax'$ (where $\vee$ denotes pointwise maximum). Then $Ax' \leq b'$, and (2.7) implies that for some x satisfying $Ax \leq b$, $\|x' - x\|_\infty \leq K(A)\|b' - b\|_\infty = K(A)\|(Ax' - b)^+\|_\infty$.

The following example demonstrates that the function mapping a matrix $A \in R^{m\times n}$ to the corresponding coefficient $K(A)$ defined in Theorem 2.1 is not continuous and is not bounded in terms of the coefficients of the matrix.

Example. For $\varepsilon \geq 0$, let

$$A^\varepsilon \equiv \begin{vmatrix} 1+\varepsilon & -1 \\ -1 & 1+\varepsilon \end{vmatrix}.$$

Then the set $\sigma(A^\varepsilon)$ consists of the set of all nonnegative vectors $\lambda \in R^2$ that satisfy the inequalities

$$\begin{aligned} \varepsilon\lambda_1 + \varepsilon\lambda_2 &= [(1+\varepsilon)\lambda_1 - \lambda_2] + [-\lambda_1 + (1+\varepsilon)\lambda_2] \leq 1, \\ (2+\varepsilon)(\lambda_1 - \lambda_2) &= [(1+\varepsilon)\lambda_1 - \lambda_2] - [-\lambda_1 + (1+\varepsilon)\lambda_2] \leq 1, \\ (2+\varepsilon)(\lambda_2 - \lambda_1) &= -[(1+\varepsilon)\lambda_1 - \lambda_2] + [-\lambda_1 + (1+\varepsilon)\lambda_2] \leq 1 \\ -\varepsilon\lambda_1 - \varepsilon\lambda_2 &= -[(1+\varepsilon)\lambda_1 - \lambda_2] - [-\lambda_1 + (1+\varepsilon)\lambda_2] \leq 1. \end{aligned}$$

As the extreme points of these sets are $[0,0], [(2+\varepsilon)^{-1}, 0], [0, (2+\varepsilon)^{-1}], \varepsilon^{-1}(2+\varepsilon)^{-1}$ $[(1+\varepsilon), 1]$, and $\varepsilon^{-1}(2+\varepsilon)^{-1}[1, (1+\varepsilon)]$ when $\varepsilon > 0$ and $[0,0], [0, 2^{-1}], [2^{-1}, 0]$ when $\varepsilon = 0$, we have that

$$K(A) = \varepsilon^{-1} \quad \text{if } \varepsilon > 0 \quad \text{and} \quad K(A) = 2^{-1} \quad \text{if } \varepsilon = 0.$$

In particular, we see that $K(\cdot)$ is not a continuous function; in fact, $K(A)$ is not bounded in terms of the elements of A; see Luo and Tseng [14] for a more detailed study of the behavior of $K(A)$ under various perturbations of the data (A, b). □

Recall that for every pair of norms $\|\ \|$ and $\|\ \|'$ on R^n there exists a positive constant C such that for every $x \in R^n$, $\|x\|' \leq C\|x\|$. This fact combines with standard arguments to yield the following corollary of Theorem 2.1.

COROLLARY 2.3. *Let $A \in R^{m\times n}$ and let $K(A)$ be the constant defined as in Theorem* 2.1. *Also, let $\|\ \|_\alpha$ and $\|\ \|_\beta$ be norms on R^n and on R^m, respectively, and let L and M be positive scalars such that $\|z\|_\alpha \leq L\|z\|_\infty$ for $z \in R^n$ and $\|y\|_\infty \leq M\|y\|_\beta$ for all $y \in R^m$. Then for each $b \in R^m$ for which the set $\{x \in R^n : Ax \leq b\} \neq \emptyset$ and for each $x' \in R^n$,*

$$\min_{Ax\leq b} \|x - x'\|_\alpha \leq K_{\alpha\beta}(A)\|(Ax' - b)^+\|_\beta, \tag{2.10}$$

where $K_{\alpha\beta}(A) \equiv LMK(A)$. In particular, the minimum on the left-hand side of (2.10) *is attained.*

Robinson [21] established Theorem 1.1 by first considering the l_2 norm. He then used the equivalence of all norms over a finite dimensional normed space and the arguments of the above proof of Corollary 2.3 to establish the general case. But Robinson's representation of the coefficient $K_{\alpha\beta}(A)$ for a given matrix A in the specific case where $\|\ \|_\alpha$ and $\|\ \|_\beta$ are the l_2 norms was not explicit. In §4 we obtain explicit bounds for the case where $\|\ \|_\alpha$ and $\|\ \|_\beta$ are the l_2 norm.

Corollary 2.3 and the arguments used to deduce Corollary 2.2 from Theorem 2.1 can next be used to extend Corollary 2.2 to arbitrary norms $\|\ \|_\alpha$ and $\|\ \|_\beta$ by replacing (2.7) with

$$\|x' - x\|_\alpha \leq K_{\alpha\beta}(A)\|b' - b\|_\beta. \tag{2.11}$$

3. Bounds for general norms. For a norm $\|\ \|$ on R^k, let $\|\ \|^*$ be the *dual norm* defined for each $\lambda \in R^k$ by $\|\lambda\|^* \equiv \max\{\lambda^T x : x \in R^k, \|x\| \leq 1\}$. A norm $\|\ \|$ on R^n is called *polyhedral* if the unit ball $\{z \in R^n : \|z\| \leq 1\}$ is a polyhedral set. It is well known that a norm $\|\ \|$ is polyhedral if and only if its dual norm $\|\ \|^*$ is polyhedral. Now, if $\|\ \|$ is a polyhedral norm and the unit ball $\{z \in R^n : \|z\|^* \leq 1\}$ of its dual norm $\|\ \|^*$ has the form $\{z \in R^n : Bz \leq b\}$ for some matrix $B \in R^{q\times n}$ and vector $b \in R^q$, where q is a positive integer, then for every $u \in R^n$, $\|u\| = \|u\|^{**} = \max\{u^T f : Bf \leq b\}$. It follows that the arguments used in the proofs of Theorem 2.1 and Corollary 2.2 (that rely only on linear programming duality) can be carried through and the conclusion of Theorem 1.1 holds with

$$K_{\alpha\beta}(A) \equiv \max\{\|\lambda\|_\beta^* : \lambda \text{ is an extreme point of } \sigma_\alpha(A)\},$$

where

$$\sigma_\alpha(A) \equiv \{\lambda \in R^m : \lambda \geq 0, \|\lambda^T A\|_\alpha^* \leq 1\}.$$

So, a direct proof to Theorem 1.1 is obtained with a coefficient $K_{\alpha\beta}(A)$, which obviously yields tighter bounds than those derived in Corollary 2.3.

In this section we establish corresponding improvements when $\|\ \|_\alpha$ and $\|\ \|_\beta$ are arbitrary (not necessarily polyhedral) norms. In this situation, linear programming duality, as employed in §2 is not enough to obtain the desired results. Overlapping results were obtained, independently, by Bergthaller and Singer [3] and Li [11].

The next lemma extends (2.5) from the l_∞ norm to arbitrary norms. The proof we provide augments the use of the duality theorem of linear programming, used to establish (2.5), with a standard separation theorem for convex sets.

LEMMA 3.1. *Let $A \in R^{m\times n}$, $a \in R^m$, $\| \quad \|_\alpha$ be a norm on R^n and*

$$\sigma_\alpha(A) = \{\lambda \in R^m : \lambda \geq 0, \|\lambda^T A\|_\alpha^* \leq 1\}. \tag{3.1}$$

Suppose, $\{z \in R^n : Az \geq a\} \neq \emptyset$. Then

$$\min_{Az\geq a} \|z\|_\alpha = \max_{\lambda\in\sigma_\alpha(A)} \lambda^T a, \tag{3.2}$$

in particular, the min *and the* max *in* (3.2) *are attained.*

Proof. Compactness arguments show that the min on the left-hand side of (3.2) is attained, say in $\bar{z}$. Furthermore, if $z \in R^n$ and $\lambda \in R^m$ satisfy $Az \geq a, \lambda \geq 0$, and $\|\lambda^T A\|_\alpha^* \leq 1$, then $\lambda^T a \leq \lambda^T Az \leq \|\lambda^T A\|_\alpha^* \|z\|_\alpha \leq \|z\|_\alpha$, establishing the inequality

$$\min_{Az\geq a} \|z\|_\alpha \geq \sup_{\lambda\in\sigma_\alpha(A)} \lambda^T a.$$

We complete our proof by establishing the existence of $\lambda \in \sigma_\alpha(A)$ for which $\lambda^T a \geq \|\bar{z}\|_\alpha$.

If $\bar{z} = 0, \lambda = 0$ is in $\sigma_\alpha(A)$ and $\lambda^T a = 0 = \|\bar{z}\|_\alpha$. Next assume that $\bar{z} \neq 0$. In this case, $\|\bar{z}\|_\alpha \neq 0$ and the interior of $\{z \in R^n : \|z\|_\alpha \leq \|\bar{z}\|_\alpha\}$ is nonempty and equals $\{z \in R^n : \|z\|_\alpha < \|\bar{z}\|_\alpha)$. As the minimality of $\bar{z}$ assures that there is no $z \in R^n$ with $Az \geq a$ and $\|z\|_\alpha < \|\bar{z}\|_\alpha$, i.e., the intersection of $\{z \in R^n : Az \geq a\}$ and the interior of $\{z \in R^n : \|z\|_\alpha \leq \|\bar{z}\|_\alpha\}$ are empty. It now follows from the Eidelheit Separation Theorem (see Luenberger [12, Thm. 3, p. 133]) that for some nonzero vector $\mu \in R^n$

$$\sup\{\mu^T z : \|z\|_\alpha \leq \|\bar{z}\|_\alpha\} \leq \inf\{\mu^T z : Az \geq a\}.$$

By possibly normalizing μ, we may and do assume that $\|\mu\|_\alpha^* = 1$, implying that $\sup\{\mu^T z : \|z\|_\alpha \leq \|\bar{z}\|_\alpha\} = \|\bar{z}\|_\alpha$. Furthermore, as $\bar{z} \in \{z \in R^n : \|z\|_\alpha \leq \|\bar{z}\|_\alpha\} \cap \{z \in R^n : Az \geq a\}$, we conclude that the above sup and inf are attained and that

$$\|\bar{z}\|_\alpha = \max\{\mu^T z : \|z\|_\alpha \leq \|\bar{z}\|_\alpha\} = \mu^T \bar{z} = \min\{\mu^T z : Az \geq a\}.$$

Next, using the duality theorem of linear programming and the established fact that $\min\{\mu^T z : Az \geq a\}$ is attained, we have that

$$\min\{\mu^T z : Az \geq a\} = \max\{\lambda^T a : \lambda \geq 0, \lambda^T A = \mu^T\}.$$

Let λ be an optimal solution of the latter maximization problem. Then $\lambda \geq 0$, $\|A^T\lambda\|_\alpha^* = \|\mu\|_\alpha^* = 1$ and $\lambda^T a = \min\{\mu^T z : Az \geq a\} = \|\bar{z}\|_\alpha$. So, $\lambda \in \sigma_\alpha(A)$ and $\lambda^T a = \|\bar{z}\|_\alpha$. □

We are now ready to establish the main result of this section. Its proof requires the following additional notation. For a matrix $A \in R^{m\times n}$ and a subset J of $\{1, \ldots, m\}$, let A_J be the submatrix of A consisting of the rows indexed by J, and for a subset J of $\{1, \ldots, n\}$ let A^J be the submatrix of A consisting of the columns indexed by J. Furthermore, if J consists of a single element j, we use the notation A_j for $A_{\{j\}}$ and A^j for $A^{\{j\}}$.

THEOREM 3.2. *Let $A \in R^{m\times n}$ and $b \in R^m$, where $\{x \in R^n : Ax \leq b\} \neq \emptyset$, let $\|\ \|_\alpha$ and $\|\ \|_\beta$ be norms on R^n and R^m, respectively, let $\sigma_\alpha(A)$ be defined by* (3.1) *and let*

$$K_{\alpha\beta}(A) \equiv \sup\{\|\lambda\|_\beta^*: \lambda \text{ is an extreme point of } \sigma_\alpha(A)\}. \tag{3.3}$$

Then for each $x' \in R^n$,

$$\min_{Ax\leq b} \|x - x'\|_\alpha \leq K_{\alpha\beta}(A)\|(Ax' - b)^+\|_\beta; \tag{3.4}$$

in particular, the supremum on the right-hand side of (3.3) *is finite, and the minimum on the left-hand side of* (3.4) *is attained.*

Proof. Let $x' \in R^n$ be given and $a \equiv Ax' - b$. Compactness arguments show that the minimum of the left-hand side of (3.4) is attained. Furthermore,

$$\min_{Ax\leq b} \|x - x'\|_\alpha = \min_{A(x-x')\leq b-Ax'} \|x - x'\|_\alpha = \min_{Az\geq Ax'-b} \|z\|_\alpha = \min_{Az\geq a} \|z\|_\alpha; \tag{3.5}$$

in particular, $\{z \in R^n : Az \geq a\} \neq \emptyset$ and all minima in (3.5) are finite and are attained.

Next, by Lemma 3.1, $\min\{\|z\|_\alpha: Az \geq a\} = \max\{\lambda^T a: \lambda \in \sigma_\alpha(A)\}$, where the max on the right-hand side of the equality is attained. So, the linear functional mapping each λ in the closed convex set $\sigma_\alpha(A)$ into $\lambda^T a$ attains a maximum. Also, as $\sigma_\alpha(A)$ is a subset of $\{\lambda \in R^n : \lambda \geq 0\}$, it contains no lines. Thus, standard arguments imply that the above linear functional attains a maximum over $\alpha_\alpha(A)$ at an extreme point. For each such extreme point λ,

$$\lambda^T a \leq \lambda^T a^+ \leq \|\lambda\|_\beta^* \|a^+\|_\beta. \tag{3.6}$$

These observations, (3.5), (3.2), (3.6), and the definitions of $K_{\alpha\beta}(A)$ and a, imply that

$$\begin{aligned}
\min_{Ax\leq b} \|x - x'\|_\alpha &= \min_{Az\geq a} \|z\|_\alpha = \max_{\lambda\in\sigma_\alpha(A)} \lambda^T a \\
&= \max\{\lambda^T a: \lambda \text{ is an extreme point of } \sigma_\alpha(A)\} \\
&\leq \sup\{\|\lambda\|_\beta^*: \lambda \text{ is an extreme point of } \sigma_\alpha(A)\}\|a^+\|_\beta \\
&= K_{\alpha\beta}(A)\|(Ax' - b)^+\|_\beta.
\end{aligned}$$

It remains to show that the supremum on the right-hand side of (3.3) is finite.

Let $U(A)$ be the set of subsets of $\{1, \ldots, m\}$ for which the corresponding rows of A are linearly independent. For $J \in U(A)$ there exists a matrix $B(J) \in R^{n\times m}$ such that $A_J B(J)^J$ is the identity matrix in $R^{|J|\times|J|}$ and $B(J)^j = 0 (\in R^m)$ for each $j \in \{1, \ldots, n\} \setminus J$. As the unit ball $\{v \in R^n : \|v\|_\alpha^* \leq 1\}$ is bounded, for each $J \in U(A), K(J) \equiv \sup\{\|B(J)^T v\|_\beta^*: v \in R^n, \|v\|_\alpha^* \leq 1\}$ is finite. Let λ be an extreme point of $\sigma_\alpha(A)$ and let $v \equiv A^T\lambda$. Then $\lambda \geq 0, \|v\|_\alpha^* \leq 1$, and standard arguments show that $\{A_i: \lambda_i > 0\}$ are linearly independent, i.e., $J \equiv \{i = 1, \ldots, m : \lambda_i > 0\} \in U(A)$. As $v^T = \lambda^T A = (\lambda_J)^T A_J$, we have that, $v^T B(J)^J = (\lambda_J)^T A_J B(J)^J = (\lambda_J)^T$ and therefore $v^T B(J) = \lambda^T$. Thus,

$$\|\lambda\|_\beta^* \leq \sup_{\|v'\|_\alpha^*\leq 1} \|B(J)^T v'\|_\beta^* = K(J) \leq \max_{J'\in U'(A)} K(J') < \infty$$

and the finiteness of the right-hand side of (3.3) follows. □

Mangasarian and Shiau [18] establish (3.4), where $\| \ \|_\alpha$ is the l_1 norm and

$$K_{1\beta}(A) = \sup\{\|\lambda\|_\beta^*: \|\lambda^T A\|_1 = 1, \lambda \geq 0, \{i = 1, \ldots, m : \lambda_i > 0\} \in U(A)\}.$$

Our proof of Theorem 3.2 shows that our bound yields that of Mangasarian and Shiau.

The following corollary of Theorem 3.2 shows that inequality (2.11) of the generalization of Corollary 2.2 can be improved. Its proof follows directly from Theorem 3.2 by the arguments used to establish Corollary 2.2 from Theorem 2.1.

COROLLARY 3.3. *Let $A \in R^{m\times n}$, let $\| \ \|_\alpha$ and $\| \ \|_\beta$ be norms on R^n and on R^m, respectively, and let $K_{\alpha\beta}(A)$ be the constant defined through* (3.1) *and* (3.3). *Then for each $x' \in R^n$ and $b' \in R^m$ satisfying $Ax' \leq b'$ and for each $b \in R^m$ satisfying $\{x \in R^n : Ax \leq b\} \neq \emptyset$, there exists a vector $x \in R^n$ satisfying $Ax \leq b$ with*

$$\|x' - x\|_\alpha \leq K_{\alpha\beta}(A)\|b' - b\|_\beta. \tag{3.7}$$

4. Bounds and singular values. For a matrix $A \in R^{m\times n}$, let $K_2(A)$ be the coefficient $K_{\alpha\beta}(A)$ defined through (3.1) and (3.3) when the norms $\| \ \|_\alpha$ and $\| \ \|_\beta$ are the l_2 norms. In the current section we bound $K_2(A)$ by expressions that depend on the singular values of submatrices of A.

Let $A \in R^{m\times n}$. Recall that the *singular values* of A are the square roots of the nonzero eigenvalues of the matrix $A^T A$. The smallest singular value of the matrix A is denoted $\underline{\rho}(A)$. The following result is well known; see Golub and Van Loan [6].

PROPOSITION 4.1. *Let $A \in R^{m\times n}$. Then for every nonzero vector $b \in R^m$ for which $Ax = b$ is feasible,*

$$\min_{\{x\in R^n : Ax=b\}} \|x\|_2/\|b\|_2 \leq 1/\underline{\rho}(A). \tag{4.1}$$

THEOREM 4.2. *Let $A \in R^{m\times n}$, let $U(A)$ be the set of subsets of $\{1, \ldots, m\}$ for which the corresponding rows of A are linearly independent and let $U^*(A)$ be the maximal elements in $U(A)$. Then*

$$K_2(A) \leq \max_{J\in U^*(A)} 1/\underline{\rho}(A_J). \tag{4.2}$$

Proof. Let λ be an extreme point of $\sigma_2(A)$ and let $w \equiv A^T\lambda$. Then $\|w\|_2 \leq 1$ and, as was shown in the proof of Theorem 3.2, for some $J \in U(A)$, $\lambda_i = 0$ for every $i \in \{1, \ldots, m\} \setminus J$. Let J^* be a maximal set in $U(A)$ that contains J. In particular, $w^T = \lambda^T A = (\lambda_{J^*})^T A_{J^*}$ and the linear independence of the rows of A_{J^*} assures that λ_{J^*} is the unique solution of the system $(A_{J^*})^T x = w$. So, Proposition 4.1 implies that

$$\|\lambda\|_2 = \|\lambda_{J^*}\|_2 \leq \|\lambda_{J^*}\|_2/\|w\|_2 = \min_{\{x\in R^{|J^*|} : (A_{J^*})^T x=w\}} \|x\|_2/\|w\|_2 \leq 1/\underline{\rho}[(A_{J^*})^T].$$

As the singular values of a matrix are known to coincide with those of its transpose, we conclude that

$$\|\lambda\|_2 \leq 1/\underline{\rho}[(A_{J^*})^T] = 1/\underline{\rho}(A_{J^*}) \leq \max_{J'\in U^*(A)} 1/\underline{\rho}(A_{J'}),$$

and (4.2) follows. □

Theorem 4.2 and the equivalence of all norms over a finite dimensional normed space yield another representation for the bounding coefficient; see the arguments of Robinson [21] and the proof of Corollary 2.3.

Acknowledgments. The authors wish to acknowledge useful comments of Ivan Singer, Steve Robinson, Paul Tseng, and anonymous referees.

REFERENCES

[1] S. AGMON, *The relaxation method for linear inequalities*, Canad. J. Math., 6 (1954), pp. 382–392.

[2] A. A. AUSLENDER AND J.-P. CROUZEIX, *Global regularity theorems*, Math. Oper. Res., 13 (1988), pp. 243–253.

[3] C. BERGTHALLER AND I. SINGER, *The distance to a polyhedron*, Linear Algebra Appl., 169 (1992), pp. 111–129.

[4] W. COOK, A. M. H. GERARDS, A. SCHRIJVER, AND E. TARDOS, *Sensitivity theorems in integer linear programming*, Math. Programming, 34 (1986), pp. 251–264.

[5] J. L. GOFFIN, *The relaxation method for solving systems of linear inequalities*, Math. Oper. Res., 5 (1980), pp. 388–414.

[6] G. H. GOLUB AND C. VAN LOAN, *Matrix Computations*, The John Hopkins University Press, Baltimore, MD, 1983.

[7] O. GÜLER, *Augmented Lagrangian algorithms for linear programming*, J. Optim. Theory and Appl., 75 (1992), pp. 445–470.

[8] A. J. HOFFMAN, *On approximate solutions of systems of linear inequalities*, J. Res. Nat. Bur. Stand., 49 (1952), pp. 263–265.

[9] H. HU AND Q. WANG, *On approximate solutions of infinite systems of linear inequalities*, Linear Algebra and Appl., 114/115 (1989), pp. 429–438.

[10] A. N. IUSEM AND A. R. DE PIERRO, *On the convergence properties of Hildreth's quadratic programming algorithm*, Math. Programming, 47 (1990), pp. 37–51.

[11] W. LI, *The best error bounds for feasible and optimal solutions of a perturbed linear program*, Linear Algebra Appl., 187 (1993), pp. 15–40.

[12] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley and Sons, New York, 1969.

[13] Z.-Q. LUO AND P. TSENG, *On the linear convergence of descent methods for convex essentially smooth minimization*, SIAM J. Cont. Optim., 30 (1992), pp. 408–425.

[14] Z.-Q. LUO AND P. TSENG, *Perturbation analysis of a condition number for linear systems*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 636–660.

[15] O. L. MANGASARIAN, *A condition number for linear inequalities and linear programs*, in Methods of Operations Research 43, Proc. 6th Symposium uber Operations Research, Universitat Augsburg, Sept. 7–9 1981, G. Bamberg and O. Opitz, eds., Verlagsgruppe Athenaum/Hain/Scriptor/Hanstein, Konigstein, 1981, pp. 3–15.

[16] ———, *A condition number for differentiable convex inequalities*, Math. Oper. Res., 10 (1985), pp. 175–179.

[17] O. L. MANGASARIAN AND T.-H. SHIAU, *Error bounds for monotone linear complementarity problems*, Math. Programming, 36 (1986), pp. 81–89.

[18] ———, *Lipschitz continuity of solutions of linear inequalities programs and complementarity problems*, SIAM J. Control Optim., 25 (1987), pp. 583–595.

[19] T. S. MOTZKIN AND I. J. SCHOENBERG, *The relaxation method for linear equalities*, Canad. J. Math., 6 (1954), pp. 393–404.

[20] B. T. POLYAK AND N. V. TRETI'AKOV, *An iterative method for linear programming and its economic interpretation*, Matekon 8 (1972), pp. 81–100.

[21] S. M. ROBINSON, *Bounds for the error in the solution set of a perturbed linear program*, Linear Algebra Appl., 6 (1973), pp. 69–81.

[22] ———, *An application of error bounds for convex programming in linear space*, SIAM J. Control Optim., 13 (1975), pp. 271–273.

[23] Y. YE, O. GÜLER, R. A. TAPIA, AND Y. ZHANG, *A quadratically convergent $n^{1/2L}$ iteration algorithm for linear programming*, Math. Programming, to appear.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 697–718, July 1995

SECOND DERIVATIVES FOR OPTIMIZING EIGENVALUES OF SYMMETRIC MATRICES*

MICHAEL L. OVERTON† AND ROBERT S. WOMERSLEY‡

Abstract. Let A denote an $n \times n$ real symmetric matrix-valued function depending on a vector of real parameters, $x \in \Re^m$. Assume that A is a twice continuously differentiable function of x, with the second derivative satisfying a Lipschitz condition. Consider the following optimization problem: minimize the largest eigenvalue of $A(x)$. Let x^* denote a minimum. Typically, the maximum eigenvalue of $A(x^*)$ is multiple, so the objective function is not differentiable at x^*, and straightforward application of Newton's method is not possible. Nonetheless, the formulation of a method with local quadratic convergence is possible. The main idea is to minimize the maximum eigenvalue subject to a constraint that this eigenvalue has a certain multiplicity. The manifold Ω of matrices with such multiple eigenvalues is parameterized using a matrix exponential representation, leading to the definition of an appropriate Lagrangian function. Consideration of the Hessian of this Lagrangian function leads to the second derivative matrix used by Newton's method. The convergence proof is nonstandard because the parameterization of Ω is explicitly known only in the limit. In the special case of multiplicity one, the maximum eigenvalue is a smooth function and the method reduces to a standard Newton iteration.

Key words. nonsmooth optimization, multiple eigenvalues

AMS subject classifications. 15A18, 65F15, 65K10, 90C25

1. Introduction.

Let A denote an $n \times n$ real symmetric matrix-valued function depending on a vector of real parameters, $x \in \Re^m$. Assume that A depends smoothly on x, specifically that it is at least twice continuously differentiable, with the second derivative satisfying a Lipschitz condition in x. Denote the eigenvalues of $A(x)$ by

$$\lambda_1(x) \geq \cdots \geq \lambda_n(x).$$

The eigenvalues λ_i are Lipschitz continuous functions of x [7] and, in any region where they are distinct from one another, it is well known that they are (Fréchet) differentiable; in fact, they inherit the C^2 smoothness of the function $A(x)$ [7, p. 134]. Let $\widehat{x}$ be given, with

$$A(\widehat{x}) = \widehat{Q}\widehat{\Lambda}\widehat{Q}^T, \qquad \widehat{Q}^T\widehat{Q} = I, \tag{1.1}$$

where

$$\widehat{\Lambda} = \mathrm{Diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_n), \qquad \widehat{Q} = [\widehat{q}_1, \ldots, \widehat{q}_n]. \tag{1.2}$$

Thus, $\{\widehat{\lambda}_i\}$ and $\{\widehat{q}_i\}$ are, respectively, the eigenvalues and an orthonormal set of eigenvectors of $A(\widehat{x})$. Assume that $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_n$, so that $\widehat{\lambda}_i = \lambda_i(\widehat{x})$. Then formulas for the first and second partial derivatives of the eigenvalues λ_i at $x = \widehat{x}$, assuming that the $\widehat{\lambda}_i$ are distinct, are

$$\frac{\partial \lambda_i(\widehat{x})}{\partial x_k} = \widehat{q}_i^T \frac{\partial A(\widehat{x})}{\partial x_k} \widehat{q}_i \tag{1.3}$$

* Received by the editors March 23, 1993; accepted for publication (in revised form) by G. A. Watson April 27, 1994.

† Computer Science Department, Courant Institute of Mathematical Sciences, New York University, New York 10012 (overton@cs.nyu.edu). The work of this author was supported in part by National Science Foundation grant CCR-9101649.

‡ School of Mathematics, University of New South Wales, Australia (rsw@hydra.maths.unsw.edu.au).

and

$$\frac{\partial^2}{\partial x_k \partial x_j}\lambda_i(\widehat{x}) = \widehat{q}_i^T \frac{\partial^2 A(\widehat{x})}{\partial x_k \partial x_j}\widehat{q}_i + 2\sum_{s\neq i} \frac{\widehat{q}_i^T \frac{\partial A(\widehat{x})}{\partial x_k}\widehat{q}_s \widehat{q}_i^T \frac{\partial A(\widehat{x})}{\partial x_j}\widehat{q}_s}{\widehat{\lambda}_i - \widehat{\lambda}_s}. \tag{1.4}$$

The first of these formulas is well known, and the second may be found in a variety of sources; see [8], [9], as well as (in a somewhat less accessible form) [7, p. 95]. Both will follow as special cases of the results given in this paper.

However, if $A(x)$ has multiple eigenvalues at a point $x = \widehat{x}$, its eigenvalues, while still Lipschitz continuous, may not generally be written as differentiable functions of several variables at $x = \widehat{x}$. For example, consider

$$A(x) = \begin{bmatrix} 1 + x_1 & x_2 \\ x_2 & 1 - x_1 \end{bmatrix}.$$

The eigenvalues are

$$\lambda_{1,2} = 1 \pm \sqrt{x_1^2 + x_2^2}.$$

Thus λ_1, the largest eigenvalue of $A(x)$, is generally not a smooth function of x; furthermore, it cannot even be written as the maximum of n smooth functions of x, if x has two or more components. Also, the eigenvectors of $A(x)$ cannot generally be written as continuous functions of x; this is a consequence of the fact that eigenvectors corresponding to simple eigenvalues are unique (up to sign and normalization) while those corresponding to multiple eigenvalues are not.

Generally speaking, applications involving eigenvalues of matrices depending on free parameters fall into one of two categories. In the first, it is specified that some or all of the eigenvalues $\lambda_i(x)$ achieve some given values λ_i^*; this is known as an *inverse eigenvalue problem.* If these given values are distinct, the inverse eigenvalue problem may be formulated as a differentiable system of nonlinear equations, and, if the number of free parameters and the number of equations is the same, the application of Newton's method is straightforward, using (1.3). In [4] it was shown how, even in the multiple eigenvalue case, the inverse eigenvalue problem may be formulated as a differentiable system of nonlinear equations, so that Newton methods, with generic quadratic convergence, are applicable.

In the second class of applications, the eigenvalues are not required to have particular values, but rather it is desired to solve some *optimization problem involving the eigenvalues.* A particularly common case is the min-max problem

$$\min_{x \in \Re^m} \phi(x), \tag{1.5}$$

where $\phi(x) = \lambda_1(x)$, the largest eigenvalue of $A(x)$. Let x^* be a locally unique minimizer of ϕ. If x^* has the property that the eigenvalue $\lambda_1(x^*)$ is simple, i.e., has multiplicity one, then the function to be minimized, λ_1, is twice continuously differentiable in a neighborhood of x^*, and Newton's method for unconstrained minimization may be applied, using the Hessian matrix defined by (1.4). However, it is more often the case that $A(x^*)$ has multiple eigenvalues; this is a consequence of the optimization objective, which in driving all the eigenvalues down as much as possible usually forces the coalescence of some of them. In such a case λ_1 is generally not differentiable at $x = x^*$.

This paper is concerned with the formulation of a method to solve optimization problems involving eigenvalues in exactly this case, where multiple eigenvalues occur at the solution. We shall show that the correct problem formulation leads to a method with generic quadratic convergence. This method was first given by [10], inspired in part by [3], [4]. Quadratic convergence was demonstrated by numerical examples. The purpose of this paper is primarily to prove the quadratic convergence property for the method presented in [10], justifying the Hessian matrix formulas given there, which were originally derived only formally and stated without any derivation or proof. The ideas of this paper can be applied to other classes of eigenvalue and singular value optimization problems, e.g., those discussed in [1], [6], [11], [12], [14], [18], [19], as well as many other references which can be found in these papers. However, we concentrate on the model problem (1.5). We consider only the issue of local convergence. For details of how to use the method and related methods in practice, see [11].

2. Tensor notation. We shall have frequent need to refer to the first and second derivatives, with respect to several variables, of matrix-valued functions. Such objects are, respectively, tensors in three and four dimensions, a matrix being a tensor in two dimensions. We shall use subscripts to denote differentiation: thus A_x and A_{xx} refer to the first and second derivatives of the matrix-valued function A, with respect to the variable $x \in \Re^m$. Rather than attempt to describe the elements of a tensor, however, we shall describe its action as a linear operator, the result having the same dimension as the undifferentiated quantity, whether a matrix, a vector, or a scalar. For example, we write $[A_x \Delta x]$ to mean

$$\sum_{k=1}^{m} \{\Delta x\}_k \frac{\partial}{\partial x_k} A$$

and $[A_{xx} \Delta x \Delta x]$ to mean

$$\sum_{k=1}^{m} \sum_{j=1}^{m} \{\Delta x\}_k \{\Delta x\}_j \frac{\partial^2}{\partial x_k \partial x_j} A.$$

We reserve square brackets [,] for this purpose, and use parentheses (,) primarily to mean "evaluated at." We shall use braces {, } to indicate expression precedence. For example, the first and second derivatives of $\phi(x) \equiv \lambda_1(x)$ at $x = \hat{x}$, when $\lambda_1(\hat{x})$ is simple, given by (1.3)–(1.4), are written in tensor notation as

$$[\phi_x(\hat{x}) \Delta x] = \hat{q}_1^T [A_x(\hat{x}) \Delta x] \hat{q}_1$$

and

$$[\phi_{xx}(\hat{x}) \Delta x \Delta x] = \hat{q}_1^T [A_{xx} \Delta x \Delta x] \hat{q}_1 + 2 \sum_{s \neq 1} \frac{\{\hat{q}_1^T [A_x \Delta x] \hat{q}_s\}^2}{\hat{\lambda}_1 - \hat{\lambda}_s}.$$

Because the second derivative of a twice continuously differentiable function is symmetric with respect to its two arguments of differentiation, there is no ambiguity in this notation. There should be no confusion between those subscripts indicating differentiation and those indicating components.

We shall use $\| \cdot \|$ to denote the Euclidean vector norm. The expression $A \bullet B$, where A and B are symmetric matrices of the same dimension, means the matrix inner product

$$A \bullet B = \operatorname{tr} AB.$$

The operator "vec" maps the set of symmetric matrices of dimension t into the corresponding vector space $\Re^{t(t+1)/2}$, multiplying the off-diagonal components by the factor $\sqrt{2}$ so that

$$(\text{vec } A)^T(\text{vec } B) = A \bullet B.$$

Consequently,

$$\|\text{vec } A\| = \|A\|_F,$$

the Frobenius norm of A.

3. The matrix exponential formulation. Let x^* be a locally unique minimizer of $\phi \equiv \lambda_1$, and let $\lambda_i^* = \lambda_i(x^*)$, $i = 1, \ldots, n$. Suppose that

$$\lambda_1^* = \cdots = \lambda_t^* > \lambda_{t+1}^* > \cdots > \lambda_n^* \tag{3.1}$$

i.e., the maximum eigenvalue of $A(x^*)$ has multiplicity t, but all other eigenvalues are simple. The latter assumption usually holds in practice; it could be relaxed, at the cost of more complex notation. Let

$$\Lambda_1^* = \lambda_1^* I, \qquad \Lambda_2^* = \text{Diag}(\lambda_{t+1}^*, \ldots, \lambda_n^*), \tag{3.2}$$

the identity block having order t, and let $Q^* = [q_1^*, \ldots, q_n^*]$ be a corresponding orthogonal basis of eigenvectors, with

$$Q_1^* = [q_1^* \ldots q_t^*], \qquad Q_2^* = [q_{t+1}^* \ldots q_n^*]. \tag{3.3}$$

The matrix Q_2^* is unique, up to the choice of signs for its columns, but the matrix Q_1^* is not, since any particular choice of basis may be rotated by postmultiplying by a $t \times t$ orthogonal matrix.

It was shown in [11] that a necessary condition for x^* to minimize $\phi(x)$ is that there exist a t by t symmetric matrix V^*, with V^* positive semidefinite, such that

$$\text{tr } V^* = 1, \qquad V^* \bullet \{Q_1^*\}^T[A_x(x^*)\Delta x]Q_1^* = 0 \tag{3.4}$$

for all Δx. In the case $t = 1$, when Q_1^* consists of a single column q_1^*, this reduces to the statement that $\{q_1^*\}^T[A_x(x^*)\Delta x]q_1^* = 0$, equivalently $[\phi_x(x^*)\Delta x] = 0$ for all Δx, i.e., the gradient of $\phi(x^*)$ is zero. If $A(x)$ is an affine function, the necessary condition is also sufficient for optimality.

We wish to consider the correct local formulation of a Newton-based method so that quadratic convergence to x^* is obtained generically. We assume that the optimal multiplicity t is known. This is not the case in practice, and must be determined during the course of the computation, as explained in [10], [11]. If t is set incorrectly, the method to be described would converge locally to a minimizer of ϕ subject to the wrong multiplicity constraint, which might not be a minimizer of ϕ. This can be avoided by computing an approximation to V^* and verifying that the necessary conditions for optimality, including the positive semidefinite condition on V^*, are satisfied. See [11] for discussion of the case where all optimality conditions except the positive semidefinite condition are satisfied.

Assuming, then, that the optimal value of t is known, the local minimizer x^* of ϕ clearly also locally solves the constrained problem

$$\min_{x,\omega} \qquad \omega \tag{3.5}$$

$$\text{subject to } A(x) \in \Omega(t, \omega), \tag{3.6}$$

where $x \in \Re^m$, ω is a real parameter, and $\Omega(t,\omega)$ is the *set of matrices whose greatest eigenvalue has multiplicity t and value ω*. The set $\Omega(t,\omega)$ is an analytic manifold contained in the space of n by n symmetric matrices. The structure of this manifold is well known. It was observed as early as 1929 [17] that the number of conditions imposed on the space of symmetric matrices by the restriction that a matrix lie on this manifold is $\frac{t(t+1)}{2}$. In other words, the codimension of the manifold $\Omega(t,\omega)$ is $\frac{t(t+1)}{2}$. Formulas for the tangent space to the manifold $\Omega(t,\omega)$ at any point can be computed using standard techniques in differential geometry [13], [15]. Much less obvious, however, is how to parameterize a description of the manifold that is suitable for the application of Newton methods. This is really the main point of the paper.

The key idea, following [4], is to parameterize the orthogonal matrix of eigenvectors using a *matrix exponential*. Any orthogonal matrix P with $\det P = 1$ can be represented by

$$P = e^Y = I + Y + \tfrac{1}{2}Y^2 + \cdots,$$

where Y is skew-symmetric, i.e., $Y = -Y^T$. Since eigenvector signs are arbitrary, the assumption that $\det P = 1$ is not a restriction. A proof that this representation is always possible and locally unique is given in the Appendix.

Let $\widehat{x}$ be a given point, with the eigenvalues and eigenvectors of $A(\widehat{x})$ given by (1.1)–(1.2). Let

$$\widehat{\Lambda}_1 = \mathrm{Diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_t), \quad \widehat{\Lambda}_2 = \mathrm{Diag}(\widehat{\lambda}_{t+1}, \ldots, \widehat{\lambda}_n), \tag{3.7}$$

and let

$$\widehat{Q}_1 = [\widehat{q}_1 \ldots \widehat{q}_t], \quad \widehat{Q}_2 = [\widehat{q}_{t+1} \ldots \widehat{q}_n]. \tag{3.8}$$

Define the twice continuously differentiable $n \times n$ symmetric matrix-valued function

$$\widehat{F}(x, Y, \omega, \Theta) = \begin{bmatrix} \omega I & 0 \\ 0 & \Theta \end{bmatrix} - e^{-Y}\widehat{Q}^T A(x)\widehat{Q}e^Y, \tag{3.9}$$

where $x \in \Re^m$, ω is a real scalar, $\Theta = \mathrm{Diag}(\theta_1, \ldots, \theta_{n-t})$ is a real diagonal matrix of order $n-t$, and Y is a real $n \times n$ skew-symmetric matrix. From the context, it is clear that I is used to mean the identity matrix of order t. Subsequent block matrices will have dimensions conforming with those of $\widehat{F}$. We shall find it useful to write

$$Y = \begin{bmatrix} Y_{11} & Y_{12} \\ -Y_{12}{}^T & Y_{22} \end{bmatrix}, \tag{3.10}$$

where Y_{11} and Y_{22} are skew-symmetric but Y_{12} is not. Note that the definition of $\widehat{F}$ depends on $\widehat{x}$ through $\widehat{Q}$. Of course, $\widehat{Q}$ could be removed from $\widehat{F}$ by absorbing it into e^Y. The reason for the explicit inclusion of $\widehat{Q}$ in the definition of $\widehat{F}$ is so that the function e^Y can always be expanded about $Y = 0$.

Now consider the nonlinear program

$$\min_{x,Y,\omega,\Theta} \quad \omega \tag{3.11}$$

$$\text{subject to } \widehat{F}(x, Y, \omega, \Theta) = 0. \tag{3.12}$$

It is clear that if $\{x, Y, \omega, \Theta\}$ solves (3.12), with $\omega > \theta_i$, $i = 1, \ldots, n-t$, then $\{x, \omega\}$ satisfies the constraint (3.6), with $A(x)$ having eigenvalues $\omega, \ldots, \omega, \theta_1, \ldots \theta_{n-t}$, and

eigenvectors given by the columns of $\widehat{Q}e^Y$. Conversely, if x, ω satisfy (3.6), then, regardless of $\widehat{Q}$, (3.12) has a solution $\{x, Y, \omega, \Theta\}$, with $\theta_i = \lambda_{t+i}(x)$ and $e^Y = \widehat{Q}^T Q$, where Q is an orthogonal matrix of eigenvectors for $A(x)$.

The number of equations in (3.12) is $\frac{n(n+1)}{2}$. Formulation (3.11)–(3.12) introduces additional variables Y, Θ which are not present in (3.5)–(3.6), with corresponding space dimension $\frac{n(n-1)}{2} + n - t = \frac{n(n+1)}{2} - t$. The difference between the number of equations and number of extra variables is t, which is *not the codimension of* $\Omega(t, \omega)$. This shows that there is a difficulty with regularity in the parameterization of $\Omega(t, \omega)$ given by (3.12).

This difficulty is clarified by a key observation. Consider

$$F^*(x, Y, \omega, \Theta) = \begin{bmatrix} \omega I & 0 \\ 0 & \Theta \end{bmatrix} - e^{-Y}(Q^*)^T A(x) Q^* e^Y$$

and the associated nonlinear program

$$\min_{x, Y, \omega, \Theta} \quad \omega \tag{3.13}$$

$$\text{subject to } F^*(x, Y, \omega, \Theta) = 0. \tag{3.14}$$

The functions $\widehat{F}$ and F^* coincide if $\widehat{x} = x^*$ and the same basis $\widehat{Q} = Q^*$ is used in both definitions. We have $A(x^*) \in \Omega(t, \lambda_1^*)$ and

$$(Q^*)^T A(x^*) Q^* = \begin{bmatrix} \lambda_1^* I & 0 \\ 0 & \Lambda_2^* \end{bmatrix}$$

so Y satisfying (3.14) is not unique if $t > 1$. Specifically, any Y of the form

$$Y = \begin{bmatrix} Y_{11} & 0 \\ 0 & 0 \end{bmatrix}$$

solves (3.14) with $x = x^*$, $\omega = \lambda_1^*$, $\Theta = \Lambda_2^*$. Consequently, to obtain regularity in (3.13)–(3.14), the additional condition

$$Y_{11} = 0 \tag{3.15}$$

should be imposed in (3.10). The number of equations in (3.14) reduced by the dimension of the space of variables Y, Θ is then

$$\frac{n(n+1)}{2} - \left(\frac{n(n-1)}{2} - \frac{t(t-1)}{2} \right) - (n - t) = \frac{t(t+1)}{2},$$

which *is* the codimension of $\Omega(t, \omega)$. Ideally then, we would like to parameterize (3.5)–(3.6), not by (3.11)–(3.12), but by (3.13)–(3.14) together with (3.10), (3.15). However, this is not possible in practice, because Q^* is *known only in the limit.* The best we can do is to use (3.11)–(3.12), where $\widehat{Q}$ is the matrix of eigenvectors for $\widehat{x}$, the current best approximation to the solution x^*. Thus, we shall work with *a different function* $\widehat{F}$ *at each step of the iteration.*

But now a second key point must be emphasized. Although the Y_{11} variables are redundant in (3.14), they are *not* redundant in (3.12) if $\widehat{x} \neq x^*$, or more specifically if $A(\widehat{x}) \notin \Omega(t, \widehat{\lambda}_1)$. On the contrary, the *freedom in* Y_{11} *is necessary* to ensure that a feasible solution to (3.12) exists in general. Clearly, the closer $\widehat{x}$ is to x^*, i.e., the

closer $A(\hat{x})$ is to $\Omega(t, \hat{\lambda}_1)$, the closer the Y_{11} variables come to being redundant. This observation is quantified by the following theorem, which follows directly from [4], Corollary 3.1, and subsequent remarks. It will be convenient to denote the variables $\{x, Y, \omega, \Theta\}$ collectively by a single variable Z, which lies in a space of dimension $\frac{n(n+1)}{2} + m + 1 - t$.

THEOREM 1. *There exist $\epsilon > 0$, $C < \infty$ such that, if $\|\hat{x} - x^*\| \leq \epsilon$, then $\hat{F}(Z) = 0$ has a solution $\hat{Z}^* = \{x^*, \hat{Y}^*, \lambda_1^*, \Lambda_2^*\}$ with*

$$\|\hat{Y}^*\| \leq C\|\hat{x} - x^*\|$$

and with the leading t by t block of $\hat{Y}^*$ satisfying

$$\|\hat{Y}_{11}^*\| \leq C\|\hat{x} - x^*\|^2.$$

Here $\hat{Y}^*$ and $\hat{Z}^*$ are so denoted because, unlike x^*, they depend on the choice of function $\hat{F}$.

Roughly speaking, the Y variables describe the rotation of the eigenvectors $\hat{q}_i$ needed to transform them to eigenvectors of $A(x^*)$, while Y_{11} describes the rotation of the first t of these eigenvectors within the t-dimensional space they span. The rotation of the latter kind becomes relatively unimportant, as $\hat{x} \to x^*$, because of the nonuniqueness of the eigenvectors of $A(x^*)$.

Straightforward application of Newton's method to solve (3.11)–(3.12) is not satisfactory, since inclusion of the Y_{11} variables, which are redundant in the limit, prevents rapid convergence. On the other hand, setting $Y_{11} = 0$ in (3.11)–(3.12) makes (3.12) infeasible in general. We shall see that the solution to these difficulties is to remove Y_{11} from each linearization step, but include Y_{11} in the convergence analysis of this procedure. Thus, our convergence analysis is nonstandard.

Let us calculate the derivatives of $\hat{F}$. The appearance of the matrix exponential function in the definition makes this an easy task. We obtain

$$[\hat{F}_x \Delta x] = -e^{-Y}\hat{Q}^T[A_x(x)\Delta x]\hat{Q}e^Y; \tag{3.16}$$

$$[\hat{F}_Y \Delta Y] = -B - B^T, \quad \text{where} \tag{3.17}$$

$$B = \{-\Delta Y + \tfrac{1}{2}\{\Delta Y\}Y + \tfrac{1}{2}Y\{\Delta Y\} + O(Y^2)\}\hat{Q}^T A(x)\hat{Q}\{I + Y + O(Y^2)\};$$

$$[\hat{F}_\omega \Delta\omega] = \begin{bmatrix} \Delta\omega & I & 0 \\ 0 & & 0 \end{bmatrix}; \tag{3.18}$$

$$[\hat{F}_\Theta \Delta\Theta] = \begin{bmatrix} 0 & 0 \\ 0 & \Delta\Theta \end{bmatrix}. \tag{3.19}$$

Here $\Delta x, \Delta Y, \Delta\omega, \Delta\Theta$ are variables with the same dimensions as x, Y, ω, Θ, respectively; for example ΔY, like Y, is an n by n skew-symmetric matrix, with

$$\Delta Y = \begin{bmatrix} \{\Delta Y\}_{11} & \{\Delta Y\}_{12} \\ -\{\Delta Y\}_{12}^T & \{\Delta Y\}_{22} \end{bmatrix}, \tag{3.20}$$

where ΔY_{11} and ΔY_{22} are skew-symmetric (but ΔY_{12} is not). We shall use ΔZ to

denote $\{\Delta x, \Delta Y, \Delta\omega, \Delta\Theta\}$.

Now let us evaluate $\widehat{F}$ and its derivatives $\widehat{F}_x$, $\widehat{F}_Y$ at the point

$$\widehat{Z} = \{\widehat{x}, \widehat{Y}, \widehat{\lambda}_1, \widehat{\Lambda}_2\}, \tag{3.21}$$

where

$$\widehat{Y} = 0,$$

this equation being essential to keep the formulas simple. The derivatives $\widehat{F}_\omega$ and $\widehat{F}_\Theta$ are constant. We have

$$\widehat{F}(\widehat{Z}) = \begin{bmatrix} \widehat{\lambda}_1 I - \widehat{\Lambda}_1 & 0 \\ 0 & 0 \end{bmatrix}; \tag{3.22}$$

$$[\widehat{F}_x(\widehat{Z})\Delta x] = -\widehat{Q}^T[A_x(\widehat{x})\Delta x]\widehat{Q}; \tag{3.23}$$

$$[\widehat{F}_Y(\widehat{Z})\Delta Y] = -\widehat{\Lambda}\{\Delta Y\} + \{\Delta Y\}\widehat{\Lambda} = \tag{3.24}$$

$$\begin{bmatrix} -\widehat{\Lambda}_1\{\Delta Y\}_{11} + \{\Delta Y\}_{11}\widehat{\Lambda}_1 & -\widehat{\Lambda}_1\{\Delta Y\}_{12} + \{\Delta Y\}_{12}\widehat{\Lambda}_2 \\ \widehat{\Lambda}_2\{\Delta Y\}_{12}^T - \{\Delta Y\}_{12}^T\widehat{\Lambda}_1 & -\widehat{\Lambda}_2\{\Delta Y\}_{22} + \{\Delta Y\}_{22}\widehat{\Lambda}_2 \end{bmatrix}.$$

Notice that the leading t by t block of this matrix is zero if and only if $\widehat{\Lambda}_1$ is a multiple of the identity matrix, i.e., $A(\widehat{x}) \in \Omega(t, \widehat{\lambda}_1)$.

An immediate consequence of Theorem 1 which we shall need later is

$$\|\widehat{Z} - \widehat{Z}^*\| = O(\|\widehat{x} - x^*\|), \tag{3.25}$$

using (3.21) and the Lipschitz continuity of the eigenvalues.

The rest of the paper is organized as follows. In the next section, we analyze the special case $\frac{t(t+1)}{2} = m + 1$, when the dimension of the variable space matches the number of conditions imposed by the multiple eigenvalue, and hence quadratic convergence to a local solution of (1.5) can be achieved by a method that only uses first derivative information. In the subsequent section, we consider the general case, where second derivative information is necessary.

4. A special case. In this section we assume $\frac{t(t+1)}{2} = m+1$, where t, as before, is the multiplicity of λ_1^*. This is the case when the number of variables equals the number of conditions imposed by the multiple eigenvalue, and hence x^* is a locally unique solution of (3.14), given a nonsingularity condition to be defined shortly. Consider the following iteration.

ITERATION 1. *Given an initial value* $\widehat{x}$*:*

1. *Define* $\widehat{\Lambda}$, $\widehat{Q}$ *by* (1.1)–(1.2), *and* $\widehat{F}$ *by* (3.9). *Let* $\widehat{Z} = \{\widehat{x}, 0, \widehat{\lambda}_1, \widehat{\Lambda}_2\}$.
2. *Solve the* n *by* n *symmetric matrix equation*

$$[\widehat{F}_Z(\widehat{Z})\Delta Z] = -\widehat{F}(\widehat{Z}) \tag{4.1}$$

for ΔZ, *imposing also the condition*

$$\{\Delta Y\}_{11} = 0. \tag{4.2}$$

Set $\overline{Z} = \widehat{Z} + \Delta Z$.

3. *Replace* $\widehat{x}$ *by* $\overline{x}$, *the* x *component of* $\overline{Z}$. *Go to Step* 1.

Iteration 1 consists of a *Newton iteration applied to a varying function*, since the function that is differentiated, $\widehat{F}$, changes at each step. Such a situation is not unusual; see [5], [16]. The linear system (4.1) is equivalent to

$$[\widehat{F}_x(\widehat{Z})\Delta x] + [\widehat{F}_Y(\widehat{Z})\Delta Y] + [\widehat{F}_\omega(\widehat{Z})\Delta\omega] + [\widehat{F}_\Theta(\widehat{Z})\Delta\Theta] = -\widehat{F}(\widehat{Z}). \tag{4.3}$$

Because of the assumption that $\frac{t(t+1)}{2} = m+1$, together with the fact that Y_{11} is constrained to be zero, this is a system of $\frac{n(n+1)}{2}$ equations in the same number of variables. Examining (3.18)–(3.24), we see that it separates very conveniently. Imposing the condition $\{\Delta Y\}_{11} = 0$, the 1,1 block of (4.3) reduces to the t by t symmetric matrix equation

$$\Delta\omega\, I - \widehat{Q}_1^T[A_x(\widehat{x})\Delta x]\widehat{Q}_1 = \widehat{\Lambda}_1 - \widehat{\lambda}_1 I. \tag{4.4}$$

Let us denote this system of linear equations by

$$\widehat{K}\begin{bmatrix} \Delta\omega \\ \Delta x \end{bmatrix} = \widehat{b}, \tag{4.5}$$

where

$$\widehat{K} = \left[\text{vec } I,\ -\text{vec}\left(\widehat{Q}_1^T\frac{\partial A(\widehat{x})}{\partial x_1}\widehat{Q}_1\right), \ldots, -\text{vec}\left(\widehat{Q}_1^T\frac{\partial A(\widehat{x})}{\partial x_m}\widehat{Q}_1\right)\right] \tag{4.6}$$

and

$$\widehat{b} = \text{vec}\,(\widehat{\Lambda}_1 - \widehat{\lambda}_1 I). \tag{4.7}$$

Note that $\widehat{K}$ has dimension $t(t+1)/2$ by $m+1$, i.e., it is square under the assumptions of this section. (The operator "vec" was defined at the end of §2.)

The 1,2 block of (4.3) is the t by $n-t$ matrix equation

$$-\widehat{Q}_1^T[A_x(\widehat{x})\Delta x]\widehat{Q}_2 - \widehat{\Lambda}_1\{\Delta Y\}_{12} + \{\Delta Y\}_{12}\widehat{\Lambda}_2 = 0, \tag{4.8}$$

which can be solved for $\{\Delta Y\}_{12}$ in terms of Δx by

$$\Delta y_{ij} = \frac{\widehat{q}_i^T[A_x(\widehat{x})\Delta x]\widehat{q}_j}{\widehat{\lambda}_j - \widehat{\lambda}_i}, \tag{4.9}$$

for $1 \le i \le t$, $t < j \le n$; the denominator is bounded away from zero for $\widehat{x}$ in a small enough neighborhood of x^*. The 2,1 block of (4.3) contains the same information as the 1,2 block. The 2,2 block of (4.3) is

$$\Delta\Theta - \widehat{Q}_2^T[A_x(\widehat{x})\Delta x]\widehat{Q}_2 - \widehat{\Lambda}_2\{\Delta Y\}_{22} + \{\Delta Y\}_{22}\widehat{\Lambda}_2 = 0. \tag{4.10}$$

The off-diagonal equations of this symmetric system can be solved for $\{\Delta Y\}_{22}$ in a manner similar to equation (4.9), while the diagonal equations, which vanish in the last two terms, can be solved for $\Delta\Theta$.

In fact, though, we see that each step of Iteration 1 actually requires solving *only* one linear system for $\Delta\omega$ and Δx, namely (4.5), a system of $\frac{t(t+1)}{2}$ linear equations in $m+1$ variables and therefore square by assumption. The variables ΔY and $\Delta\Theta$ are *not required* to continue with the next iteration; their only purpose is their use in the problem formulation and convergence analysis. Iteration 1 is therefore equivalent to:

ITERATION 2. *Given an initial value* $\widehat{x}$*:*

1. *Define* $\widehat{\Lambda}$, $\widehat{Q}$ *by* (1.1)–(1.2).

2. *Solve the linear system* $\widehat{K}\left[\begin{smallmatrix}\delta\omega\\ \delta x\end{smallmatrix}\right] = \widehat{b}$, *defined in* (4.6)–(4.7), *for* $\Delta\omega$, Δx. *Set* $\overline{x} = \widehat{x} + \Delta x$.

3. *Replace* $\widehat{x}$ *by* $\overline{x}$, *and go to Step* 1.

Let us analyze the rate of convergence of Iteration 1, equivalently Iteration 2. We first need the following theorem.

THEOREM 2. *Define*

$$K^* = \left[\text{vec } I, \quad -\text{vec}\left(Q_1^{*T}\frac{\partial A(x^*)}{\partial x_1}Q_1^*\right), \ldots, -\text{vec}\left(Q_1^{*T}\frac{\partial A(x^*)}{\partial x_m}Q_1^*\right)\right]. \tag{4.11}$$

Then the smallest singular value of K^* *is independent of the choice of basis* Q_1^*.

Proof. The freedom in Q_1^* is that it may be postmultiplied by any t by t orthogonal matrix. The smallest singular value of K^* is, by definition,

$$\min_{\Delta\omega^2+\|\Delta x\|^2=1} \|K^*\begin{bmatrix}\Delta\omega\\ \Delta x\end{bmatrix}\|. \tag{4.12}$$

The vector norm being minimized is in fact

$$\|\Delta\omega\ I - \{Q_1^*\}^T[A_x(x^*)\Delta x]Q_1^*\|_F$$

(see the discussion at the end of §2). This quantity is not changed if Q_1^* is postmultiplied by an orthogonal matrix. □

Using this result, we can speak unambiguously about whether or not K^* is singular. The convergence result may now be stated.

THEOREM 3. *Suppose* K^* *is nonsingular. Then there exist constants* ϵ *and* C *such that, if* $\|\widehat{x} - x^*\| \le \epsilon$, *then*

$$\|\overline{x} - x^*\| \le C\|\widehat{x} - x^*\|^2.$$

Consequently, Iteration 1, *equivalently Iteration* 2, *generates points* $\widehat{x}$ *that converge quadratically to the solution* x^*.

Proof. That Iterations 1 and 2 generate the same point $\widehat{x}$ follows from the equivalence of (4.1)–(4.2) with (4.5), (4.8), (4.10). Expanding $\widehat{F}$ in a Taylor series about $\widehat{Z}$, using the point $\widehat{Z}^*$ whose existence is guaranteed by Theorem 1, gives

$$0 = \widehat{F}(\widehat{Z}^*) = \widehat{F}(\widehat{Z}) + [\widehat{F}_Z(\widehat{Z})\{\widehat{Z}^* - \widehat{Z}\}] + O(\|\widehat{Z} - \widehat{Z}^*\|^2). \tag{4.13}$$

By definition of Iteration 1, we also have

$$0 = \widehat{F}(\widehat{Z}) + [\widehat{F}_Z(\widehat{Z})\{\overline{Z} - \widehat{Z}\}], \tag{4.14}$$

noting that the Y_{11} component of $\overline{Z}$ is zero. The difference of these two equations gives

$$[\widehat{F}_Z(\widehat{Z})\{\overline{Z} - \widehat{Z}^*\}] = O(\|\widehat{Z} - \widehat{Z}^*\|^2). \tag{4.15}$$

Some comments here will be helpful. As usual, the proof of convergence of Newton's method involves three points: the current iterate, the new iterate, and the solution

point. Here, these are, respectively, $\widehat{Z}$, $\overline{Z}$, and $\widehat{Z}^*$, the subtlety being that $\widehat{Z}^*$ is the solution to $\widehat{F}(Z) = 0$, an equation whose definition depends on $\widehat{Z}$. Equation (4.15) states that

$$\begin{aligned}(4.16)\qquad &[\widehat{F}_x\{\overline{x} - x^*\}] + [\widehat{F}_{\{Y_{11}\}}\{-\widehat{Y}_{11}^*\}] + [\widehat{F}_{\{Y_{12}\}}\{\{\Delta Y\}_{12} - \widehat{Y}_{12}^*\}] \\ &\quad + [\widehat{F}_{\{Y_{22}\}}\{\{\Delta Y\}_{22} - \widehat{Y}_{22}^*\}] + [\widehat{F}_\omega\{\widehat{\lambda}_1 + \Delta\omega - \lambda_1^*\}] \\ &\quad + [\widehat{F}_\Theta\{\widehat{\Lambda}_2 + \Delta\Theta - \Lambda_2^*\}] = O(\|\widehat{x} - x^*\|^2),\end{aligned}$$

all of the derivatives being evaluated at $\widehat{Z}$, the appearance of $O(\|\widehat{x} - x^*\|^2)$ instead of $O(\|\widehat{Z} - \widehat{Z}^*\|^2)$ on the right-hand side being justified by (3.25). By Theorem 1, the $\widehat{F}_{\{Y_{11}\}}$ term on the left-hand side can be absorbed into the right-hand side, reducing (4.17) to a linear system of $\frac{n(n+1)}{2}$ equations in $\frac{n(n+1)}{2}$ variables. By precisely the argument which showed the equivalence of (4.1)–(4.2) with (4.5), (4.8), (4.10), this system can be reduced to $\frac{t(t+1)}{2}$ equations in $\frac{t(t+1)}{2}$ unknowns, namely

$$(4.17)\qquad \widehat{K}\begin{bmatrix} \widehat{\lambda}_1 + \Delta\omega - \lambda_1^* \\ \overline{x} - x^* \end{bmatrix} = O(\|\widehat{x} - x^*\|^2).$$

The proof is then complete if we can assert that the norm of the inverse of $\widehat{K}$ is bounded for $\widehat{x}$ in a neighborhood of x^*. Theorem 1 shows that there is an orthonormal basis of eigenvectors for $A(x^*)$, namely $Q^* = \widehat{Q}e^{\widehat{Y}^*}$, for which

$$(4.18)\qquad \|\widehat{Q} - Q^*\| = \|\widehat{Q}^T(\widehat{Q} - Q^*)\| = \|I - e^{\widehat{Y}^*}\| = O(\|\widehat{Y}^*\|) = O(\|\widehat{x} - x^*\|).$$

Using this choice of Q^* in (4.11), we have

$$(4.19)\qquad \|\widehat{K} - K^*\| = O(\|\widehat{x} - x^*\|).$$

Since K^* is nonsingular by assumption, and this nonsingularity is independent of the basis choice, the boundedness of the inverse of $\widehat{K}$ follows from the standard Banach lemma. □

Note that the use of the notation $O(\|\cdot\|^2)$ to denote neglected terms in the Taylor expansion is valid even though a family of functions $\widehat{F}$ is being considered, for a sequence of values $\widehat{Q}$ defining $\widehat{F}$. This is because the definition of $\widehat{F}$ in (3.9) shows that second and higher derivatives cannot blow up regardless of $\widehat{Q}$, given the corresponding smoothness assumptions on the matrix function $A(x)$, together with the orthogonality of $\widehat{Q}$.

5. The general case. In this section we assume that $\frac{t(t+1)}{2} \leq m + 1$. Since the codimension of $\Omega(t,\omega)$ is $\frac{t(t+1)}{2}$, and the dimension of the x,ω variable space is $m+1$, the opposite inequality can hold only nongenerically. Equality can be expected to hold only occasionally since relatively few of the integers have the form $\frac{t(t+1)}{2}$. In the general case, the constraints (3.12) are not enough to define x^* locally, so minimization of (3.11) must also be considered.

Define the Lagrangian function for (3.11)–(3.12) by

$$(5.1)\qquad \widehat{L}(Z,U) = \omega - U \bullet \widehat{F}(Z),$$

where U is an $n \times n$ symmetric matrix of Lagrange multipliers corresponding to the $n \times n$ symmetric matrix constraint (3.12). The matrix U is called the *dual matrix* since

its components are dual variables. The Frobenius inner product $A \bullet B$ was defined at the end of §2. Assuming a full rank condition to be discussed in detail later, the first-order necessary conditions for Z to minimize (3.11) subject to (3.12) are that, in addition to the satisfaction of (3.12) by Z, there exists U satisfying

$$\widehat{L}_Z(Z,U) = 0, \tag{5.2}$$

that is,

$$U \bullet \widehat{F}_x(Z) = 0, \tag{5.3}$$

$$U \bullet \widehat{F}_Y(Z) = 0, \tag{5.4}$$

$$U \bullet \widehat{F}_\omega = 1, \tag{5.5}$$

and

$$U \bullet \widehat{F}_\Theta = 0. \tag{5.6}$$

Here (5.3), for example, is understood to mean $U \bullet [\widehat{F}_x(Z)\Delta x] = 0$ for all Δx, i.e., $U \bullet \frac{\partial \widehat{F}(Z)}{\partial x_k} = 0$, $1 \le k \le m$. A pair Z, U, which satisfies conditions (5.3)–(5.6), is denoted $\widehat{Z}^*, \widehat{U}^*$.

In the following Newton iteration we shall, as in the previous section, impose the additional condition that $\{\Delta Y\}_{11} = 0$, and we shall therefore also *relax* the corresponding dual condition $U \bullet \widehat{F}_{\{Y_{11}\}}(Z) = 0$, replacing (5.4) by

$$U \bullet \widehat{F}_{\{Y_{12}\}}(Z) = 0, \qquad U \bullet \widehat{F}_{\{Y_{22}\}}(Z) = 0. \tag{5.7}$$

Each step of the iteration requires a dual matrix *estimate* $\widehat{U}$, which is necessary to define the Lagrangian function. It is important to note that a dual matrix estimate from the *previous* step of the iteration *cannot* be used, since the function $\widehat{F}$ changes from one iteration to the next, with the basis $\widehat{Q}$, which defines $\widehat{F}$, not converging in general.

ITERATION 3. *Given an initial value* $\widehat{x}$*:*

1. *Define* $\widehat{\Lambda}$, $\widehat{Q}$ *by* (1.1)–(1.2), *and* $\widehat{F}$ *by* (3.9). *Let* $\widehat{Z} = \{\widehat{x}, 0, \widehat{\lambda}_1, \widehat{\Lambda}_2\}$.

2. *Define* $\widehat{U}$ *to be any* $n \times n$ *symmetric matrix such that the norm of the residual of* (5.3), (5.7), (5.5), (5.6), *with* $Z = \widehat{Z}$, $U = \widehat{U}$, *is* $O(\|\widehat{Z} - \widehat{Z}^*\|)$.

3. *Solve the quadratic program*

$$\min_{\Delta Z} [\widehat{L}_Z(\widehat{Z},\widehat{U})\Delta Z] + \tfrac{1}{2}[\widehat{L}_{ZZ}(\widehat{Z},\widehat{U})\Delta Z \Delta Z] \tag{5.8}$$

$$\text{subject to} \qquad [\widehat{F}_Z(\widehat{Z})\Delta Z] = -\widehat{F}(\widehat{Z}) \tag{5.9}$$

with the restriction also that

$$\{\Delta Y\}_{11} = 0. \tag{5.10}$$

Set $\overline{Z} = \widehat{Z} + \Delta Z$.

4. *Replace* $\widehat{x}$ *by* $\overline{x}$, *the* x *component of* $\overline{Z}$. *Go to Step* 1.

Like Iteration 1, Iteration 3 can be substantially simplified using the structure of the problem. We begin with a closer look at the dual matrix. Suppose we choose

$$\widehat{U} = \begin{bmatrix} \widehat{U}_{11} & 0 \\ 0 & 0 \end{bmatrix} \tag{5.11}$$

and consider (5.3)–(5.6) with $Z = \widehat{Z}$, $U = \widehat{U}$. We see then that, for $U = \widehat{U}$, (3.19) implies (5.6) and (3.24) implies (5.4). In order to satisfy the condition in Step 2, then, we see from (3.18) and (3.23) that we need only ensure that

$$\operatorname{tr} \widehat{U}_{11} = 1 + O(\|\widehat{x} - x^*\|) \tag{5.12}$$

and

$$\widehat{U}_{11} \bullet \widehat{Q}_1^T \frac{\partial A(\widehat{x})}{\partial x_k} \widehat{Q}_1 = O(\|\widehat{x} - x^*\|), \quad 1 \le k \le m. \tag{5.13}$$

This is a system of $m+1$ equations in $\frac{t(t+1)}{2}$ unknowns, which can also be written

$$\widehat{K}^T\{\operatorname{vec} \widehat{U}_{11}\} = e_1 + O(\|\widehat{x} - x^*\|). \tag{5.14}$$

As we shall see in Theorem 6, this can be achieved by solving the least squares problem

$$\min_{\widehat{U}_{11}} \|\widehat{K}^T\{\operatorname{vec} \widehat{U}_{11}\} - e_1\|. \tag{5.15}$$

The constraints (5.9)–(5.10) are identical to the condition in Step 2 of Iteration 1, the only difference being that the system of linear equations is underdetermined rather than square. The same argument given following Iteration 1 therefore shows that (5.9)–(5.10) is equivalent to the constraint (4.5) on Δx, $\Delta\omega$ together with (4.8),(4.10) defining $\{\Delta Y\}_{12}$, $\{\Delta Y\}_{22}$.

It is instructive to consider the special case $t = 1$ at this point: in this case the max eigenvalue function $\phi(x)$ is differentiable at x^*. Then $\widehat{Q}_1$ consists of a single column $\widehat{q}_1$, $\widehat{U}_{11}$ is a scalar that can be taken to be the number 1, (5.13) states that the gradient of ϕ at $x = \widehat{x}$ is $O(\|\widehat{x} - x^*\|)$, and the constraint (4.5) states that

$$\Delta\omega = [\phi_x \Delta x]. \tag{5.16}$$

Now let us consider the quadratic objective function (5.8). The linear term may be replaced by $\Delta\omega$, since the rest of this term is fixed by the constraint (5.9). To evaluate the quadratic term in (5.8), we need to calculate the second derivatives of $\widehat{F}$. Clearly, all terms involving ω or Θ are zero. Differentiating (3.16)–(3.17) we obtain

$$\begin{aligned}
[\widehat{F}_{xx}(\widehat{Z})\Delta x \Delta x] &= -\widehat{Q}^T[A_{xx}(\widehat{x})\Delta x \Delta x]\widehat{Q}; \\
[\widehat{F}_{xY}(\widehat{Z})\Delta x \Delta Y] &= [\widehat{F}_{Yx}(\widehat{Z})\Delta Y \Delta x] \\
&= \{\Delta Y\}\widehat{Q}^T[A_x(\widehat{x})\Delta x]\widehat{Q} - \widehat{Q}^T[A_x(\widehat{x})\Delta x]\widehat{Q}\{\Delta Y\}; \\
[\widehat{F}_{YY}(\widehat{Z})\Delta Y \Delta Y] &= \Delta Y\{\widehat{\Lambda}\{\Delta Y\} - \{\Delta Y\}\widehat{\Lambda}\} - \{\widehat{\Lambda}\{\Delta Y\} - \{\Delta Y\}\widehat{\Lambda}\}\Delta Y.
\end{aligned}$$

Since $\widehat{U}$ satisfies (5.11), we need only the 1,1 block of each of these terms. Using (5.10) and (3.20), we obtain

$$\begin{aligned}
[\widehat{F}_{xx}(\widehat{Z})\Delta x \Delta x]_{11} &= -\widehat{Q}_1^T[A_{xx}(\widehat{x})\Delta x \Delta x]\widehat{Q}_1; \\
[\widehat{F}_{xY}(\widehat{Z})\Delta x \Delta Y]_{11} &= \{\Delta Y\}_{12}\widehat{Q}_2^T[A_x(\widehat{x})\Delta x]\widehat{Q}_1 + \widehat{Q}_1^T[A_x(\widehat{x})\Delta x]\widehat{Q}_2\{\Delta Y\}_{12}^T; \\
[\widehat{F}_{YY}(\widehat{Z})\Delta Y \Delta Y]_{11} &= \{\Delta Y\}_{12}\{-\widehat{\Lambda}_2\{\Delta Y\}_{12}^T + \{\Delta Y\}_{12}^T\widehat{\Lambda}_1\} \\
&\quad + \{\widehat{\Lambda}_1\{\Delta Y\}_{12} - \{\Delta Y\}_{12}\widehat{\Lambda}_2\}\{\Delta Y\}_{12}^T.
\end{aligned}$$

But since ΔY must satisfy the constraint (5.9), whose 1,2 block is (4.8), we see that

$$[\widehat{F}_{YY}(\widehat{Z})\Delta Y\Delta Y]_{11} = -[\widehat{F}_{xY}(\widehat{Z})\Delta x\Delta Y]_{11}. \tag{5.17}$$

We therefore have

$$\begin{aligned}[\widehat{F}_{ZZ}(\widehat{Z})\Delta Z\Delta Z]_{11} &= [\widehat{F}_{xx}(\widehat{Z})\Delta x\Delta x]_{11} + [\widehat{F}_{xY}(\widehat{Z})\Delta x\Delta Y]_{11} \\ &\quad +[\widehat{F}_{Yx}(\widehat{Z})\Delta Y\Delta x]_{11} + [\widehat{F}_{YY}(\widehat{Z})\Delta x\Delta Y]_{11} \\ &= -\widehat{Q}_1^T[A_{xx}(\widehat{x})\Delta x\Delta x]\widehat{Q}_1 + \{\Delta Y\}_{12}\widehat{Q}_2^T[A_x(\widehat{x})\Delta x]\widehat{Q}_1 \\ &\quad +\widehat{Q}_1^T[A_x(\widehat{x})\Delta x]\widehat{Q}_2\{\Delta Y\}_{12}^T.\end{aligned}$$

Let us denote the right-hand side of this equation by $-\widehat{M}$; then we see that, under the constraints (5.9)–(5.10),

$$[\widehat{L}_{ZZ}(\widehat{Z},\widehat{U})\Delta Z\Delta Z] = \widehat{U}_{11} \bullet \widehat{M}.$$

Using (4.8) we see that the elements of the $t \times t$ matrix $\widehat{M}$ are given by

$$\widehat{M}_{ij} = \widehat{q}_i^T[A_{xx}(\widehat{x})\Delta x\Delta x]\widehat{q}_j + \sum_{s=t+1}^{n} \gamma_{ijs}\widehat{q}_i^T[A_x\Delta x]\widehat{q}_s\ \widehat{q}_j^T[A_x\Delta x]\widehat{q}_s, \tag{5.18}$$

where $1 \le i \le t$, $1 \le j \le t$ and

$$\gamma_{ijs} = \frac{1}{\widehat{\lambda}_i - \widehat{\lambda}_s} + \frac{1}{\widehat{\lambda}_j - \widehat{\lambda}_s} = \frac{2}{\widehat{\lambda}_1 - \widehat{\lambda}_s} + O(\|\widehat{x} - x^*\|). \tag{5.19}$$

Writing out the double sums in the square brackets explicitly we see that, under the constraints (5.9)–(5.10),

$$[\widehat{L}_{ZZ}(\widehat{Z},\widehat{U})\Delta Z\Delta Z] = \widehat{U}_{11} \bullet \widehat{M} = \{\Delta x\}^T\widehat{W}\{\Delta x\}, \tag{5.20}$$

where $\widehat{W}$ is an m by m symmetric matrix whose k,l element satisfies

$$\widehat{W}_{kl} = \widehat{U}_{11} \bullet \widehat{G}^{kl} \tag{5.21}$$

with $\widehat{G}^{kl}$ defined to be the t by t symmetric matrix with elements

$$\{\widehat{G}^{kl}\}_{ij} = \widehat{q}_i^T\frac{\partial^2 A(\widehat{x})}{\partial x_k \partial x_l}\widehat{q}_j + \sum_{s=t+1}^{n} \gamma_{ijs}\widehat{q}_i^T\frac{\partial A(\widehat{x})}{\partial x_k}\widehat{q}_s\ \widehat{q}_j^T\frac{\partial A(\widehat{x})}{\partial x_l}\widehat{q}_s. \tag{5.22}$$

Again, the case $t = 1$ is instructive: then, since $\widehat{U}_{11} = 1$, $\widehat{G}^{kl}$ is the scalar quantity (1.4) (with $i = 1$), i.e., the second partial derivative of ϕ at $x = \widehat{x}$, and $\widehat{W}$ is the Hessian matrix of ϕ at $x = \widehat{x}$.

Therefore, Iteration 3, with $\widehat{U}$ satisfying (5.11), reduces to the following iteration.

ITERATION 4. *Given an initial value* $\widehat{x}$*:*

1. *Define* $\widehat{\Lambda}$, $\widehat{Q}$ *by* (1.1)–(1.2).
2. *Define* $\widehat{U}_{11}$ *by any* t *by* t *symmetric matrix such that* (5.14) *holds.*
3. *Define* $\widehat{W}$ *by* (5.19)–(5.22). *Solve the following quadratic program:*

$$\min_{\Delta\omega,\Delta x} \Delta\omega + \tfrac{1}{2}\{\Delta x\}^T\widehat{W}\{\Delta x\} \tag{5.23}$$

$$\textit{subject to} \quad \widehat{K}\begin{bmatrix} \Delta\omega \\ \Delta x \end{bmatrix} = \widehat{b} \tag{5.24}$$

where the latter constraint is defined by (4.6)–(4.7). *Set* $\overline{x} = \widehat{x} + \Delta x$.

4. *Replace* $\widehat{x}$ *by* $\overline{x}$ *and go to Step* 1.

In the case $t = 1$, we see from (5.16) that (5.23)–(5.24) reduces to the ordinary Newton iteration

$$\min_{\Delta x}[\phi_x(\widehat{x})\Delta x] + \tfrac{1}{2}[\phi_{xx}(\widehat{x})\Delta x \Delta x].$$

Iteration 4 is the method given by [10], with two exceptions: (i) [10] addresses a slightly different problem, namely, minimizing $\max(\lambda_1(x), -\lambda_n(x))$, with A assumed to be an affine matrix function; (ii) the method of [10] substitutes the quantities $2/\{\widehat{\lambda}_1 - \widehat{\lambda}_s\}$ for γ_{ijs}, dropping the last term on the right-hand side of (5.19). With this simplification, the corresponding formulas for (5.18), (5.22) can be written conveniently using matrix notation as

$$\widetilde{M} = \widehat{Q}_1^T[A_{xx}(\widehat{x})\Delta x \Delta x]\widehat{Q}_1 + 2\widehat{Q}_1^T[A_x(\widehat{x})\Delta x]\widehat{Q}_2 D^{-1}\widehat{Q}_2^T[A_x(\widehat{x})\Delta x]\widehat{Q}_1 \tag{5.25}$$

with $D = \widehat{\lambda}_1 I - \widehat{\Lambda}_2$,

$$\widetilde{G}^{kl} = \widehat{Q}_1^T \frac{\partial^2 A(\widehat{x})}{\partial x_k \partial x_l}\widehat{Q}_1 + 2\widehat{Q}_1^T \frac{\partial A(\widehat{x})}{\partial x_k}\widehat{Q}_2 D^{-1}\widehat{Q}_2^T \frac{\partial A(\widehat{x})}{\partial x_l}\widehat{Q}_1 \tag{5.26}$$

and

$$\widetilde{W}_{kl} = \widehat{U}_{11} \bullet \widetilde{G}^{kl}. \tag{5.27}$$

The use of $\widetilde{W}$ instead of $\widehat{W}$ does not affect the convergence rate of Iteration 4, but the advantage of the latter formula is that it leads to the following observation, due to M.K.H. Fan [2].

THEOREM 4. *Suppose A is an affine function, i.e., $A_{xx} = 0$. Then if $\widehat{U}_{11}$ is positive semidefinite, $\widetilde{W}$ is also positive semidefinite, regardless of the magnitude of $\widehat{x} - x^*$.*

Proof. Since $A_{xx} = 0$, it is clear that, for any choice of Δx, $\widetilde{M}$ is positive semidefinite. Since $\widehat{U}_{11}$ is positive semidefinite, the inner product $\widehat{U}_{11} \bullet \widetilde{M}$ is nonnegative for all Δx, which is equivalent to the condition $\{\Delta x\}^T \widetilde{W}\{\Delta x\} \geq 0$ for all Δx. □

Clearly, the same result holds if $[A_{xx}(\widehat{x})\Delta x \Delta x]$ is positive semidefinite for all Δx. Furthermore, if $\widehat{x}$ is close enough to x^*, and $\widetilde{W}$ is positive definite, then $\widehat{W}$ is positive definite. However, even if A is affine, $\widehat{W}$ is not positive semidefinite in general. For example, suppose $n = 3$, $t = 2$, and $\widehat{Q} = I$. The condition that $\widehat{M}$ is positive semidefinite then reduces to the condition $\gamma_{113}\gamma_{223} \geq \gamma_{123}^2$, regardless of A_x. Choosing $\widehat{\Lambda} = \mathrm{Diag}(2, 1, 0)$ gives

$$\gamma_{113} = 1, \quad \gamma_{123} = \gamma_{213} = 1.5, \quad \gamma_{223} = 2,$$

so that $\widehat{M}$ is indefinite. Then $\widehat{U}_{11}$ can be chosen positive semidefinite such that (5.20) is negative. However, substituting $2/(\widehat{\lambda}_1 - \widehat{\lambda}_3)$ for γ_{ijs} results in the matrices $\widetilde{M}$ and $\widetilde{W}$, which are positive semidefinite.

The positive semidefinite condition on $\widehat{U}_{11}$ is a natural one, because, as indicated by the next two theorems, $\widehat{U}_{11}$ is an approximation to the matrix V^* given in (3.4). Specifically, note that (5.30) defining U_{11}^* in the following theorem is identical to (3.4) defining V^*. There is no condition on the definiteness of U_{11}^*, because in the formulation of the nonlinear program (3.13)–(3.14) we *assumed* that the optimal multiplicity t is known; consequently, indefiniteness of U_{11}^* indicates that t *was chosen incorrectly* and hence that x^* does *not* minimize ϕ.

THEOREM 5.

1. *Consider the r by $(m+1)$ matrix K^*, defined by (4.11), where $r = t(t+1)/2$. Then the rth singular value of K^* does not depend on the choice of basis Q_1^*.*

2. *Suppose that the rth singular value of K^* is nonzero, i.e., K^* has linearly independent rows. Consider the nonlinear program (3.13)–(3.15), noting that the latter constraint removes Y_{11} from the variable set. Let*

$$L^*(Z,U) = \omega - U \bullet F^*(Z).$$

A necessary condition for $Z^* = (x^*, 0, \lambda_1^*, \Lambda_2^*)$ to solve (3.13)–(3.15) is that there exists an $n \times n$ symmetric matrix U^*, satisfying

$$L_Z^*(Z^*, U^*) = 0. \tag{5.28}$$

Furthermore, U^* is unique, with

$$U^* = \begin{bmatrix} U_{11}^* & 0 \\ 0 & 0 \end{bmatrix}, \tag{5.29}$$

where the t by t block U_{11}^* satisfies

$$\{K^*\}^T\{\text{vec } U_{11}^*\} = e_1. \tag{5.30}$$

3. *Define W^* to be the m by m symmetric matrix with elements,*

$$W_{kl}^* = U_{11}^* \bullet G^{*kl}$$

where G^{*kl} is the t by t symmetric matrix with elements

$$G^{*kl} = Q_1^{*T}\frac{\partial^2 A(x^*)}{\partial x_k \partial x_l}Q_1^* + 2\{Q_1^*\}^T\frac{\partial A(x^*)}{\partial x_k}Q_2^*\{\lambda_1^* I - \Lambda_2^*\}^{-1}\{Q_2^*\}^T\frac{\partial A(x^*)}{\partial x_l}Q_1^*.$$

Then W^* is independent of the choice of basis Q_1^*.

4. *The null space of K^* is independent of the choice of basis Q_1^*. Consequently, if N^* is a matrix with orthonormal columns spanning the null space of K^*, the eigenvalues of the reduced Hessian matrix*

$$\{N^*\}^T\begin{bmatrix} 0 & 0 \\ 0 & W^* \end{bmatrix}N^* \tag{5.31}$$

are independent of the choice of bases Q_1^*, N^*. (The matrix in the center of this expression has dimension $m+1$ by $m+1$.)

Proof. 1. The rth singular value of K^* can be written

$$\min_{\|S\|_F=1} \|\{K^*\}^T\{\text{vec } S\}\|,$$

where S is a t by t symmetric matrix. (The quantity (4.12) is zero in the general case that K^* has more columns than rows.) The quantity being minimized is

$$\left(\{\text{tr } S\}^2 + \sum_{k=1}^{m}\left\{S \bullet \{Q_1^*\}^T\frac{\partial A(x^*)}{\partial x_k}Q_1^*\right\}^2\right)^{\frac{1}{2}}.$$

This minimum value is independent of the choice of basis Q_1^*, since any rotation of the basis can be absorbed into S.

2. Let

$$U^* = \begin{bmatrix} U_{11}^* & U_{12}^* \\ \{U_{12}^*\}^T & U_{22}^* \end{bmatrix}.$$

We claim that (5.28) is equivalent to the two conditions (5.29)–(5.30). To see that (5.28) implies (5.29)–(5.30), observe, by analogy with (5.3)–(5.7) and (3.18)–(3.24), that $U^* \bullet F_\Theta^* = 0$ implies the diagonal elements of U_{22}^* are zero, while $U^* \bullet F_{Y_{22}}^*(Z^*) = 0$ and $U^* \bullet F_{Y_{12}}^*(Z^*) = 0$, together with (3.1), imply respectively that the off-diagonal elements of U_{22}^* and all elements of U_{12}^* are zero. The conditions $U^* \bullet F_\omega^* = 1$ and $U^* \bullet F_x^*(Z^*) = 0$ then reduce to (5.30). Conversely, if (5.29)–(5.30) hold, it is easily verified that (5.28) holds. The linear independence of the columns of $\{K^*\}^T$, equivalently the columns of the coefficient matrix of the linear system (5.28), provides a constraint qualification guaranteeing the existence and uniqueness of U^*.

3. Let M^* be defined by (5.25) with $\widehat{x}, \widehat{\Lambda}, \widehat{Q}$ replaced respectively by x^*, Λ^*, Q^*. (This is equivalent to (5.18) in this case since $\lambda_1^* = \cdots = \lambda_t^*$.) When Q_1^* is postmultiplied by a t by t orthogonal matrix P, it has the following effect: the first column of K^* is unchanged and the others are replaced by vec $P^T Q_1^* \frac{\partial A(x^*)}{\partial x_k} Q_1^* P$; the matrix M^* is replaced by $P^T M^* P$; the matrix U_{11}^* is replaced by $P^T U_{11}^* P$. By analogy with (5.20), $\{\Delta x\}^T W^* \{\Delta x\} = U_{11}^* \bullet M^*$ for all $\{\Delta x\}$, so it follows that W^* is independent of the choice of basis Q_1^*.

4. The null space of K^* is

$$\{v: \quad K^* v = 0\}$$

that is,

$$\left\{ v = (v_0 \; v_1 \ldots v_m)^T : \quad v_0 I + \sum_{k=1}^{m} v_k \{Q_1^*\}^T \frac{\partial A(x^*)}{\partial x_k} Q_1^* = 0 \right\},$$

which is unchanged if Q_1^* is postmultiplied by an orthogonal matrix. □

The previous theorem was concerned only with quantities involving x^* and F^*. In order to prove convergence of Iterations 3 and 4, however, we need to quantify the relationship between $\widehat{U}$ and $\widehat{U}^*$, the latter quantity being the dual matrix associated with the solution of (3.11)–(3.12).

THEOREM 6. *Suppose K^* has linearly independent rows and that $\widehat{x}$ is sufficiently close to x^*. Consider the nonlinear program* (3.11)–(3.12), *which has no constraint that $Y_{11} = 0$. A necessary condition for $\widehat{Z}^* = (x^*, \widehat{Y}^*, \lambda_1^*, \Lambda_2^*)$ to solve* (3.11)–(3.12) *is that there exists an $n \times n$ symmetric matrix $\widehat{U}^*$ satisfying*

$$\widehat{L}_Z(\widehat{Z}^*, \widehat{U}^*) = 0, \tag{5.32}$$

i.e., (5.3)–(5.6) *hold for $Z = \widehat{Z}^*$, $U = \widehat{U}^*$. Furthermore, $\widehat{U}^*$ is unique. Now assume that the discrepancy in* (5.3)–(5.6), *with $Z = \widehat{Z}$, $U = \widehat{U}$ is $O(\|\widehat{Z} - \widehat{Z}^*\|)$, as required by Iteration* 3. *Then*

$$\|\widehat{U} - \widehat{U}^*\| = O(\|\widehat{Z} - \widehat{Z}^*\|). \tag{5.33}$$

Furthermore, such a matrix $\widehat{U}$ is obtained by using the block structure (5.11) *and solving the least squares problem* (5.15).

Proof. From Theorem 5, the independence of the rows of K^* and the independence of the columns of the coefficient matrix defining the linear system (5.28) are equivalent. Using (4.18)–(4.19), it follows that if $\|\widehat{x} - x^*\|$ is sufficiently small, the columns of the linear system (5.32) are also independent. (The fact that the columns of the latter system have more rows than the columns of the former, because of the presence of the additional variables Y_{11}, does not affect the linear independence.) This rank condition provides a constraint qualification guaranteeing the existence and uniqueness of $\widehat{U}^*$, satisfying (5.32), i.e.,

$$\widehat{U}^* \bullet \widehat{F}_Z(\widehat{Z}^*) = v, \tag{5.34}$$

where v is a vector with one nonzero element, namely 1, in the position corresponding to the variable ω. By definition, $\widehat{U}$ satisfies

$$\widehat{U} \bullet \widehat{F}_Z(\widehat{Z}) = v + O(\|\widehat{Z} - \widehat{Z}^*\|),$$

which has no equations corresponding to Y_{11}. Subtracting this equation from the corresponding equations in (5.34), ignoring the Y_{11} equations in (5.34), and noting that $\widehat{F}_Z$ is Lipschitz, gives

$$\{\widehat{U} - \widehat{U}^*\} \bullet \widehat{F}_Z(\widehat{Z}) = O(\|\widehat{Z} - \widehat{Z}^*\|).$$

The independence of the columns of the coefficient matrix defining this system then gives (5.33).

The proof of the final statement of the theorem is as follows. From (5.30),

$$K^*\{K^*\}^T\{\text{vec } U_{11}^*\} = K^* e_1$$

and, from (5.15),

$$\widehat{K}\widehat{K}^T\{\text{vec } \widehat{U}_{11}\} = \widehat{K}^T e_1.$$

It follows as a consequence, using (4.19) and the fact that K^* is full rank, that

$$\|\widehat{U}_{11} - U_{11}^*\| = O(\|\widehat{x} - x^*\|).$$

Combining this equation with (4.19) and (5.30) gives

$$\widehat{K}^T\{\text{vec } \widehat{U}_{11}\} = e_1 + O(\|\widehat{x} - x^*\|)$$

from which the result follows. □

We are now ready to prove the main convergence theorem.

THEOREM 7. *Suppose that K^* has independent rows and that the reduced Hessian* (5.31) *is positive definite. Then there exist constants ϵ and C such that, if $\|\widehat{x} - x^*\| \leq \epsilon$, then*

$$\|\overline{x} - x^*\| \leq C\|\widehat{x} - x^*\|^2$$

for both Iterations 3 *and* 4. *Consequently, both iterations generate points $\widehat{x}$ that converge quadratically to the solution x^*.*

Proof. From Theorem 6, assuming that $\widehat{x}$ is sufficiently close to x^*, a necessary condition for a pair $\widehat{Z}^*, \widehat{U}^*$ to solve the nonlinear program (3.11)–(3.12) (*without*

the condition $Y_{11} = 0$ imposed), is that, in addition to (3.12), the equation (5.32) holds. Theorem 1 shows that we can take the $\hat{Y}^*$ component of $\hat{Z}^*$ to satisfy $\|\hat{Y}^*\| = O(\|\hat{x} - x^*\|)$ and $\|\hat{Y}^*_{11}\| = O(\|\hat{x} - x^*\|^2)$. Furthermore, we can expand $\hat{F}$ in a Taylor series just as in the proof of Theorem 3, obtaining all of (4.13)–(4.17) exactly as before, the only difference being that these equations are not square systems. Specifically, (4.17), with its Y_{11} terms absorbed into the right-hand side, gives

$$[\hat{F}_Z(\hat{Z})\{\overline{Z} - \hat{Z}^*\}] = O(\|\hat{x} - x^*\|^2). \tag{5.35}$$

Now let us expand (5.32) in a Taylor series. We have

$$\begin{aligned} 0 = \hat{L}_Z(\hat{Z}^*, \hat{U}^*) = \hat{L}_Z(\hat{Z}, \hat{U}) + [\hat{L}_{ZZ}(\hat{Z}, \hat{U})\{\hat{Z}^* - \hat{Z}\}] + [\hat{L}_{ZU}(\hat{Z}, \hat{U})\{\hat{U}^* - \hat{U}\}] \\ + O(\|\hat{Z} - \hat{Z}^*\|^2 + \|\hat{Z} - \hat{Z}^*\|\|\hat{U} - \hat{U}^*\|) \end{aligned}$$

using the linearity of $\hat{L}(Z, U)$ in U. Note that the terms in square brackets, although involving second-order differentiation, are summed over only one argument and are therefore vectors of length $n(n+1)/2 + m + 1 - t$, the number of variables in Z. This system of equations has a row and a column corresponding to each element of $Z = (x, Y, \omega, \Theta)$. Let us *discard* the *rows* corresponding to Y_{11}, and *absorb* the *columns* corresponding to Y_{11} into the O term, which is permissible since $\hat{Y} = 0$, $\hat{Y}^*_{11} = O(\|\hat{x} - x^*\|^2)$. Using the fact that $\hat{L}_{ZU} = -\hat{F}_Z$, this gives

$$0 = \hat{L}_Z(\hat{Z}, \hat{U}) + [\hat{L}_{ZZ}(\hat{Z}, \hat{U})\{\hat{Z}^* - \hat{Z}\}] - \{\hat{U}^* - \hat{U}\} \bullet \hat{F}_Z(\hat{Z}) + O(\|\hat{x} - x^*\|^2) \tag{5.36}$$

with the understanding that all Y_{11} terms are omitted. The $O(\|\hat{x} - x^*\|^2)$ term on the right-hand side is justified by (3.25) and (5.33).

The necessary condition for a pair $\Delta Z, \Delta U$ to solve the quadratic program defining a step of Iteration 3 is, in addition to the constraints (5.9)–(5.10), that there exists a dual matrix ΔU such that

$$\Delta U \bullet \hat{F}_Z(\hat{Z}) = \hat{L}_Z(\hat{Z}, \hat{U}) + [\hat{L}_{ZZ}(\hat{Z}, \hat{U})\Delta Z], \tag{5.37}$$

where rows and columns of the coefficient matrix corresponding to Y_{11} have been omitted because of (5.10). Noting that $\Delta Z = \overline{Z} - \hat{Z}$ and subtracting (5.36) from (5.37) gives

$$[\hat{L}_{ZZ}(\hat{Z}, \hat{U})\{\overline{Z} - \hat{Z}^*\}] - \{\overline{U} - \hat{U}^*\} \bullet \hat{F}_Z(\hat{Z}) = O(\|\hat{x} - x^*\|^2), \tag{5.38}$$

where $\overline{U} = \hat{U} + \Delta U$.

Equations (5.35), (5.38) state the first-order optimality conditions for the quadratic program

$$\min_{\overline{Z} - \hat{Z}^*} \quad h \cdot \{\overline{Z} - \hat{Z}^*\} + \tfrac{1}{2}[\hat{L}_{ZZ}(\hat{Z}, \hat{U})\{\overline{Z} - \hat{Z}^*\}\{\overline{Z} - \hat{Z}^*\}] \tag{5.39}$$

$$\text{subject to} \qquad [\hat{F}_Z(\hat{Z})\{\overline{Z} - \hat{Z}^*\}] = O(\|\hat{x} - x^*\|^2), \tag{5.40}$$

where the first term in (5.39) is an inner product, with h (which has the same structure as Z) satisfying $h = O(\|\hat{x} - x^*\|^2)$. It is understood that there are no Y_{11} terms in $\overline{Z}$, $\hat{Z}^*$. Note that the Hessian and constraint coefficients of this quadratic program are identical to those of (5.8)–(5.9). We shall now simplify this quadratic program,

using an argument similar to that which reduced (5.8)–(5.9) to (5.23)–(5.24). First consider the linear term in (5.39). We have

$$h \cdot \{\overline{Z} - \widehat{Z}^*\} = \widetilde{h} \cdot \begin{bmatrix} \widehat{\lambda}_1 + \Delta\omega - \lambda_1^* \\ \overline{x} - x^* \end{bmatrix} + \psi, \tag{5.41}$$

where $\widetilde{h} \in \Re^{m+1}$ and $\psi \in \Re$ satisfy $\widetilde{h} = O(\|\widehat{x} - x^*\|^2)$ and $\psi = O(\|\widehat{x} - x^*\|^4)$. This equation holds because of the constraint (5.40), which defines the Y and Θ elements of $\overline{Z} - \widehat{Z}^*$ in terms of the x and ω components, by analogy with (4.8)–(4.10). Now consider the quadratic term in (5.39). The argument that showed that the quadratic form in (5.8) reduces to that in (5.23) uses (5.17), which follows from the 1,2 block of (5.9), namely, (4.8). We now use a similar argument to simplify the quadratic term in (5.39). Instead of (4.8), we have, from the 1,2 block of (5.40),

$$-\widehat{Q}_1^T[A_x(\widehat{x})\{\overline{x} - x^*\}]\widehat{Q}_2 - \widehat{\Lambda}_1\{\Delta Y - \widehat{Y}^*\}_{12} + \{\Delta Y - \widehat{Y}^*\}_{12}\widehat{\Lambda}_2 = O(\|\widehat{x} - x^*\|^2).$$

Instead of (5.17), we conclude that

$$[\widehat{F}_{YY}(\widehat{Z})\{\Delta Y - \widehat{Y}^*\}\{\Delta Y - \widehat{Y}^*\}]_{11} + [\widehat{F}_{xY}(\widehat{Z})\{\overline{x} - x^*\}\{\Delta Y - \widehat{Y}^*\}]_{11}$$

$$= O(\|\widehat{x} - x^*\|^2 \|\Delta Y - \widehat{Y}^*\|).$$

Again using (5.40) to define $\Delta Y - \widehat{Y}^*$ in terms of the x and ω components of $\overline{Z} - \widehat{Z}^*$, we see that the right-hand side consists of two terms, of which one can be absorbed into the first term of (5.41), and the other into the second. We therefore see that, just as the quadratic form in (5.8) reduces to that in (5.23), the quadratic form in (5.39) reduces to

$$\psi + \widetilde{h} \cdot \begin{bmatrix} \widehat{\lambda}_1 + \Delta\omega - \lambda_1^* \\ \overline{x} - x^* \end{bmatrix} + \tfrac{1}{2}\{\overline{x} - x^*\}^T \widehat{W}\{\overline{x} - x^*\}, \tag{5.42}$$

where $\widetilde{h} = O(\|\widehat{x} - x^*\|^2)$. The constraint (5.40) reduces to (4.17), i.e.,

$$\widehat{K}\begin{bmatrix} \widehat{\lambda}_1 + \Delta\omega - \lambda_1^* \\ \overline{x} - x^* \end{bmatrix} = O(\|\widehat{x} - x^*\|^2). \tag{5.43}$$

The optimality conditions for the quadratic program defined by (5.42)–(5.43) are

$$\begin{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \widehat{W} \end{bmatrix} & -\widehat{K}^T \\ \widehat{K} & 0 \end{bmatrix} \begin{bmatrix} \begin{bmatrix} \widehat{\lambda}_1 + \Delta\omega - \lambda_1^* \\ \overline{x} - x^* \end{bmatrix} \\ \operatorname{vec}\{\overline{U}_{11} - \widehat{U}_{11}^*\} \end{bmatrix} = O(\|\widehat{x} - x^*\|^2), \tag{5.44}$$

By assumption, K^* has full rank and (5.31) is positive definite, so

$$\begin{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & W^* \end{bmatrix} & \{K^*\}^T \\ K^* & 0 \end{bmatrix}$$

is nonsingular. Therefore using (4.18)–(4.19) and noting that $\|\widehat{W} - W^*\| = O(\|\widehat{x} - x^*\|)$, we see that the inverse of the coefficient matrix of (5.44) is bounded for $\widehat{x}$ near x^*. The desired quadratic contraction is therefore proved. □

6. Concluding remarks. The convergence proof just given is complicated, because of the disparity in the number of free parameters in the equations $\widehat{F} = 0$ and $F^* = 0$, even as $\widehat{x} \to x^*$. An alternative analysis of the same method has been given recently by Shapiro and Fan [15] in contemporary, independent work. Our results and those of [15] complement each other nicely. The analysis in [15] is shorter than ours but rests on several nontrivial results. The principal idea is that although eigenvectors are not smooth, eigenprojections are differentiable, and indeed derivative formulas are known (Kato [7]). Shapiro and Fan show how to construct a smoothly varying orthonormal basis for the eigenprojection, which agrees with a given orthonormal basis of eigenvectors at a point, though not in a neighborhood of the point. Neither the results from Kato nor the construction of the eigenprojection basis could be said to be elementary, though both are powerful. By contrast, our convergence proof is completely self-contained. The Hessian formulas arise simply from differentiating the function $\widehat{F}$ and do not require any machinery from Kato. The only outside result that is needed is Theorem 1, whose proof is elementary [4].

Appendix. The following shows that any real orthogonal matrix P with $\det P = 1$ may be written in the form $P = e^Y$, where $Y = -Y^T$. This derivation was suggested by J.-P. Haeberly. It is undoubtedly well known, though we lack a standard reference.

An orthogonal matrix has eigenvalues of the form ± 1 and $\cos\theta \pm i \sin\theta$, with a corresponding orthogonal set of eigenvectors. Thus, there exists an orthogonal matrix V such that

$$V^T P V = \mathrm{Diag}(D_1, \ldots, D_k),$$

where each D_j is either the number ± 1, or a 2×2 matrix of the form

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

Since $\det P = 1$, the number of -1's that occur must be even, so we may assume that the D_j's are either the number $+1$ or a 2×2 matrix as above. But $1 = e^0$, and

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = \exp \begin{bmatrix} 0 & -\theta \\ \theta & 0 \end{bmatrix}.$$

Hence, $\mathrm{Diag}(D_1, \ldots, D_k) = e^X$ for some block diagonal matrix X with nonzero diagonal blocks of the form

$$\begin{bmatrix} 0 & -\theta \\ \theta & 0 \end{bmatrix}.$$

Note that $X = -X^T$. Defining $Y = VXV^T$, we have

$$P = V\mathrm{Diag}(D_1, \ldots, D_k)V^T = Ve^X V^T = e^Y.$$

It remains to show that Y is skew-symmetric:

$$(VXV^T)^T = VX^T V^T = V(-X)V^T = -VXV^T.$$

The matrix Y is not unique, since incrementing θ by multiples of 2π does not change e^Y, but the solution set consists of isolated points in matrix space. In our local convergence analysis, we are concerned only with $P = e^Y$ in a neighborhood of the identity matrix and the corresponding Y in a neighborhood of the zero matrix (see Theorem 3.1).

Acknowledgments. We thank Jean-Pierre Haeberly and the anonymous referees for helping us to improve the discussion in §3. The first author would also like to acknowledge the kind hospitality of the Centre for Process Systems Engineering, Imperial College, London, where part of this work was conducted with support from the United Kingdom Science and Engineering Research Council.

REFERENCES

[1] J. CULLUM, W. DONATH, AND P. WOLFE, *The minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices*, Math. Programming Study, 3 (1975), pp. 35–55.
[2] M. K. H. FAN, private communication, 1988
[3] R. FLETCHER, *Semi-definite constraints in optimization*, SIAM J. Control Optim., 23 (1985), pp. 493–513.
[4] S. FRIEDLAND, J. NOCEDAL, AND M. OVERTON, *The formulation and analysis of numerical methods for inverse eigenvalue problems*, SIAM J. Numer. Anal., 24 (1987), pp. 634–667.
[5] J. GOODMAN, *Newton's method for constrained optimization*, Math. Programming, 33 (1985), pp. 162–171.
[6] J.-P. A. HAEBERLY AND M. OVERTON, *A hybrid algorithm for optimizing eigenvalues of symmetric definite pencils*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1141–1156.
[7] T. KATO, *A Short Introduction to Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1982.
[8] P. LANCASTER, *On eigenvalues of matrices dependent on a parameter*, Numer. Math., 6 (1964), pp. 377–387.
[9] J. NOCEDAL AND M. OVERTON, *Numerical methods for solving inverse eigenvalue problems*, in Lecture Notes in Mathematics 1005, V. Pereyra and A. Reinoza, eds., Springer-Verlag, New York, 1983.
[10] M. OVERTON, *On minimizing the maximum eigenvalue of a symmetric matrix*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 256–268.
[11] ———, *Large-scale optimization of eigenvalues*, SIAM J. Optim., 2 (1992), pp. 88–120.
[12] M. OVERTON AND R. WOMERSLEY, *Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices*, Math. Programming, 62 (1993), pp. 321–357.
[13] M. OVERTON AND X.-J. YE, *Towards second-order methods for structured nonsmooth optimization*, in Advances in Optimization and Numerical Analysis, S. Gomez and J.-P. Hennart, eds., Kluwer, The Netherlands, pp. 97–110, 1994.
[14] E. POLAK AND Y. WARDI, *A nondifferentiable optimization algorithm for structural problems with eigenvalue inequality constraints*, J. Structural Mechanics, 11 (1983), pp. 561–577.
[15] A. SHAPIRO AND M. FAN, *On eigenvalue optimization*, SIAM J. Optim., 5 (1995), pp. 552–568.
[16] R. TAPIA, *A stable approach to Newton's method for general mathematical programming problems in R^n*, J. Optim. Theory Appl., 14 (1974), pp. 453–476.
[17] J. VON NEUMANN AND E. WIGNER, *Über das Verhalten von Eigenwerten bei adiabatischen Prozessen*, Physik. Zeitschr., 30 (1929), pp. 467–470.
[18] G. WATSON, *Algorithms for minimum trace factor analysis*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1039–1053.
[19] ———, *Computing the structured singular value*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1054–1066.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 719–734, July 1995

002

A MINIMAX THEOREM AND A DULMAGE–MENDELSOHN TYPE DECOMPOSITION FOR A CLASS OF GENERIC PARTITIONED MATRICES*

SATORU IWATA† AND KAZUO MUROTA†

Abstract. This paper discusses an extension of the Dulmage–Mendelsohn decomposition for a certain class of matrices whose row-set and column-set are divided into couples or singletons. A genericity assumption is imposed and an admissible transformation is defined in respect of this partition structure. Extensions of the König–Egerváry theorem and the Hall–Ore theorem are established. The latter states that the rank of such a matrix is characterized by the minimum value of a submodular function, of which the set of minimizers yields a canonical block-triangularization under the admissible transformations.

Key words. combinatorial matrix theory, Dulmage–Mendelsohn decomposition, generic partitioned matrix, minimax theorem, submodular function

AMS subject classifications. 15A21, 05C50

1. Introduction.

A *generic matrix*, which is a basic concept in combinatorial matrix theory, is a matrix whose nonzero elements are indeterminates (independent parameters). The rank and other properties of a generic matrix are determined by the zero-nonzero pattern. The *Dulmage–Mendelsohn decomposition* (or the DM-*decomposition*)[4] of the generic matrix is a canonical *block-triangularization* by means of independent permutations of the row-set and the column-set. The DM-decomposition is characterized (see §3) by the set of minimizers of a certain submodular function, say p_{DM}, whose minimum value in turn characterizes the rank of the generic matrix. The practical significance of the DM-decomposition is now widely recognized in numerical computation and systems analysis.

This paper discusses an extension of the DM-decomposition. Specifically, we are concerned with a class of matrices whose row-set and column-set are independently divided into couples or singletons as follows:

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1\nu} \\ A_{21} & A_{22} & \cdots & A_{2\nu} \\ \vdots & \vdots & \ddots & \vdots \\ A_{\mu 1} & A_{\mu 2} & \cdots & A_{\mu\nu} \end{pmatrix}.$$

We assume that the matrix A is generic in the sense that $A_{\alpha\beta} = t_{\alpha\beta} A^{\natural}_{\alpha\beta}$ for $\alpha = 1, \ldots, \mu$; $\beta = 1, \ldots, \nu$, where $\mathcal{T} = \{t_{\alpha\beta} \mid \alpha = 1, \ldots, \mu; \beta = 1, \ldots, \nu\}$ is the set of independent parameters (indeterminates) and $A^{\natural}_{\alpha\beta}$ are constant matrices of size 2×2 or smaller. We call such a matrix A a GP(2)-*matrix*. Equivalence transformations of

* Received by the editors September 20, 1993; accepted for publication by R. Brualdi April 16, 1994.

† Research Institute for Mathematical Sciences, Kyoto University, Kyoto 606, Japan (iwata@kurims.kyoto-u.ac.jp) and (murota@kurims.kyoto-u.ac.jp). The work of the first author was supported by JSPS Fellowship for Japanese Junior Scientists. The work of the second author was supported by Grant-in-Aid for Scientific Research (C) of the Ministry of Education, Science and Culture Grant 05650064.

the form

$$(1)\quad \begin{pmatrix} S_1 & O & \cdots & O \\ O & S_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & O \\ O & \cdots & O & S_\mu \end{pmatrix}^{-1} \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1\nu} \\ A_{21} & A_{22} & \cdots & A_{2\nu} \\ \vdots & \vdots & \ddots & \vdots \\ A_{\mu 1} & A_{\mu 2} & \cdots & A_{\mu\nu} \end{pmatrix} \begin{pmatrix} T_1 & O & \cdots & O \\ O & T_2 & \ddots & O \\ \vdots & \ddots & \ddots & O \\ O & \cdots & O & T_\nu \end{pmatrix},$$

with S_α and T_β being constant matrices, preserve the partition structure as well as the genericity in the above sense and are called *admissible transformations* for GP(2)-matrices. Note that the resulting matrix is also a GP(2)-matrix.

It will be shown that the rank of a GP(2)-matrix is characterized by the minimum value of (a variant of) the submodular function p introduced in [7]. This result can be understood as a minimax theorem since the rank is a maximum size of nonsingular submatrices. In fact, this is an extension of the König–Egerváry theorem and the Hall–Ore theorem for generic matrices (bipartite graphs). The set of minimizers of p yields the finest proper block-triangularization of a GP(2)-matrix under admissible transformations, just as the submodular function p_{DM} induces the DM-decomposition of a generic matrix. The uniqueness of this decomposition is established by using a result of [7].

In the literature, a number of extensions of the DM-decomposition have been considered in different directions. Murota [10] has introduced the concept of *layered mixed matrices* and *multilayered matrices*, and established the *combinatorial canonical forms* for them (See also Murota [11] and Murota, Iri, and Nakamura [12]). All these cases have a common feature that the rank of a matrix is characterized by the minimum value of a certain submodular function on a boolean lattice. More general framework is investigated in Ito, Iwata, and Murota [7] under the name of *partitioned matrices* without reference to the genericity. It has been shown that the rank of a partitioned matrix is bounded from above by the minimum value of a submodular function defined on a modular (nonboolean) lattice and that a DM-type decomposition exists if and only if the bound is tight. Partitioned matrices with a genericity assumption (to be defined in §2) will be named *generic partitioned matrices*, of which our GP(2)-matrix is a special case. The minimax theorem for the rank, however, does not hold in the general framework even under this genericity assumption, as will be shown by a counterexample in §6. Apart from linear algebra, Iri [6] has discussed a decomposition principle for combinatorial systems characterized by submodular functions.

The outline of this paper is as follows. The concept of GP(2)-matrix and the block-triangularization is described in §2. Section 3 affords preliminaries on the DM-decomposition. The rank identity for GP(2)-matrices, which is the main result of this paper, is proven in §4. Section 5 is devoted to the existence and the uniqueness of the block-triangularization of GP(2)-matrices. The extension of our result to generic partitioned matrices of general types is discussed in §6.

2. Generic partitioned matrix. Let $\boldsymbol{K}$ be a field, e.g., the rational numbers $\mathbf{Q}$, and $A^{\natural}_{\alpha\beta}$ be an $m_\alpha \times n_\beta$ matrix over the field $\boldsymbol{K}$ for $\alpha = 1, \ldots, \mu$; $\beta = 1, \ldots, \nu$. Put $A_{\alpha\beta} = t_{\alpha\beta} A^{\natural}_{\alpha\beta}$, where $\mathcal{T} = \{t_{\alpha\beta} \mid \alpha = 1, \ldots, \mu;\ \beta = 1, \ldots, \nu\}$ is the set of

independent parameters (indeterminates). Then

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1\nu} \\ A_{21} & A_{22} & \cdots & A_{2\nu} \\ \vdots & \vdots & \ddots & \vdots \\ A_{\mu 1} & A_{\mu 2} & \cdots & A_{\mu\nu} \end{pmatrix}$$

is a matrix over the field $\boldsymbol{K}(\mathcal{T})$ of rational functions in $\mathcal{T}$ over $\boldsymbol{K}$. Such a matrix A is named here a *generic partitioned matrix* (or GP-*matrix*) of type $(m_1, \ldots, m_\mu; n_1, \ldots, n_\nu)$ with base field $\boldsymbol{K}$. In particular, A is called a GP(2)-*matrix* if $m_\alpha \leq 2$ for $\alpha = 1, \ldots, \mu$ and $n_\beta \leq 2$ for $\beta = 1, \ldots, \nu$. An equivalence transformation $S^{-1}AT$ is said to be an *admissible transformation* for a generic partitioned matrix A, if $S = \bigoplus_{\alpha=1}^{\mu} S_\alpha$ and $T = \bigoplus_{\beta=1}^{\nu} T_\beta$ with m_α-dimensional nonsingular matrices S_α and n_β-dimensional nonsingular matrices T_β over the field $\boldsymbol{K}$ (see expression (1) in §1).

For a matrix A in general we denote by $\mathrm{Row}(A)$ and $\mathrm{Col}(A)$ the row-set of A and the column-set of A, respectively. We also denote by $A[R', C']$ the submatrix of A with row indices in $R' \cap \mathrm{Row}(A)$ and column indices in $C' \cap \mathrm{Col}(A)$.

We now define precisely the notion of a block-triangular form, following Murota [11] and Ito, Iwata, and Murota [7]. Let $\widetilde{A}$ be a generic partitioned matrix. We say that $\widetilde{A}$ is in a *block-triangular form* or *block-triangularized* if the row-set $R = \mathrm{Row}(\widetilde{A})$ and the column-set $C = \mathrm{Col}(\widetilde{A})$ are split into a certain number of disjoint blocks: $(R_0; R_1, \ldots, R_b; R_\infty)$ and $(C_0; C_1, \ldots, C_b; C_\infty)$ in such a way that

$$\begin{array}{lll} |R_0| < |C_0| & \text{or} & |R_0| = |C_0| = 0, \\ |R_k| = |C_k| > 0 & \text{for} & k = 1, \ldots, b, \\ |R_\infty| > |C_\infty| & \text{or} & |R_\infty| = |C_\infty| = 0 \end{array}$$

and

$$\widetilde{A}[R_k, C_l] = O \quad \text{if} \quad 0 \leq l < k \leq \infty.$$

$\widetilde{A}$ is said to be *properly* block-triangularized, if, in addition,

$$\mathrm{rank}\widetilde{A}[R_k, C_k] = \min(|R_k|, |C_k|) \quad \text{for} \quad k = 0, 1, \ldots, b, \infty$$

is satisfied. $\widetilde{A}[R_0, C_0]$ and $\widetilde{A}[R_\infty, C_\infty]$ are called *horizontal tail* and *vertical tail* of $\widetilde{A}$, respectively. It is clear that if $\widetilde{A}$ is block-triangularized in the above sense, we can put it into an explicit upper block-triangular form $\bar{A} = P\widetilde{A}Q$ in the usual sense by using certain permutation matrices P and Q.

A partial order is induced among the blocks $\{C_k \mid k = 1, \ldots, b\}$ in a natural manner by the zero-nonzero structure of a block-triangular matrix $\widetilde{A}$. The partial order $\preceq$ is the reflexive and transitive closure of the relation defined by: C_k is "smaller" than or equal to C_l if $\widetilde{A}[R_k, C_l] \neq O$. We denote this poset $(\{C_1, \ldots, C_b\}, \preceq)$ by $\mathcal{P}(\widetilde{A})$.

A generic partitioned matrix A is said to be GP-*irreducible* if $\mathrm{rank}A = \min(m, n)$ and it can never be transformed into a proper block-triangular form with two or more nonempty blocks by any admissible transformation. If $\widetilde{A}$ is a proper block-triangular matrix obtained from A by an admissible transformation and if, in addition, all the diagonal blocks $\widetilde{A}[R_k, C_k]$ for $k = 0, 1, \ldots, b, \infty$ are GP-irreducible, we say that $\widetilde{A}$ is a GP-*irreducible decomposition* of A and $\widetilde{A}[R_k, C_k]$ are the GP-*irreducible components* of A.

Example 1. Consider the following 6×6 GP(2)-matrix:

$$A = \left(\begin{array}{cc|cc|cc} 2t_{11} & t_{11} & t_{12} & t_{12} & t_{13} & t_{13} \\ & & & & t_{13} & \\ \hline t_{21} & t_{21} & t_{22} & & t_{23} & \\ & & & & t_{23} & \\ \hline t_{31} & t_{31} & t_{32} & -t_{32} & & \\ & & & t_{32} & t_{33} & \end{array}\right).$$

Using admissible transformation matrices:

$$S = \left(\begin{array}{cc|cc|cc} 1 & 0 & & & & \\ 0 & 1 & & & & \\ \hline & & 1 & 0 & & \\ & & 0 & 1 & & \\ \hline & & & & 1 & -1 \\ & & & & 0 & 1 \end{array}\right), \quad T = \left(\begin{array}{cc|cc|cc} 1 & 0 & & & & \\ -1 & 1 & & & & \\ \hline & & 1 & 0 & & \\ & & 0 & 1 & & \\ \hline & & & & 1 & 0 \\ & & & & 0 & 1 \end{array}\right),$$

we obtain

$$\widetilde{A} = S^{-1}AT = \left(\begin{array}{cc|cc|cc} t_{11} & t_{11} & t_{12} & t_{12} & t_{13} & t_{13} \\ & & & & t_{13} & \\ \hline & t_{21} & t_{22} & & t_{23} & \\ & & & & t_{23} & \\ \hline & t_{31} & t_{32} & & t_{33} & \\ & & & t_{32} & t_{33} & \end{array}\right).$$

This is a block-triangular matrix since it can be put into an explicitly upper block-triangular form:

$$\bar{A} = P\widetilde{A}Q = \left(\begin{array}{cccccc} t_{13} & t_{11} & t_{11} & t_{12} & t_{12} & t_{13} \\ & & t_{21} & t_{22} & & t_{23} \\ & & t_{31} & t_{32} & & t_{33} \\ & & & & t_{32} & t_{33} \\ & & & & & t_{13} \\ & & & & & t_{23} \end{array}\right)$$

by using permutation matrices

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}, \quad Q = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Thus $\widetilde{A}$ is a GP-irreducible decomposition with the horizontal tail $\widetilde{A}[R_0, C_0] = \begin{pmatrix} t_{13} & t_{11} \end{pmatrix}$, the vertical tail $\widetilde{A}[R_\infty, C_\infty] = \begin{pmatrix} t_{13} \\ t_{23} \end{pmatrix}$, square diagonal blocks $\widetilde{A}[R_1, C_1] = \begin{pmatrix} t_{21} & t_{22} \\ t_{31} & t_{32} \end{pmatrix}$ and $\widetilde{A}[R_2, C_2] = \begin{pmatrix} t_{32} \end{pmatrix}$. Note that the partial order in $\mathcal{P}(\widetilde{A})$ is trivial, i.e., neither $C_1 \preceq C_2$ nor $C_1 \succeq C_2$ since $\widetilde{A}[R_1, C_2] = O$.

The GP-irreducible decomposition $\widetilde{A}$ is finer than the DM-decomposition (see §3) of A. In fact, by using permutation matrices

$$\breve{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}, \quad \breve{Q} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

we obtain the DM-decomposition of A:

$$\breve{A} = \breve{P}A\breve{Q} = \begin{pmatrix} t_{13} & 2t_{11} & t_{11} & t_{12} & t_{12} & t_{13} \\ & t_{21} & t_{21} & t_{22} & & t_{23} \\ & t_{31} & t_{31} & t_{32} & -t_{32} & \\ & & & & t_{32} & t_{33} \\ & & & & & t_{13} \\ & & & & & t_{23} \end{pmatrix}.$$

3. Recapitulation on DM-decomposition. We review a construction of the DM-decomposition (cf., e.g., [6], [8], [10]), which will serve as a prototype of the argument in §§4 and 6 for GP(2)-matrices.

A matrix A is said to be a generic matrix [3] if its nonzero elements are indeterminates (independent parameters). Since the determinants of submatrices of a generic matrix are free from numerical cancellations, the rank and other related properties are determined by its zero-nonzero pattern.

With a generic matrix $A = (a_{ij})$ is associated a submodular function p_{DM} defined as follows. Set $R = \mathrm{Row}(R)$, $C = \mathrm{Col}(A)$ and put

$$\begin{aligned} \Gamma(J) &= \bigcup_{j \in J} \{i \in R \mid a_{ij} \neq 0\}, \\ \gamma(J) &= |\Gamma(J)|, \\ p_{\mathrm{DM}}(J) &= \gamma(J) - |J| \end{aligned}$$

for $J \subseteq C$, where a_{ij} denotes the (i,j)-entry of A. As is well known, p_{DM} is submodular, i.e.,

$$p_{\mathrm{DM}}(J_1) + p_{\mathrm{DM}}(J_2) \geq p_{\mathrm{DM}}(J_1 \cup J_2) + p_{\mathrm{DM}}(J_1 \cap J_2), \quad J_h \subseteq C \ \ (h = 1, 2).$$

The rank of a generic matrix A is characterized by the minimum value of this submodular function:

$$\text{(2)} \qquad \mathrm{rank}\, A = \min\{p_{\mathrm{DM}}(J) \mid J \subseteq C\} + |C|,$$

which follows from the genericity and the Hall–Ore theorem for bipartite graphs [9]. The rank identity (2) is called here the Hall–Ore theorem for generic matrices.

Since the set of minimizers of a submodular function on a lattice forms a sublattice, the set of minimizers of p_{DM}, i.e.,

$$\mathcal{D}(p_{\mathrm{DM}}) = \{J \subseteq C \mid p_{\mathrm{DM}}(J) = \min_{J' \subseteq C} p_{\mathrm{DM}}(J')\},$$

is a distributive lattice. Consider a maximal chain

$$\mathcal{C}_{\mathrm{DM}} : J_0 \subset J_1 \subset \cdots \subset J_b,$$

of $\mathcal{D}(p_{\mathrm{DM}})$ and put

$$\begin{array}{llll} C_0 = J_0, & & R_0 = \Gamma(J_0), & \\ C_h = J_h - J_{h-1}, & & R_h = \Gamma(J_h) - \Gamma(J_{h-1}), & \text{for} \quad h = 1, \ldots, b, \\ C_\infty = C - J_b, & & R_\infty = R - \Gamma(J_b). & \end{array}$$

Then A is in a block-triangular form with respect to the blocks $(R_0; R_1, \ldots, R_b; R_\infty)$ and $(C_0; C_1, \ldots, C_b; C_\infty)$. Furthermore, it follows from (2) that A is in a proper block-triangular form. This block-triangularization is the Dulmage–Mendelsohn decomposition (or DM-decomposition).

4. A minimax theorem. In this section we introduce a submodular function p for a GP(2)-matrix, similar to p_{DM} in §3, and establish an analogue of the Hall–Ore theorem (2).

Let A be an $m \times n$ generic partitioned matrix and V be an n-dimensional vector space over the field $\boldsymbol{K}$ given by $V = \bigoplus_{\beta=1}^{\nu} V_\beta$, where V_β is an n_β-dimensional vector space for each β. Similarly, set $U = \bigoplus_{\alpha=1}^{\mu} U_\alpha$, where $\dim U_\alpha = m_\alpha$ for each α. We denote by $\mathcal{W}$ the modular lattice which consists of all the subspaces of V that can be represented as the direct sums of subspaces of V_β's.

Regarding $A_\alpha = (\, A_{\alpha 1} \quad A_{\alpha 2} \quad \cdots \quad A_{\alpha\nu} \,)$ as a linear map, we define

$$p(W) := \sum_{\alpha=1}^{\mu} \dim A_\alpha W - \dim W, \quad W \in \mathcal{W}.$$

Then $p : \mathcal{W} \to \mathbf{Z}$ is a submodular function, i.e.,

$$p(W_1) + p(W_2) \geq p(W_1 + W_2) + p(W_1 \cap W_2), \quad W_h \in \mathcal{W} \ \ (h = 1, 2).$$

The following lemma as well as its proof shows that the function p is relevant in dealing with the rank of generic partioned matrices. Though this is essentially the same as Lemma 3.14 in [7], we include here a simple direct proof. Note also that the genericity does not play a rôle.

LEMMA 4.1. *For an $m \times n$ generic partitioned matrix A of any type,*

$$\operatorname{rank} A \leq \min\{p(W) \mid W \in \mathcal{W}\} + n. \tag{3}$$

Proof. For any $W \in \mathcal{W}$, A can be transformed, by a certain admissible transformation, to another matrix

$$\widetilde{A} = \begin{array}{c} \begin{array}{cc} J \simeq W & C - J \end{array} \\ \begin{pmatrix} \widetilde{A}_1[R, J] & \widetilde{A}_1[R, C - J] \\ \widetilde{A}_2[R, J] & \widetilde{A}_2[R, C - J] \\ \vdots & \vdots \\ \widetilde{A}_\mu[R, J] & \widetilde{A}_\mu[R, C - J] \end{pmatrix} \end{array}$$

such that the set of basis vectors corresponding to J spans the subspace $W \in \mathcal{W}$ (as indicated by "$J \simeq W$"). Then we have

$$\begin{aligned}\operatorname{rank} A &= \operatorname{rank} \widetilde{A} \\ &\leq \sum_{\alpha=1}^{\mu} \operatorname{rank} \widetilde{A}_\alpha[R, J] + \operatorname{rank} \widetilde{A}[R, C - J] \\ &\leq \sum_{\alpha=1}^{\mu} \dim A_\alpha W + n - \dim W \\ &= p(W) + n.\end{aligned}$$

Hence we obtain the inequality (3). □

One of the main objectives of this paper is to show that the equality holds in (3) if A is of type $(2, \ldots, 2; 2, \ldots, 2)$.

We denote by $\mathcal{Y}$ the modular lattice consisting of such subspaces of U that can be represented as direct sums of subspaces of U_α's. Put

$$\lambda(Y, W) = \dim(AW/Y) = \dim AW - \dim AW \cap Y,$$

and then we have the following lemmas.

LEMMA 4.2. *The function $\lambda : \mathcal{Y} \times \mathcal{W} \to \mathbf{Z}$ is submodular, i.e.,*

$$\lambda(Y_1, W_1) + \lambda(Y_2, W_2) \geq \lambda(Y_1 + Y_2, W_1 + W_2) + \lambda(Y_1 \cap Y_2, W_1 \cap W_2)$$

for $Y_h \in \mathcal{Y}$, $W_h \in \mathcal{W}$ $(h = 1, 2)$.

Proof. It is clear from the following manipulation:

$$\begin{aligned}\lambda(Y_1, W_1) + \lambda(Y_2, W_2) &= \dim AW_1 + \dim AW_2 - \dim AW_1 \cap Y_1 - \dim AW_2 \cap Y_2 \\ &= \dim A(W_1 + W_2) + \dim(AW_1 \cap AW_2)/(Y_1 \cap Y_2) \\ &\quad - \dim((AW_1 \cap Y_1) + (AW_2 \cap Y_2)) \\ &\geq \dim A(W_1 + W_2) + \dim A(W_1 \cap W_2)/(Y_1 \cap Y_2) \\ &\quad - \dim A(W_1 + W_2) \cap (Y_1 + Y_2) \\ &= \lambda(Y_1 + Y_2, W_1 + W_2) + \lambda(Y_1 \cap Y_2, W_1 \cap W_2). \qquad \square\end{aligned}$$

LEMMA 4.3. *For an $m \times n$ generic partitioned matrix A of any type, there exists a pair $Y^* \in \mathcal{Y}$ and $W^* \in \mathcal{W}$ such that*

(i) $\dim W^* - \dim Y^* - \lambda(Y^*, W^*) = n - \operatorname{rank} A$.

(ii) $\lambda(Y', W') = \lambda(Y^*, W^*)$ *for any $Y' \supset Y^*$ and $W' \subset W^*$ such that* $\dim Y' = \dim Y^* + 1$ *and* $\dim W' = \dim W^* - 1$.

Proof. Consider a pair (Y^*, W^*) which minimizes $\dim W^* - \dim Y^*$ subject to (i). Such (Y^*, W^*) certainly exists since (i) is satisfied by $(\{\mathbf{0}\}, V)$. Then for any $Y' \supset Y^*$ and $W' \subset W^*$ such that $\dim Y' = \dim Y^* + 1$ and $\dim W' = \dim W^* - 1$, it follows from Lemma 4.2 that

$$\lambda(Y^*, W^*) + \lambda(Y', W') \geq \lambda(Y', W^*) + \lambda(Y^*, W').$$

Because of the minimality of $\dim W^* - \dim Y^*$ we have $\lambda(Y', W^*) = \lambda(Y^*, W^*)$ since otherwise (Y', W^*) would satisfy (i) with $\dim W^* - \dim Y' < \dim W^* - \dim Y^*$. Likewise we have $\lambda(Y^*, W') = \lambda(Y^*, W^*)$. Therefore $\lambda(Y', W') \geq \lambda(Y^*, W^*)$. On

the other hand, it is clear that $\lambda(Y', W') \leq \lambda(Y^*, W^*)$ since $Y' \supset Y^*$ and $W' \subset W^*$. Hence (Y^*, W^*) satisfies (ii). □

Whereas the above three lemmas are valid for generic partitioned matrices of any type, the following theorem (Theorem 4.4) states a key property valid for GP(2)-matrices (and not for generic partitioned matrices of general type). A special case of Theorem 4.4 with A being a generic matrix (i.e., GP(2)-matrix type $(1, \ldots, 1; 1, \ldots, 1)$) is nothing but the König–Egerváry theorem for bipartite graphs.

THEOREM 4.4. *For an $m \times n$ GP(2)-matrix A, there exists a pair $Y^* \in \mathcal{Y}$ and $W^* \in \mathcal{W}$ such that*

(i) $\dim W^* - \dim Y^* = n - \operatorname{rank} A$.

(ii) $\lambda(Y^*, W^*) = 0$.

In other words, there exists an admissible transformation $\widetilde{A} = S^{-1}AT$ and subsets $R^* \subseteq \operatorname{Row}(\widetilde{A})$ and $C^* \subseteq \operatorname{Col}(\widetilde{A})$ such that

(i′) $|R^*| + |C^*| = m + n - \operatorname{rank} A$,

(ii′) $\operatorname{rank} \widetilde{A}[R^*, C^*] = 0$.

Proof. Given a pair (Y^*, W^*) of Lemma 4.3, consider an admissible transformation $\widetilde{A} = S^{-1}AT$ such that a subset of the column vectors of S spans Y^* and a subset of the column vectors of T spans W^*. We denote by R^* the complement of the subset of $\operatorname{Row}(\widetilde{A})$ corresponding to Y^* and by C^* the subset of $\operatorname{Col}(\widetilde{A})$ corresponding to W^*. Note that $\operatorname{Row}(\widetilde{A})$ and $\operatorname{Col}(\widetilde{A})$ have natural one-to-one correspondences with $\operatorname{Col}(S)$ and $\operatorname{Col}(T)$, respectively, and that $|R^*| = m - \dim Y^*$ and $|C^*| = \dim W^*$.

We claim that $\operatorname{rank} \widetilde{A}_{\alpha\beta}[R^*, C^*] \neq 1$ for each (α, β). Assume to the contrary that $\widetilde{A}_{\alpha\beta}[R^*, C^*]$ has rank 1 for some (α, β). We may further assume that $\widetilde{A}_{\alpha\beta}[R^*, C^*]$ is in its rank normal form (i.e., $\left(\begin{smallmatrix} t_{\alpha\beta} & 0 \\ 0 & 0 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} t_{\alpha\beta} \\ 0 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} t_{\alpha\beta} & 0 \end{smallmatrix}\right)$ or $\left(t_{\alpha\beta}\right)$) and that $\widetilde{A}_{\alpha\beta}[R^*, C^*]$ has the only nonzero element at (i, j)-entry of $\widetilde{A}$. Let $\widetilde{A}[I^*, J^*]$ be a maximum-size nonsingular submatrix of $\widetilde{A}[R^* - \{i\}, C^* - \{j\}]$, and then $\widetilde{A}[I^* \cup \{i\}, J^* \cup \{j\}]$ is nonsingular since the nonzero terms arising from $t_{\alpha\beta} \det \widetilde{A}[I^*, J^*]$ would not vanish in the determinant expansion of $\widetilde{A}[I^* \cup \{i\}, J^* \cup \{j\}]$ because of the genericity. Therefore we have

$$\operatorname{rank} \widetilde{A}[R^*, C^*] > \operatorname{rank} \widetilde{A}[R^* - \{i\}, C^* - \{j\}],$$

which contradicts the condition (ii) of Lemma 4.3. Hence $\operatorname{rank} \widetilde{A}_{\alpha\beta}[R^*, C^*]$ is 0 or 2.

Consider a generic matrix $B = (b_{\alpha\beta})$ with $\operatorname{Row}(B) = \{1, \ldots, \mu\}$ and $\operatorname{Col}(B) = \{1, \ldots, \nu\}$ defined by

$$b_{\alpha\beta} = \begin{cases} t_{\alpha\beta} & \text{if } \operatorname{rank} \widetilde{A}_{\alpha\beta}[R^*, C^*] = 2, \\ 0 & \text{if } \operatorname{rank} \widetilde{A}_{\alpha\beta}[R^*, C^*] = 0. \end{cases}$$

Note the correspondence between the entry $b_{\alpha\beta}$ of B and the submatrix $A_{\alpha\beta}$ of A. The DM-decomposition of B splits $\overline{R} = \operatorname{Row}(B)$ and $\overline{C} = \operatorname{Col}(B)$ into blocks $(\overline{R}_0; \overline{R}_1, \ldots, \overline{R}_b; \overline{R}_\infty)$ and $(\overline{C}_0; \overline{C}_1, \ldots, \overline{C}_b; \overline{C}_\infty)$, respectively. Accordingly, R^* and C^* are split into blocks $(R_0^*; R_1^*, \ldots, R_b^*; R_\infty^*)$ and $(C_0^*; C_1^*, \ldots, C_b^*; C_\infty^*)$, respectively. Since $\operatorname{rank} \widetilde{A}_{\alpha\beta}[R^*, C^*]$ is either 2 or 0, it follows from the genericity that

$$\operatorname{rank} \widetilde{A}[R^*, C^*] = 2 \operatorname{rank} B.$$

Moreover, $\widetilde{A}[R^*, C^*]$ is in a proper block-triangular form with respect to the blocks $(R_0^*; R_1^*, \ldots, R_b^*; R_\infty^*)$ and $(C_0^*; C_1^*, \ldots, C_b^*; C_\infty^*)$. For any $i \in R^* - R_\infty^*$, we would have

$$\operatorname{rank} \widetilde{A}[R^* - \{i\}, C^*] < \operatorname{rank} \widetilde{A}[R^*, C^*],$$

which contradicts the condition (ii) in Lemma 4.3. Similarly, for any $j \in C^* - C_0^*$, we would have

$$\operatorname{rank} \widetilde{A}[R^*, C^* - \{j\}] < \operatorname{rank} \widetilde{A}[R^*, C^*],$$

which also contradicts the condition (ii) in Lemma 4.3. Therefore $R^* = R_\infty^*$ and $C^* = C_0^*$. That is to say, $\widetilde{A}[R^*, C^*] = O$, i.e., $\operatorname{rank} \widetilde{A}[R^*, C^*] = 0$. □

We now state the main result of this paper, namely the rank identity for GP(2)-matrices which is an extension of the Hall–Ore theorem for generic matrices.

THEOREM 4.5. *For an $m \times n$ GP(2)-matrix A,*

$$\operatorname{rank} A = \min\{p(W) \mid W \in \mathcal{W}\} + n. \tag{4}$$

Proof. Let (Y^*, W^*) be the pair of Theorem 4.4. From (ii) it follows that $A_\alpha W^* \subseteq Y^* \cap U_\alpha$. Using this and (i) we obtain

$$\begin{aligned} \operatorname{rank} A &= \dim Y^* - \dim W^* + n \\ &\geq \sum_{\alpha=1}^{\mu} \dim A_\alpha W^* - \dim W^* + n \\ &= p(W^*) + n. \end{aligned}$$

The other direction of the inequality is given in Lemma 4.1. □

REMARK 1. Lemma 4.3 has been inspired by a recent result of Bapat [2], which gives a matroid-theoretic abstraction of the König–Egerváry theorem and its extensions (Theorem 1 of Hartfiel and Loewy [5], Theorem 16 of Murota [11]) to mixed matrices. However, neither Lemma 4.3 nor Theorem 4.4 follows from these previous results.

5. Dulmage–Mendelsohn type decomposition.

5.1. Construction of the decomposition. In this section we consider a DM-type decomposition of GP(2)-matrices. The minimax result of Theorem 4.5 immediately yields such a decomposition (cf. Theorem 3.15 of [7]). In fact, the following construction is essentially the same as the one in [7] except that we consider here the linear subspaces over the subfield $\boldsymbol{K}$.

It is well known that the set of the minimizers of a submodular function on a lattice is a sublattice and that a sublattice of a modular lattice is also modular [1]. Therefore

$$\mathcal{L}(p) := \{W \in \mathcal{W} \mid p(W) = \min_{W' \in \mathcal{W}} p(W')\}$$

is a modular lattice.

Let $\mathcal{C}$ be a maximal chain of $\mathcal{L}(p)$:

$$\mathcal{C} : W_0 \subset W_1 \subset \cdots \subset W_b.$$

Denoting $W_h \cap V_\beta$ by $W_{h\beta}$, we obtain from $\mathcal{C}$ a family of increasing chains

$$\mathcal{C}_\beta : W_{0\beta} \subseteq W_{1\beta} \subseteq \cdots \subseteq W_{b\beta}$$

for $\beta = 1, \ldots, \nu$. Let $\Psi_{h\beta}$ be a set of linearly independent column vectors spanning $W_{h\beta}$ for $h = 0, 1, \ldots, b$ and $\Psi_{\infty\beta}$ spanning V_β such that

$$\Psi_{0\beta} \subseteq \Psi_{1\beta} \subseteq \cdots \subseteq \Psi_{b\beta} \subseteq \Psi_{\infty\beta}.$$

Then $\Psi_h = \bigcup_{\beta=1}^{\nu} \Psi_{h\beta}$ spans W_h for $h = 0, 1, \ldots, b$, and $\Psi = \bigcup_{\beta=1}^{\nu} \Psi_{\infty\beta}$ becomes a basis of V. Order the n column vectors of Ψ as $[\Psi_{\infty 1}, \Psi_{\infty 2}, \ldots, \Psi_{\infty\nu}]$ to get a nonsingular matrix $T = \bigoplus_{\beta=1}^{\nu} T_\beta$.

Similarly, we obtain from $\mathcal{C}$ another family of increasing chains

$$A_\alpha\mathcal{C} : A_\alpha W_0 \subseteq A_\alpha W_1 \subseteq \cdots \subseteq A_\alpha W_b$$

for $\alpha = 1, \ldots, \mu$. Let $\Phi_{h\alpha}$ be a set of linearly independent column vectors spanning $A_\alpha W_h$ for $h = 0, 1, \ldots, b$ and $\Phi_{\infty\alpha}$ spanning U_α such that

$$\Phi_{0\alpha} \subseteq \Phi_{1\alpha} \subseteq \cdots \subseteq \Phi_{b\alpha} \subseteq \Phi_{\infty\alpha}.$$

Then $\Phi_h = \bigcup_{\alpha=1}^{\mu} \Phi_{h\alpha}$ spans AW_h for $h = 0, 1, \ldots, b$, and $\Phi = \bigcup_{\alpha=1}^{\mu} \Phi_{\infty\alpha}$ becomes a basis of U. Order the m column vectors of Φ as $[\Phi_{\infty 1}, \Phi_{\infty 2}, \ldots, \Phi_{\infty\mu}]$ to get a nonsingular matrix $S = \bigoplus_{\alpha=1}^{\mu} S_\alpha$.

Put $\widetilde{A} := S^{-1}AT$. Let $C_h \subseteq \mathrm{Col}(\widetilde{A})$ be the column subset corresponding to $\hat{\Psi}_h$, and $R_h \subseteq \mathrm{Row}(\widetilde{A})$ the row subset corresponding to $\hat{\Phi}_h$, where

$$\begin{aligned} &\hat{\Psi}_0 = \Psi_0, && \hat{\Phi}_0 = \Phi_0, \\ &\hat{\Psi}_h = \Psi_h - \Psi_{h-1}, && \hat{\Phi}_h = \Phi_h - \Phi_{h-1}, \quad \text{for} \quad h = 1, \ldots, b, \\ &\hat{\Psi}_\infty = \Psi_\infty - \Psi_b, && \hat{\Phi}_\infty = \Phi_\infty - \Phi_b. \end{aligned}$$

Then we have

$$\widetilde{A}[R_k, C_l] = O \quad \text{if} \quad 0 \le l < k \le \infty.$$

Since

$$p(W_k) = \sum_{h=0}^{k} |R_h| - \sum_{h=0}^{k} |C_h|, \quad \text{for } k = 0, 1, \ldots, b,$$

it holds that

$$|R_l| = |C_l| \quad \text{for } l = 1, \ldots, b.$$

It follows from Theorem 4.5 that

$$\mathrm{rank}\widetilde{A}[R_k, C_k] = \min(|R_k|, |C_k|) \quad \text{for} \quad k = 0, 1, \ldots, b, \infty.$$

That is to say, $\widetilde{A}$ is in a proper block-triangular form, where the number of square blocks b is given by the length of $\mathcal{C}$. The maximality of $\mathcal{C}$ guarantees that $\widetilde{A}$ is a GP-irreducible decomposition (cf. §2) of A. Thus we have the following theorem.

THEOREM 5.1. *For a* GP(2)*-matrix A, there exists a proper block-triangular matrix $\widetilde{A}$ with* GP*-irreducible diagonal blocks which is obtained from A by an admissible transformation (of the form* (1)*).*

5.2. Uniqueness of the decomposition. In this section, we discuss the uniqueness of the GP-irreducible diagonal blocks in the decomposition of a GP(2)-matrix.

Example 2. Consider an 8×8 GP(2)-matrix

$$A = \left(\begin{array}{cc|cc|cc|cc} t_{11} & & & & & & & t_{14} \\ & 2t_{11} & & & & & t_{14} & \\ \hline & & t_{22} & & t_{23} & -2t_{23} & & \\ & & & t_{22} & & t_{23} & & \\ \hline & & t_{32} & & t_{33} & -2t_{33} & & \\ & & & t_{32} & & t_{33} & & \\ \hline & t_{41} & t_{42} & 2t_{42} & t_{43} & & & t_{44} \\ t_{41} & & 2t_{42} & 4t_{42} & t_{43} & & t_{44} & \end{array}\right)$$

with the base field $\mathbf{Q}$. Let S and T be nonsingular matrices:

$$S = \left(\begin{array}{cc|cc|cc} 1 & 0 & & & & \\ 0 & 1 & & & & \\ \hline & & 1 & 0 & & \\ & & 0 & 1 & & \\ \hline & & & & 1 & 0 \\ & & & & 0 & 1 \\ \hline & & & & & \end{array}\right), \quad T = \left(\begin{array}{cc|cc|cc} 1 & 0 & & & & \\ 0 & 1 & & & & \\ \hline & & 1 & 0 & & \\ & & 0 & 1 & & \\ \hline & & & & 1 & 2 \\ & & & & 0 & 1 \\ \hline & & & & & \end{array}\right).$$

Then we have

$$\widetilde{A} = S^{-1}AT = \left(\begin{array}{cc|cc|cc} t_{11} & & & & & t_{14} \\ & 2t_{11} & & & & & t_{14} \\ \hline & & t_{22} & & t_{23} & \\ & & & t_{22} & & t_{23} & \\ \hline & & t_{32} & & t_{33} & \\ & & & t_{32} & & t_{33} & \\ \hline t_{41} & & 2t_{42} & 4t_{42} & t_{43} & 2t_{43} & & t_{44} \\ & t_{41} & t_{42} & 2t_{42} & t_{43} & 2t_{43} & t_{44} & \end{array}\right).$$

By using permutation matrices

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}, \quad Q = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix},$$

we obtain an explicitly upper block-triangular matrix:

$$\bar{A} = P\widetilde{A}Q = \begin{pmatrix} t_{11} & & t_{14} & & & & & \\ & 2t_{11} & & t_{14} & & & & \\ t_{41} & & & t_{44} & 2t_{42} & t_{43} & 4t_{42} & 2t_{43} \\ & t_{41} & t_{44} & & t_{42} & t_{43} & 2t_{42} & 2t_{43} \\ & & & & t_{22} & t_{23} & & \\ & & & & t_{32} & t_{33} & & \\ & & & & & & t_{22} & t_{23} \\ & & & & & & t_{32} & t_{33} \end{pmatrix}.$$

Thus $\widetilde{A}$ is a GP-irreducible decomposition of A with empty tails and square diagonal blocks

$$\widetilde{A}[R_1, C_1] = \begin{pmatrix} t_{11} & & t_{14} & \\ & 2t_{11} & & t_{14} \\ t_{41} & & & t_{44} \\ & t_{41} & t_{44} & \end{pmatrix},$$

$$\widetilde{A}[R_2, C_2] = \begin{pmatrix} t_{22} & t_{23} \\ t_{32} & t_{33} \end{pmatrix} \text{ and } \widetilde{A}[R_3, C_3] = \begin{pmatrix} t_{22} & t_{23} \\ t_{32} & t_{33} \end{pmatrix}.$$

On the other hand, let S' and T' be nonsingular matrices:

$$S' = \left(\begin{array}{cc|cc|cc|cc} 1 & 0 & & & & & & \\ 0 & 1 & & & & & & \\ \hline & & 1 & -2 & & & & \\ & & 0 & 1 & & & & \\ \hline & & & & 1 & -2 & & \\ & & & & 0 & 1 & & \\ \hline & & & & & & 1 & 0 \\ & & & & & & 0 & 1 \end{array}\right), \quad T' = \left(\begin{array}{cc|cc|cc|cc} 1 & 0 & & & & & & \\ 0 & 1 & & & & & & \\ \hline & & 1 & -2 & & & & \\ & & 0 & 1 & & & & \\ \hline & & & & 1 & 0 & & \\ & & & & 0 & 1 & & \\ \hline & & & & & & 1 & 0 \\ & & & & & & 0 & 1 \end{array}\right).$$

Then we have

$$\widetilde{A}' = S'^{-1}AT' = \left(\begin{array}{cc|cc|cc|cc} t_{11} & & & & & & & t_{14} \\ & 2t_{11} & & & & & t_{14} & \\ \hline & & t_{22} & & t_{23} & & & \\ & & & t_{22} & & t_{23} & & \\ \hline & & t_{32} & & t_{33} & & & \\ & & & t_{32} & & t_{33} & & \\ \hline & t_{41} & t_{42} & & t_{43} & & & t_{44} \\ t_{41} & & 2t_{42} & & t_{43} & & t_{44} & \end{array}\right).$$

By using permutation matrices

$$P' = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}, \quad Q' = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix},$$

we obtain an explicitly upper block-triangular matrix:

$$\bar{A}' = P'\widetilde{A}'Q' = \begin{pmatrix} t_{22} & t_{23} & & & & & & \\ t_{32} & t_{33} & & & & & & \\ & & t_{11} & & & t_{14} & & \\ & & & 2t_{11} & t_{14} & & & \\ & & & t_{41} & & t_{44} & t_{42} & t_{43} \\ & & t_{41} & & t_{44} & & 2t_{42} & t_{43} \\ & & & & & & t_{22} & t_{23} \\ & & & & & & t_{32} & t_{33} \end{pmatrix}.$$

Thus $\widetilde{A}'$ is a GP-irreducible decomposition of A with empty tails and square diagonal blocks

$$\widetilde{A}'[R'_1, C'_1] = \begin{pmatrix} t_{22} & t_{23} \\ t_{32} & t_{33} \end{pmatrix},$$

$$\widetilde{A}'[R'_2, C'_2] = \begin{pmatrix} t_{11} & & & t_{14} \\ & 2t_{11} & t_{14} & \\ & t_{41} & & t_{44} \\ t_{41} & & t_{44} & \end{pmatrix} \text{ and } \widetilde{A}'[R'_3, C'_3] = \begin{pmatrix} t_{22} & t_{23} \\ t_{32} & t_{33} \end{pmatrix}.$$

It is easy to see that we have a permutation σ of $\{1,2,3\}$ such that $\widetilde{A}[R_k, C_k]$ and $\widetilde{A}'[R'_{\sigma(k)}, C'_{\sigma(k)}]$ are connected by an admissible transformation for each $k = 1,2,3$. In the following we prove that such is always the case.

The following argument is not directly affected by the type of GP-matrices and is valid for all GP-matrices for which the inequality (4) holds.

Let $\boldsymbol{F}$ be a field in general and A be a matrix over $\boldsymbol{F}$ of the form

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1\nu} \\ A_{21} & A_{22} & \cdots & A_{2\nu} \\ \vdots & \vdots & \ddots & \vdots \\ A_{\mu 1} & A_{\mu 2} & \cdots & A_{\mu\nu} \end{pmatrix}.$$

Then A is said to be a partitioned matrix. An equivalence transformation of the form (s1) is called a partition-respecting equivalence transformation (or PE-transformation), where the matrices S and T are matrices over the field $\boldsymbol{F}$ in this case. It has been shown by a module-theoretic argument that, provided a PE-irreducible decomposition exists, the set of PE-irreducible components of a partitioned matrix is unique to within PE-transformations of each component [7].

A GP(2)-matrix is a special kind of partitioned matrix over the field $\boldsymbol{F} = \boldsymbol{K}(\mathcal{T})$. Admissible transformations for a GP(2)-matrix, with S and T being matrices over $\boldsymbol{K}$, are more restricted than PE-transformations, which employ transformation matrices over $\boldsymbol{F}$. Hence a GP-irreducible decomposition can possibly be coarser than a PE-irreducible decomposition. However, a GP-irreducible decomposition has the same number of components as a PE-irreducible decomposition if the field $\boldsymbol{K}$ is infinite, as follows.

LEMMA 5.2. *Suppose $\boldsymbol{K}$ is an infinite field and let A be a* GP(2)*-matrix with base field $\boldsymbol{K}$. Then a* GP*-irreducible decomposition of A is a* PE*-irreducible decomposition of A over the field $\boldsymbol{K}(\mathcal{T})$.*

Proof. Let A be a GP(2)-matrix and $\widehat{A} = \widehat{S}^{-1}A\widehat{T}$ be its PE-irreducible decomposition with $\widehat{S}$ and $\widehat{T}$ over $\boldsymbol{K}(\mathcal{T})$. Since $\boldsymbol{K}$ is an infinite field, we may choose a suitable set of parameter values from $\boldsymbol{K}$ such that $\widehat{S}$ and $\widehat{T}$ remain nonsingular when those values are substituted for the parameters in $\mathcal{T}$. Let S and T be the matrices obtained from $\widehat{S}$ and $\widehat{T}$ by this substitution. Put $\widetilde{A} = S^{-1}AT$, and then (i,j)-entry of $\widetilde{A}$ remains zero if (i,j)-entry of $\widehat{A}$ is zero. Therefore $\widetilde{A}$ is in a block-triangular form with the same number of blocks as $\widehat{A}$ has. Furthermore this is a proper block-triangularization since $\operatorname{rank}\widetilde{A} = \operatorname{rank}\widehat{A}$. Since a GP-irreducible decomposition can not be finer than a PE-irreducible decomposition, we conclude that $\widetilde{A}$ is a GP-irreducible decomposition of A. □

Before establishing the uniqueness of a GP-irreducible decomposition, we claim the following lemma.

LEMMA 5.3. *Suppose $\boldsymbol{K}$ is an infinite field and let A and $\widehat{A}$ be* GP(2)*-matrices with base field $\boldsymbol{K}$ connected by a* PE*-transformation over $\boldsymbol{K}(\mathcal{T})$. Then A and $\widehat{A}$ are connected by an admissible transformation (over $\boldsymbol{K}$) for* GP(2)*-matrices.*

Proof. Suppose $\widehat{A} = \widehat{S}^{-1}A\widehat{T}$, where $\widehat{S} = \bigoplus_{\alpha=1}^{\mu}\widehat{S}_\alpha$ and $\widehat{T} = \bigoplus_{\beta=1}^{\nu}\widehat{T}_\beta$ are over $\boldsymbol{K}(\mathcal{T})$. Note that for any α and β we have

$$\widehat{A}^{\natural}_{\alpha\beta} = \widehat{S}_\alpha^{-1}A^{\natural}_{\alpha\beta}\widehat{T}_\beta,$$

where $\widehat{A}^{\natural}_{\alpha\beta}$ and $A^{\natural}_{\alpha\beta}$ are matrices over $\boldsymbol{K}$. Let $S = \bigoplus_{\alpha=1}^{\mu}S_\alpha$ and $T = \bigoplus_{\beta=1}^{\nu}T_\beta$ be nonsingular matrices over $\boldsymbol{K}$ obtained from $\widehat{S}$ and $\widehat{T}$, respectively, by substituting

suitable values to the independent parameters. Such parameter values exist since $\boldsymbol{K}$ is an infinite field. Then clearly we have

$$\widehat{A}^{\natural}_{\alpha\beta} = S_{\alpha}^{-1} A^{\natural}_{\alpha\beta} T_{\beta}$$

for any α and β. Hence $\widehat{A} = S^{-1}AT$. □

Suppose we have two GP-irreducible decompositions $\widetilde{A}$ and $\widetilde{A}'$ of A (cf. Example 2). Both of them are also PE-irreducible decompositions by Lemma 5.2. We denote by D_k and D'_k the kth irreducible component of $\widetilde{A}$ and $\widetilde{A}'$, respectively, for $k = 0, 1, \ldots, b, \infty$. Then, by Theorem 3.7 of [7], there exists a certain permutation σ of $\{1, \ldots, b\}$ such that D_k is PE-equivalent to $D'_{\sigma(k)}$. It follows from Lemma 5.3 that D_k and $D'_{\sigma(k)}$ are connected by an admissible transformation. Thus we have the following theorem.

THEOREM 5.4. *The set of* GP*-irreducible components of a* GP(2)*-matrix is unique to within admissible transformations of each component.*

The DM-decomposition of a generic matrix is uniquely determined with respect not only to irreducible components but also to the partial order defined by the zero/nonzero pattern. However, for a GP(2)-matrix, Theorem 5.4 shows the former uniqueness only. The latter uniqueness does not hold in this case as is seen in Example 2 above. In fact, $\bar{A}$ has the partial order

$$\begin{array}{ccc} \{C_2\} & & \{C_3\} \\ & \searrow\!\!\swarrow & \\ & \{C_1\} & \end{array}$$

whereas $\bar{A}'$ has

$$\begin{array}{ccc} & & \{C'_3\} \\ \{C'_1\} & & | \\ & & \{C'_2\} \end{array}.$$

6. Discussions.

6.1. Summary on generic partitioned matrices. From the previous works [7], [10], it has been known that the rank identity (4) holds for the following types of generic partitioned matrices.

Generic matrix: $m_\alpha = 1$ for $\alpha = 1, \ldots, \mu$ and $n_\beta = 1$ for $\beta = 1, \ldots, \nu$.

Generic multilayered matrix: $n_\beta = 1$ for $\beta = 1, \ldots, \nu$.

Transposed generic multilayered matrix: $m_\alpha = 1$ for $\alpha = 1, \ldots, \mu$.

Our present result (Theorem 4.5) added another type.

GP(2)-matrix: $m_\alpha \leq 2$ for $\alpha = 1, \ldots, \mu$ and $n_\beta \leq 2$ for $\beta = 1, \ldots, \nu$.

Hence, it seems natural to expect that Theorem 4.5 is valid not only for GP(2)-matrices but also for generic partitioned matrices in general. However, this is not the case, as we see in the following counterexample.

Example 3. Consider a 6×6 generic partitioned matrix of type $(3, 3; 2, 2, 2)$:

$$A = \left(\begin{array}{cc|cc|cc} t_{11} & & t_{12} & & t_{13} & \\ & t_{11} & & t_{12} & & t_{13} \\ \hline & & t_{22} & & t_{23} & \\ t_{21} & & & & & t_{23} \\ & t_{21} & & t_{22} & & \end{array}\right).$$

It can be easily verified that $\operatorname{rank} A = 5 < \min p(W) + n = 6$.

6.2. Remark on the base field. As long as we are concerned with the DM-decomposition of generic matrices, it does not matter what the base field is. However, when it comes to the GP(2)-matrices, it depends on the base field how fine a given matrix can be decomposed. Let us see such a situation in the following example.

Example 4. Consider a 4×4 GP(2)-matrix:

$$A = \left(\begin{array}{cc|cc} t_{11} & & & t_{12} \\ & 2\,t_{11} & t_{12} & \\ \hline & t_{21} & & t_{22} \\ t_{21} & & t_{22} & \end{array}\right).$$

Regarding A as a GP(2)-matrix with the base field $\mathbf{Q}$, we can easily verify that A is GP-irreducible. If A is with the base field $\mathbf{R}$, on the other hand, we have the following block-triangularization of A:

$$\widetilde{A} = S^{-1}AT = \left(\begin{array}{cc|cc} t_{11} & & -t_{12} & \\ & t_{11} & & t_{12} \\ \hline t_{21} & & \sqrt{2}\,t_{22} & \\ & t_{21} & & \sqrt{2}\,t_{22} \end{array}\right)$$

with

$$S = \left(\begin{array}{cc|cc} \sqrt{2} & \sqrt{2} & & \\ -2 & 2 & & \\ \hline & & -1 & 1 \\ & & \sqrt{2} & \sqrt{2} \end{array}\right), \quad T = \left(\begin{array}{cc|cc} \sqrt{2} & \sqrt{2} & & \\ -1 & 1 & & \\ \hline & & 2 & 2 \\ & & -\sqrt{2} & \sqrt{2} \end{array}\right).$$

By using permutation matrices

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad Q = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

we obtain an explicitly upper block-triangular (block-diagonal) matrix:

$$\bar{A} = P\widetilde{A}Q = \begin{pmatrix} t_{11} & -t_{12} & & \\ t_{21} & \sqrt{2}\,t_{22} & & \\ & & t_{11} & t_{12} \\ & & t_{21} & \sqrt{2}\,t_{22} \end{pmatrix}.$$

Thus $\widetilde{A}$ is in a block-triangular (block-diagonal) form with two square blocks and empty tails.

7. Conclusion. We have introduced the notion of generic partitioned matrices and proved a Hall–Ore type minimax theorem on the rank of a subclass of generic partitioned matrices, called GP(2)-matrices. This minimax relation provides a decomposition (or proper block-triangularization) of GP(2)-matrices which is an extension of the DM-decomposition of generic matrices.

REFERENCES

[1] M. AIGNER, *Combinatorial Theory*, Springer-Verlag, Berlin, 1979.
[2] R. B. BAPAT, *König's theorem and bimatroids*, Linear Algebra Appl., 212/213 (1994), pp. 353–365.
[3] R. A. BRUALDI AND H. J. RYSER, *Combinatorial Matrix Theory*, Cambridge University Press, London, 1991.
[4] A. L. DULMAGE AND N. S. MENDELSOHN, *A structure theory of bipartite graphs of finite exterior dimension*, Trans. Roy. Soc. Canada, 53 (1959), pp. 1–13.
[5] D. J. HARTFIEL AND R. LOEWY, *A determinantal version of the Frobenius–König theorem*, Linear Multilinear Algebra, 16 (1984), pp. 155–165.
[6] M. IRI, *Structural theory for the combinatorial systems characterized by submodular functions*, Progress in Combinatorial Optimization, W. R. Pulleyblank, ed., Academic Press, New York, 1984, pp. 197–219.
[7] H. ITO, S. IWATA, AND K. MUROTA, *Block-triangularizations of partitioned matrices under similarity/equivalence transformations*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1226–1255.
[8] L. LOVÁSZ AND M. PLUMMER, *Matching Theory*, North-Holland, Amsterdam, 1986.
[9] L. MIRSKY, *Transversal Theory*, Academic Press, New York, 1977.
[10] K. MUROTA, *Systems Analysis by Graphs and Matroids — Structural Solvability and Controllability*, Springer-Verlag, Berlin, 1987.
[11] ———, *Mixed matrices — Irreducibility and decomposition*, Combinatorial and Graph-Theoretical Problems in Linear Algebra, R. A. Brualdi, S. Friedland, and V. Klee, eds., The IMA Volumes in Mathematics and Its Applications, Vol. 50, Springer-Verlag, Berlin, 1993, pp. 39–71.
[12] K. MUROTA, M. IRI, AND M. NAKAMURA, *Combinatorial canonical form of layered mixed matrices and its application to block-triangularization of systems of equations*, SIAM J. Algebraic Discrete Methods, 8 (1987), pp. 123–149.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 735–742, July 1995

003

ON A SPECIAL CLASS OF GENERALIZED DOUBLY STOCHASTIC MATRICES AND ITS RELATION TO BÉZIER POLYGONS*

MIROSLAV FIEDLER†

Abstract. We introduce and study a new set $\mathcal{B}$ of $m \times n$ nonnegative integral matrices B_{mn}, $m \geq n \geq 1$. They have column-rhomboidal form and are closely related to the confluent Vandermonde matrices V_{mn} with n-tuple node 1 and to Bézier polygons.

Key words. doubly stochastic matrices, Vandermonde matrix, Hadamard product, combinatorial identity, interpolation, Bézier curve

AMS subject classifications. 15A36, 65D10, 05A19

1. Introduction.

In the present note we intend to investigate the set $\mathcal{B}$ of nonnegative integral matrices B_{mn}, m, n being positive integers, $m \geq n$; the matrix B_{mn} is an $m \times n$ matrix with entries

$$(1) \qquad (B_{mn})_{ik} = \binom{i}{k}\binom{m-1-i}{n-1-k}, \quad i = 0, 1, \ldots, m-1, \quad k = 0, 1, \ldots, n-1.$$

Example 1.1. We have

$$B_{53} = \begin{pmatrix} 6 & 0 & 0 \\ 3 & 3 & 0 \\ 1 & 4 & 1 \\ 0 & 3 & 3 \\ 0 & 0 & 6 \end{pmatrix}, \quad B_{54} = \begin{pmatrix} 4 & 0 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 0 & 2 & 2 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 4 \end{pmatrix}.$$

It is immediate that for $m = n$, B_{mn} is the identity matrix of order n; for $m > n$, it is *centrosymmetric* and has a *column-rhomboidal form.* Here, we say that an $m \times n$ matrix $A = (a_{ik})$ is in column-rhomboidal form if its transpose is in the row-rhomboidal form in the sense of [4], i.e., if $m > n$, $a_{ij} = 0$ for $1 \leq i < j \leq n$ as well as for $1 \leq j < i+n-m \leq n$, and all entries $a_{ii}, i = 1, \ldots, n$ and $a_{m-j,n-j}, j = 0, \ldots, n-1$ are different from zero.

We shall investigate properties of $\mathcal{B}$, show connections with the *confluent Vandermonde matrix* $V_{mn}(1)$ with entries

$$(2) \qquad (V_{mn}(1))_{ik} = \left(\!\binom{i}{k}\!\right), \quad i = 0, 1, \ldots, m-1, \quad k = 0, 1, \ldots, n-1,$$

and also with the *Bézier polygons.*

For simplicity, we shall write V_{mn} instead of $V_{mn}(1)$. We also denote by J_t the *flip matrix* of order t, i.e., the $t \times t$ matrix

$$(3) \qquad J_t = \begin{pmatrix} 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & \ldots & 1 & 0 \\ . & . & \ldots & . & . \\ 1 & 0 & \ldots & 0 & 0 \end{pmatrix}.$$

* Received by the editors November 23, 1993; accepted for publication (in revised form) by R. A. Horn April 20, 1994.

† Academy of Sciences of the Czech Republic, Institute of Computer Science, Pod vodárenskou věží 2, 182 07 Praha 8, Czech Republic (hojek@csearn.bitnet). This research was supported by grant GAAVCR-130407.

Using this notation, we can express the centrosymmetry of B_{mn} by

$$B_{mn} = J_m B_{mn} J_n. \tag{4}$$

As usual, we denote by $\circ$ the *Hadamard product* of matrices, the binary operation of entrywise multiplication of matrices having the same dimensions.

2. Properties of $\mathcal{B}$.

THEOREM 2.1. *If $B_{mn} \in \mathcal{B}$, if V_{mn} is the Vandermonde confluent matrix in* (2), *and J_m, J_n the flip matrices from* (3), *then*

$$B_{mn} = V_{mn} \circ J_m V_{mn} J_n.$$

Proof. The proof follows immediately from (1) and the fact that $(J_m V_{mn} J_n)_{ik} = \binom{m-1-i}{n-1-k}$. □

The main property of $\mathcal{B}$ is the following theorem.

THEOREM 2.2. *Let m, n, s be integers, $m \geq s \geq n \geq 1$. Then,*

$$B_{mn} = \frac{1}{\binom{m-n}{s-n}} B_{ms} B_{sn}. \tag{5}$$

Also,

$$B_{mn} = \frac{1}{(m-n)!} B_{m,m-1} B_{m-1,m-2} \cdots B_{n+1,n}. \tag{6}$$

Proof. It suffices to prove (5) for $s = n+1$ and use induction. To prove

$$B_{mn} = \tfrac{1}{m-n} B_{m,n+1} B_{n+1,n} \tag{7}$$

for $m > n \geq 1$, we shall use the formula

$$\binom{p}{q} = \frac{p-q+1}{q} \binom{p}{q-1} \quad \text{for} \quad p \geq q \geq 1 \tag{8}$$

and the fact that

$$(B_{n+1,n})_{jj} = n-j,\ (B_{n+1,n})_{j+1,j} = j+1 \quad \text{for} \quad j = 0, \ldots, n-1, \tag{9}$$

and zero otherwise. The (i,k)-entry of the product $B_{m,n+1} B_{n+1,n}$ in (7) is

$$(B_{m,n+1})_{ik}(n-k) + (B_{m,n+1})_{i,k+1}(k+1)$$

$$= \binom{i}{k} \binom{m-1-i}{n-k} (n-k) + \binom{i}{k+1} \binom{m-1-i}{n-k-1} (k+1)$$

$$= \binom{i}{k} \binom{m-1-i}{n-k-1} \left[\frac{m-1-i-n+k+1}{n-k} (n-k) + \frac{i-k}{k+1} (k+1) \right]$$

which is easily seen to be $(m-n)(B_{mn})_{ik}$. □

THEOREM 2.3. *Every matrix in $\mathcal{B}$ is generalized doubly stochastic, i.e., all the row-sums are mutually equal as well as all the column-sums are mutually equal. In B_{mn}, the row sums are equal to $\binom{m-1}{n-1}$, the column-sums are equal to $\binom{m}{n}$.*

Proof. Follows from (6) by postmultiplication (resp. premultiplication) by the vector e_n (resp. e_m^T) with all n (resp. m) coordinates equal to one and by an easy induction since the formulae are true for $m = n+1$. □

Remark. The row-sums condition is the *convolution identity* of *Vandermonde* [3, p. 154].

The following basic theorem shows that the column-spaces of B_{mn} and V_{mn} from (2) coincide.

THEOREM 2.4. *Let m, n be integers, $m \geq n \geq 1$. Then*

$$B_{mn} = V_{mn} D_{mn} (V_{nn})^{-1}, \tag{10}$$

where D_{mn} is the diagonal matrix of order n

$$D_{mn} = \operatorname{diag}\left(\binom{m-k-1}{m-n}\right), \qquad k = 0, \ldots, n-1. \tag{11}$$

Remark. The matrix V_{nn}^{-1} is the Vandermonde confluent matrix with n-fold node -1, i.e., the matrix $\left((-1)^{i+k}\binom{i}{k}\right)$.

Proof. Let us show first that

$$B_{n+1,n} = V_{n+1,n} D_{n+1,n} V_{nn}^{-1}.$$

Indeed, by (2) and (9)

$$(B_{n+1,n} V_{nn})_{ik} = (n-i)\binom{i}{k} + i\binom{i-1}{k},$$

which can be written by (8) as $(n-k)\binom{i}{k}$.

$(V_{n+1,n} D_{n+1,n})_{ik}$ yields the same result.

By (7), it suffices for induction to show that

$$V_{mn} D_{mn} V_{nn}^{-1} = \tfrac{1}{m-n} V_{m,n+1} D_{m,n+1} V_{n+1,n+1}^{-1} V_{n+1,n} D_{n+1,n} V_{nn}^{-1},$$

or equivalently, that

$$V_{mn} D_{mn} = \tfrac{1}{m-n} V_{m,n+1} D_{m,n+1} V_{n+1,n+1}^{-1} V_{n+1,n} D_{n+1,n}. \tag{12}$$

Since $V_{n+1,n}$ is obtained from $V_{n+1,n+1}$ by removing the last column, the matrix $V_{n+1,n+1}^{-1} V_{n+1,n}$ is the identity matrix of order $n+1$ with the last column removed. The product $D_{m,n+1} V_{n+1,n+1}^{-1} V_{n+1,n} D_{n+1,n}$ is thus the diagonal matrix

$$\operatorname{diag}\left((n-k)\binom{m-k-1}{m-n-1}\right), \qquad k = 0, \ldots, n,$$

with the last column removed.

It is then easily checked that both sides of (12) are equal, their (i,k) entry being $\binom{i}{k}\binom{m-k-1}{m-n}$. □

From this theorem, another characteristic property of the matrix B_{mn} for $m > n$ follows.

THEOREM 2.5. *Let m, n be integers, $m > n \geq 1$. Then B_{mn} is the unique matrix of the form $V_{mn}X$ where X is square nonsingular, which has column-rhomboidal form and all column-sums equal to* $\binom{m}{n}$.

Proof. The matrix B_{mn} satisfies the mentioned property. Suppose that C_{mn} also has the property. Theorem 2.1 of [4] asserts that if two matrices, both in the row-rhomboidal form, are left multiples of an $m \times n$ matrix then they differ by premultiplication by a nonsingular diagonal matrix. Applying the transpose of this theorem, we get that C_{mn} is obtained from B_{mn} by postmultiplication by a nonsingular diagonal matrix. Comparing the column sums, equality follows. □

Another consequence of Theorem 2.4 is a property of V_{mn}.

THEOREM 2.6. *Let m, n be integers, $m \geq n \geq 1$. Then, if J_m is the flip matrix from* (3), *there exists a unique square matrix Z_{mn} such that*

$$J_m V_{mn} = V_{mn} Z_{mn}. \tag{13}$$

This matrix Z_{mn} is

$$Z_{mn} = D_{mn} V_{nn}^{-1} J_n V_{nn} D_{mn}^{-1}, \tag{14}$$

where D_{mn} is defined in (11).

Proof. If we substitute from (10) into (4), we obtain

$$J_m V_{mn} D_{mn} V_{nn}^{-1} J_n = V_{mn} D_{mn} V_{nn}^{-1}.$$

This implies the result. Uniqueness follows from the fact that V_{mn} has full column rank. □

THEOREM 2.7. *Let m, n be integers satisfying $m > n \geq 1$. The matrix B_{mn} is the unique $m \times n$ matrix which has column-rhomboidal form, all column-sums equal to* $\binom{m}{n}$, *and satisfies*

$$F_{mn} B_{mn} = 0, \tag{15}$$

where for

$$f_k = (-1)^{n-k} \binom{n}{k}, \quad k = 0, \ldots, n, \tag{16}$$

F_{mn} is the $(m-n) \times m$ matrix

$$F_{mn} = \begin{pmatrix} f_0 & f_1 & \ldots & f_n & 0 & \ldots & 0 \\ 0 & f_0 & \ldots & \cdot & f_n & \ldots & 0 \\ \cdot & \cdot & \ldots & \cdot & \cdot & \ldots & \cdot \\ 0 & 0 & \ldots & f_0 & f_1 & \ldots & f_n \end{pmatrix}. \tag{17}$$

Proof. The matrix V_{mn} from (2) satisfies $F_{mn} V_{mn} = 0$ since f_k are the coefficients in the polynomial $(x-1)^n$. Since V_{mn} has full column rank, every $m \times n$ matrix W_{mn}, which has full column rank and satisfies $F_{mn} W_{mn} = 0$ is of the form $V_{mn}X$ where X is nonsingular.

By (10), B_{mn} satisfies (15), and also the remaining conditions. By Theorem 2.5, it is the unique matrix with these properties. □

THEOREM 2.8. *Let $m > n \geq 1$. Then the matrix B_{mn} is the unique $m \times n$ matrix $C = (c_{ij})$ which satisfies the following four properties:*

(i) $c_{ij} = 0$ *for* $1 \leq i < j \leq n$;

(ii) $c_{ij} = 0$ *for* $1 \leq j < i + n - m \leq n$;

(iii) $F_{mn}C = 0$ *for* F_{mn} *defined in* (17) *and* (16);

(iv) *the first n row-sums of C are equal to* $\binom{m-1}{n-1}$.

Proof. The matrix B_{mn} clearly satisfies (i) – (iv). Let now $C = (c_{ij})$ satisfy (i) – (iv). By (iii), $C = B_{mn}X$ for some $n \times n$ matrix $X = (x_{ij})$ since B_{mn} has a full column rank. By (i), $x_{ij} = 0$ for $1 \leq i < j \leq n$. By (ii), $x_{ij} = 0$ for $1 \leq j < i \leq n$. Thus X is diagonal. By (iv), it follows easily that X is the identity matrix and $C = B_{mn}$. □

We conclude this section by finding the formula for the maximal singular value of B_{mn}.

THEOREM 2.9. *The maximal singular value σ_1 of $B_{mn} \in \mathcal{B}$ is*

$$\sigma_1(B_{mn}) = \sqrt{\binom{m}{n}\binom{m-1}{n-1}}.$$

Proof. Since B_{mn} is generalized doubly stochastic by Theorem 2.3, the vector e_m with all coordinates one is the Perron eigenvector of the nonnegative matrix $B_{mn}B_{mn}^T$ with eigenvalue $\binom{m}{n}\binom{m-1}{n-1}$. This is the square of the maximal singular value of B_{mn}. □

3. Application to combinatorial identities. We now present a few combinatorial identities that follow from the formulae in the previous section and which seem to be new.

THEOREM 3.1. *Let m, n, s, i, k be integers satisfying $m \geq s \geq n \geq 1$, $m - n \geq i - k \geq 0$, $n - 1 \geq k$. Then,*

$$\sum_{j=0}^{s-1} \binom{i}{j}\binom{j}{k}\binom{m-1-i}{s-1-j}\binom{s-1-j}{n-1-k} = \binom{i}{k}\binom{m-1-i}{n-1-k}\binom{m-n}{s-n}.$$

In particular, for $s = n + 1$,

$$\sum_{j=0}^{n} \binom{i}{j}\binom{j}{k}\binom{m-1-i}{n-j}\binom{n-j}{n-1-k} = (m-n)\binom{i}{k}\binom{m-1-i}{n-1-k}.$$

Proof. Follows immediately from the formula for the (i, k)-entry of both sides of (5). □

THEOREM 3.2. *Let m, n, i, k be integers satisfying $m \geq n \geq 1$, $0 \leq i \leq n-1$, $0 \leq k \leq n - 1$. Then,*

$$\sum_{j=0}^{n-1} (-1)^{j+k} \binom{i}{j}\binom{j}{k}\binom{m-1-j}{m-n} = \binom{i}{k}\binom{m-1-i}{n-1-k},$$

and also

$$\sum_{j=0}^{n-1} \binom{i}{j}\binom{j}{k}\binom{m-1-i}{n-1-j} = \binom{i}{k}\binom{m-1-k}{m-n}.$$

Proof. The first identity follows from the matrix equality (10) by considering the (i,k)-entry, the second identity in a similar way from the same equality postmultiplied by V_{nn}. □

THEOREM 3.3. *Let m,n be integers satisfying $m > n \geq 1$. Then for $i = 0,1,\ldots,n-1$ and $k = 0,1,\ldots,n-1$ the following identities hold:*

$$\sum_{j=0}^{m-1}(-1)^{i+j}\binom{n}{j-i+1}\binom{j}{k}\binom{m-1-j}{n-1-k} = 0.$$

Proof. The proof follows from explicitly expressing the (i,k)-entry of (15) using (17) and (16). □

As a final remark of this section let us mention the following. Since the matrix B_{mn} can be completed (compare Cor. 2.6 of [4]) by an $m \times (m-n)$ matrix P_{mn} and the matrix F_{mn} by an $n \times m$ matrix Q_{mn} in such a way that the matrices $\begin{pmatrix} Q_{mn} \\ F_{mn} \end{pmatrix}$ and $(B_{mn}\ P_{mn})$ are inverse to each other, the determinants formed by any n rows of B_{mn} and the determinants formed by the complementary $m-n$ columns of F_{mn} to the previous rows are proportional up to a factor of $(-1)^s$ where s is the sum of indices of the chosen rows in B_{mn}. The constant of proportionality can easily be found from some particular choice of the rows. This observation allows evaluation of the mentioned determinants in B_{mn} by using the simpler determinants of F_{mn}.

4. Bézier polygons and Bézier curves. Let $(A_0, A_1, \ldots, A_{n-1})$ be an ordered n-tuple of points in a Euclidean (or, affine) point space $\mathcal{E}_s$. We shall associate with it a polygon P_n consisting of the n points $A_0,\ldots,A_{n-1}$ and of the $n-1$ segments A_0A_1, $A_1A_2,\ldots,A_{n-2}A_n$ and denote it in the same way. The dimension s of $\mathcal{E}_s$ is irrelevant and all points considered in the sequel are in the linear hull of the given points.

We call *Bézier simple refinement* of P_n the polygon

$$P_{n+1} = (A_0^{(1)}, A_1^{(1)}, \ldots, A_n^{(1)}),$$

where

$$A_0^{(1)} = A_0, \quad A_n^{(1)} = A_{n-1}, \quad A_k^{(1)} = \frac{k}{n}A_{k-1} + \frac{n-k}{n}A_k, \quad k = 1,\ldots,n-1. \tag{18}$$

We write $P_{n+1} = R(P_n)$ and allow continuation to the *second refinement* $P_{n+2} = R(P_{n+1})$, etc.

This recurrent process has been studied ([1], [5]) and it is well known that the continuation to infinity leads to a parametric curve P_∞, called *Bézier curve* [2, p. 120].

We intend to show the connection of the matrices B_{mn} with the mentioned process. In particular, we find explicit (not just recurrent) formulae for the rth refined polygon and introduce an (apparently new) family of Bézier curves.

THEOREM 4.1. *Let $n \geq 2$, let $P_n = (A_0, A_1, \ldots, A_{n-1})$ be a polygon in $\mathcal{E}_s$. Then the rth Bézier refinement P_{n+r} is obtained as follows:*

The kth point $A_k^{(r)}$ is the linear combination of the points $A_0,\ldots,A_{n-1}$ whose coefficients are the entries of the $(k+1)$st row of $B_{r+n,n}$ divided by $\binom{n+r-1}{n-1}$.

Remark. Observe that the sum of the *parametric barycentric* coordinates of $A_k^{(r)}$ is one by Theorem 2.3.

Proof. By (18) and (9), the result is correct for $r = 1$. The proof is then completed by induction using (6). □

In the next theorem we present a geometric characterization of the rth Bézier refinement P_{n+r} of a polygon P_n which is, in a sense, a counterpart of the algebraic Theorem 2.8. To formulate the result, let us say that an ordered set of $N+1$ points $(B_0, B_1, \dots, B_N)$ in an affine space is in the (*affine*) $(N-1)$-*parabolic position* if the vector (the sum of the coefficients is zero)

$$\sum_{t=0}^{N} (-1)^t \binom{N}{t} B_s$$

is the zero vector.

THEOREM 4.2. *Let $P_n = (A_0, A_1, \dots, A_{n-1})$ be a polygon in $\mathcal{E}_{n-1}$ with linearly independent points, let r be a positive integer. Then the rth Bézier refinement $P_{n+r} = (A_0^{(r)}, A_1^{(r)}, \dots, A_{n+r-1}^{(r)})$ of P_n is the unique polygon in $\mathcal{E}_{n-1}$ with $n+r$ points having the following properties:*

(i) *For $j = 0, \dots, n-2$, the point $A_j^{(r)}$ belongs to the linear hull of the points $A_0, \dots, A_j$.*

(ii) *For $j = 0, \dots, n-2$, the point $A_{n+r-1-j}^{(r)}$ belongs to the linear hull of the points $A_{n-1}, \dots, A_{n-1-j}$.*

(iii) *Any $n+1$ consecutive points in P_{n+r} are in the (affine) $(n-1)$-parabolic position.*

Proof. It is easily seen that the rth Bézier refinement of P_n has all three properties. In particular, the property (iii) follows from (15), (16), and (17).

If, conversely, the properties (i), (ii), and (iii) are satisfied for some polygon P_{n+r} in $\mathcal{E}_{n-1}$ then the $\binom{n+r-1}{n-1}$-multiple of the $(n+r) \times n$ matrix formed by the parametric barycentric coordinates of the points $A_k^{(r)}$ with respect to the points $A_0, \dots, A_{n-1}$ satisfies the properties (i)–(iv) in Theorem 2.8. By Theorem 4.1, P_{n+r} is the rth Bézier refinement P_n. □

Another geometric property of the Bézier refinements is the following theorem.

THEOREM 4.3. *The centroid of the points of every Bézier refinement of a polygon P coincides with the centroid of the polygon P.*

Proof. The proof follows from the doubly stochasticity of B_{mn} and the algebraic equivalence in Theorem 4.1. □

The explicit formula (1) allows us to formulate a continuous version. If we set $i = (m-1)t$, $0 \le t \le 1$, we obtain instead of the $m \times n$ matrix B_{mn} the parametric curve

$$X_m(t) = \frac{1}{\binom{m-1}{n-1}} \sum_{k=0}^{n-1} \binom{(m-1)t}{k} \binom{(m-1)(1-t)}{n-1-k} A_k, \quad t \in [0,1]. \tag{19}$$

We call this curve the $(m-n+1)$*st Bézier curve.* For $m = n$, the first Bézier curve has the equation

$$X_1(t) = \sum_{k=0}^{n-1} \binom{(n-1)t}{k} \binom{(n-1)(1-t)}{n-1-k} A_k, \quad t \in [0,1]. \tag{20}$$

THEOREM 4.4. *Let $P_n = (A_0, \dots, A_{n-1})$ be a polygon in $\mathcal{E}_s$, P_{n+r} its rth Bézier refinement. Then the $(r+1)$st Bézier curve of P_n contains all $n+r$ points of P_{n+r}; their parameters are $t_k = \frac{k}{n+r-1}$, $k = 0, \dots, n+r-1$.*

Proof. The proof is immediate. □

Remark. In the case that the points $A_k = (x_k, y_k)$ are in the plane (x, y) and satisfy $x_k = \frac{k}{n-1}$, the first Bézier curve has parametric equation

$$X_1(t) = (t, y(t)), \quad t \in [0, 1],$$

where $y(t)$ is the interpolation polynomial with nodes (x_k, y_k), $k = 0, \dots, n-1$.

The well-known parametric equation of the (classical) Bézier curve follows now easily for $r \to \infty$.

THEOREM 4.5. *Let $P_n = (A_0, A_1, \dots, A_{n-1})$ be a polygon in $\mathcal{E}_s$. Then the Bézier curve corresponding to the polygon P_n has the parametric equation*

$$X(t) = \sum_{k=0}^{n-1} \binom{n-1}{k} (1-t)^{n-1-k} t^k A_k, \quad t \in [0, 1]. \tag{21}$$

Proof. Using the two formulae

$$\lim_{N \to \infty} \binom{Nu}{j} \Big/ \binom{N}{j} = u^j \quad \text{for} \quad u \neq 0 \quad \text{and nonnegative integer} \quad j$$

and

$$\lim_{N \to \infty} \binom{N}{j} \binom{N}{k} \Big/ \binom{N}{j+k} = \binom{j+k}{j} \text{ for nonnegative integers } j, k,$$

we obtain (21) by an easy limiting procedure from (19). □

REFERENCES

[1] P. J. BARRY AND R. N. GOLDMAN, *De Casteljan-type subdivision is peculiar to Bézier curves*, Computer-Aided Design, 20 (1988), pp. 114–116.

[2] P. BÉZIER, *Numerical Control Mathematics and Applications*, J.Wiley & Sons, London, New York, Sydney, Toronto, 1972.

[3] L. COMTET, *Advanced Combinatorics*, D. Reidel Publ. Co., Dordrecht, Boston, 1974.

[4] M. FIEDLER, *A note on the row-rhomboidal form of a matrix*, Linear Algebra Appl., in print.

[5] L. PIEGL, *Recursive algorithms for the representation of parametric curves and surfaces*, Computer Aided Design, 17 (1985), pp. 225–229.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 743–759, July 1995

004

BOUNDING THE SUBSPACES FROM RANK REVEALING TWO-SIDED ORTHOGONAL DECOMPOSITIONS*

RICARDO. D. FIERRO† AND JAMES R. BUNCH‡

Abstract. The singular value decomposition (SVD) is a widely used computational tool in various applications. However, in some applications the SVD is viewed as computationally demanding or difficult to update. The rank revealing QR (RRQR) decomposition and the recently proposed URV and ULV decompositions are promising alternatives for determining the numerical rank k of an $m \times n$ matrix and approximating its fundamental numerical subspaces whenever $k \approx \min(m, n)$. In this paper we prove a posteriori bounds for assessing the quality of the subspaces obtained by two-sided orthogonal decompositions. In particular, we show that the quality of the subspaces obtained by the URV or ULV algorithm depends on the quality of the condition estimator and not on a gap condition. From our analysis we conclude that these decompositions may be more accurate alternatives to the SVD than the RRQR decomposition. Finally, we implement the algorithms in an adaptive manner, which is particularly useful for applications where the "noise" subspace must be computed, such as in signal processing or total least squares.

Key words. rank revealing, orthogonal decomposition, URV, subspaces, subspace angle, numerical rank

AMS subject classifications. 65F25, 65F05, 65F30

1. Introduction.

The singular value decomposition (SVD) is a widely used computational tool and is the most reliable tool for detecting near rank-deficiency in a matrix [14, p. 246]. It has important applications, for instance, in matrix approximation, subset selection, spectral estimation, direction of arrival estimation, optimization, rank-deficient least squares (LS), total least squares (TLS), etc. [1], [3], [12], [20], [21], [22].

The SVD of A (see [14, §2.3]) is denoted

$$A = U\Sigma V^T \tag{1}$$

where, for $m \geq n$,

$$U = [U_k \; U_0 \; U_\perp], \qquad V = [V_k \; V_0]$$

and

$$\Sigma = \begin{array}{c} \begin{array}{cc} k & n-k \end{array} \\ \begin{bmatrix} \Sigma_k & 0 \\ 0 & \Sigma_0 \\ 0 & 0 \end{bmatrix} \begin{array}{l} k \\ n-k \\ m-n. \end{array} \end{array}$$

Therefore

$$A = U_k \Sigma_k V_k^T + U_0 \Sigma_0 V_0^T. \tag{2}$$

The nonnegative diagonal elements of Σ, denoted σ_i, are the singular values of A and are arranged in decreasing order. The numerical rank of A is k. Also, $\eta \equiv \sigma_{k+1}/\sigma_k$, and the "gap" in the singular values of A is large when $1 - \eta$ is close to 1.

* Received by the editors March 23, 1993; accepted for publication (in revised form) by C. Van Loan April 20, 1994.

† Department of Mathematics, California State University, San Marcos, California 92096 (fierro@thunder.csusm.edu).

‡ Department of Mathematics, University of California, San Diego, La Jolla, California 92093 (jbunch@ucsd.edu).

In particular, the SVD can be used to characterize the solutions or solve the matrix approximation problems in the LS and TLS methods. These methods are used to solve the overdetermined system of linear equations

$$AX = B, \tag{3}$$

where $A \in \Re^{m\times n}$, $B \in \Re^{m\times d}$, and $m \geq n$. Due to various sources of errors, the system usually lacks a solution and the relationship between the columns of A and B must be estimated. To estimate the relationship, many direct methods solve a *nearby* compatible system $CX = D$, and this is a plausible strategy when $\|[A\ B] - [C\ D]\|$ is small. For stability reasons, when A is ill conditioned with numerical rank k, it makes sense to require that C be of rank k. In some applications, due to the potential instability, the LS problem

$$\min_{x\in\Re^n} \|AX - B\|_2 \tag{4}$$

is replaced by the "stabilized" LS problem

$$\min_{x\in\Re^n} \|A_kX - B\|_2, \tag{5}$$

where A_k is the nearest rank-k matrix approximation to A in the 2-norm. This is essentially the same as solving the compatible system $A_kX = B_k$ where B_k is the orthogonal projection of B on the range (column space) of A_k. The TLS approach to (3) requires the minimum norm solution of

$$\hat{A}X = \hat{B}, \tag{6}$$

where $[\hat{A}\ \hat{B}]$ is the nearest rank-k matrix approximation to $[A\ B]$ in the 2-norm. The SVD is a convenient tool in solving the matrix approximation problem associated with these two methods, as well as providing elegant formulas for the solutions. A comprehensive treatment of the full rank TLS problem is given in [27]. The LS and TLS methods can be viewed as orthogonal projection methods, and a sensitivity analysis for the LS and TLS solutions is presented in [11]. The subspace angle is a key factor in the analysis.

For applications where a sequence of LS or TLS problems must be solved, or a "noise" subspace must be adaptively estimated, the SVD is viewed as computationally demanding or difficult to update. Therefore, alternative decompositions have been considered that yield the numerical rank, subspace information, or matrix approximations that are nearly as reliable as the powerful but computationally demanding SVD.

The rank revealing QR (RRQR) decomposition of Chan [4] is a potentially useful alternative to the SVD. In this algorithm the rectangular matrix A is preprocessed by an initial QR factorization, and then condition estimation, careful column pivoting, and plane rotations on the left side are employed to produce a rank revealing decomposition. Some applications of RRQR are discussed in [2], [6], and [17].

G. W. Stewart [24], [25] introduced rank-revealing "two-sided" orthogonal (or complete) decompositions, so-called URV and ULV decompositions, as alternatives. While complete orthogonal decompositions have been around for some time (e.g., see [16]), Stewart's technique is quite promising because it is guaranteed to reveal the numerical rank. In this algorithm the rectangular matrix A is preprocessed by a QR factorization, and then condition estimation and plane rotations on both sides are employed to produce a rank revealing decomposition.

The RRQR, URV, and ULV algorithms are designed for the case $k \approx \min(m,n)$, where k is the numerical rank of the $m \times n$ matrix. For the low-rank case $k \ll \min(m,n)$, more efficient algorithms are available: Chan and Hansen [7] present and analyze the L-RRQR algorithm, and Fierro and Hansen [10] present and analyze low-rank URV and ULV algorithms. We also mention that a perturbation analysis for two-sided orthogonal (or complete) decompositions is given in [9].

For a URV decomposition of A, there exist orthogonal matrices $U_R \in \Re^{m\times m}$ and $V_R \in \Re^{n\times n}$ such that

$$\begin{aligned} A &= U_R R V_R^T \\ &= [U_{Rk}\ U_{R0}\ U_{R\perp}]\, R\, [V_{Rk}\ V_{R0}]^T, \end{aligned} \tag{7}$$

where

$$R = \begin{array}{c} \begin{matrix} k & n-k \end{matrix} \\ \left[\begin{array}{cc} R_k & F \\ 0 & G \\ 0 & 0 \end{array}\right] \begin{array}{l} k \\ n-k \\ m-n \end{array} \end{array}$$

is upper triangular and $k \le n$. Based on this decomposition,

$$A = U_{Rk} R_k V_{Rk}^T + U_{Rk} F V_{R0}^T + U_{R0} G V_{R0}^T. \tag{8}$$

Also, $A_{Rk} \equiv U_{Rk} R_k V_{Rk}^T$ is a rank-k matrix approximation to A satisfying

$$\|A - A_{Rk}\| = \left\| \begin{bmatrix} F \\ G \end{bmatrix} \right\|.$$

$\|\cdot\| = \|\cdot\|_2$ unless otherwise indicated. For a ULV decomposition of A, there exist orthogonal matrices $U_L \in \Re^{m\times m}$ and $V_L \in \Re^{n\times n}$ such that

$$\begin{aligned} A &= U_L L V_L^T \\ &= [U_{Lk}\ U_{L0}\ U_{L\perp}]\, L\, [V_{Lk}\ V_{L0}]^T, \end{aligned} \tag{9}$$

where

$$L = \begin{array}{c} \begin{matrix} k & n-k \end{matrix} \\ \left[\begin{array}{cc} L_k & 0 \\ H & E \\ 0 & 0 \end{array}\right] \begin{array}{l} k \\ n-k \\ m-n \end{array} \end{array}$$

is lower triangular. Based on this decomposition,

$$A = U_{Lk} L_k V_{Lk}^T + U_{L0} H V_{Lk}^T + U_{L0} E V_{L0}^T. \tag{10}$$

$A_{Lk} \equiv U_{Lk} L_k V_{Lk}^T$ denotes a rank-k matrix approximation to A satisfying

$$\|A - A_{Lk}\| = \|[H\ E]\|.$$

In [24] it is shown how so-called "left" and "right" iterations may be used to iteratively refine the decompositions (in the sense that the norm of the off-diagonal block decreases). Based on this refinement strategy, error bounds for estimating the singular values of the matrix A are also provided in [18] and [24].

Ideally, it would be useful for the user to have some kind of a diagnostic measure to assess the quality of subspaces obtained by a two-sided orthogonal decomposition as compared to the reliable SVD. This has important applications in areas where the SVD might be used, for instance, in matrix approximation, subset selection, signal processing, LS, and TLS problems.

The objectives of this paper are to:

- provide error bounds for the subspaces determined by any two-sided orthogonal decomposition;
- show the importance of a good condition estimator in the high-rank revealing URV and ULV algorithms; and
- show how these decompositions may be more accurate alternatives to the SVD than RRQR.

The paper is organized as follows. In §2 we derive a posteriori error bounds for decompositions of the form (7) and (9). These bounds, which are independent of the numerical rank, suggest that if $\|H\| \approx \|F\|$, a ULV decomposition may yield a more accurate estimate of the numerical nullspace than a URV decomposition, while the URV decomposition may yield a better estimate of the numerical range. In §3 our theoretical results show that the quality of the subspaces obtained by Stewart's high-rank revealing URV and ULV algorithms depend on the condition estimator, not on the gap in the singular values. This is illustrated in our numerical simulations. We implement the rank revealing two-sided orthogonal decompositions in an adaptive manner. In our simulations the refinement procedure is based on the repeated estimation of the singular vectors using the Cline–Conn–Van Loan (CCVL) condition estimator [8]. Our experimental evidence shows that this process has the tendency to reduce the nearest off-diagonal elements when estimating a small singular value in a cluster of small singular values. This improves the subsequent estimation step by the CCVL condition estimator. In §4 we compare the subspaces obtained by the RRQR decomposition and the URV and ULV decomposition. The analysis implies that rank revealing two-sided orthogonal decompositions may be more accurate alternatives to the SVD than the RRQR decomposition. In §5 we summarize our conclusions.

Finally, $C(1:i, 1:i)$ denotes the leading submatrix of C of order i, and superscripts $\dagger$ and T denote the pseudoinverse and transpose, respectively.

2. Subspace bounds. It is well known [23], [28] that a singular vector corresponding to a singular value in a cluster is extremely sensitive to small perturbations, but that the span of the singular vectors corresponding to the cluster is well determined, i.e., relatively insensitive to small perturbations. Thus we provide bounds for the error in approximating the span. The following definition defines subspaces associated with the SVD.

DEFINITION [23]. Let $A \in \Re^{m\times n}$ and let $\mathbf{X} \subset \Re^n$ and $\mathbf{Y} \subset \Re^m$ be subspaces of dimension l. Then $\mathbf{X}$ and $\mathbf{Y}$ form a pair of singular subspaces for A if

$$\begin{aligned} &\text{(i)} \quad A\mathbf{X} \subset \mathbf{Y} \\ &\text{(ii)} \quad A^T\mathbf{Y} \subset \mathbf{X}. \end{aligned}$$

$\mathcal{R}(V_k)$ and $\mathcal{R}(U_k)$ are subspace pairs of dimension k, and $\mathcal{R}(V_0)$ and $\mathcal{R}(U_0)$ are subspace pairs of dimension $n-k$, where $\mathcal{R}(C)$ denote the range of matrix C. $\mathcal{R}(V_0)$ is termed the numerical nullspace of A, sometimes referred to as the "noise" subspace. $\mathcal{R}(U_k)$ is termed the numerical range of A.

For the (nonunique) URV decomposition in (7), there exist matrices Q and P (see [23]) such that $\mathcal{R}(U_{Rk} + U_{R0}Q)$ and $\mathcal{R}(V_{Rk} + V_{R0}P)$ form a pair of singular subspaces for A with

$$\|[Q\ P]\|_F \leq \frac{2\|F\|_F}{\sigma_{\min}(R_k) - \|G\|},$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\sigma_{\min}(C)$ denotes the smallest singular value of the matrix C. If $F = 0$ then $\mathcal{R}(U_{Rk})$ and $\mathcal{R}(V_{Rk})$ form a pair of singular subspaces for A as well as $\mathcal{R}(U_{R0})$ and $\mathcal{R}(V_{R0})$. If $\|F\|$ is small then $\mathcal{R}(U_i)$ and $\mathcal{R}(V_i)$ ($i = Rk, R0$) nearly form a pair of singular subspaces for A (similar statements can be made for the ULV).

However, we wish to determine the "distance" between the subspaces $\mathcal{R}(V_0)$ and $\mathcal{R}(V_{R0})$ (and $\mathcal{R}(V_0)$ and $\mathcal{R}(V_{L0})$), as well as $\mathcal{R}(U_k)$ and $\mathcal{R}(U_{Rk})$ (and $\mathcal{R}(U_k)$ and $\mathcal{R}(U_{Lk})$). We will need the following definition for the distance between two subspaces.

DEFINITION [14, p. 76]. Let $W = [W_1\ W_2]$ and $Z = [Z_1\ Z_2]$ be orthogonal matrices, where W_1, $Z_1 \in \Re^{p\times(p-q)}$ and W_2, $Z_2 \in \Re^{p\times q}$. If $\mathcal{S}_1 = \mathcal{R}(W_1)$ and $\mathcal{S}_2 = \mathcal{R}(Z_1)$ then $\operatorname{dist}(\mathcal{S}_1, \mathcal{S}_2) = \|W_1^T Z_2\|$.

2.1. Subspace bounds for the URV and SVD. If $\sin\theta \equiv \operatorname{dist}(\mathcal{S}_1, \mathcal{S}_2)$ then θ is the subspace angle between $\mathcal{S}_1$ and $\mathcal{S}_2$. Let $\sin\theta_{\text{URV}} \equiv \operatorname{dist}(\mathcal{R}(V_0), \mathcal{R}(V_{R0}))$ and let $\sin\phi_{\text{URV}} \equiv \operatorname{dist}(\mathcal{R}(U_k), \mathcal{R}(U_{Rk}))$. Based on the definition, it easily follows that

$$\sin\theta_{\text{URV}} = \|V_k^T V_{R0}\| = \|V_{Rk}^T V_0\| \quad \text{and} \quad \sin\phi_{\text{URV}} = \|{U_k^\perp}^T U_{Rk}\| = \|{U_{Rk}^\perp}^T U_k\|,$$

where $U_k^\perp \equiv [U_0\ U_\perp]$ and $U_{Rk}^\perp \equiv [U_{R0}\ U_{R\perp}]$. However, for our analysis in §2 a more useful expression for $\sin\phi_{\text{URV}}$ is needed and requires some preliminary work. We will use the following result throughout the paper.

LEMMA 2.1. *Given the usual* SVD *and* URV *factorizations of A, then*

$$\sin\phi_{\text{URV}} = \|U_k^T U_{R0}\| = \|U_{Rk}^T U_0\|.$$

Proof. First we will find an expression for ${U_k^\perp}^T U_{Rk}$ and show $\|{U_k^\perp}^T U_{Rk}\| = \|U_{Rk}^T U_0\|$.

$$\begin{aligned}
{U_k^\perp}^T U_{Rk} &= {U_k^\perp}^T A V_{Rk} R_k^{-1} \\
&= \begin{bmatrix} U_0^T \\ {U_\perp}^T \end{bmatrix} [U_k\ U_0\ U_\perp] \begin{bmatrix} \Sigma_k & 0 \\ 0 & \Sigma_0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_k^T \\ V_0^T \end{bmatrix} V_{Rk} R_k^{-1} \\
&= \begin{bmatrix} 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \Sigma_k V_k^T V_{Rk} R_k^{-1} \\ \Sigma_0 V_0^T V_{Rk} R_k^{-1} \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} \Sigma_0 V_0^T V_{Rk} R_k^{-1} \\ 0 \end{bmatrix},
\end{aligned}$$

and thus $\|{U_k^\perp}^T U_{Rk}\| = \|\Sigma_0 V_0^T V_{Rk} R_k^{-1}\|$. Now, from the URV factorization of A, $U_{Rk}^T = R_k^{-T} V_{Rk}^T A^T$. Thus, $U_{Rk}^T U_0 = R_k^{-T} V_{Rk}^T A^T U_0 = R^{-T} V_{Rk}^T V_0 \Sigma_0$ and therefore $\|U_{Rk}^T U_0\| = \|(U_{Rk}^T U_0)^T\| = \|{U_k^\perp}^T U_{Rk}\|$. Similarly, one can show $\|{U_{Rk}^\perp}^T U_k\| = \|G V_{R0}^T V_k \Sigma_k^{-1}\| = \|U_k^T U_{R0}\|$.

A corresponding lemma can be proven for the ULV decomposition. Now we are ready for the main result of this section.

THEOREM 2.2 (URV error bounds). *Let $A \in R^{m\times n}$ have the usual* SVD *and* URV *decomposition, and define $\eta = \sigma_{k+1}/\sigma_k$. Then the distance between the numerical nullspace $\mathcal{R}(V_0)$ and the* URV *approximate nullspace $\mathcal{R}(V_{R0})$, and the distance between the numerical range $\mathcal{R}(U_k)$ and its* URV *estimate $\mathcal{R}(U_{Rk})$ are bounded by*

$$\text{(a)}\quad \frac{\|F\|}{\|R\| + \eta\,\|G\|} \le \operatorname{dist}(\mathcal{R}(V_0), \mathcal{R}(V_{R0})) \le \frac{\sigma_{\min}(R_k)\,\|F\|}{\sigma^2_{\min}(R_k) - \sigma^2_{k+1}},$$

$$\text{(b)}\quad \operatorname{dist}(\mathcal{R}(U_k), \mathcal{R}(U_{Rk})) \le \frac{\|F\|\,\|G\|}{\sigma_k^2 - \|G\|^2}.$$

Proof. To prove the upper bound in (a),

$$\begin{aligned} V_{Rk}^T V_0 &= R_k^{-1}(U_{Rk}^T A - F V_{R0}^T)V_0 \\ &= R_k^{-1}(U_{Rk}^T U_0 \Sigma_0 - F V_{R0}^T V_0). \end{aligned} \tag{11}$$

Now we need to find an expression for $U_{Rk}^T U_0$. First, $AV_{Rk} = U_{Rk}R_k$ implies $U_{Rk}^T = R_k^{-T} V_{Rk}^T A^T$. Second, $U_{Rk}^T U_0 = R_k^{-T} V_{Rk}^T V_0 \Sigma_0$. Substituting into (11),

$$\begin{aligned} V_{Rk}^T V_0 &= R_k^{-1}(R_k^{-T} V_{Rk}^T V_0 \Sigma_0^2 - F V_{R0}^T V_0) \\ &= R_k^{-1} R_k^{-T} V_{Rk}^T V_0 \Sigma_0^2 - R_k^{-1} F V_{R0}^T V_0. \end{aligned}$$

Taking norms,

$$\sin\theta_{\text{URV}} \le \|R_k^{-1}\|^2 \sin\theta_{\text{URV}} \sigma_{k+1}^2 + \|R_k^{-1}\|\,\|F\|,$$

and solving for $\sin\theta_{\text{URV}}$ yields

$$\sin\theta_{\text{URV}} \le \frac{\|R_k^{-1}\|\,\|F\|}{1 - \|R_k^{-1}\|^2\,\sigma_{k+1}^2} = \frac{\sigma_{\min}(R_k)\,\|F\|}{\sigma^2_{\min}(R_k) - \sigma^2_{k+1}}.$$

To prove (b), we use an argument similar to (a). Using $U_k^T U_{R0} = \Sigma_k^{-1}(V_k^T V_{R0})G^T$, it remains to find an expression for $V_k^T V_{R0}$:

$$V_k^T V_{R0} = \Sigma_k^{-1} U_k^T A V_{R0} = \Sigma_k^{-1}(U_k^T U_{Rk} F + U_k^T U_{R0} G).$$

Upon substitution, $U_k^T U_{R0} = \Sigma_k^{-2}(U_k^T U_{Rk} F + U_k^T U_{R0} G)G^T$. It follows

$$\sin\phi_{\text{URV}} \le \sigma_k^{-2}(\|F\|\,\|G\| + \sin\phi_{\text{URV}}\|G\|^2).$$

Solving for $\sin\phi_{\text{URV}}$ yields

$$\sin\phi_{\text{URV}} \le \frac{\sigma_k^{-2}\|F\|\,\|G\|}{1 - \|G\|^2\sigma_k^{-2}} = \frac{\|F\|\,\|G\|}{\sigma_k^2 - \|G\|^2}.$$

To prove the lower bound in (a), it is straightforward to show

$$F = U_{Rk}^T U_1 \Sigma_1 V_k^T V_{R0} + U_{Rk}^T U_2 \Sigma_2 V_0^T V_{R0}.$$

Taking norms in an obvious manner,

$$\begin{aligned} \|F\| &\le \|R\|\,\|V_k^T V_{R0}\| + \|U_{Rk}^T U_2\|\,\|\Sigma_2\| \\ &= \|R\|\,\sin\theta_{\text{URV}} + \|U_k^T U_{R0}\|\,\sigma_{k+1} \\ &\le \|R\|\,\sin\theta_{\text{URV}} + \|\Sigma_k^{-1}(V_k^T V_{R0})G^T\|\,\|\Sigma_2\| \\ &\le (\|R\| + \eta\|G\|)\sin\theta_{\text{URV}}, \end{aligned}$$

hence the lower bound follows. This completes the proof of the theorem.

It can also be shown

$$\operatorname{dist}(\mathcal{R}(V_0),\mathcal{R}(V_{R0})) \le \frac{\|F\|}{\sigma_{\min}(R_k)-\eta\|G\|} \quad \text{and} \quad \operatorname{dist}(\mathcal{R}(U_k),\mathcal{R}(U_{Rk})) \le \frac{\|F\|\,\sigma_{k+1}}{\sigma_{\min}^2(R_k)-\sigma_{k+1}^2}.$$

Note that it is possible to find a posteriori upper bounds by using Theorem 2.2 together with the facts $\sigma_{\min}(R_k) \le \sigma_k$, $\eta \le 1$, and $\sigma_{k+1} \le \|G\|$. Given the above, the following corollary is immediate.

COROLLARY 2.3 (A posteriori bounds for URV). *Under the assumptions of Theorem 2.2, the following a posteriori bounds hold:*

$$\text{(a)} \quad \frac{\|F\|}{\|R\|+\|G\|} \le \operatorname{dist}(\mathcal{R}(V_0),\mathcal{R}(V_{R0})) \le \frac{\|F\|\,\sigma_{\min}(R_k)}{\sigma_{\min}^2(R_k)-\|G\|^2},$$

$$\text{(b)} \quad \operatorname{dist}(\mathcal{R}(U_k),\mathcal{R}(U_{Rk})) \le \frac{\|F\|\,\|G\|}{\sigma_{\min}^2(R_k)-\|G\|^2}.$$

These bounds show explicitly that when $\|F\|$ is small then the subspaces nearly coincide. In §3 we discuss a way to achieve a small $\|F\|$ so that high-quality subspaces are obtained.

2.2. Subspace bounds for the ULV and SVD. In this section we are concerned with the ULV decomposition. Let A have the usual SVD and ULV as in §1. Let $\sin\theta_{\text{ULV}} \equiv \operatorname{dist}(\mathcal{R}(V_0),\mathcal{R}(V_{L0}))$ and $\sin\phi_{\text{ULV}} \equiv \operatorname{dist}(\mathcal{R}(U_k),\mathcal{R}(U_{Lk}))$. After the ULV factorization is complete, we wish to determine upper bounds on the errors in the approximate subspaces.

THEOREM 2.4 (ULV error bounds). *Let $A \in R^{m\times n}$ have the usual* SVD *and* ULV *decomposition, and define $\eta = \sigma_{k+1}/\sigma_k$. Then the distance between the numerical nullspace $\mathcal{R}(V_0)$ and the* ULV *approximate nullspace $\mathcal{R}(V_{L0})$, and the distance between the numerical range $\mathcal{R}(U_k)$ and its* ULV *estimate $\mathcal{R}(U_{Lk})$ are bounded by*

$$\text{(a)} \quad \operatorname{dist}(\mathcal{R}(V_0),\mathcal{R}(V_{L0})) \le \frac{\sigma_{k+1}\|H\|}{\sigma_{\min}^2(L_k)-\sigma_{k+1}^2},$$

$$\text{(b)} \quad \frac{\|H\|}{\|L\|+\eta\|E\|} \le \operatorname{dist}(\mathcal{R}(U_k),\mathcal{R}(U_{Lk})) \le \frac{\|H\|\sigma_{\min}(L_k)}{\sigma_{\min}^2(L_k)-\sigma_{k+1}^2}.$$

Proof. To prove the upper bound in (a), we have

$$V_{Lk}^T V_0 = L_k^{-1} U_{Lk}^T A V_0 = L_k^{-1} U_{Lk}^T U_0 \Sigma_0. \tag{12}$$

Now to find an expression for $U_{Lk}^T U_0$, $A = [U_{Lk}\ U_{L0}]L[V_{Lk}\ V_{L0}]^T$ implies $U_{Lk}^T = L_k^{-T}(V_{Lk}^T A^T - H^T U_{L0}^T)$. Hence,

$$\begin{aligned} U_{Lk}^T U_0 &= L_k^{-T}(V_{Lk}^T A^T - H^T U_{L0}^T)U_0 \\ &= L_k^{-T}(V_{Lk}^T A^T - H^T U_{L0}^T)U_0 \qquad (13) \\ &= L_k^{-T}(V_{Lk}^T V_0 \Sigma_0 - H^T U_{L0}^T U_0). \qquad (14) \end{aligned}$$

Substituting (14) into (12), it follows that

$$V_{Lk}^T V_0 = L_k^{-1}(L_k^{-T}(V_{Lk}^T V_0 \Sigma_0 - H^T U_{L0}^T U_0)\Sigma_0),$$

and taking norms,

$$\|V_{Lk}^T V_0\| \le \|L_k^{-1}\|^2 \|\Sigma_0\|\,\|V_{Lk}^T V_0\| + \|L_k^{-1}\|^2 \|H\|\,\|\Sigma_0\|.$$

Solving for $\|V_{Lk}^T V_0\| = \sin\theta_{\mathrm{ULV}}$, we get

$$\sin\theta_{\mathrm{ULV}} \leq \frac{\|H\|\,\|L_k^{-1}\|^2\,\sigma_{k+1}}{1-\|L_k^{-1}\|^2\sigma_{k+1}^2} = \frac{\|H\|\,\sigma_{k+1}}{\sigma_{\min}^2(L_k)-\sigma_{k+1}^2}.$$

To prove (b), substituting $V_{Lk}^T V_0 = L_k^{-1}(U_{Lk}^T U_0)\Sigma_0$ into

$$U_{Lk}^T U_0 = L_k^{-1}(V_{Lk}^T V_0 \Sigma_0 - H^T U_{L0}^T U_0),$$

it follows that

$$\begin{aligned} U_{Lk}^T U_0 &= L_k^{-T}(L_k^{-1}(U_{Lk}^T U_0)\Sigma_0^2 - H^T U_{L0}^T U_0) \\ &= L_k^{-T}L_k^{-1}(U_{Lk}^T U_0)\Sigma_0^2 - L_k^{-T}H^T U_{L0}^T U_0. \end{aligned}$$

Taking norms in the obvious way yields

$$\sin\phi_{\mathrm{ULV}} \leq \|L_k^{-1}\|^2\sigma_{k+1}^2 \sin\phi_{\mathrm{ULV}} + \|L_k^{-1}\|\,\|H\|,$$

and solving for $\sin\phi_{\mathrm{ULV}}$ yields

$$\sin\phi_{\mathrm{ULV}} \leq \frac{\|L_{11}^{-1}\|\,\|H\|}{1-\sigma_{k+1}^2\|L_{11}^{-1}\|^2} = \frac{\sigma_{\min}(L_k)\|H\|}{\sigma_{\min}^2(L_k)-\sigma_{k+1}^2}.$$

The lower bound for $\sin\phi_{\mathrm{ULV}}$ follows analogously to proof for the lower bound for $\sin\theta_{URV}$. This completes the proof of the theorem.

It can also be shown that

$$\begin{aligned} \mathrm{dist}(\mathcal{R}(V_0),\mathcal{R}(V_{L0})) &\leq \frac{\sigma_{k+1}\|H\|}{\sigma_{\min}^2(L_k)-\sigma_{k+1}^2} \quad \text{and} \\ \mathrm{dist}(\mathcal{R}(U_k),\mathcal{R}(U_{Lk})) &\leq \frac{\|H\|\sigma_{\min}(L_k)}{\sigma_k\sigma_{\min}(L_k)-\sigma_{k+1}\|E\|}. \end{aligned}$$

As before, it is possible to generate a posteriori upper bounds for the subspace angles in terms of computed ULV factors. The necessary facts are $\sigma_{\min}(L_k) \leq \sigma_k$, $\eta < 1$, and $\sigma_{k+1} \leq \|E\|$, as well as Theorem 2.4.

COROLLARY 2.5 (A posteriori bounds for ULV). *Under the assumptions of Theorem 2.4, the following a posteriori bounds hold:*

$$\text{(a)}\quad \mathrm{dist}(\mathcal{R}(V_0),\mathcal{R}(V_{L0})) \leq \frac{\|H\|\|E\|}{\sigma_{\min}^2(L_k)-\|E\|^2},$$

$$\text{(b)}\quad \frac{\|H\|}{\|L\|+\|E\|} \leq \mathrm{dist}(\mathcal{R}(U_k),\mathcal{R}(U_{Lk})) \leq \frac{\sigma_{\min}(L_k)\,\|H\|}{\sigma_{\min}^2(L_k)-\|E\|^2}.$$

This shows that a large $\|H\|$ translates to a large $\sin\phi_{\mathrm{ULV}}$. In the next section we discuss a way to produce a small $\|H\|$ for the high-rank case $k \approx \min(m,n)$. Comparing the a posteriori bounds for the URV and ULV, we may conclude that the ULV can be expected to yield a higher quality estimate of the numerical nullspace than the URV. Tables 1–4 summarize the results of some typical experiments that verify this conclusion. We remark that in a few cases the bounds "appear" violated, but we emphasize that it is due to round-off errors; hence in those cases they are numerically satisfied.

TABLE 1

Results of typical experiments verifying the a posteriori error bounds for the URV *and* ULV *subspace angles θ. The matrices $\{A_i\}$ have various singular value spectrums (cf. §3). The numerical rank of the matrices was $k = 7$ for* tol $= 0.003$ *and* max_iter $= 1$.

A_i	σ_8/σ_7	$\frac{\|F\|}{\|R\|+\|G\|}$	$\sin\theta_{\text{URV}}$	$\frac{\|F\|\sigma_{\min}(R_k)}{\sigma^2_{\min}(R_k)-\|G\|^2}$	$\sin\theta_{\text{ULV}}$	$\frac{\|H\|\|E\|}{\sigma^2_{\min}(L_k)-\|E\|^2}$
A_1	10^{16}	6.1329e-17	1.3910e-15	6.1329e-15	1.4571e-15	2.5784e-43
A_2	10^4	3.2106e-11	3.2008e-09	3.2106e-09	1.0951e-15	7.0547e-17
A_3	10^3	3.2105e-09	3.2008e-07	3.2106e-07	4.0770e-13	7.0546e-13
A_4	10^2	3.2100e-07	3.2009e-05	3.2107e-05	4.0752e-09	7.0503e-09
A_5	10^1	3.1848e-05	3.2104e-03	3.2202e-03	3.7756e-05	6.6495e-05
A_6	10^1	1.3264e-05	1.1913e-03	1.3304e-03	1.8277e-06	2.6570e-06

TABLE 2

Results of typical experiments verifying the a posteriori error bounds for the URV *and* ULV *subspace angles ϕ. The matrices $\{A_i\}$ have various singular value spectrums (cf. §3). The numerical rank of the matrices was $k = 7$ for* tol $= 0.003$ *and* max_iter $= 1$.

A_i	σ_8/σ_7	$\sin\phi_{\text{URV}}$	$\frac{\|F\|\|G\|}{\sigma^2_{\min}(R_k)-\|G\|^2}$	$\frac{\|H\|}{\|L\|+\|E\|}$	$\sin\phi_{\text{ULV}}$	$\frac{\|H\|\sigma_{\min}(L_k)}{\sigma^2_{\min}(L_k)-\|E\|^2}$
A_1	10^{16}	2.9483e-15	3.5002e-29	4.5179e-31	2.8743e-15	4.5179e-29
A_2	10^4	3.1667e-13	3.2106e-13	7.0547e-15	4.8397e-13	7.0547e-13
A_3	10^3	3.1980e-10	3.2106e-10	7.0546e-12	4.8157e-10	7.0546e-10
A_4	10^2	3.1981e-07	3.2107e-07	7.0489e-09	4.8119e-07	7.0503e-07
A_5	10^1	3.2077e-04	3.2202e-04	6.5765e-06	4.4795e-04	6.6495e-04
A_6	10^1	5.7760e-05	6.6519e-05	5.2981e-07	4.0374e-05	5.3141e-05

3. Algorithm and numerical simulations. As mentioned earlier, any URV or ULV decomposition may be refined iteratively using orthogonal transformations, cf. [24]. The purpose of refinement procedures is to concentrate less "energy" of R (or L) in the 1,2 (or 2,1) position so as to decouple the matrix as much as possible. Recall from §2 that when the triangular matrix is decoupled (i.e., the off-diagonal block is a zero matrix) then we have obtained singular subspaces for the matrix, and when the off-diagonal block is small then we have obtained good singular subspace approximations. In [24] it is shown how so-called "left" and "right" ("shiftless" QR) iterations may be used to iteratively reduce the norm of the off-diagonal block, and hence refine the subspaces. Based on this particular refinement strategy, error bounds for estimating the singular values of the matrix A are provided [18], [24]. In this section we show that a small off-diagonal block is achieved in the high-rank revealing URV and ULV algorithms by using a good condition estimator.

3.1. Algorithm. Now we turn our attention to a brief but important discussion of the algorithms by Stewart. At the ith step of the URV algorithm, we work with the upper triangular matrix

$$\begin{array}{c} \begin{array}{cc} i & n-i \end{array} \\ \begin{bmatrix} R_i & F_i \\ 0 & G_i \end{bmatrix} \begin{array}{l} i \\ n-i \end{array} \end{array}$$

and a unit estimate v^i_{est} of the exact $i \times 1$ right singular vector $v^i_{\min}$ corresponding to $\sigma_{\min}(R_i)$. Using plane rotations, find an $i \times i$ orthogonal matrix Q^i such that $(Q^i)^T v^i_{\text{est}} = (0, \ldots, 0, 1)^T$ and $R_i Q^i$ is upper Hessenberg. Then determine an $i \times i$ orthogonal matrix P^i such that $(P^i)^T (R_i Q^i)$ is upper triangular. Partition the updated

TABLE 3

Results of typical experiments verifying the a posteriori error bounds for the URV *and* ULV *subspace angles θ. The matrices $\{A_i\}$ have various singular value spectrums (cf.* §3*). The numerical rank of the matrices was $k = 7$ for* tol = 0.003, max_iter = 2, $\epsilon =$ 1e − 09.

A_i	σ_8/σ_7	$\frac{\|F\|}{\|R\|+\|G\|}$	$\sin\theta_{\text{URV}}$	$\frac{\|F\|\sigma_{\min}(R_k)}{\sigma^2_{\min}(R_k)-\|G\|^2}$	$\sin\theta_{\text{ULV}}$	$\frac{\|H\|\|E\|}{\sigma^2_{\min}(L_k)-\|E\|^2}$
A_1	10^{16}	6.1329e-17	1.3910e-15	6.1329e-15	1.4571e-15	2.5784e-43
A_2	10^{4}	2.8124e-16	2.8124e-14	2.8516e-14	1.0951e-15	7.0547e-17
A_3	10^{3}	5.3222e-14	5.0362e-12	5.3222e-12	4.0770e-13	7.0546e-13
A_4	10^{2}	2.7995e-12	2.7914e-10	2.8001e-1	8.1686e-13	8.2493e-12
A_5	10^{1}	1.0899e-10	5.4685e-09	1.1020e-08	7.3859e-08	2.9776e-07
A_6	10^{1}	3.6423e-09	3.0300e-07	3.6532e-07	6.1356e-10	6.8873e-10

TABLE 4

Results of typical experiments verifying the a posteriori error bounds for the URV *and* ULV *subspace angles ϕ. The matrices $\{A_i\}$ have various singular value spectrums (cf.* §3*). The numerical rank of the matrices was $k = 7$ for* tol = 0.003, max_iter = 2, *and* $\epsilon =$ 1e-09.

A_i	σ_8/σ_7	$\sin\phi_{URV}$	$\frac{\|F\|\|G\|}{\sigma^2_{\min}(R_k)-\|G\|^2}$	$\frac{\|H\|}{\|L\|+\|E\|}$	$\sin\phi_{\text{ULV}}$	$\frac{\|H\|\sigma_{\min}(L_k)}{\sigma^2_{\min}(L_k)-\|E\|^2}$
A_1	10^{16}	2.9483e-15	3.5002e-29	4.5179e-31	2.8743e-15	4.5179e-29
A_2	10^{4}	3.9864e-15	2.8124e-18	7.0547e-15	4.8397e-13	7.0547e-13
A_3	10^{3}	1.6643e-15	5.3222e-15	7.0546e-12	4.8157e-10	7.0546e-10
A_4	10^{2}	2.7838e-13	2.8001e-12	8.2477e-12	8.1642e-10	8.2493e-10
A_5	10^{1}	2.7313e-10	1.1020e-09	2.9449e-08	1.4772e-06	2.9776e-06
A_6	10^{1}	2.8756e-09	1.8266e-08	1.3733e-10	1.2276e-08	1.3775e-08

triangular matrix by

$$\begin{bmatrix} (P^i)^T(R_iQ^i) & (P^i)^TF_i \\ 0 & G_i \end{bmatrix} = \begin{array}{c} \begin{matrix} i-1 & 1 & n-i \end{matrix} \\ \begin{bmatrix} R_{(i-1)} & f_i & F_i' \\ 0 & g_{ii} & g_i^T \\ 0 & 0 & G_i \end{bmatrix} \end{array} \begin{matrix} i-1 \\ 1 \\ n-i. \end{matrix}$$

Analogously, at the ith step of the ULV algorithm we work with the lower triangular matrix

$$\begin{array}{c} \begin{matrix} i & n-i \end{matrix} \\ \begin{bmatrix} L_i & 0 \\ H_i & E_i \end{bmatrix} \end{array} \begin{matrix} i \\ n-i \end{matrix}$$

and a unit estimate u^i_{est} of the exact $i \times 1$ left singular vector $u^i_{\min}$ corresponding to $\sigma_{\min}(L_i)$. Using plane rotations, find an $i \times i$ orthogonal matrix P^i such that $P^iu^i_{\text{est}} = (0, \ldots, 0, 1)^T$ and P^iL_i is lower Hessenberg. Then use plane rotations to determine an $i \times i$ orthogonal matrix Q^i such that $(P^iL_i)(Q^i)^T$ is lower triangular. Partition the updated triangular matrix by

$$\begin{bmatrix} (P^iL_i)(Q^i)^T & 0 \\ H_i(Q^i)^T & E_i \end{bmatrix} = \begin{array}{c} \begin{matrix} i-1 & 1 & n-i \end{matrix} \\ \begin{bmatrix} L_{(i-1)} & 0 & 0 \\ h_i^T & e_{ii} & 0 \\ H_i' & e_i & E_i \end{bmatrix} \end{array} \begin{matrix} i-1 \\ 1 \\ n-i. \end{matrix}$$

The following result shows how the accuracy of the estimate v^i_{est} is related to the norm of the subcolumn f_i, the approximation of $\sigma^2_{\min}(R_i)$ by r^2_{ii}, and $\|F\|$, as well as how the accuracy of the estimate u^i_{est} is related to the size of the subrow h_i^T, the approximation of $\sigma^2_{\min}(L_i)$ by l^2_{ii}, and $\|H\|$.

THEOREM 3.1. *Using the notation above, let v^i_{est} with unit 2-norm denote an estimate of $v^i_{\min}$, the right singular vector of R_i corresponding to $\sigma_{\min}(R_i)$. If θ^i_{URV} denotes the angle between v^i_{est} and $v^i_{\min}$, then*

$$\frac{\|f_i\|}{\|R_i\|} \le \sin\theta^i_{\text{URV}} \quad \text{and} \quad \frac{\sqrt{r_{ii}^2 - \sigma^2_{\min}(R_i)}}{\|R_i\|} \le \sin\theta^i_{\text{URV}}.$$

Analogously, using the notation above, let u^i_{est} with unit 2-norm denote an estimate of $u^i_{\min}$, the left singular vector of L_i corresponding to $\sigma_{\min}(L_i)$. If ϕ^i_{ULV} denotes the angle between u^i_{est} and $u^i_{\min}$, then

$$\frac{\|h_i\|}{\|L_i\|} \le \sin\phi^i_{\text{ULV}} \quad \text{and} \quad \frac{\sqrt{l_{ii}^2 - \sigma^2_{\min}(L_i)}}{\|L_i\|} \le \sin\phi^i_{\text{ULV}}.$$

Moreover,

$$\frac{\|F\|}{\|R\|} \le \sqrt{n-k}\,\sin\theta^{\max}_{\text{URV}} \quad \text{and} \quad \frac{\|H\|}{\|L\|} \le \sqrt{n-k}\,\sin\phi^{\max}_{\text{ULV}},$$

where $\theta^{\max}_{\text{URV}} \equiv \max\{\theta^n_{\text{URV}}, \ldots, \theta^{k+1}_{\text{URV}}\}$ and $\phi^{\max}_{\text{ULV}} \equiv \max\{\phi^n_{\text{ULV}}, \ldots, \phi^{k+1}_{\text{ULV}}\}$.

Proof. We begin with

$$\begin{aligned}\left\|R_i v^i_{\text{est}}\right\|^2 &= \left\|P^{iT}\left(R_i Q^i\right)\left(Q^i\right)^T v^i_{\text{est}}\right\|^2 \\ &= \|f_i\|^2 + r_{ii}^2.\end{aligned}$$

Let R_i have the SVD

$$R_i = U_{R_i}\Sigma_{R_i}V_{R_i}^T + u^i_{\min}\sigma_{\min}(R_i)\left(v^i_{\min}\right)^T,$$

where Σ_{R_i} is an $i-1 \times i-1$ diagonal matrix. Then it follows

$$\begin{aligned}\left\|R_i v^i_{\text{est}}\right\|^2 &= \left\|U_{R_i}\Sigma_{R_i}V_{R_i}^T v^i_{\text{est}}\right\|^2 + \sigma^2_{\min}(R_i)\left|\left(v^i_{\min}\right)^T v^i_{\text{est}}\right|^2 \\ &\le \|R_i\|^2\left(\sin\theta^i_{\text{URV}}\right)^2 + \sigma^2_{\min}(R_i).\end{aligned}$$

Hence,

$$\|f_i\|^2 + r_{ii}^2 \le \|R_i\|^2\left(\sin\theta^i_{\text{URV}}\right)^2 + \sigma^2_{\min}(R_i),$$

which implies

$$\frac{\|f_i\|^2}{\|R_i\|^2} + \frac{r_{ii}^2 - \sigma^2_{\min}(R_i)}{\|R_i\|^2} \le \left(\sin\theta^i_{\text{URV}}\right)^2.$$

Since each term on the left is nonnegative, then

$$\frac{\|f_i\|^2}{\|R_i\|^2} \le \left(\sin\theta^i_{\text{URV}}\right)^2 \quad \text{and} \quad \frac{r_{ii}^2 - \sigma^2_{\min}(R_i)}{\|R_i\|^2} \le \left(\sin\theta^i_{\text{URV}}\right)^2,$$

and the desired result follows immediately. Now we bound $\|F\|/\|R\|$. We use the fact that the Frobenius norm $\|\cdot\|_F$ is orthogonally invariant.

$$\|F\|^2 \le \|F\|_F^2$$

$$\leq \sum_{i=k+1}^{n} \|f_i\|^2$$
$$\leq \|R\|^2 \sum_{i=k+1}^{n} (\sin\theta_{\text{URV}}^i)^2$$
$$\leq \|R\|^2 (n-k)\ (\sin\theta_{\text{URV}}^{\max})^2,$$

hence $\|F\|/\|R\| \leq \sqrt{n-k}\ \sin\theta_{\text{URV}}^{\max}$. This proves the URV results. Since the ULV results can be shown analogously, the theorem is proved.

This means good estimates of the right (left) singular vectors of R_i (L_i) for $i = n, \ldots, k+1$ lead to a small $\|F\|$ ($\|H\|$). By Corollaries 2.3 and 2.5 this means the quality of the subspaces depends on the quality of the estimated singular vectors. Theorem 3.1 generalizes [26, Theorem 1], where it is proven that if all the estimates v_{est}^i (or u_{est}^i) are indeed singular vectors, then $F = 0$ (or $H = 0$) and the relevant subspaces coincide.

We consider a refinement strategy that monitors the columns (rows) of the off-diagonal block of the triangular matrix as it is generated one column (row) at a time. We proceed to the next step (a deflation step) provided the newly generated column (row) of the off-diagonal block is sufficiently small. Our refinement step is based on the repeated estimation of the singular vector using the CCVL condition estimator [8]. The following algorithm slightly modifies Stewart's algorithm [24], [25] by giving it an adaptive flavor.

AN ADAPTIVE VERSION OF THE URV DECOMPOSITION

Input: An $m \times n$ data matrix A, parameter tol for numerical rank determination, integer max_iter to control the maximum number of singular vector estimates per step, and ϵ to bound $\|F\|$.

Step 1. Compute a QR factorization $A = Q\binom{R}{0}$, where Q is orthogonal and $R \in \Re^{n\times n}$ is upper triangular.

Step 2. Initialize $i \leftarrow n$, $V \leftarrow I$, $U \leftarrow Q$.

Step 3. Compute a unit estimate v_{est}^n of the right singular vector corresponding to $\sigma_{\min}(R)$ using condition estimator CCVL.

Step 4. While ($\|R(1:i, 1:i)\, v_{\text{est}}^i\| <$ tol and $i > 1$) do

Set t=1.
While $t \leq$ max_iter
(4a) Compute a sequence of plane rotations $Q_1, \ldots, Q_{i-1}$ so that $Q_{i-1}^T \cdots Q_1^T v_{\text{est}}^i = (0, 0, .., 0, 1)^T$. Update V and R

$$V \leftarrow V \begin{bmatrix} Q^i & 0 \\ 0 & I_{n-i} \end{bmatrix} \text{ and } R \leftarrow R \begin{bmatrix} Q^i & 0 \\ 0 & I_{n-i} \end{bmatrix},$$

where $Q^i \equiv Q_1 \cdots Q_{i-1}$.
(4b) Determine a sequence of plane rotations $P_1, \ldots, P_{i-1}$ so that $P_{i-1}^T \cdots P_1^T R(1:i, 1:i)$ is upper triangular. Update U and R

$$U \leftarrow U \begin{bmatrix} P^i & 0 \\ 0 & I_{m-i} \end{bmatrix} \text{ and } R \leftarrow \begin{bmatrix} (P^i)^T & 0 \\ 0 & I_{n-i} \end{bmatrix} R,$$

where $P^i \equiv P_1 \cdots P_{i-1}$.

(4c) If ($t = \text{max\_iter}$ or $\|R(1:i-1,i)\| < \epsilon/\sqrt{n}$)
Deflate: set $i \leftarrow i-1$ and $t = \text{max\_iter} + 1$
else
t=t+1.

(4d) Compute a unit estimate v^i_{est} of the right singular vector corresponding to $\sigma_{\min}(R(1:i,1:i))$ using condition estimator CCVL.

End while

End while.

END.

It is important to remember that when max_iter =1 then this algorithm is the URV algorithm with condition estimator CCVL. When max_iter > 1, our experiments show this refinement process has the tendency to reduce the nearest off-diagonal elements when working in a cluster of small singular values, which improves the subsequent estimation step by the CCVL condition estimator. See [15] for an excellent survey on condition estimators.

For the ULV algorithm we implemented, the following changes were made in the URV algorithm described above.

- Given an initial QR factorization of A, initialize $U \leftarrow QZ$, $V \leftarrow Z$, and $L \leftarrow ZRZ$ where Z is the anti-identity matrix (zeros everywhere except for ones in the $j, n-j+1$ position).
- Compute an estimate $u^i_{\min}$ of the left singular vector corresponding to the smallest singular value of $L(1:i,1:i)$. Transform $u^i_{\min}$ to $(0,\ldots,0,1)$, thereby transforming $L(1:i,1:i)$ to lower Hessenberg using $i-1$ plane rotations on the left. Accumulate the left plane rotations. Then restore lower triangularity to L using $i-1$ plane rotations on the right. Accumulate the right plane rotations.
- We monitored $\|L(i,1:i-1)\|$ each step.

The numerical results herein indeed verify that the quality of the subspaces is independent of the gap in the singular value spectrum of A, but instead depend on the quality of the estimated singular vector as proved in Theorem 3.1 (cf. Tables 4 and 5).

However, the gap may have an indirect effect in a practical implementation—it can affect the ability of the condition estimator to deliver good singular vector estimates. A poor gap can hamper the rate of convergence when improving the estimated vector using inverse iteration or using repeated condition estimation as described in the algorithm, and a few extra singular vector estimates may be needed before deflation. This is illustrated by the following numerical simulations.

3.2. Numerical simulations. The matrices $\{A_i\}_{i=1}^6$ of dimensions $m = 25$, $n = 10$ with numerical rank $k = 7$ are created by generating random matrices and replacing the singular values by the following values as in [5, §7]. The first seven singular values are fixed at

$$\sigma_1 = 1,\ \sigma_2 = 0.5,\ \sigma_3 = 0.2,\ \sigma_4 = 0.1,\ \sigma_5 = 0.05,\ \sigma_6 = 0.02,\ \sigma_7 = 0.01,$$

with the remaining singular values varying as follows:

σ	A_1	A_2	A_3	A_4	A_5	A_6
σ_8	$1.0 \cdot 10^{-18}$	$1.0 \cdot 10^{-6}$	$1.0 \cdot 10^{-5}$	$1.0 \cdot 10^{-4}$	$1.0 \cdot 10^{-3}$	$5.0 \cdot 10^{-4}$
σ_9	$1.0 \cdot 10^{-18}$	$1.0 \cdot 10^{-7}$	$1.0 \cdot 10^{-6}$	$1.0 \cdot 10^{-5}$	$1.0 \cdot 10^{-4}$	$5.0 \cdot 10^{-4}$
σ_{10}	$1.0 \cdot 10^{-18}$	$1.0 \cdot 10^{-8}$	$1.0 \cdot 10^{-7}$	$1.0 \cdot 10^{-6}$	$1.0 \cdot 10^{-5}$	$1.0 \cdot 10^{-4}$.

We implemented the adaptive algorithms (as described earlier) in MATLAB [19] with machine precision $\approx 2.2 \times 10^{-16}$. We

- compared the quality of the approximate subspaces and
- verified the a posteriori bounds.

The a posteriori bounds in Corollaries 2.3 and 2.5 and Theorem 3.1 prove that the two-sided orthogonal decompositions do not depend on the gap and that the ULV decomposition yields a more accurate estimate of the numerical nullspace, while the URV decomposition yields a more accurate estimate of the numerical range. This is verified in the numerical simulations (see Tables 1–7) with max_iter = 1 and 2, where the experiments also illustrate the quality of the a posteriori bounds.

The bounds are not overly pessimistic and in most cases provide an extremely accurate indication of the quality of the subspace. Note that when max_iter = 2 then CCVL provides improved estimates of the singular vectors resulting in higher quality subspaces, which confirms that this strategy is a good refinement strategy. In addition, Table 5 confirms the theoretical implications of [26, Theorem 1] and Theorem 3.1, which proves that exact estimates of the singular vectors yield singular subspaces ($\theta_{\text{URV}}^{\max} = 0$ and $\phi_{\text{ULV}}^{\max} = 0$).

TABLE 5

Results of typical experiments for the error bounds for the RRQR, URV, *and* ULV *subspace angles θ and ϕ. The matrices $\{A_i\}$ have various singular value spectrums (see §3). The numerical rank of the matrices was $k = 7$ for* tol = 0.003, max_iter = 1. *The exact singular vectors were supplied in this experiment. No subspace refinement for any factorization.*

A_i	$\sin\theta_{\text{RRQR}}$	$\sin\theta_{\text{URV}}$	$\sin\theta_{\text{ULV}}$	$\sin\phi_{\text{RRQR}}$	$\sin\phi_{\text{URV}}$	$\sin\phi_{\text{ULV}}$
A_1	5.6861e-15	2.7846e-15	3.6910e-15	2.7010e-14	1.6871e-15	2.9126e-15
A_2	5.2541e-10	3.7177e-15	3.3077e-15	1.8372e-04	2.4332e-15	5.0195e-15
A_3	5.2541e-08	2.3045e-15	3.2493e-15	1.8372e-03	1.5029e-15	3.1826e-15
A_4	5.2541e-06	4.1955e-15	2.3565e-15	1.8369e-02	1.5318e-15	6.4979e-15
A_5	5.2598e-04	4.5027e-15	4.5145e-15	1.8070e-01	1.8536e-15	5.8246e-15
A_6	2.4305e-04	1.6924e-15	1.3442e-15	9.4604e-02	8.4914e-16	5.1436e-15

4. Rank revealing QR factorizations. An RRQR factorization of A (with numerical rank k) is any factorization

$$A\Pi = QR \tag{15}$$

$$= [Q_1\ Q_2] \begin{bmatrix} k & n-k \\ R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} \begin{matrix} k \\ n-k, \end{matrix} \tag{16}$$

where R is upper triangular, $\|R_{22}\|$ is order σ_{k+1}, R_{11} is well conditioned, and Π is a permutation matrix. In this factorization the approximate nullspace is not exhibited explicitly since the 1,2 position of R, namely, R_{12} is generally not small. In [4] Chan presents an RRQR algorithm where the matrix A is preprocessed by an initial QR

factorization followed by condition estimation, strategic column pivoting, and plane rotations applied on the left to restore triangularity. When the algorithm runs to completion, it can be shown [5] that

$$\|A(\Pi W)\| \leq \sqrt{n-k}\,\sigma_{k+1}, \tag{17}$$

where the columns of

$$W = \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} \begin{matrix} k \\ n-k \end{matrix}$$

store the estimated $n-k$ right singular vectors. Hence $\mathcal{R}(\Pi W)$ is an approximate nullspace for A. More rigorously, from [5, Theorem 4.1], if θ_{QR} denotes the subspace angle between $\mathcal{R}(\Pi W)$ and $\mathcal{R}(V_0)$ and $\sin\theta_{\mathrm{QR}} \equiv \mathrm{dist}(\mathcal{R}(\Pi W), \mathcal{R}(V_0))$, then

$$\mathrm{dist}(\mathcal{R}(\Pi W), \mathcal{R}(V_0)) \leq \eta\,(1 + \|W_2^{-1}\|\sqrt{n-k}\,). \tag{18}$$

In addition, it can be shown [6] that if ϕ_{RRQR} denotes the subspace angle between $\mathcal{R}(U_k)$ and $\mathcal{R}(Q_1)$ and $\sin\phi_{\mathrm{RRQR}} \equiv \mathrm{dist}(\mathcal{R}(Q_1), \mathcal{R}(U_k))$, then

$$\mathrm{dist}(\mathcal{R}(Q_1), \mathcal{R}(U_k)) \leq \eta\,\|W_2^{-1}\|\sqrt{n-k}. \tag{19}$$

The RRQR bounds show that if

- η is small (large gap in the singular value spectrum of A),
- A has low rank-deficiency (k not too much smaller than n), and
- W_2 is well conditioned ($\|W_2^{-1}\|$ not too large),

then $\mathcal{R}(\Pi W)$ ($\mathcal{R}(Q_1)$) is a good approximation to the numerical nullspace (range) in the sense that the subspace angle is small. Hence, from these bounds and the numerical evidence in [5] the quality of the RRQR-based subspaces depends on a gap condition. This contrasts with the upper bounds we derived in Theorems 2.2 and 2.4 for rank revealing URV and ULV factorizations; for these factorizations η is not required to be small (cf. Theorem 3.1 in §3). Table 5 demonstrates this property. In practice, we need only a good estimate of the singular vector to which the plane rotations are applied! The condition estimator CCVL is well suited for this purpose as demonstrated in §4; so is inverse iteration as demonstrated in [26].

The numerical results in Tables 5–7 illustrate that the rank revealing URV and ULV factorizations may be more accurate alternatives to the SVD than RRQR. Using the same simulation setup as described in §4, Tables 5–7 compare the subspace angles θ_{RRQR}, θ_{URV}, θ_{ULV} and ϕ_{RRQR}, ϕ_{URV}, ϕ_{ULV} as computed by the URV and ULV algorithms we implemented and by Chan's RRQR algorithm.

The results in Table 5 confirm that the URV and ULV decompositions do not depend on the gap, unlike the RRQR factorization. For these results the exact singular vectors were supplied to the algorithms. However, in Table 6 we supply the algorithms with estimates obtained from CCVL. In the absence of refinement, the results show θ_{URV} is comparable to θ_{RRQR}. θ_{URV} could be better if we used better estimates; however, in fairness we supply both of these factorizations with the same estimated right singular vector. Note that θ_{ULV} is superior to both. Both ϕ_{URV} and ϕ_{ULV} are conspicuously better than ϕ_{RRQR}. Note that ϕ_{RRQR} exhibits the dependence on η as in (19) and that decreasing the gap (i.e., decreasing $1/\eta$) affects the ability of CCVL to deliver good estimates.

TABLE 6

Results of typical experiments for the error bounds for the RRQR, URV, *and* ULV *subspace angles* θ *and* ϕ. *The matrices* $\{A_i\}$ *have various singular value spectrums (see* §3). *The numerical rank of the matrices was* $k = 7$ *for* tol $= 0.003$, max_iter $= 1$. *The singular vectors were estimated by the* CCVL *condition estimator. No subspace refinement for any factorization.*

A_i	$\sin\theta_{\text{RRQR}}$	$\sin\theta_{\text{URV}}$	$\sin\theta_{\text{ULV}}$	$\sin\phi_{\text{RRQR}}$	$\sin\phi_{\text{URV}}$	$\sin\phi_{\text{ULV}}$
A_1	5.6861e-15	1.3910e-15	1.4571e-15	2.7010e-14	2.9483e-15	2.8743e-15
A_2	5.2541e-10	3.2008e-09	1.0951e-15	1.8372e-04	3.1667e-13	4.8397e-13
A_3	5.2541e-08	3.2008e-07	4.0770e-13	1.8372e-03	3.1980e-10	4.8157e-10
A_4	5.2541e-06	3.2009e-05	4.0752e-09	1.8369e-02	3.1981e-07	4.8119e-07
A_5	5.2598e-04	3.2104e-03	3.7756e-05	1.8070e-01	3.2077e-04	4.4795e-04
A_6	2.4305e-04	1.1913e-03	1.8277e-06	9.4604e-02	5.7760e-05	4.0374e-05

TABLE 7

Results of typical experiments for the error bounds for the RRQR, URV, *and* ULV *subspace angles* θ. *The matrices* $\{A_i\}$ *have various singular value spectrums (see* §3). *The numerical rank of the matrices was* $k = 7$ *for* tol $= 0.003$, max_iter $= 2$ *and* $\epsilon =$ 1e-09. *The singular vectors were estimated by the* CCVL *condition estimator. The* RRQR *approximate nullspace was improved by one step of simultaneous inverse iteration.*

A_i	$\sin\theta_{\text{RRQR}}$	$\sin\theta_{\text{URV}}$	$\sin\theta_{\text{ULV}}$
A_1	5.2048e-15	1.3910e-15	3.7295e-15
A_2	5.8684e-13	2.8516e-14	1.0951e-15
A_3	4.6403e-13	5.0362e-12	4.0770e-13
A_4	1.4888e-10	2.7914e-10	8.1686e-13
A_5	1.4931e-06	5.4685e-09	7.3859e-08
A_6	9.6430e-08	3.0300e-07	6.1356e-10

However, when CCVL makes an additional pass per iteration (max_iter $= 2$), it delivers a markedly improved estimate of the singular vectors, as indicated by the smaller subspace angles in Table 7. For max_iter $= 2$, the URV and ULV decompositions make more appreciable gains in estimating the singular subspaces of A than the RRQR factorization with one step of simultaneous inverse iteration.

5. Conclusion. In this paper we derived a posteriori error bounds (§2) for assessing the quality of subspaces obtained by a rank revealing two-sided orthogonal decomposition, which is a product of an orthogonal matrix, a triangular matrix, and another orthogonal matrix. The bounds are independent of both the numerical rank of the matrix and the algorithm used to compute the decomposition. The theoretical results show how the quality of the subspaces (compared to the singular subspaces of the matrix) depend on the size of the off-diagonal block of the triangular factor, and that the ULV provides a better estimate of the numerical nullspace than the URV decomposition. Specifically, we considered the promising high-rank ($k \approx \min(m, n)$) revealing URV and ULV decompositions introduced by G. W. Stewart. The theoretical analysis shows that the quality of the subspaces depend on the quality of the estimated singular vectors and not on a gap condition (cf. §2 and Theorem 3.1). We implemented the algorithms in an adaptive manner. Based on our analysis in §4, we conclude that the URV and ULV decompositions may be more accurate alternatives to the SVD than the RRQR factorization. Finally, we provided numerical examples to illustrate our conclusions.

Acknowledgment. The authors wish to thank Sabine Van Huffel for suggesting the comparison in §4 and Chris Bischof for supplying the CCVL routine.

REFERENCES

[1] E. BIGLIERI AND K. YAO, *Some properties of singular value decomposition and their applications to digital signal processing*, Signal Processing, 18 (1989), pp. 277–289.

[2] C. H. BISCHOF AND P. C. HANSEN, *Structure-preserving and rank revealing QR factorizations*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 1332–1350.

[3] C. H. BISCHOF AND G. M. SHROFF, *On updating signal subspaces*, IEEE Trans. Signal Processing, 40 (1992), pp. 96–105.

[4] T. F. CHAN, *Rank revealing QR factorizations*, Linear Algebra Appl., 88 (1987), pp. 67–82.

[5] T. F. CHAN AND P. C. HANSEN, *Computing truncated singular value decomposition least squares solutions by rank revealing QR factorizations*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 519–530.

[6] ———, *Some applications of the rank revealing QR factorization*, SIAM J. Sci. Stat. Comput., 13 (1992), pp. 727–741.

[7] ———, *Low-rank revealing QR factorizations*, Numer. Linear Algebra Appl., 1 (1994), pp. 33–44.

[8] A. K. CLINE, A. R. CONN, AND C. F. VAN LOAN, *Generalizing the LINPACK condition estimator*, Lecture Notes in Math. 909, Springer-Verlag, Berlin, New York, 1982, pp. 73–83.

[9] R. D. FIERRO, *Perturbation analysis for two-sided (or complete) orthogonal decompositions*, SIAM J. Matrix Anal. Appl., to appear.

[10] R. D. FIERRO AND P. C. HANSEN, *Low-Rank Revealing Two-Sided Orthogonal Decompositions*, PAM Tech. Report 94-09, Dept. of Mathematics, California State University, San Marcos.

[11] R. D. FIERRO AND J. R. BUNCH, *Perturbation theory for orthogonal projection methods with applications to least squares and total least squares*, Linear Algebra Appl., to appear.

[12] P. E. GILL, W. MURRAY, AND M. H. WRIGHT, Numerical Linear Algebra and Optimization, Addison-Wesley, Redwood City, CA, 1991.

[13] G. H. GOLUB, V. KLEMA, AND G.W. STEWART, *Rank Degeneracy and Least Squares Problems*, Tech. Report TR-456, Dept. of Computer Science, University of Maryland, College Park, 1976.

[14] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., John Hopkins University Press, 1989.

[15] N. J. HIGHAM, *A survey of condition number estimation for triangular matrices*, SIAM Rev., 29 (1987), pp. 575–596.

[16] C. L. LAWSON AND R. J. HANSEN, *Solving Least Squares Problems*, Prentice Hall, Englewood Cliffs, NJ, 1974.

[17] F. LORENZELLI, P. C. HANSEN, T. F. CHAN, AND K. YAO, *The* RRQR *factorization and its systolic implementation*, UCLA Tech. Report, Dept. of Mathematics, University of California, Los Angeles, 1992.

[18] R. MATHIAS AND G. W. STEWART, *A block QR algorithm and the singular value decomposition*, Linear Algebra Appl., 182 (1993), pp. 91–100.

[19] C. B. MOLER, J. LITTLE, AND S. BANGERT, *Pro-Matlab User's Guide*, The Math Works, Sherborn, MA, 1987.

[20] M. A. RAHMAN AND K. YU, *Total least squares approach for frequency estimation using linear prediction*, IEEE Trans. Acoustics, Speech, Signal Processing, 35 (1987), pp. 1442–1454.

[21] R. ROY AND T. KAILATH, *ESPIRIT—Estimation of signal parameters via rotational invariance techniques*, IEEE Trans. Acoustics, Speech, Signal Processing, 37 (1989), pp. 984–995.

[22] G. W. STEWART, *An updating algorithm for subspace tracking*, IEEE Trans. Signal Processing, 40 (1992), pp. 1535–1541.

[23] ———, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.

[24] ———, *On an Algorithm for Refining a Rank-Revealing* URV *Decomposition and a Perturbation Theorem for Singular Values*, Tech. Report CS-TR 2626, Dept. of Computer Science, University of Maryland, 1991.

[25] ———, *Updating a rank revealing* ULV *decomposition*, SIAM J. Matrix Anal. Appl., 4 (1993), pp. 494–499.

[26] S. VAN HUFFEL AND H. ZHA, *An efficient total least squares algorithm based on a rank revealing two-sided orthogonal decomposition*, Numer. Algorithms, 4 (1993), pp. 101–133.

[27] S. VAN HUFFEL AND J. VANDEWALLE, *The Total Least Squares Problem: Computational Aspects and Analysis*, Society for Industrial and Applied Mathematics, Philadelphia, 1991.

[28] P.Å. WEDIN, *Perturbation bounds in connection with the singular value decomposition*, BIT, 12 (1972), pp. 99–111.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 760–775, July 1995

005

PERTURBATION ANALYSIS OF THE CHOLESKY DOWNDATING AND QR UPDATING PROBLEMS*

JI-GUANG SUN†

Dedicated to Professor Åke Björk on his 60th birthday.

Abstract. Given upper triangular matrices R, G and column vectors x, f such that $R^TR - xx^T$ and $(R+G)^T(R+G) - (x+f)(x+f)^T$ are positive definite, let U and $U+T$ be the corresponding Cholesky factors. In this paper, upper bounds on $\|T\|$ in terms of $\|G\|$ and $\|f\|$ and upper bounds on $\|T\|/\|U\|$ in terms of $\|G\|/\|R\|$ and $\|f\|/\|x\|$ are given, and the first order perturbation expansions of $\|T\|$ and $\|T\|/\|U\|$ are derived. Moreover, a perturbation analysis of the QR updating problem is also given.

Key words. Cholesky factorization, Cholesky downdating problem, QR factorization, QR updating problem, perturbation bound, condition number

AMS subject classifications. 15A23, 65F99

1. Introduction.

Let $A \in \mathcal{R}^{n\times n}$ be a symmetric positive definite matrix. Then there exists a unique upper triangular matrix $R \in \mathcal{R}^{n\times n}$ with positive diagonal elements such that $A = R^TR$. This factorization is known as the Cholesky factorization of A, and R is called the Cholesky factor [7, p. 141]. Let $A \in \mathcal{R}^{m\times n}$ with $\operatorname{rank}(A) = n$. The QR factorization of A is a unique decomposition of the form $A = QR$, in which $R \in \mathcal{R}^{n\times n}$ is an upper triangular matrix with positive diagonal elements, and $Q \in \mathcal{R}^{m\times n}$ satisfies $Q^TQ = I$, the identity matrix. The matrices Q and R are referred to as the QR factors of A [7, p. 211].

In this paper we consider the following problems. (i) Given an upper triangular matrix $R \in \mathcal{R}^{n\times n}$ and a vector $x \in \mathcal{R}^n$ such that $R^TR - xx^T$ is positive definite, find an upper triangular matrix $U \in \mathcal{R}^{n\times n}$ with positive diagonal elements such that

$$R^TR - xx^T = U^TU.$$

This problem is called the Cholesky downdating problem, and the matrix U is referred to as the downdated Cholesky factor [7], [10]. (ii) Given $A \in \mathcal{R}^{m\times n}, x \in \mathcal{R}^m$, and $y \in \mathcal{R}^n$ such that $\operatorname{rank}(A) = \operatorname{rank}(A + xy^T) = n$, find an upper triangular matrix $U \in \mathcal{R}^{n\times n}$ with positive diagonal elements and $P \in \mathcal{R}^{m\times n}$ satisfying $P^TP = I$ such that

$$A + xy^T = PU.$$

This problem is called the QR updating problem, and the matrices P and U are referred to as the updated QR factors [4], [7]. The Cholesky downdating and QR updating problems have many important applications, and there are several stable algorithms (see [1], [3]–[8], [10] and the references contained therein).

Recently, Pan [8] presented a first order perturbation analysis of the Cholesky downdating problem. In this paper, we shall give upper bounds on the perturbation of U in terms of the perturbations of R and x, and then derive a first order perturbation

* Received by the editors May 17, 1993; accepted for publication (in revised form) by R. J. Plemmons May 6, 1994. This subject was supported by the Swedish Natural Science Research Council contract F-FU 6952-300 and the Department of Computing Science, Umeå University.

† Department of Computing Science, Umeå University, S-901 87 Umeå, Sweden (jisun@cs.umu.se).

bound of U, which is simpler and in most cases is sharper than those of [8]. Moreover, we shall give a perturbation analysis of the QR updating problem by the same way.

We start with the simplest case. Let μ be a positive number and $\rho, \xi \in \mathcal{R}$ such that $\rho^2 - \xi^2 = \mu^2$. Let $\nu = \xi/\rho$. Then for any $\gamma, \phi, \epsilon \in \mathcal{R}$ satisfying

$$|\epsilon\gamma|/\rho < 1 \quad \text{and} \quad \frac{\nu + |\epsilon\phi|/\rho}{1 - |\epsilon\gamma|/\rho} < 1$$

there is a unique number $\tau(\epsilon)$ such that $\mu + \tau(\epsilon)$ is positive and

$$(\rho + \epsilon\gamma)^2 - (\xi + \epsilon\phi)^2 = (\mu + \tau(\epsilon))^2. \tag{1.1}$$

From (1.1) we get

$$\begin{aligned} \frac{\tau(\epsilon)}{\mu} &= \left[1 + \frac{2\rho(\gamma - \nu\phi)\epsilon + (\gamma^2 - \phi^2)\epsilon^2}{\mu^2}\right]^{1/2} - 1 \\ &= \frac{1}{1 - \nu^2}\left(\frac{\gamma}{\rho} - \nu^2\frac{\phi}{\xi}\right)\epsilon + O(\epsilon^2) \qquad \epsilon \to 0 \end{aligned}$$

and

$$\frac{|\tau(\epsilon)|}{\mu} \leq \frac{1}{1 - \nu^2}\left(\frac{|\gamma|}{\rho} + \nu^2\frac{|\phi|}{\xi}\right)|\epsilon| + O(\epsilon^2) \qquad \epsilon \to 0. \tag{1.2}$$

Moreover, taking ρ, ξ, and μ as variables and differentiating $\rho^2 - \xi^2 = \mu^2$, we get

$$d\mu = \frac{1}{\sqrt{(1 - \nu^2)}}(d\rho - \nu d\xi)$$

and

$$|d\mu| \leq \frac{1}{\sqrt{(1 - \nu^2)}}(|d\rho| + |\nu||d\xi|). \tag{1.3}$$

In §2 we generalize the relation (1.3) to the Cholesky downdating problem, and apply the elementary calculus (ref. [2], [12], [13]) to get perturbation bounds for the downdated Cholesky factor U, and then derive first order perturbation bounds of U. Moreover, we present results of numerical tests, and give remarks. In §3 we derive first order perturbation bounds of the updated QR factors.

The symbol $\| \|_2$ will be used for the Euclidean vector norm and spectral matrix norm, and $\| \|_F$ the Frobenius norm. $A^\dagger$ denotes the Moore–Penrose inverse of a matrix A. $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the largest and smallest eigenvalues of a symmetric matrix A, respectively. For a full rank matrix A we define

$$\kappa(A) = \|A\|_2\|A^\dagger\|_2. \tag{1.4}$$

If A is nonsingular, then $\kappa(A) = \|A\|_2\|A^{-1}\|_2$.

2. The Cholesky downdating problem.

2.1. Perturbation theorems. For a matrix $A = (\alpha_{ij})$ we define the differential of A by $dA = (d\alpha_{ij})$.

We first derive some differential inequalities.

THEOREM 2.1. *Let $R \in \mathcal{R}^{n\times n}$ be upper triangular and $x \in \mathcal{R}^n$ such that $R^TR - xx^T$ is positive definite, and let*

$$v = R^{-T}x. \tag{2.1}$$

Moreover, let

$$R^TR - xx^T = U^TU \tag{2.2}$$

be the Cholesky factorization of $R^TR - xx^T$. Then

$$\begin{aligned}\|dU\|_F &\le \frac{\sqrt{2}\kappa(U)}{(1-\|v\|_2^2)^{1/2}}(\|dR\|_F + \|v\|_2\|dx\|_2)\\ &\le \frac{\sqrt{2}\kappa(R)}{1-\|v\|_2^2}(\|dR\|_F + \|v\|_2\|dx\|_2),\end{aligned} \tag{2.3}$$

where $\kappa(\cdot)$ is defined by (1.4).

Proof. It is known that the elements of U are differentiable functions of the elements of R and x. Differentiating the relation (2.2) we get

$$dR^TR + R^TdR - dxx^T - xdx^T = dU^TU + U^TdU$$

and

$$\begin{aligned}&(dRU^{-1})^TRU^{-1} + (RU^{-1})^TdRU^{-1} - U^{-T}dx(U^{-T}x)^T - U^{-T}x(U^{-T}dx)^T\\ &= (dUU^{-1})^T + dUU^{-1}.\end{aligned} \tag{2.4}$$

Since dUU^{-1} is upper triangular, we have [9]

$$\|(dUU^{-1})^T + dUU^{-1}\|_F \ge \sqrt{2}\|dUU^{-1}\|_F \ge \sqrt{2}\|U\|_2^{-1}\|dU\|_F.$$

Combining it with (2.4) we get

$$\|dU\|_F \le \sqrt{2}\|U\|_2\|U^{-1}\|_2(\|RU^{-1}\|_2\|dR\|_F + \|U^{-T}x\|_2\|dx\|_2). \tag{2.5}$$

Observe the following facts.

(i) By the hypothesis the matrix $R^TR - xx^T$ is positive definite. It is easy to see that this hypothesis is equivalent to $\|v\|_2 < 1$. From (2.1) and (2.2)

$$RU^{-1}(RU^{-1})^T = (I - vv^T)^{-1}. \tag{2.6}$$

Thus, we have

$$\|RU^{-1}\|_2 = \|(I - vv^T)^{-1}\|_2^{1/2} = \frac{1}{(1-\|v\|_2^2)^{1/2}}. \tag{2.7}$$

(ii) From (2.1)

$$U^{-T}x = (RU^{-1})^Tv,$$

which combined with (2.7) gives

$$\|U^{-T}x\|_2 \le \|RU^{-1}\|_2\|v\|_2 = \frac{\|v\|_2}{(1-\|v\|_2^2)^{1/2}}. \tag{2.8}$$

Hence, substituting (2.7) and (2.8) into (2.5) we get the first inequality of (2.3).

Notice that

$$\begin{aligned}\kappa(U) &= (\|R^TR - xx^T\|_2\|(R^TR - xx^T)^{-1}\|_2)^{1/2} \\ &= (\|R^T(I-vv^T)R\|_2\|R^{-1}(I-vv^T)^{-1}R^{-T}\|_2)^{1/2} \qquad (2.9)\\ &\le \kappa(R)/(1-\|v\|_2^2)^{1/2}.\end{aligned}$$

Substituting it into the first inequality of (2.3) we derive the second inequality of (2.3). □

We now use the differential inequalities of (2.3) to derive perturbation bounds for the downdated Cholesky factor U. The main results are Theorems 2.2 and 2.3.

THEOREM 2.2. *Given an upper triangular $R \in \mathcal{R}^{n\times n}$ and an $x \in \mathcal{R}^n$. Assume that the Cholesky factorization* (2.2) *exists, and define $v = R^{-T}x$. Moreover, given an upper triangular $G \in \mathcal{R}^{n\times n}$ and an $f \in \mathcal{R}^n$. If G and f satisfy*

$$\|R^{-1}\|_2\|G\|_2 < 1 \quad \text{and} \quad \nu \equiv \frac{\|v\|_2 + \|R^{-1}\|_2\|f\|_2}{1-\|R^{-1}\|_2\|G\|_2} < 1, \tag{2.10}$$

then there is a unique Cholesky factorization

$$(R+G)^T(R+G) - (x+f)(x+f)^T = (U+T)^T(U+T), \tag{2.11}$$

and

$$\begin{aligned}\|T\|_F &\le \tfrac{\sqrt{2}}{1-\nu^2}[\omega(\|R^{-1}\|_2\|G\|_2)(\kappa(R)+1)-1](\|G\|_F + \nu\|f\|_2) \qquad (2.12)\\ &\le \frac{\sqrt{2}\kappa(R)}{(1-\nu^2)(1-\|R^{-1}\|_2\|G\|_2)}(\|G\|_F + \nu\|f\|_2),\end{aligned}$$

$$\begin{aligned}&\|T\|_F/\|U\|_p \qquad (2.13)\\ &\le \frac{\sqrt{2}[\omega(\|R^{-1}\|_2\|G\|_2)(\kappa(R)+1)-1]}{(1-\nu^2)(1-\|v\|_2^2)^{1/2}}\left(\frac{\|G\|_F}{\|R\|_p} + \nu\|v\|_2\frac{\|f\|_2}{\|x\|_2}\right)\\ &\le \frac{\sqrt{2}\kappa(R)}{(1-\nu^2)(1-\|v\|_2^2)^{1/2}(1-\|R^{-1}\|_2\|G\|_2)}\left(\frac{\|G\|_F}{\|R\|_p} + \nu\|v\|_2\frac{\|f\|_2}{\|x\|_2}\right) \equiv \beta_p,\end{aligned}$$

where $p = 2, F$, and the function $\omega(t)$ is defined by

$$\omega(t) = \frac{1}{t}\ln\frac{1}{1-t}, \quad 0 < t < 1. \tag{2.14}$$

Proof. Observe that for $\epsilon \in [0,1]$ we have

$$\begin{aligned}&(R+\epsilon G)^T(R+\epsilon G) - (x+\epsilon f)(x+\epsilon f)^T = \\ &(R+\epsilon G)^T[I - (I+\epsilon GR^{-1})^{-T}(v+\epsilon R^{-T}f)(v+\epsilon R^{-T}f)^T(I+\epsilon GR^{-1})^{-1}](R+\epsilon G)\end{aligned}$$

and

$$\|(I+\epsilon GR^{-1})^{-T}(v+\epsilon R^{-1}f)\|_2 \le \|(I+\epsilon GR^{-1})^{-1}\|_2\|v+\epsilon R^{-1}f\|_2 \le \frac{\|v\|_2+\|R^{-1}\|_2\|f\|_2}{1-\|R^{-1}\|_2\|G\|_2} = \nu.$$

Hence, if G and f satisfy (2.10), then

$$(R+\epsilon G)^T(R+\epsilon G)-(x+\epsilon f)(x+\epsilon f)^T$$

is positive definite, and consequently there is a unique Cholesky factorization

$$(R+\epsilon G)^T(R+\epsilon G)-(x+\epsilon f)(x+\epsilon f)^T = (U+T(\epsilon))^T(U+T(\epsilon)) \tag{2.15}$$

for each $\epsilon \in [0,1]$.

Write

$$R(\epsilon)=R+\epsilon G,\quad x(\epsilon)=x+\epsilon f,\quad U(\epsilon)=U+T(\epsilon),\quad v(\epsilon)=R(\epsilon)^{-T}x(\epsilon) \tag{2.16}$$

for $\epsilon \in [0,1]$. Let $\sigma_1(\epsilon) \ge \cdots \ge \sigma_n(\epsilon)$ be the singular values of $R(\epsilon)$, and let $\sigma_j = \sigma_j(0)$, $j=1,\dots,n$. Then by the second inequality of (2.3)

$$\begin{aligned}\|T\|_F &= \|U(1)-U(0)\|_F = \|\textstyle\int_0^1 dU(\epsilon)\|_F \le \int_0^1 \|dU(\epsilon)\|_F \\ &\le \sqrt{2}\int_0^1 \frac{\kappa(R(\epsilon))}{1-\|v(\epsilon)\|_2^2}(\|dR(\epsilon)\|_F+\|v(\epsilon)\|_2\|dx(\epsilon)\|_2) \\ &= \sqrt{2}\int_0^1 \frac{\sigma_1(\epsilon)}{\sigma_n(\epsilon)}\cdot\frac{1}{1-\|v(\epsilon)\|_2^2}(\|G\|_F+\|v(\epsilon)\|_2\|f\|_2)d\epsilon.\end{aligned} \tag{2.17}$$

Observe that by Mirsky's theorem on perturbation bounds for singular values (ref. [11, p. 204]) we have

$$\frac{\sigma_1(\epsilon)}{\sigma_n(\epsilon)} \le \frac{\sigma_1+\|G\|_2\epsilon}{\sigma_n-\|G\|_2\epsilon},$$

and from

$$v(\epsilon)=R(\epsilon)^{-T}x(\epsilon)=(I+\epsilon GR^{-1})^{-T}(v+\epsilon R^{-T}f)$$

we get

$$\|v(\epsilon)\|_2 \le \frac{\|v\|_2+\epsilon\|R^{-1}\|_2\|f\|_2}{1-\epsilon\|R^{-1}\|_2\|G\|_2} \equiv \nu(\epsilon). \tag{2.18}$$

Hence, from (2.17)

$$\|T\|_F \le \sqrt{2}\int_0^1 \frac{\sigma_1+\|G\|_2\epsilon}{\sigma_n-\|G\|_2\epsilon}\cdot\frac{1}{1-\nu(\epsilon)^2}(\|G\|_F+\nu(\epsilon)\|f\|_2)d\epsilon. \tag{2.19}$$

For simplicity, we use the relation $\nu(\epsilon) \le \nu$ for $\epsilon \in [0,1]$, where ν and $\nu(\epsilon)$ are defined by (2.10) and (2.18), respectively. Then from (2.19)

$$\|T\|_F \le \frac{\sqrt{2}(\|G\|_F+\nu\|f\|_2)}{1-\nu^2}\int_0^1 \frac{\sigma_1+\|G\|_2\epsilon}{\sigma_n-\|G\|_2\epsilon}d\epsilon. \tag{2.20}$$

By integral techniques [12]

$$\begin{aligned}\int_0^1 \frac{\sigma_1+\|G\|_2\epsilon}{\sigma_n-\|G\|_2\epsilon}d\epsilon &= \frac{\sigma_1+\sigma_n}{\|G\|_2}\ln\frac{\sigma_n}{\sigma_n-\|G\|_2} - 1 \\ &= (\kappa(R)+1)\omega(\|R^{-1}\|_2\|G\|_2) - 1,\end{aligned} \tag{2.21}$$

where $\omega(\cdot)$ is defined by (2.14). Combining it with (2.20) we get the first inequality of (2.12).

Moreover, the inequality

$$\frac{\sigma_1 + \|G\|_2\epsilon}{\sigma_n - \|G\|_2\epsilon} \le \frac{\sigma_1\sigma_n}{(\sigma_n - \|G\|_2\epsilon)^2}$$

gives

$$\begin{aligned}\int_0^1 \frac{\sigma_1+\|G\|_2\epsilon}{\sigma_n-\|G\|_2\epsilon}d\epsilon &\le \int_0^1 \frac{\sigma_1\sigma_n}{(\sigma_n-\|G\|_2\epsilon)^2}d\epsilon \\ &= \frac{\sigma_1}{\sigma_n - \|G\|_2} = \frac{\kappa(R)}{1 - \|R^{-1}\|_2\|G\|_2}.\end{aligned} \tag{2.22}$$

Combining it with (2.20) we get the second inequality of (2.12).

Finally, observe the following facts.

(i) From (2.7) it follows that for $p = 2, F$

$$\begin{aligned}\|U\|_p &= \|(RU^{-1})^{-1}R\|_p \ge \|RU^{-1}\|_2^{-1}\|R\|_p \\ &= (1 - \|v\|_2^2)^{1/2}\|R\|_p.\end{aligned} \tag{2.23}$$

(ii) From (2.1)

$$\|R\|_p \ge \|x\|_2/\|v\|_2. \tag{2.24}$$

Hence, combining (2.23) and (2.24) with (2.12) we derive (2.13). □

THEOREM 2.3. *Let R, x, U, v, G, f, ν be as in Theorem 2.2, in which G and f satisfy (2.10). Let (2.11) be the unique Cholesky factorization of $(R+G)^T(R+G) - (x+f)(x+f)^T$. Moreover, let*

$$\gamma_1 = \|G^TR + R^TG - fx^T - xf^T\|_2, \quad \gamma_2 = \|G^TG - ff^T\|_2, \quad \gamma = \gamma_1 + \gamma_2 \tag{2.25}$$

and

$$\lambda_n = \lambda_{\min}(R^TR - xx^T). \tag{2.26}$$

If $\gamma < \lambda_n$, then

$$\|T\|_F \le \frac{\sqrt{2}\kappa(U)}{(1-\nu^2)^{1/2}}\omega\left(\frac{\gamma}{\lambda_n}\right)(\|G\|_F + \nu\|f\|_2) \tag{2.27}$$

and

$$\frac{\|T\|_F}{\|U\|_p} \le \frac{\sqrt{2}\kappa(U)}{(1-\nu^2)^{1/2}(1-\|v\|_2^2)^{1/2}}\omega\left(\frac{\gamma}{\lambda_n}\right)\left(\frac{\|G\|_F}{\|R\|_p} + \nu\|v\|_2\frac{\|f\|_2}{\|x\|_2}\right) \equiv b_p, \tag{2.28}$$

where $p = 2, F$, and the function $\omega(t)$ is defined by (2.14).

Proof. On the basis of the proof of Theorem 2.2 there is a unique Cholesky factorization (2.15) for each $\epsilon \in [0,1]$. Define $R(\epsilon), x(\epsilon), U(\epsilon)$ and $v(\epsilon)$ by (2.16), and let

$$\lambda_1(\epsilon) = \lambda_{\max}(U(\epsilon)^T U(\epsilon)), \quad \lambda_n(\epsilon) = \lambda_{\min}(U(\epsilon)^T U(\epsilon)).$$

Then by the first inequality of (2.3)

$$\begin{aligned} \|T\|_F &= \|\textstyle\int_0^1 dU(\epsilon)\|_F \le \int_0^1 \|dU(\epsilon)\|_F \\ &\le \sqrt{2}\int_0^1 \frac{\kappa(U(\epsilon))}{(1-\|v(\epsilon)\|_2^2)^{1/2}}(\|dR(\epsilon)\|_F + \|v(\epsilon)\|_2\|dx(\epsilon)\|_2) \\ &\le \sqrt{2}\int_0^1 \left(\frac{\lambda_1(\epsilon)}{\lambda_n(\epsilon)}\right)^{1/2} \frac{1}{(1-\|v(\epsilon)\|_2^2)^{1/2}}(\|G\|_F + \|v(\epsilon)\|_2\|f\|_2)d\epsilon. \end{aligned} \tag{2.29}$$

Moreover, by Weyl's theorem on perturbation bounds for eigenvalues (ref. [11, p. 203]) we have

$$\lambda_1(\epsilon) \le \lambda_1 + \gamma\epsilon, \quad \lambda_n(\epsilon) \ge \lambda_n - \gamma\epsilon,$$

where $\lambda_n = \lambda_n(0)$ (by (2.26)), and we define $\lambda_1 = \lambda_1(0)$. Combining it with (2.29) and (2.18) we get

$$\|T\|_F \le \sqrt{2}\int_0^1 \left(\frac{\lambda_1+\gamma\epsilon}{\lambda_n-\gamma\epsilon}\right)^{1/2} \frac{1}{(1-\nu(\epsilon)^2)^{1/2}}(\|G\|_F + \nu(\epsilon)\|f\|_2)d\epsilon. \tag{2.30}$$

For simplicity, we use the relations $\nu(\epsilon) \le \nu$ and

$$\left(\frac{\lambda_1+\gamma\epsilon}{\lambda_n-\gamma\epsilon}\right)^{1/2} \le \frac{(\lambda_1\lambda_n)^{1/2}}{\lambda_n-\gamma\epsilon} \quad \forall \epsilon \in [0,1].$$

Then from (2.30)

$$\|T\|_F \le \frac{\sqrt{2}(\|G\|_F + \nu\|f\|_2)}{(1-\nu^2)^{1/2}}(\lambda_1\lambda_n)^{1/2}\int_0^1 \frac{d\epsilon}{\lambda_n-\gamma\epsilon}.$$

Combining it with

$$(\lambda_1\lambda_n)^{1/2}\int_0^1 \frac{d\epsilon}{\lambda_n-\gamma\epsilon} = \left(\frac{\lambda_1}{\lambda_n}\right)^{1/2}\omega\left(\frac{\gamma}{\lambda_n}\right)$$

we get the inequality (2.27).

Finally, from (2.27) and (2.23)–(2.24) we derive (2.28). □

The following two results, as corollaries of Theorems 2.2 and 2.3, present first order perturbation expansions for the downdated Cholesky factor U.

COROLLARY 2.4. *Let R, x, G, f be as in Theorem 2.2, and assume that the Cholesky factorization (2.2) exists. Define $v = R^{-T}x$. Let $\epsilon_0 > 0$ be small enough so that the Cholesky factorization*

$$(R+\epsilon G)^T(R+\epsilon G) - (x+\epsilon f)(x+\epsilon f)^T = (U+T(\epsilon))^T(U+T(\epsilon))$$

always exists for $\epsilon \in (-\epsilon_0, \epsilon_0)$. *Then we have*

$$\|T(\epsilon)\|_F \leq \frac{\sqrt{2}\kappa(R)}{1-\|v\|_2^2}(\|G\|_F + \|v\|_2\|f\|_2)|\epsilon| + O(\epsilon^2) \tag{2.31}$$

and

$$\frac{\|T(\epsilon)\|_F}{\|U\|_p} \leq \beta_p^{(1)} + O(\epsilon^2) \quad \epsilon \to 0, \tag{2.32}$$

where $p = 2, F$, *and*

$$\beta_p^{(1)} \equiv \frac{\sqrt{2}\kappa(R)}{(1-\|v\|_2^2)^{3/2}}\left(\frac{\|G\|_F}{\|R\|_p} + \|v\|_2^2\frac{\|f\|_2}{\|x\|_2}\right)|\epsilon|. \tag{2.33}$$

COROLLARY 2.5. *Let* $R, x, G, f, v, \epsilon_0$ *be as in Corollary* 2.4. *Then we have*

$$\|T(\epsilon)\|_F \leq \frac{\sqrt{2}\kappa(U)}{(1-\|v\|_2^2)^{1/2}}(\|G\|_F + \|v\|_2\|f\|_2)|\epsilon| + O(\epsilon^2) \tag{2.34}$$

and

$$\frac{\|T(\epsilon)\|_F}{\|U\|_p} \leq b_p^{(1)} + O(\epsilon^2) \quad \epsilon \to 0, \tag{2.35}$$

where $p = 2, \mathrm{F}$, *and*

$$b_p^{(1)} \equiv \frac{\sqrt{2}\kappa(U)}{1-\|v\|_2^2}\left(\frac{\|G\|_F}{\|R\|_p} + \|v\|_2^2\frac{\|f\|_2}{\|x\|_2}\right)|\epsilon|. \tag{2.36}$$

We now cite the main result of [8], which also presents a first order perturbation expansion for the downdated Cholesky factor U.

THEOREM 2.6 (Pan). *Let* $R, x, G, f, v, \alpha, \epsilon, U, T(\epsilon)$ *be as in Corollary* 2.4, *and define*

$$C(v) = \frac{\|v\|_2^2}{(1-\|v\|_2^2)^{1/2}}$$

and

$$\begin{aligned}\beta_{\mathrm{Pan}}^{(1)} = \kappa(R)\Big(&[2n^{3/2}\kappa(R)C^2(v) + 2n\kappa(R)C(v)]\frac{\|f\|_2}{\|x\|_2} \\ &+ [2n^{3/2}\kappa(R)C^2(v) + 2n\kappa(R)C(v) + 1]\frac{\|G\|_2}{\|R\|_2}\Big)|\epsilon|.\end{aligned} \tag{2.37}$$

Then

$$\frac{\|T(\epsilon)\|_2}{\|U\|_2} \leq \beta_{\mathrm{Pan}}^{(1)} + O(\epsilon^2) \quad \epsilon \to 0. \tag{2.38}$$

Comparing Theorem 2.6 (Pan) with Theorems 2.2 and 2.3 and Corollaries 2.4 and 2.5 we see that the perturbation bounds of this paper are simpler. Moreover, numerical tests show that in most cases, especially for an ill-conditioned downdating problem (i.e., $\kappa(R) \gg 1$ and/or $1 - \|v\|_2^2 \approx 0$), the perturbation bounds of this paper are sharper than the bound of (2.37)–(2.38).

2.2. Condition numbers. The estimates (2.32)–(2.33) and (2.35)–(2.36) show that the quantities

$$\kappa(R,x) \equiv \frac{\kappa(R)}{(1-\|v\|_2^2)^{3/2}}, \quad c(R,x) \equiv \frac{\kappa(U)}{1-\|v\|_2^2}$$

can be used to measure the (relative) sensitivity of the Cholesky downdating problem, where $v = R^{-T}x$ satisfies $0 < \|v\|_2 < 1$, and U is the Cholesky factor of $R^TR - xx^T$. Note that from (2.9) we have $c(R,x) \le \kappa(R,x)$.

In view of the estimate (2.37)–(2.38) Pan [8] suggests using the quantity

$$\chi(R,x) \equiv \kappa(R)\left(\frac{\kappa(R)\|v\|_2^2}{1-\|v\|_2^2}+1\right)$$

to assess the condition of the downdating problem. Comparing $\chi(R,x)$ with $\kappa(R,x)$ and $c(R,x)$ we see that when $\|v\|_2$ is near zero, the three quantities are approximately equal; otherwise it is difficult to assess which one is the smallest. But numerical tests show that in most cases the quantity $c(R,x)$ is the smallest. Especially, when $\kappa(R) \gg 1$ and/or $1-\|v\|_2^2 \approx 0$ the quantity $c(R,x)$ is, in general, much smaller than $\chi(R,x)$. Consequently, we suggest using $c(R,x)$ as the (relative) condition number of the Cholesky downdating problem.

2.3. A numerical example. In this section we present some results of numerical tests.

Example 1. Let

$$R_c = \begin{pmatrix} 1 & -c & -c & -c & -c \\ 0 & 1 & -c & -c & -c \\ 0 & 0 & 1 & -c & -c \\ 0 & 0 & 0 & 1 & -c \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad x_0 = \begin{pmatrix} 0.240 \\ -0.899 \\ 0.899 \\ 1.560 \\ -2.390 \end{pmatrix},$$

$$G_0 = \begin{pmatrix} 0.2113 & -0.4649 & 0.6174 & 0.4857 & -0.6167 \\ 0 & 0.4524 & 0.8441 & -0.7382 & 0.4374 \\ 0 & 0 & -0.6538 & 0.6630 & 0.5072 \\ 0 & 0 & 0 & 0.7469 & 0.1891 \\ 0 & 0 & 0 & 0 & -0.1167 \end{pmatrix}, \quad f_0 = \begin{pmatrix} -0.4237 \\ 0.2190 \\ -0.8531 \\ 0.1194 \\ 0.4762 \end{pmatrix}$$

and

$$R = \operatorname{diag}(1, s, s^2, s^3, s^4)R_c, \quad x = \tau x_0, \quad G = \epsilon G_0, \quad f = \epsilon f_0,$$

where $c = 0.95$, $s = \sqrt{1-c^2}$ and $\epsilon = 1.0\text{e}-10$. Computation gives $\kappa(R) \approx 1901$.

Assume that U and $U+T(\epsilon)$ are the downdated Cholesky factors of $R^TR - xx^T$ and $(R+G)^T(R+G) - (x+f)(x+f)^T$, respectively. Then by Theorem 2.6 (Pan)

$$\frac{\|T(\epsilon)\|_2}{\|U\|_2} \lesssim \beta_{\text{Pan}}^{(1)} \text{ (by (2.37) – (2.38))},$$

and by Theorems 2.2 and 2.3 and Corollaries 2.4 and 2.5

$$\frac{\|T(\epsilon)\|_F}{\|U\|_2} \lesssim \beta_2^{(1)}, \quad \frac{\|T(\epsilon)\|_F}{\|U\|_2} \lesssim b_2^{(1)}, \quad \frac{\|T(\epsilon)\|_F}{\|U\|_2} \le \beta_2, \quad \frac{\|T(\epsilon)\|_F}{\|v\|_2} \le b_2$$

(see (2.32)–(2.33), (2.35)–(2.36), (2.13), and (2.28), respectively), in which $\beta_{\text{Pan}}^{(1)}$, $\beta_2^{(1)}$, and $b_2^{(1)}$ are the first order perturbation bounds. Numerical tests show that in most cases the bounds $b_2^{(1)}$ and b_2 are better than the others.

Some numerical results obtained by using MATLAB are listed in Table 1, in which $v = R^{-T}x$, and we take

$$\tau_1 = 1.004015006005433e-02, \quad \tau_2 = 1.003021021209640e-02,$$

$$\tau_3 = 9.036225416303058e-03$$

and $\tau_4 = \tau_3 \cdot e-01$, $\tau_5 = \tau_3 \cdot e-03$, $\tau_6 = \tau_3 \cdot e-05$.

TABLE 1

τ	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
$\|v\|_2$	0.99999	0.999	0.9	0.09	0.0009	0.000009
$\beta_{\text{Pan}}^{(1)}$	1.44e+04	1.46e+02	1.36e+00	1.16e−02	1.13e−04	1.27e−06
$\beta_2^{(1)}$	1.07e+02	1.07e−01	1.05e−04	1.12e−06	2.73e−07	2.65e−07
$b_2^{(1)}$	8.04e+01	8.05e−02	8.41e−05	1.12e−06	2.73e−07	2.65e−07
β_2	1.10e+02	1.07e−01	1.05e−04	1.12e−06	2.73e−07	2.65e−07
b_2	———	8.25e−02	8.41e−05	1.12e−06	2.73e−07	2.65e−07
$\chi(R,x)$	1.81e+11	1.81e+09	1.54e+07	3.14e+04	1.90e+03	1.90e+03
$\kappa(R,x)$	2.13e+10	2.13e+07	2.30e+04	1.92e+03	1.90e+03	1.90e+03
$c(R,x)$	1.59e+10	1.59e+07	1.84e+04	1.92e+03	1.90e+03	1.90e+03

2.4. Block Cholesky downdating problem. Given an upper triangular matrix $R \in \mathcal{R}^{n\times n}$ and a matrix $X \in \mathcal{R}^{n\times r}$ such that $R^TR - XX^T$ is positive definite, find an upper triangular matrix $U \in \mathcal{R}^{n\times n}$ with positive diagonal elements such that

$$R^TR - XX^T = U^TU.$$

This problem can be called the block Cholesky downdating problem. We note that there is no difficulty in applying the technique of this paper to get perturbation bounds for the block downdated Cholesky factor U. For example, we have the following result.

Let $V = R^{-T}X$. Given an upper triangular $G \in \mathcal{R}^{n\times n}$ and a matrix $F \in \mathcal{R}^{n\times r}$. Let ϵ_0 be small enough so that the Cholesky factorization

$$(R+\epsilon G)^T(R+\epsilon G) - (X+\epsilon F)(X+\epsilon F)^T = (U+T(\epsilon))^T(U+T(\epsilon))$$

always exists for $\epsilon \in (-\epsilon_0, \epsilon_0)$. Then

$$\frac{\|T(\epsilon)\|_F}{\|U\|_p} \le \beta_p^{(1)} + O(\epsilon^2) \quad \text{and} \quad \frac{\|T(\epsilon)\|_F}{\|U\|_p} \le b_p^{(1)} + O(\epsilon^2) \quad \epsilon \to 0, \tag{2.39}$$

where $p = 2, F$, and the first order perturbation bounds $\beta_p^{(1)}$ and $b_p^{(1)}$ are

$$\beta_p^{(1)} = \frac{\sqrt{2}\kappa(R)}{[\lambda_{\min}(I-VV^T)]^{3/2}}\left(\frac{\|G\|_F}{\|R\|_p} + \|V\|_2^2\frac{\|F\|_F}{\|X\|_p}\right)|\epsilon| \tag{2.40}$$

and

$$b_p^{(1)} = \frac{\sqrt{2}\kappa(U)}{\lambda_{\min}(I-VV^T)}\left(\frac{\|G\|_F}{\|R\|_p} + \|V\|_2^2\frac{\|F\|_F}{\|X\|_p}\right)|\epsilon|. \tag{2.41}$$

The relations (2.39)–(2.41) show that the quantities

$$\kappa(R,X) \equiv \frac{\kappa(R)}{[\lambda_{\min}(I-VV^T)]^{3/2}}, \quad c(R,X) \equiv \frac{\kappa(U)}{\lambda_{\min}(I-VV^T)}$$

can be used to measure the (relative) sensitivity of the block Cholesky downdating problem. Note that $c(R,X) \le \kappa(R,X)$.

Recently, Eldén and Park [5] presented a perturbation analysis of the block Cholesky downdating problem. Perturbation bounds for the block downdated Cholesky factor U are given when only R or X is perturbed. It follows from [5, Theorem 3.2 and Corollary 3.3] that

$$\frac{\|T(\epsilon)\|_F}{\|U\|_2} \le \frac{2\sqrt{2}n\kappa^2(R)}{\lambda_{\min}(I-VV^T)} \frac{\|G\|_2}{\|R\|_2}|\epsilon| + O(\epsilon^2) \quad \text{if} \quad F=0 \tag{2.42}$$

and

$$\frac{\|T(\epsilon)\|_F}{\|U\|_2} \le \frac{2\sqrt{2}n\kappa^2(R)}{\lambda_{\min}(I-VV^T)} \frac{\|F\|_2}{\|R\|_2}|\epsilon| + O(\epsilon^2) \quad \text{if} \quad G=0. \tag{2.43}$$

Therefore, Eldén and Park [5] take the quantity

$$\kappa_{\text{down}} \equiv \frac{\kappa^2(R)}{\lambda_{\min}(I-VV^T)}$$

as a condition number for the block downdating problem. But numerical tests show that in most cases, especially when $\kappa(R) \gg 1$ and/or $\lambda_{\min}(I-VV^T) \approx 0$, the quantity $c(R,X)$ is much smaller than κ_{down}. Consequently, we suggest using $c(R,X)$ as the (relative) condition number of the block downdating problem.

3. The QR updating problem. We first derive some differential inequalities.

THEOREM 3.1. *Let* $A \in \mathcal{R}^{m\times n}, x \in \mathcal{R}^m$ *and* $y \in \mathcal{R}^n$ *such that* $\operatorname{rank}(A) = \operatorname{rank}(A+xy^T) = n$. *Let* $A = QR$ *and*

$$A + xy^T = PU \tag{3.1}$$

be the QR *factorizations of* A *and* $A+xy^T$, *respectively. Moreover, let*

$$u = Q^T x, \quad v = R^{-T} y, \tag{3.2}$$

$$S(u,v,x) = I + uv^T + vu^T + \|x\|_2^2 vv^T \tag{3.3}$$

and

$$\delta(u,v,x) = \lambda_{\min}(S(u,v,x)), \quad \rho(u,v,x) = \lambda_{\max}(S(u,v,x)). \tag{3.4}$$

Then

(3.5)

$$\begin{aligned}\|dU\|_F &\le \sqrt{2}\left[\kappa(A+xy^T)(\|dA\|_F + \|x\|_2\|dy\|_2) + \frac{\|A+xy^T\|_2\|v\|_2}{(\delta(u,v,x))^{1/2}}\|dx\|_2\right] \\ &\le \sqrt{2}\left(\frac{\rho(u,v,x)}{\delta(u,v,x)}\right)^{1/2}[\kappa(A)(\|dA\|_F + \|x\|_2\|dy\|_2) + \|A\|_2\|v\|_2\|dx\|_2],\end{aligned}$$

$$\begin{aligned} \|dP\|_F &\le \sqrt{2}\left[\|(A+xy^T)^\dagger\|_2(\|dA\|_F+\|x\|_2\|dy\|_2)+\tfrac{\|v\|_2}{(\delta(u,v,x))^{1/2}}\|dx\|_2\right] \\ &\le \sqrt{2}\frac{1}{(\delta(u,v,x))^{1/2}}\left[\|A^\dagger\|_2(\|dA\|_F+\|x\|_2\|dy\|_2)+\|v\|_2\|dx\|_2\right]. \end{aligned} \tag{3.6}$$

Proof. We first prove the differential inequalities of (3.5). It is known that the elements of P and U are differentiable functions of the elements of A, x, and y. Differentiating the relation (3.1), we get

$$dA + dxy^T + xdy^T = dPU + PdU \tag{3.7}$$

and

$$P^T dP + dUU^{-1} = P^T dAU^{-1} + P^T dx(U^{-T}y)^T + P^T xdy^T U^{-1} \equiv \Phi. \tag{3.8}$$

Notice that

$$dP^T P + P^T dP = 0. \tag{3.9}$$

Therefore, from (3.8)

$$dUU^{-1} + (dUU^{-1})^T = \Phi + \Phi^T.$$

Combining it with

$$\|dUU^{-1} + (dUU^{-1})^T\|_F \ge \sqrt{2}\|U\|_2^{-1}\|dU\|_F,$$

we get

$$\|dU\|_F \le \sqrt{2}\|U\|_2\left[\|U^{-1}\|_2(\|dA\|_F+\|P^T x\|_2\|dy\|_2)+\|U^{-T}y\|_2\|dx\|_2\right]. \tag{3.10}$$

Observe the following facts.

(i) $\operatorname{rank}(A+xy^T) = n$ if and only if $\delta(u,v,x) > 0$.

This is an important fact. We now give a proof. Take $Q_\perp \in \mathcal{R}^{m\times(m-n)}$ so that $\hat{Q} = (Q, Q_\perp)$ is orthogonal. Then x can be expressed as

$$x = \hat{Q}\begin{pmatrix} u \\ w \end{pmatrix},$$

where u is defined by (3.2), and $w = Q_\perp^T x$. Thus

$$A + xy^T = \hat{Q}\begin{pmatrix} R \\ 0 \end{pmatrix} + \hat{Q}\begin{pmatrix} u \\ w \end{pmatrix} y^T = \hat{Q}CR, \tag{3.11}$$

where

$$C = \begin{pmatrix} I \\ 0 \end{pmatrix} + \begin{pmatrix} u \\ w \end{pmatrix} v^T, \tag{3.12}$$

and v is defined by (3.2). The relation (3.11) shows that $\operatorname{rank}(A+xy^T) = n$ if and only if C^TC is positive definite. From

$$C^TC = (I+uv^T)^T(I+uv^T) + vw^Twv^T, \quad w^Tw = x^Tx - u^Tu$$

it follows that

$$C^TC = S(u,v,x), \tag{3.13}$$

where $S(u,v,x)$ is defined by (3.3). Consequently, $\operatorname{rank}(A+xy^T) = n$ if and only if $\lambda_{\min}(S(u,v,x)) > 0$, i.e., $\delta(u,v,x) > 0$ (by (3.4)).

(ii) From (3.1)

$$\|U\|_2 = \|A+xy^T\|_2, \quad \|U^{-1}\|_2 = \|(A+xy^T)^\dagger\|_2. \tag{3.14}$$

(iii) It follows from (3.1), (3.11), (3.13), and (3.14) that

$$\begin{aligned} \|U\|_2 &= \|PU\|_2 = \|A+xy^T\|_2 = \|\hat{Q}CR\|_2 \\ &\le \|C\|_2\|R\|_2 = (\rho(u,v,x))^{1/2}\|R\|_2, \end{aligned} \tag{3.15}$$

$$\begin{aligned} \|U^{-1}\|_2 &= \|(U^TU)^{-1}\|_2^{1/2} = \|R^{-1}(C^TC)^{-1}R^{-T}\|_2^{1/2} \\ &\le \|R^{-1}\|_2\|S^{-1}\|_2^{1/2} = \|R^{-1}\|_2/(\delta(u,v,x))^{1/2} \end{aligned} \tag{3.16}$$

and

$$\begin{aligned} \|U^{-T}y\|_2 &= (y^T(U^TU)^{-1}y)^{1/2} = (y^TR^{-1}(C^TC)^{-1}R^{-T}y)^{1/2} \\ &= (v^TS(u,v,x)^{-1}v)^{1/2} \le \|v\|_2/(\delta(u,v,x))^{1/2}. \end{aligned} \tag{3.17}$$

Hence, combining (3.10) with (3.14), (3.17), and $\|P^Tx\|_2 \le \|x\|_2$, we get the first inequality of (3.5). Furthermore, combining (3.10) with (3.15)–(3.17), $\|P^Tx\|_2 \le \|x\|_2$, $\|R\|_2 = \|A\|_2$, and $\|R^{-1}\|_2 = \|A^\dagger\|_2$, we get the second inequality of (3.5).

Now we are going to prove the differential inequalities of (3.6). Take $Y \in \mathcal{R}^{m\times(m-n)}$ so that $V = (P,Y)$ is orthogonal, and let

$$\delta X = V^TdP = \begin{pmatrix} \delta X^{(1)} \\ \delta X^{(2)} \end{pmatrix}, \quad \delta X^{(1)} \in \mathcal{R}^{n\times n}. \tag{3.18}$$

Then from (3.7) and (3.9)

$$\delta X = V^T(dA + dxy^T + xdy^T)U^{-1} - \begin{pmatrix} dUU^{-1} \\ 0 \end{pmatrix} \tag{3.19}$$

and

$$(\delta X^{(1)})^T + \delta X^{(1)} = 0. \tag{3.20}$$

Observe that any matrix X can be split uniquely as

$$X = X_L + X_D + X_U,$$

where X_L is strictly lower triangular, X_D is diagonal, and X_U is strictly upper triangular, i.e., $(X_L)_{ij} = 0$ for all $i \le j$, $(X_D)_{ij} = 0$ for all $i \ne j$, and $(X_U)_{ij} = 0$ for all $i \ge j$. Thus, the relation (3.20) implies $(\delta X)_D = 0$. Moreover, from (3.18)–(3.20) we get

$$(\delta X)_L = \left(V^T[(dA + xdy^T)U^{-1} + dx(U^{-T}y)^T]\right)_L \tag{3.21}$$

and

$$(\delta X)_U = \begin{pmatrix} (\delta X^{(1)})_U \\ 0 \end{pmatrix} = \begin{pmatrix} -[(\delta X^{(1)})_L]^T \\ 0 \end{pmatrix}$$

$$= \begin{pmatrix} -\left[\left(P^T[(dA + xdy^T)U^{-1} + dx(U^{-T}y)^T]\right)_L\right]^T \\ 0 \end{pmatrix}. \tag{3.22}$$

Hence, from (3.18), (3.21), and (3.22)

$$\begin{aligned} \|dP\|_F &= \|\delta X\|_F = \sqrt{\|(\delta X)_L\|_F^2 + \|(\delta X)_U\|_F^2} \\ &\le \sqrt{2}\left[\|U^{-1}\|_2(\|dA\|_F + \|x\|_2\|dy\|_2) + \|U^{-T}y\|_2\|dx\|_2\right]. \end{aligned} \tag{3.23}$$

Combining (3.23) with (3.14) and (3.17) we get the first inequality of (3.6). Combining (3.23) with (3.16) and (3.17) we get the second inequality of (3.6). □

Remark 1. The differential inequalities of (3.5)–(3.6) show that the quantities $\delta(u, v, x)$ and $\rho(u, v, x)$ defined by (3.4) are important for the QR updating problem. We now give the explicit expressions of the two quantities.

Let $S(u, v, x)$ be the $n \times n$ symmetric positive definite matrix defined by (3.3). Take $V_2 \in \mathcal{R}^{n\times(n-1)}$ so that $V = (v/\|v\|_2, V_2)$ is orthogonal, and let

$$\alpha = 1 + 2u^Tv + \|v\|_2^2\|x\|_2^2, \quad a = \|v\|_2 V_2^T u. \tag{3.24}$$

Then

$$V^T S(u, v, x)V = \begin{pmatrix} \alpha & a^T \\ a & I^{(n-1)} \end{pmatrix} \equiv S_0.$$

It is easy to know that the eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n$ of the matrix S_0 (i.e., the eigenvalues of $S(u, v, x)$) are

$$\rho(u, v, x) = \lambda_1 = \frac{\alpha+1+\sqrt{(\alpha-1)^2+4\|a\|_2^2}}{2}, \qquad \delta(u, v, x) = \lambda_n = \frac{\alpha+1-\sqrt{(\alpha-1)^2+4\|a\|_2^2}}{2}, \tag{3.25}$$

and $\lambda_2 = \cdots = \lambda_{n-1} = 1$. Moreover, from (3.24)

$$\|a\|_2^2 = \|v\|_2^2 u^T V_2 V_2^T u = \|u\|_2^2\|v\|_2^2 - (u^Tv)^2. \tag{3.26}$$

Combining (3.25) with (3.24) and (3.26) we get the explicit expressions of $\delta(u, v, x)$ and $\rho(u, v, x)$.

By using the same technique of §2 and [12] and [13], from the differential inequalities of (3.5)–(3.6) we can derive perturbation bounds for the updated QR factors P and U. But for avoiding quite complicated mathematical expressions we are going to derive the first order perturbation expansions of P and U. From the expansions it is easy to see which quantities affect the condition of the updating problem. We now state the main result of this section.

THEOREM 3.2. *Given* $A \in \mathcal{R}^{m\times n}, x \in \mathcal{R}^m$ *and* $y \in \mathcal{R}^n$ *such that* $\operatorname{rank}(A) = \operatorname{rank}(A + xy^T) = n$. *Let*

$$A = QR, \quad A + xy^T = PU$$

be the QR *factorizations of* A *and* $A + xy^T$, *respectively, and define* $u = Q^Tx, v = R^{-T}y$. *Let* $\epsilon_0 > 0$ *be small enough so that the* QR *factorization*

$$A + \epsilon E = (Q + W(\epsilon))(R + G(\epsilon))$$

and

$$A+\epsilon E+(x+\epsilon f)(y+\epsilon g)^T=(P+Z(\epsilon))(U+T(\epsilon))$$

always exist for $\epsilon\in(-\epsilon_0,\epsilon_0)$. *Then*

(3.27)

$$\|T(\epsilon)\|_F\le\sqrt{2}\left[\kappa(A+xy^T)(\|E\|_F+\|x\|_2\|g\|_2)+\frac{\|A+xy^T\|_2\|v\|_2}{(\delta(u,v,x))^{1/2}}\|f\|_2\right]|\epsilon|+O(\epsilon^2),$$

(3.28)

$$\|T(\epsilon)\|_F\le\sqrt{2}\left(\frac{\rho(u,v,x)}{\delta(u,v,x)}\right)^{1/2}[\kappa(A)(\|E\|_F+\|x\|_2\|g\|_2)+\|A\|_2\|v\|_2\|f\|_2]\,|\epsilon|+O(\epsilon^2),$$

(3.29)

$$\|Z(\epsilon)\|_F\le\sqrt{2}\left[\|(A+xy^T)^\dagger\|_2(\|E\|_F+\|x\|_2\|g\|_2)+\frac{\|v\|_2}{(\delta(u,v,x))^{1/2}}\|f\|_2\right]|\epsilon|+O(\epsilon^2)$$

and

(3.30)

$$\|Z(\epsilon)\|_F\le\sqrt{2}\frac{1}{(\delta(u,v,x))^{1/2}}\left[\|A^\dagger\|_2(\|E\|_F+\|x\|_2\|g\|_2)+\|v\|_2\|f\|_2\right]|\epsilon|+O(\epsilon^2)\quad\epsilon\to0.$$

Proof. For $\epsilon\in[-\epsilon_0,\epsilon_0]$, let

$$A(\epsilon)=A+\epsilon E,\quad P(\epsilon)=P+Z(\epsilon),\quad U(\epsilon)=U+T(\epsilon),$$

$$x(\epsilon)=x+\epsilon f,\quad y(\epsilon)=y+\epsilon g,\quad u(\epsilon)=(Q+W(\epsilon))^Tx(\epsilon),\quad v(\epsilon)=(R+G(\epsilon))^{-T}y(\epsilon).$$

Then by the hypotheses and the first differential inequality of (3.5),

$$\begin{aligned}\|T(\epsilon)\|_F&=\|U(\epsilon)-U(0)\|_F=\|\textstyle\int_0^\epsilon dU(\tau)\|_F\le\int_0^{|\epsilon|}\|dU(\tau)\|_F\\&\le\sqrt{2}\int_0^{|\epsilon|}\Big[\kappa(A(\tau)+x(\tau)y(\tau)^T)(\|E\|_F+\|x(\tau)\|_2\|g\|_2)\\&\qquad+\tfrac{\|A(\tau)+x(\tau)y(\tau)^T\|_2\|v(\tau)\|_2}{(\delta(u(\tau),v(\tau),x(\tau)))^{1/2}}\|f\|_2\Big]d\tau.\end{aligned}\tag{3.31}$$

Observe that when $\tau\to0$

$$\begin{aligned}&\kappa(A(\tau)+x(\tau)y(\tau)^T)=\kappa(A+xy^T)+O(\tau),\\&\|A(\tau)+x(\tau)y(\tau)^T\|_2=\|A+xy^T\|_2+O(\tau),\\&1/(\delta(u(\tau),v(\tau),x(\tau))^{1/2}=1/(\delta(u,v,x))^{1/2}+O(\tau),\\&\|x(\tau)\|_2=\|x\|_2+O(\tau),\quad\|v(\tau)\|_2=\|v\|_2+O(\tau).\end{aligned}$$

Hence, from (3.31) we derive the perturbation expansion (3.27).

By the same way we derive (3.28)–(3.30) from the other differential inequalities of (3.5)–(3.6). □

Remark 2. The first terms of the right-hand sides of (3.27)–(3.30) give the first order perturbation bounds for $\|T(\epsilon)\|_F$ and $\|Z(\epsilon)\|_F$, respectively. These bounds show that we can take the quantities

$$c_U(A,x,y)\equiv\left[(\kappa(A+xy^T))^2+\frac{(\|A+xy^T\|_2\|v\|_2)^2}{\delta(u,v,x)}\right]^{1/2}$$

and

$$c_P(A,x,y) \equiv \left[\|(A+xy^T)^\dagger\|_2^2 + \frac{\|v\|_2^2}{\delta(u,v,x)}\right]^{1/2}$$

as the (absolute) condition numbers of the QR updating problem corresponding to U and P, respectively, where u, v, and $\delta(u,v,x)$ are defined by (3.2)–(3.4), and $\delta(u,v,x)$ has the explicit expression (3.25).

Remark 3. Given $A \in \mathcal{R}^{m\times n}, X \in \mathcal{R}^{m\times r}$ and $Y \in \mathcal{R}^{n\times r}$ such that $\text{rank}(A) = \text{rank}(A+XY^T) = n$, find an upper triangular matrix $U \in \mathcal{R}^{n\times n}$ with positive diagonal elements and $P \in \mathcal{R}^{m\times n}$ satisfying $P^TP = I$ such that

$$A + XY^T = PU.$$

This problem can be called rank-r updating of the QR factorization. We note that there is no difficulty to apply the technique of this paper and [12] and [13] to get perturbation bounds for the updated QR factors U and P.

Acknowledgment. I would like to thank the referees for helpful comments and valuable suggestions. I am also grateful to Anders Barrlund for reading and commenting on an earlier version of this paper.

REFERENCES

[1] S. T. Alexander, C.-T. Pan, and R. J. Plemmons, *Analysis of a recursive least squares hyperbolic retation algorithm for signal processing*, Linear Algebra Appl., 98 (1988), pp. 3–40.

[2] A. Barrlund, *How integrals can be used to derive matrix perturbation bounds*, Report UMINF-92.11, ISSN-0348-0542, University of Umeå, Sweden, 1992.

[3] A. W. Bojanczyk, R. P. Brent, P. Van Dooren, and F. R. de Hoog, *A note on downdating the Cholesky factorization*, SIAM J. Sci. Statist. Comput., 8(1987), pp. 210–221.

[4] J. Daniel, W. B. Gragg, L. Kaufman, and G. W. Stewart, *Reorthogonalization and stable algorithms for updating the Gram–Schmidt QR factorization*, Math. Comp., 30 (1976), pp. 772–795.

[5] L. Eldén and H. Park, *Perturbation analysis for block downdating of a Cholesky decomposition*, Report LiTH-MAT-R-93-26, Linköping University, September 1993. Also, Numer. Math., 68 (1994), pp. 457–467.

[6] P. E. Gill, G. H. Golub, W. Murray, and M. A. Saunders, *Methods for modifying matrix factorizations*, Math. Comp., 28 (1974), pp. 505–535.

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

[8] C. -T. Pan, *A perturbation analysis of the problem of downdating a Cholesky factorization*, Linear Algebra Appl., 183 (1993), pp. 103–115.

[9] G. W. Stewart, *Perturbation bounds for the QR factorization of a matrix*, SIAM J. Numer. Anal., 14 (1977), pp. 509–518.

[10] ———, *The effects of rounding error on an algorithm for downdating a Cholesky factorization*, J. Inst. Math. Appl., 23 (1979), pp. 203–213.

[11] G. W. Stewart and J. -G. Sun, *Matrix perturbation theory*, Academic Press, Boston, 1990.

[12] J. -G. Sun, *Perturbation bounds for the Cholesky and QR factorizations*, BIT, 31 (1991), pp. 341–352.

[13] ———, *On perturbation bounds for the* QR *factorization*, Report UMINF-93.07, ISSN-0348-0542, University of Umeå, Sweden, 1993. Also, Linear Algebra Appl., 215 (1995), pp. 95–112.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 776–792, July 1995

006

SMALL-SAMPLE STATISTICAL ESTIMATES FOR MATRIX NORMS*

T. GUDMUNDSSON†, C. S. KENNEY†, AND A. J. LAUB†

Abstract. This paper extends a recent statistically based vector-norm estimator to matrices. The new estimator requires only a few matrix-vector multiplications and can be applied when the matrix is not known explicitly. It is useful for efficiently estimating the sensitivity of vector-valued functions and can be applied to many problems where the power method runs into difficulties. Lower bounds for the probability that an estimate is within a given factor of the correct norm are derived. These bounds are straightforward to compute and show that a very inaccurate estimate is extremely unlikely in most cases. A conservative lower bound has been derived and a tighter bound is given in the form of a conjecture. This conjecture is true in some important special cases and the general case is supported by considerable empirical evidence.

Key words. conditioning, statistical condition estimation, matrix functions, matrix norms

AMS subject classifications. 65F35, 65F30, 15A12

1. Introduction.

A novel method for efficiently estimating the sensitivity of a scalar function at a point was recently introduced in [18]. This method is based on implicitly projecting an approximate gradient of the function onto a uniformly randomly chosen low-order subspace, and then computing the norm of the result. Properly scaled, this gives an estimate of the norm of the gradient, and thereby the condition number of the function at the estimation point. The method requires only the evaluation of the function at this point and a few nearby points.

This paper extends the results of [18] to the estimation of the Frobenius norm of the Jacobian of a general vector-valued function. Thus, let $f : \mathbb{R}^n \to \mathbb{R}^q$ be a function differentiable at a point x, and define the Jacobian at x as

$$J(x) = \frac{\partial f}{\partial x}(x).$$

The norm of the $q \times n$ matrix $J(x)$ is a measure of the sensitivity of f at x [25]. In fact, the Taylor expansion of the function about x,

$$f(x + \delta z) = f(x) + \delta J(x)z + \mathrm{O}(\delta^2),$$

where δ is a small scalar, indicates that if $J(x)$ is large in some sense, then small perturbations in x can result in large perturbations in the value of the function.

The Jacobian is usually not easy to compute explicitly if its dimensions are large, but the Taylor expansion suggests a simple approximation of the product $J(x)z$ for any vector z,

$$J(x)z \approx \frac{f(x + \delta z) - f(x)}{\delta}, \tag{1}$$

where δ is sufficiently small. The method introduced in [18] and generalized in this paper is based on this approximation.

* Received by the editors February 1, 1993; accepted for publication (in revised form) by G. H. Golub May 12, 1994. This research was supported by National Science Foundation grant ECS-9120643, Air Force Office of Scientific Research grant F49620-94-1-0104DEF, and Office of Naval Research grant N00014-92-J-1706.

† Department of Electrical and Computer Engineering, University of California, Santa Barbara, California 93106-9560 (laub@ece.ucsb.edu).

Equation (1) can be used as a basis for the power method for approximating the 2-norm of the Jacobian if f maps square matrices to square matrices of equal dimensions (see, e.g., [17]), but in no other case can this otherwise useful method be employed unless the Jacobian is known explicitly. This is because the power method requires the evaluation of the product of $J^T(x)$ and a vector, and this can usually not be approximated in the same manner as the product of $J(x)$ and a vector.

Our method inherits the benefits of the power method (when it can be applied), namely the use of the finite difference approximation (1) of the Jacobian, but eliminates the difficulty incurred by the transpose step. Moreover, this paper includes a lower bound on the probability that a sample of the estimator is within a given factor of the true norm of the Jacobian. The validity of this lower bound is contingent upon the verity of a certain conjecture, which has been shown to be true in some special cases. A conservative version of the conjecture has also been proved, and the general case is supported by considerable empirical evidence. Similar bounds for the power method have not been derived, but some discussion of the statistical properties of that method can be found in [7] and [20].

Note that the important example of the Jacobian of a function shows that methods for estimating the norm of a matrix implicitly, using only products of the matrix and a vector or approximations thereof, can be very useful. Similar situations arise when various other scalar functions of a matrix are desired, such as the largest or smallest eigenvalue of a positive definite operator [21], the size of the transient "hump" of a matrix exponential [22], the stability radius and distance to uncontrollability of a linear system [16], the norm of a Hankel operator [10], and the structured singular value [8]. Therefore, we assume throughout most of the paper that when the norm of a matrix is to be estimated, the product of the matrix and a vector can be computed exactly. We then return to the case of estimating the sensitivity of vector-valued functions when we discuss specific examples.

In the next section of the paper we introduce the estimator for the norm of a matrix, derive some of its statistical properties, state an important conjecture about the probability of an accurate estimate, and prove a conservative version of that conjecture. In §3 we derive an expression that, according to the conjecture, is a lower bound on the probability of an accurate estimate. This bound depends only on the dimensions of the randomly selected projection subspace. In §4 the estimator is applied to the computation of the sensitivity of some specific functions. Finally, some concluding remarks are made in §5.

In the remainder of this section, we review some important probability distribution functions. A beta distribution with parameters (p_1, p_2) has density function

$$f_z(z) = \frac{1}{B(p_1/2, p_2/2)} z^{p_1/2-1}(1-z)^{p_2/2-1} \quad \text{for } 0 < z < 1,$$

where $B(a_1, a_2)$ is the beta function

$$B(a_1, a_2) = \frac{\Gamma(a_1)\Gamma(a_2)}{\Gamma(a_1 + a_2)}, \tag{2}$$

and Γ is the gamma function [2], [26],

$$\Gamma(c) = \int_0^{+\infty} t^{c-1} e^{-t} dt$$

for $\mathrm{Re}(c) > 0$. If a random variable z has a beta distribution with parameters (p_1, p_2),

then its mean and variance are given by

$$\mathrm{E}(z) = \frac{p_1}{p_1 + p_2},$$
$$\mathrm{Var}(z) = \frac{2p_1p_2}{(p_1 + p_2)^2(p_1 + p_2 + 2)},$$

respectively [9]. A useful interpretation of a random variable z with a beta distribution is as the sum

$$z = \sum_{i=1}^{p_1} u_i^2,$$

where $u_1, u_2, \ldots, u_{p_1}, u_{p_1+1}, \ldots, u_{p_1+p_2}$ are the elements of a vector $u = x/\|x\|_2$, and x is a standard normal $(p_1 + p_2)$-variate normal vector.

A logical extension of the beta distribution to multiple variables exists. Let Z be an $n \times m$ matrix with orthonormal columns and denote the Riemann space[1] of such matrices by $V_{m,n}$. This space is called the *Stiefel manifold*, and studies of many different probability distributions of its elements can be found in the literature [14], [28], [23], [27], [4], [5]. We are especially interested in the uniform distribution, i.e., the Stiefel manifold with elements whose span is uniformly distributed in $\mathbb{R}^n$. A sample matrix from this distribution can be generated from m samples from the standard n-variate normal distribution by forming an orthonormal basis for their span. This can be done very efficiently, for example, by using a simplified QR decomposition [29],[18].

For future reference we note that the joint density of the elements of Z is [14]

$$f_Z(Z) = \frac{1}{\prod_{i=1}^{m} A(n-i+1)} \quad \text{on } V_{m,n},$$

where $A(k)$ is the area of the unit sphere in $\mathbb{R}^k$, given by

$$A(k) = \frac{2\pi^{k/2}}{\Gamma(k/2)}.$$

This density has an important property, which we state in the form of a lemma.

LEMMA 1.1. *The joint density of the elements of Z is invariant under rotations of the coordinate frame.*

Proof. See [23]. □

2. An estimator. In this section we introduce an estimator for the Frobenius norm of a $q \times n$ matrix L. A sample of this estimator is easily computed using a few matrix-vector multiplications, and, moreover, its statistical properties are such that the sample can be expected to be close to the norm of L in most cases.

DEFINITION 2.1. *Let $L \in \mathbb{R}^{q\times n}$ and let Z be a uniformly random matrix on the Stiefel manifold $V_{m,n}$. Define an estimator for $\|L\|_F^2$ by*

$$\psi_L(m) = \frac{n}{m}\|LZ\|_F^2.$$

[1] A Riemann space is a manifold to which a metric is attached [3].

If we denote the rank of L by $r \leq \min(q,n)$ and its nonzero singular values by σ_i, $i = 1,2,\ldots,r$, where

$$\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \cdots \geq \sigma_r > 0,$$

we have the following result.

LEMMA 2.2. *The expected value and variance of* $\psi_L(m)$ *are*

$$\mathrm{E}(\psi_L(m)) = \|L\|_F^2$$

and

$$\mathrm{Var}(\psi_L(m)) = \frac{2(n-m)}{m(n+2)(n-1)}\left(n\sum_{i=1}^{r}\sigma_i^4 - \|L\|_F^4\right),$$

respectively.

The proof is elementary, but rather lengthy, and will be deferred to the Appendix. Note that the variance is independent of the dimension q, but does depend on the rank of L. In fact, the variance has the following important property.

LEMMA 2.3. *The quantity*

$$R = \frac{\mathrm{Var}(\psi_L(m))}{\|L\|_F^4}$$

is maximized when L *is of unit rank.*

Proof. We can rewrite R as

$$R = \frac{2(n-m)}{m(n+2)(n-1)}\left(n\sum_{i=1}^{r}\left(\frac{\sigma_i^2}{\sum_{j=1}^{r}\sigma_j^2}\right)^2 - 1\right).$$

For each value of r, the single extremum of R is easily shown to be at $\sigma_1 = \sigma_2 = \cdots = \sigma_r$, and the largest of those extrema clearly occurs when $r = 1$. □

This suggests (but does not guarantee) that the probability of a sample of the estimator being close to the mean is minimized when L is of unit rank. But before we discuss this further we look more closely at the probability distribution of the estimator.

The probability density function of the estimator $\psi_L(m)$ can in principle be derived from the results of [11] and the density of Z. An expression for it is

$$\begin{aligned} f_{\psi_L}(t) &= \int_{V_{m,n}} f_Z(Z)\delta(t-\psi_L(m))\,dZ \\ &= \frac{1}{\prod_{i=1}^{m}A(n-i+1)}\int_{V_{m,n}}\delta\left(t-\frac{n}{m}\|LZ\|_F^2\right)dZ. \end{aligned} \tag{3}$$

This expression can be reduced to an elliptic integral [13], but since analytical evaluation of such integrals is generally impossible, we must use other approaches to study the statistical properties of the estimator.

We will focus on the properties of the probability that the estimator $\psi_L(m)$ is close to the correct value. By this we mean the probability that a sample is within a factor of the mean, $\Pr(\|L\|_F^2/\alpha^2 \leq \psi_L(m) \leq \alpha^2\|L\|_F^2)$ for $\alpha > 1$. For ease of notation we denote this probability function by $P_L(\alpha)$,

$$P_L(\alpha) = \Pr(\|L\|_F^2/\alpha^2 \leq \psi_L(m) \leq \alpha^2\|L\|_F^2). \tag{4}$$

To simplify the analysis in the sequel, we note that the set of admissible matrices can be reduced considerably. First we need the following.

LEMMA 2.4. *The probability $P_L(\alpha)$ is identical for all matrices L with the same relative singular values, i.e., for all $L \in \Omega$, where*

$$\Omega = \left\{ L : \frac{\sigma_i}{\|L\|_F} \text{ is fixed for } i = 1, 2, \ldots, r \right\}.$$

Proof. Using the singular value decomposition $L = U\Sigma V^T$ of a matrix L, the probability can be written as

$$P_L(\alpha) = \Pr\left(\frac{1}{\alpha^2} \leq \frac{n}{m} \left\| U \frac{\Sigma}{\|L\|_F} V^T Z \right\|_F^2 \leq \alpha^2 \right).$$

Since the Frobenius norm is unitarily invariant and the distribution of Z is invariant under rotations of the coordinate frame (Lemma 1.1), the probability is identical for all L for which $\Sigma/\|L\|_F$ is identical, i.e., for all $L \in \Omega$. □

This defines an equivalence class of matrices, and we can use any particular element to represent the whole class.

COROLLARY 2.5. *The matrix L can, without loss of generality, be taken to be square of dimension n, diagonal of the form*

$$L = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r, 0, \ldots, 0),$$

and of unit Frobenius norm,

$$\sum_{i=1}^{r} \sigma_i^2 = 1.$$

Thus the probability function $P_L(\alpha)$ depends only on the relative distribution of the singular values of L, the accuracy factor α, and the dimensions n and m of Z.

Since a computable expression for the function $P_L(\alpha)$ does not seem to be generally obtainable, our goal is to establish a computable lower bound for the function. Lemma 2.3 suggests that for fixed m and n, the probability function $P_L(\alpha)$ is smaller for the equivalence class of rank 1 matrices than for any other matrix. Moreover, as we will see later, this smallest probability function can easily be computed exactly.

DEFINITION 2.6. *For any fixed pair of integers n and m with $n \geq m$, an $n \times n$ matrix L_0 in the equivalence class of rank* 1 *matrices is called* extremal. *The corresponding estimator $\psi_{L_0}(m)$ is called the* extremal estimator *and the probability function $P_{L_0}(\alpha)$ of the extremal estimator is called the* extremal probability function.

A conservative lower bound on the probability function $P_L(\alpha)$ can be written in terms of the extremal probability function.

THEOREM 2.7. *Let $L \in \mathbb{R}^{q \times n}$ be of rank r, let $P_L(\alpha)$ be the corresponding probability function* (4) *for some value of m, and let $P_{L_0}(\alpha)$ be the extremal probability function for n and m. Then*

$$P_L(\alpha) \geq 1 - r(1 - P_{L_0}(\alpha)),$$

with equality if $r = 1$.

Proof. By Corollary 2.5, L can without loss of generality be assumed to be zero, except for the first r diagonal values. Thus we can assume that only the first r rows of the product LZ are nonzero.

For $i = 1, 2, \ldots, r$, denote the ith row of that product by $(LZ)_i$, the ith row of L by L_i, and define the estimator

$$\psi_L^i(m) = \frac{n}{m}\|(LZ)_i\|_F^2.$$

Let Υ denote the statement

$$\frac{\|LZ\|_F^2}{\alpha^2} \leq \psi_L(m) \leq \alpha^2\|LZ\|_F^2,$$

let Υ_i denote the statements

$$\frac{\|(LZ)_i\|_F^2}{\alpha^2} \leq \psi_L^i(m) \leq \alpha^2\|(LZ)_i\|_F^2$$

for $i = 1, 2, \ldots, r$, and denote the converse of each statement by an overbar. Since

$$\|LZ\|_F^2 = \sum_{i=1}^{r} \|(LZ)_i\|_F^2$$

and

$$\begin{aligned}\psi_L(m) &= \frac{n}{m}\|LZ\|_F^2 \\ &= \sum_{i=1}^{r} \frac{n}{m}\|(LZ)_i\|_F^2 \\ &= \sum_{i=1}^{r} \psi_L^i(m),\end{aligned}$$

then, clearly,

$$\begin{aligned}\Pr(\Upsilon) &\geq \Pr(\Upsilon_1 \cap \Upsilon_2 \cap \cdots \cap \Upsilon_r) \\ &\geq 1 - \sum_{i=1}^{r}(1 - \Pr(\Upsilon_i)).\end{aligned}$$

But each of the statements Υ_i involves the product of Z with a rank 1 matrix, so $\Pr(\Upsilon_i) = P_{L_0}(\alpha)$, the extremal probability, for $i = 1, 2, \ldots, r$, and thus the result follows. □

In fact, we believe that a stronger statement is true, namely, that the extremal probability function bounds all other probability functions from below for fixed n and m. This statement has been shown to be true in some special cases [13], and the general case is supported by Lemma 2.3. More importantly, the considerable empirical evidence that supports the general case shows that the statement can be assumed to be true in practice. For future reference we state this as a conjecture.

CONJECTURE. *Let $L \in \mathbb{R}^{q\times n}$, let $P_L(\alpha)$ be the corresponding probability function* (4) *for some value of m, and let $P_{L_0}(\alpha)$ be the extremal probability function for n and m. Then*

$$P_L(\alpha) \geq P_{L_0}(\alpha),$$

with equality if $r = 1$.

In §4 we discuss some of the empirical evidence supporting this conjecture. Note that since $P_{L_0}(\alpha)$ is very close to 1 in most cases, as we shall see in the following section, the difference between the conservative bound and the conjectured bound is usually very small.

3. The lower bound. This section is devoted to the probability distribution function $\Pr(\psi_{L_0}(m) \leq x)$ of the extremal estimator $\psi_{L_0}(m)$. We will show that it is easily computed from an exact analytical expression, and, therefore, that the function $P_{L_0}(\alpha)$ is easily computed. Note that some of the results in this section are proved for a slightly different estimator and in different notation in [18]. The differences are significant enough to warrant giving the proofs here also.

We assume, without loss of generality, that $L_0 = \operatorname{diag}(1, 0, \ldots, 0)$ of order n, as in Corollary 2.5.

3.1. General results. The probability distribution function of the extremal estimator $\psi_{L_0}(m)$ can be computed recursively.

THEOREM 3.1. *Let $\psi_{L_0}(m)$ be the extremal estimator for some $m \geq 3$, let $n \geq m$, and let $0 \leq b \leq 1$. Then*

$$\Pr(\psi_{L_0}(m) \leq bn/m) = \Pr(\psi_{L_0}(m-2) \leq bn/(m-2)) - \frac{2b^{(m-2)/2}(1-b)^{(n-m)/2}}{(n-m)B(m/2,(n-m)/2)}.$$

Proof. Since the matrix L_0 has only one nonzero singular value, we can write the extremal estimator as

$$\psi_{L_0}(m) = \frac{n}{m}\sum_{k=1}^{m} z_{1k}^2,$$

where z_{1k} are the elements of the first row of the matrix Z. Since this row can be augmented to form a unit n-vector,

$$\psi_{L_0}(m) = \frac{n}{m} z_m,$$

where z_m has a beta distribution with parameters $(m, n-m)$. This enables us to write $\Pr(\psi_{L_0}(m) \leq bn/m) = \Pr(z_m \leq b)$ for $0 \leq b \leq 1$, whence

$$\Pr(\psi_{L_0}(m) \leq bn/m) = \frac{1}{B(m/2,(n-m)/2)} \int_0^b t^{(m-2)/2}(1-t)^{(n-m-2)/2} dt. \tag{5}$$

Using the relationship (2) and standard relations for the gamma function [2], the result is immediate. □

By expanding the result of Theorem 3.1 and evaluating $\Pr(\psi_{L_0}(2) \leq bn/2)$ via (5), we can compute the extremal probability distribution directly for even integers $m \geq 2$ as

$$\Pr(\psi_{L_0}(m) \leq bn/m) = 1 - \sum_{i=1}^{m/2} \frac{2b^{i-1}(1-b)^{n/2-i}}{(n-2i)B(i,n/2-i)}. \tag{6}$$

Similarly, for odd $m \geq 1$, we can write

$$\Pr(\psi_{L_0}(m) \leq bn/m) = \Pr(\psi_{L_0}(1) \leq bn) - \sum_{i=1}^{(m-1)/2} \frac{2b^{i-1/2}(1-b)^{(n-2i-1)/2}}{(n-2i-1)B(i+1/2,(n-2i-1)/2)}, \tag{7}$$

TABLE 1
The probability of a sample of the extremal estimator being within a factor $\alpha = 5$ or $\alpha = 10$ of the mean for different values of n and m.

α	n	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
5	20	0.8474	0.9646	0.9912	0.9978	0.9994
	40	0.8444	0.9627	0.9903	0.9974	0.9992
	60	0.8434	0.9620	0.9900	0.9972	0.9992
	80	0.8429	0.9617	0.9898	0.9972	0.9992
	100	0.8426	0.9615	0.9897	0.9971	0.9992
10	20	0.9234	0.9910	0.9984	0.99986	0.999982
	40	0.9218	0.9905	0.9988	0.99983	0.999977
	60	0.9213	0.9904	0.9987	0.99982	0.999975
	80	0.9211	0.9903	0.9987	0.99982	0.999974
	100	0.9209	0.9902	0.9987	0.99981	0.999973

where $\Pr(\psi_{L_0}(1) \leq bn)$ can be evaluated using (5), but is left unresolved here for clarity. The probability of the extremal estimator $\psi_{L_0}(m)$ being accurate follows immediately from the density function

$$P_{L_0}(\alpha) = \Pr(\psi_{L_0}(m) \leq \alpha^2) - \Pr(\psi_{L_0}(m) \leq \alpha^{-2}) \tag{8}$$

by applying the above with b replaced by $\alpha^2 m/n$ and $m/\alpha^2 n$, respectively.

This lower bound is not necessarily very tight, but since it is usually very close to unity when α is not very small, it guarantees good performance of the estimator in most cases. In Table 1 the probabilities of the extremal estimator being within a factor of $\alpha = 5$ and $\alpha = 10$, respectively, are given for various combinations of n and m. For $m = 1$ the bound does not predict very accurate results, but for larger values of m a sample estimate is almost certainly within a factor of 10 of the correct norm.

If we allow the accuracy factor α to be larger, we can apply the results of [18]. There, it is shown that the function $P_{L_0}(\alpha)$ is very closely approximated for $\alpha > 10$ and all values of n by

$$P_{L_0}(\alpha) \approx 1 - \mu\alpha^{-m}, \tag{9}$$

where μ is on the order of unity. Thus, for example, the probability of the estimator being accurate to within a factor of 100 is approximately $1 - 10^{-2m}$, which approaches 1 very fast as m increases.

3.2. Asymptotic behavior. Using the fact that the estimator becomes less accurate as n increases, we can get a uniform lower bound on the probability of a good estimate for each m and α, independent of n. This may simplify the computation of the lower bound in some cases.

LEMMA 3.2. *Let $P_{L_0}(\alpha)$ be an extremal probability function for some value of m. Then if m is even,*

$$\lim_{n\to+\infty} P_{L_0}(\alpha) = \sum_{i=1}^{m/2} \frac{c_1^{i-1}e^{-c_1} - c_2^{i-1}e^{-c_2}}{\Gamma(i)}$$

and, if m is odd,

$$\lim_{n\to+\infty} P_{L_0}(\alpha) = \frac{2}{\sqrt{\pi}} \int_{\sqrt{c_1}}^{\sqrt{c_2}} e^{-x^2} dx + \sum_{i=1}^{(m-1)/2} \frac{c_1^{i-1/2}e^{-c_1} - c_2^{i-1/2}e^{-c_2}}{\Gamma(i+1/2)},$$

where $c_1 = m\alpha^{-2}/2$ and $c_2 = m\alpha^2/2$.

Proof. The proof is based on three standard results:

$$\lim_{n\to+\infty}\left(1-\frac{c}{n}\right)^{n/2-i} = e^{-c/2}, \tag{10}$$

$$\lim_{n\to+\infty} n^i \frac{\Gamma(n/2-i)}{\Gamma(n/2)} = 2^i, \tag{11}$$

$$\lim_{n\to+\infty} n^{i+1/2} \frac{\Gamma(n/2-i-1/2)}{\Gamma(n/2)} = 2^{i+1/2}. \tag{12}$$

The second equation can be derived using the property $\Gamma(s+1) = s\Gamma(s)$ and the third equation by using the same property, along with expressions for $\Gamma(p+1/2)$ and $\Gamma(p)$, where p is an integer, and Wallis's infinite product for $\pi/2$ [1], [2], [26].

Previously in this section we have shown that, for even m,

$$P_{L_0}(\alpha) = 2\sum_{i=1}^{m/2} \frac{(m/\alpha^2 n)^{i-1}(1-m/\alpha^2 n)^{n/2-i} - (\alpha^2 m/n)^{i-1}(1-\alpha^2 m/n)^{n/2-i}}{(n-2i)B(i,n/2-i)}.$$

Using (10) and (11), along with the relationship (2) between the beta and gamma functions, the first result of the lemma follows immediately. By employing (10) through (12) we can similarly show that the latter part of $P_{L_0}(\alpha)$ for odd m,

$$2\sum_{i=1}^{(m-1)/2} \frac{(m/\alpha^2 n)^{i-1/2}(1-m/\alpha^2 n)^{(n-2i-1)/2} - (\alpha^2 m/n)^{i-1/2}(1-\alpha^2 m/n)^{(n-2i-1)/2}}{(n-2i-1)B(i+1/2,(n-2i-1)/2)},$$

converges to the summation term in the second stated result. Thus, it remains only to derive the limit as $n \to +\infty$ for the remaining part of $P_{L_0}(\alpha)$ for odd m,

$$\Pr(\psi_{L_0}(1) \le m\alpha^2) - \Pr(\psi_{L_0}(1) \le m/\alpha^2).$$

We begin by writing the expression (5) for $\Pr(\psi L_0(1) \le c)$ for some constant c,

$$\Pr(\psi_{L_0}(1) \le c) = \frac{1}{B(1/2,(n-1)/2)} \int_0^{c/n} t^{-1/2}(1-t)^{(n-3)/2} dt.$$

The integral can be written as

$$\frac{2\sqrt{2}}{\sqrt{n}} \int_0^{\sqrt{c/2}} \left(1-\frac{2x^2}{n}\right)^{(n-3)/2} dx,$$

which behaves like

$$\frac{2\sqrt{2}}{\sqrt{n}} \int_0^{\sqrt{c/2}} e^{-x^2} dx$$

when n grows. Equation (12) gives the result as stated. □

The asymptotic bounds corresponding to Table 1 are given in Table 2. We see that those bounds are not very far from the actual lower bounds when n is large.

For larger values of α, we can again use the approximation $P_{L_0}(\alpha) \approx 1-\alpha^{-m}$, as in (9).

TABLE 2
The asymptotic probability of a sample of the extremal estimator being accurate to within a factor α.

α	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
5	0.8415	0.9608	0.9893	0.99697	0.999114
10	0.9203	0.9900	0.9986	0.99980	0.999970

4. Numerical examples. Experiments have shown that the probability of an estimate being off by a large factor is very small in most cases. This is particularly true for matrices that are well conditioned with respect to inversion.[2] In the more interesting case of ill-conditioned matrices, the probability of an accurate estimate is lower, and is in general fairly well estimated by the lower bound determined by the extremal probability function.

In this section we examine how well our estimator approximates the norms of certain matrices. For each of the three examples described we used several different values of m, the dimension of the random subspace, and for each such value we computed the "true" probability function using a large number (10,000) of samples of the estimator. We also examine how well the lower bounds derived in §3 approximate the "true" probability function.

The first example problem is the estimation of the condition number of a matrix, the second problem involves the condition of the matrix exponential map, and the third problem is concerned with the map from the coefficient matrices of the Riccati equation to the solution. The first two problems are also discussed in [18], but there the sensitivity of the entries of each function is considered, while here we estimate the sensitivity of the function itself via the norm of the Jacobian matrix.

Example 1. This example concerns the sensitivity of the solution of a linear system $Ax = b$. The condition number of A with respect to inversion, defined as

$$\kappa = \|A\| \, \|A^{-1}\|$$

for some norm $\|\cdot\|$, is a good indicator of the sensitivity. The norm of A is easy to compute, but the norm of the inverse is much harder [6]. When solving the system, however, we factor A such that x can be computed by solving triangular systems, and once this factorization is accomplished, solving the system is relatively inexpensive. This also means that solving for a few more right-hand side vectors can be done without significantly increasing the computational effort. Thus we can apply our estimator to the estimation of $\|A^{-1}\|$ in a very efficient manner. Note that we do not use the finite difference approximation (1) in this example, because we can use the matrix whose norm we want to estimate directly.

To illustrate, let A be an Ostrowski matrix of order n [24]. That is, A is an upper triangular matrix with -1 on the main diagonal and all of the upper entries equal to 1. Even though all of the eigenvalues of A are -1, this matrix is nearly singular if n is large, with the smallest singular value on the order of 2^{-n+1}.

If we take $n = 30$ as in [18] the Frobenius norm of A^{-1} is approximately 3.58×10^8. For each of four different values of m we computed the "true" probability function $P_{A^{-1}}(\alpha)$ for α between 1 and 5. Then we computed the lower bound using the extremal estimator (see (6)–(8)), as well as the asymptotic lower bound (see Lemma 3.2), and compared those to $P_{A^{-1}}(\alpha)$. The results are shown in Fig. 1, with each graph representing a different value of m. The estimator usually does very well, especially

[2] In the extreme case of an orthogonal matrix L every estimate is correct.

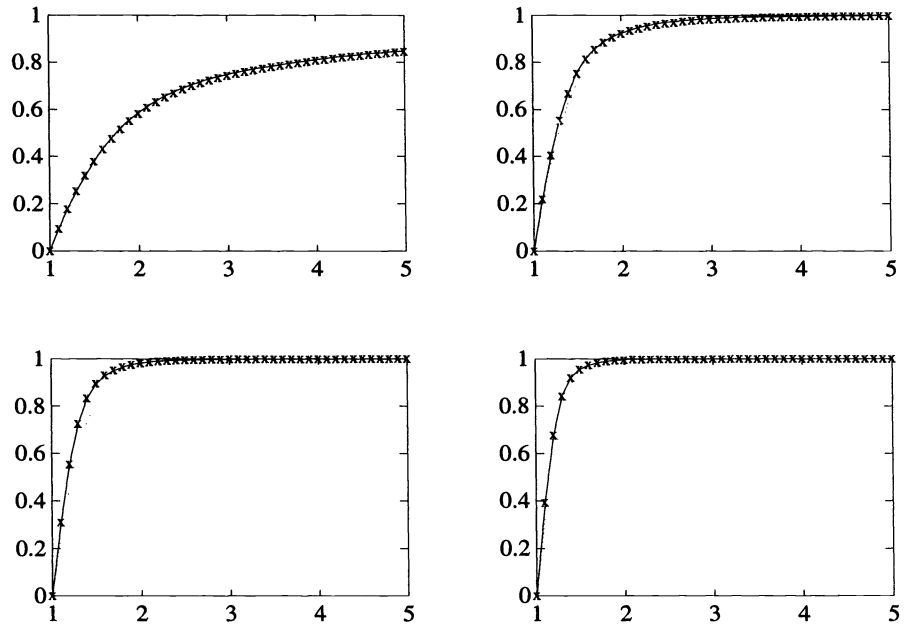


FIG. 1. *Probability functions for Example* 1. *Each graph shows three curves, the "true" probability function* $P_{A^{-1}}(\alpha)$ *(solid), the lower bound using the extremal estimator (x-marked), and the asymptotic lower bound (dotted). The four graphs represent, respectively,* $m = 1$, $m = 4$, $m = 7$, *and* $m = 10$.

for higher values of m, in effect guaranteeing a result within a factor of about 3 of the correct norm. Moreover, the lower bound given by the extremal estimator is almost indistinguishable from the "true" curve in each case, and even the asymptotic lower bound is fairly good. This is consistent with the results reported in [18].

Example 2. In this example we examine the map $X \mapsto e^X$, where X is an $n_0 \times n_0$ matrix. In accordance with our convention, we set $n = n_0^2$ and define the map $f : \mathbb{R}^n \to \mathbb{R}^n$ as

$$f(x) = \operatorname{vec}\left(e^{\operatorname{unvec} x}\right),$$

where the vec operator stacks the columns of a matrix, and the unvec operator reverses the process.

To estimate the condition of the map f at a particular matrix $X = \operatorname{unvec} x$, we form an $n \times m$ matrix Z with orthonormal columns and span randomly chosen on the unit n-sphere. Then we choose a small number $\delta > 0$, and, for each column z_i of Z, we form the difference

$$w_i = \frac{1}{\delta}\operatorname{vec}\left(e^{X+\delta \operatorname{unvec} z_i} - e^X\right).$$

Finally, we form the matrix W with columns w_i, and compute our estimate of the Frobenius norm squared of the Jacobian J of f at X as

$$\psi_J(m) = \frac{n}{m}\|W\|_F^2.$$

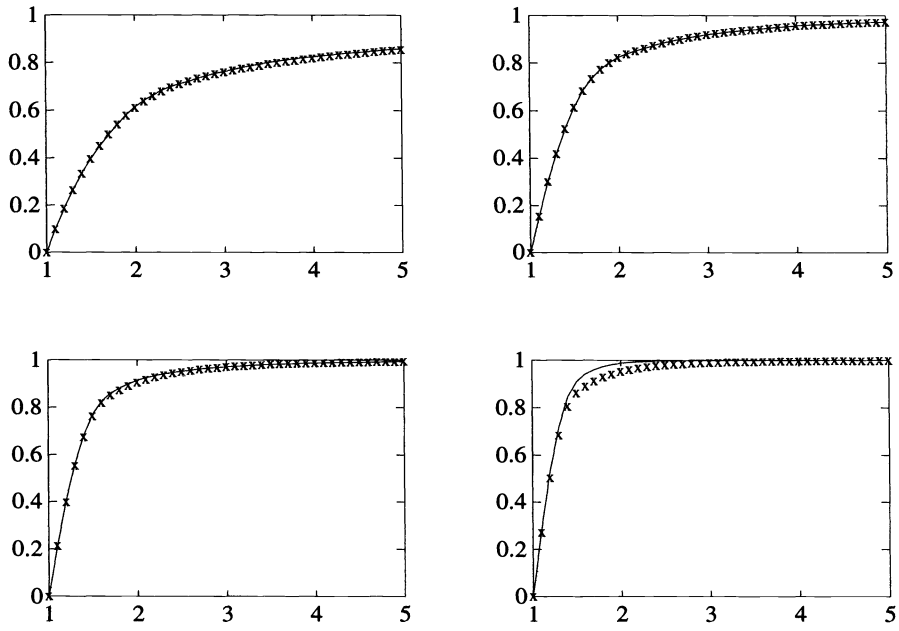


FIG. 2. *Probability functions for Example 2. Each graph shows three curves, the "true" probability function $P_J(\alpha)$ (solid), the lower bound using the extremal estimator (x-marked), and the asymptotic lower bound (dotted). The four graphs represent, respectively, $m = 1$, $m = 2$, $m = 3$, and $m = 4$.*

To test the estimator we applied it to a matrix from [18], originally reproduced from [30],

$$X = \begin{bmatrix} -131 & 19 & 18 \\ -390 & 56 & 54 \\ -387 & 57 & 52 \end{bmatrix},$$

using $\delta = 10^{-12}$ and four different values of m. We computed the "true" probability function $P_J(\alpha)$ for α from 1 to 5, using the Frobenius norm of J. We computed the latter using the power method and a trapezoidal approximation of order 20 to J [17].

The results of this experiment are shown in Fig. 2. We see similar results as in the previous example, namely that for $m > 2$ the estimator is almost guaranteed to be within a factor of 3 of the correct norm. Thus, a good estimate can be obtained with good confidence, using as few as three extra evaluations of the exponential. Again, this is consistent with the results reported in [18].

Example 3. In this last example we look at the sensitivity of solving an algebraic Riccati equation [15],

$$A^T X + XA - XFX + G = 0,$$

where all matrices are of dimension $n_0 \times n_0$, F and G are symmetric, and a unique nonnegative definite solution X is assumed to exist. We define the map $f : \mathbb{R}^{3n_0^2} \to \mathbb{R}^{n_0^2}$ from the entries of the matrices A, F, and G to the entries of the solution X, with the convention that the input vector is $((\operatorname{vec} A)^T, (\operatorname{vec} F)^T, (\operatorname{vec} G)^T)^T$ and the output vector is $\operatorname{vec} X$.

As a specific example we use the matrices

$$A = \begin{bmatrix} 0 & 10^6 \\ 0 & 0 \end{bmatrix}, \quad F = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad G = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

To compute the exact norm of the Jacobian we use the results of [19], namely that the Frobenius norm is equal to the 2-norm of the matrix $D = (D_1^T, D_2^T, D_3^T)^T$, where

$$\begin{aligned} D_1 &= \|G\|_F K^{-1}, \\ D_2 &= \|A\|_F K^{-1}(I \otimes X + X \otimes I)S, \\ D_3 &= \|F\|_F K^{-1}(X \otimes X), \\ K &= ((A - FX)^T \otimes I + I \otimes (A - FX)^T), \end{aligned}$$

S is a permutation matrix depending only on the dimension of X, and $\otimes$ is the Kronecker product operator [12]. In this case the exact norm is approximately 7.07×10^8.

To apply our estimator, we begin by generating m orthonormal vectors z_i of length $n = 3n_0^2$, with $i = 1, 2, \ldots, m$. Then we partition each vector as $z_i = (a_i^T, f_i^T, g_i^T)^T$ with each part of length n_0^2, perturb A, F, and G by the matrices

$$\begin{aligned} \Delta A_i &= \delta \|A\|_F \operatorname{unvec} a_i, \\ \Delta F_i &= \delta \|F\|_F \operatorname{unvec} f_i, \\ \Delta G_i &= \delta \|G\|_F \operatorname{unvec} g_i \end{aligned}$$

for a small scalar δ, and solve the associated perturbed Riccati equations. We use the solutions X_i to form the vectors

$$w_i = \frac{1}{\delta} \operatorname{vec}(X_i - X), \tag{13}$$

form the matrix $W = [\, w_1 \quad w_2 \quad \cdots \quad w_m \,]$, and finally compute the estimator

$$\psi_J(m) = \frac{n}{m} \|W\|_F^2.$$

Some results are shown in Fig. 3 for $\delta = 10^{-13}$ and four different values of m. Note that the lower bound is slightly higher than the theoretical norm in some cases. This is because of the error incurred by using a first-order difference (13) to approximate the product of the Jacobian and a vector. In fact, a smaller value of δ makes the approximation more accurate and eliminates the effect.

Note that in this example an $n_0 \times 3n_0$ matrix $[A \quad F \quad G]$ is mapped into an $n_0 \times n_0$ matrix X, so the power method cannot be applied to the sensitivity estimation without first evaluating the Jacobian explicitly (see the discussion in the Introduction).

5. Conclusions. We have introduced a method for estimating the Frobenius norm of a matrix that is not necessarily known explicitly. The method is based on projecting the matrix onto a randomly generated low-dimensional subspace, and requires only the ability to compute the product of the matrix and a basis for that subspace. This approach is ideal for estimating the sensitivity of vector-valued functions, at the cost of only a few function evaluations.

Numerical experiments have shown that the probability of the estimate being off by a large factor is very small in most cases, in particular when the matrix or

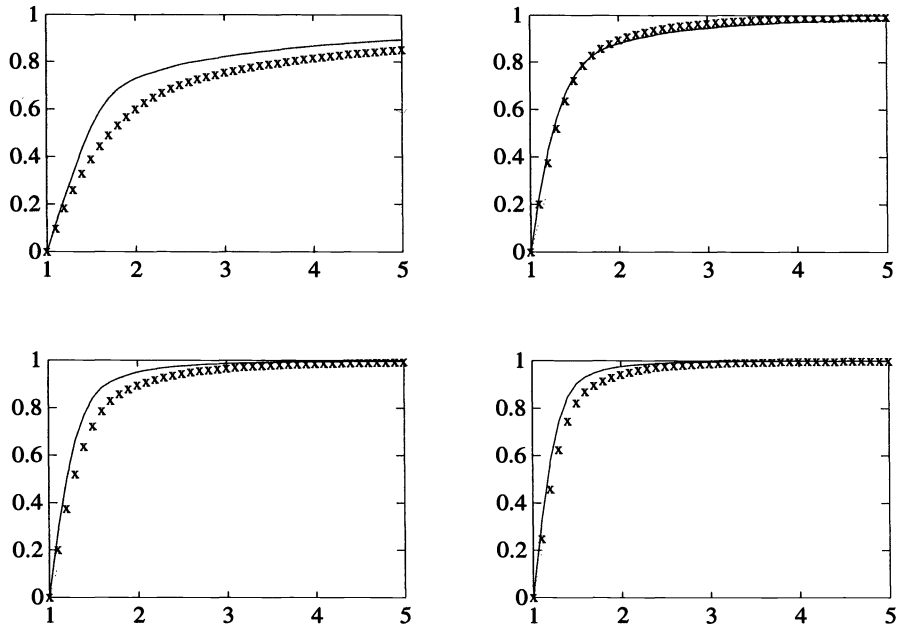


FIG. 3. *Probability functions for Example* 3. *Each graph shows three curves, the "true" probability function* $P_J(\alpha)$ (*solid*), *the lower bound using the extremal estimator* (x-*marked*), *and the asymptotic lower bound* (*dotted*). *The four graphs represent, respectively,* $m = 1$, $m = 2$, $m = 3$, *and* $m = 4$.

function is well conditioned. A conservative lower bound on the probability of an accurate estimate is derived, and a tighter bound is given in the form of a conjecture. This conjecture is supported by the results of numerical experiments, as well as some special cases. A proof of the general case is under investigation.

The lower bound stated in the conjecture is uniform for all estimators involving matrices with domains of a given dimension n and subspaces of a given dimension m. Moreover, this bound often gives a good approximation of the probability of an accurate estimate when the matrix or function is ill conditioned.

The estimator introduced in this paper is a generalization of a previously studied estimator for vector norms [18]. Like that estimator, the new one can be applied to problems where the domain and co-domain of the function are of different dimensions. This is in contrast to the power method, which is only applicable, if the Jacobian (or its transpose) is not available explicitly, to maps between square matrices of equal dimensions.

Some illustrative examples are given, including both the estimation of the condition number of a matrix and the estimation of the sensitivity of vector-valued functions. One of these functions has domain and co-domain of different dimensions.

6. Appendix. Proof of Lemma 2.2. We begin by deriving the mean of the estimator $\psi_L(m)$. Let $L = U\Sigma V^T$ be a singular value decomposition of L, and note that since the Frobenius norm is unitarily invariant, the estimator can be expressed as

$$\psi_L(m) = \frac{n}{m}\|\Sigma V^T Z\|_F^2.$$

The distribution of Z is also unitarily invariant (see Lemma 1.1), so for the purpose of statistical analysis, Z can be replaced by any product WZ, where W is unitary. In particular, Z can be replaced by VZ, giving

$$\psi_L(m) = \frac{n}{m}\|\Sigma Z\|_F^2,$$

where the columns of the new Z are uniformly random on the unit sphere in $\mathbb{R}^n$. Now, the Frobenius norm is easy to expand in terms of the nonzero singular values of L (the first r diagonal values of Σ), giving the mean as

$$\mathrm{E}(\psi_L(m)) = \frac{n}{m}\sum_{i=1}^{r}\sum_{k=1}^{m}\sigma_i^2\mathrm{E}(z_{ik}^2),$$

where z_{ik} is the (i,k)th element of Z. Since the distribution of Z is invariant under rotations, as noted before, the expected value $\mathrm{E}(z_{ik}^2)$ is the same for each pair (i,k). Thus we can replace each of those by the expected value of the first element of the first column of Z. This element has a beta distribution with parameters $(1, n-1)$ (see discussion in §1), and its mean equals $1/n$, so

$$\begin{aligned}\mathrm{E}(\psi_L(m)) &= \frac{n}{m}\sum_{i=1}^{r}\sum_{k=1}^{m}\sigma_i^2\frac{1}{n}\\ &= \|L\|_F^2.\end{aligned}$$

The derivation of the mean is based on the invariance of the distribution of Z under rotations, and a similar approach, albeit somewhat more involved, gives the variance. By expanding the Frobenius norm as before, the second moment of $\|LZ\|_F^2$ can be shown to be

$$\mathrm{E}((\|LZ\|_F^2)^2) = \sum_{i=1}^{r}\sum_{j=1}^{r}\sigma_i^2\sigma_j^2\sum_{k=1}^{m}\sum_{l=1}^{m}\mathrm{E}(z_{ik}^2 z_{jl}^2). \tag{14}$$

By the same argument as before, the terms $\mathrm{E}(z_{ik}^2 z_{jl}^2)$ can have only four different values according to the combinations of i, j, k, and l. Define

$$\begin{aligned}e_1 &= \mathrm{E}(z_{ik}^4),\\ e_2 &= \mathrm{E}(z_{ik}^2 z_{jk}^2) \quad \text{for } i \neq j,\\ e_3 &= \mathrm{E}(z_{ik}^2 z_{il}^2) \quad \text{for } k \neq l,\\ e_4 &= \mathrm{E}(z_{ik}^2 z_{jl}^2) \quad \text{for } i \neq j, k \neq l.\end{aligned}$$

Since the distribution of Z is invariant under unitary transformations, any particular column can be assumed to be the one selected first. Thus, when the other columns are not important, each column can be assumed to be randomly selected from the uniform distribution on the Stiefel manifold. Therefore each z_{ik} can be assumed to be an element of a uniformly random unit vector, and thus has a beta distribution with parameters $(1, n-1)$. Therefore e_1 is the second moment of a beta-distributed random variable

$$e_1 = \frac{3}{n(n+2)}.$$

To evaluate e_2 we expand $z_{ik}^2 z_{jk}^2$ as follows. Let w_1, w_2, and w_3 be three χ^2 variables, the first two with 1 degree of freedom and the third with $n-2$ degrees of freedom. Since z_{ik}^2 and z_{jk}^2 are two elements of the same normalized vector,

$$\begin{aligned} z_{ik}^2 z_{jk}^2 &= \frac{w_1 w_2}{(w_1+w_2+w_3)^2} \\ &= \frac{1}{2}\left(\frac{(w_1+w_2)^2}{(w_1+w_2+w_3)^2} - \frac{w_1^2}{(w_1+w_2+w_3)^2} - \frac{w_2^2}{(w_1+w_2+w_3)^2}\right). \end{aligned}$$

Each term is the square of a beta-distributed random variable, the first one with parameters $(2, n-2)$ and the others with parameters $(1, n-1)$. Thus e_2 is one half the sum of the respective second moments,

$$\begin{aligned} e_2 &= \frac{1}{2}\left(\frac{8}{n(n+2)} - \frac{3}{n(n+2)} - \frac{3}{n(n+2)}\right) \\ &= \frac{1}{n(n+2)}. \end{aligned}$$

The matrix Z can be extended to an orthogonal matrix with the elements of the extension having all the same statistical properties as the elements of Z, so the statistical relationship between the elements of Z and Z^T must be the same. In particular, this indicates that $e_3 = e_2$.

Before we look at e_4, we expand (14) as

$$m(e_1 + (m-1)e_3)\sum_{i=1}^{r}\sigma_i^4 + m(e_2 + (m-1)e_4)\sum_{i=1}^{r}\sum_{\substack{j=1 \\ j\neq i}}^{r}\sigma_i^2\sigma_j^2,$$

and replace each expected value by the correct e_i. This gives the second moment as

$$\frac{m(m+2)}{n(n+2)}\sum_{i=1}^{r}\sigma_i^4 + m\left(\frac{1}{n(n+2)} + (m-1)e_4\right)\sum_{i=1}^{r}\sum_{\substack{j=1 \\ j\neq i}}^{r}\sigma_i^2\sigma_j^2. \tag{15}$$

Since we can add a column to Z without changing the relationship between the previous columns, none of the expected values can depend on m. Therefore we can consider the special case $m = n$ for the purpose of deriving the remaining e_4. In that case

$$\|LZ\|_F^2 = \|L\|_F^2,$$

so the estimator is exact for all samples Z. Thus,

$$\begin{aligned} \mathrm{E}((\|LZ\|_F^2)^2) &= \|L\|_F^4 \\ &= \sum_{i=1}^{r}\sigma_i^4 + \sum_{i=1}^{r}\sum_{\substack{j=1 \\ j\neq i}}^{r}\sigma_i^2\sigma_j^2. \end{aligned}$$

By equating this expression and (15) with $m = n$ we obtain an expression for e_4,

$$e_4 = \frac{n+1}{n+2}\,\frac{1}{n(n-1)},$$

from which the result follows. □

Acknowledgment. We would like to thank an anonymous reviewer for many helpful comments.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, EDS., *Handbook of Mathematical Functions*, U.S. Dept. of Commerce, 1964.

[2] E. ARTIN, *The Gamma Function*, Holt, Rinehart and Winston, New York, 1964; (translated by M. Butler).

[3] É. CARTAN, *Leçons sur la Géométrie des Espaces de Riemann*, 2nd ed., Gauthier-Villars, Paris, 1951; English translation, Math. Sci. Press, Brookline, MA, 1983.

[4] Y. CHIKUSE, *Distributions of orientations on Stiefel manifolds*, J. Multivariate Anal., 33 (1990), pp. 247–264.

[5] ———, *The matrix angular central Gaussian distribution*, J. Multivariate Anal., 33 (1990), pp. 265–274.

[6] A.K. CLINE, C.B. MOLER, G.W. STEWART, AND J.H. WILKINSON, *An estimate for the condition number of a matrix*, SIAM J. Numer. Anal., 16 (1979), pp. 368–375.

[7] J.D. DIXON, *Estimating extremal eigenvalues and condition numbers of a matrix*, SIAM J. Numer. Anal., 20 (1983), pp. 812–814.

[8] J. DOYLE, *Analysis of feedback systems with structured uncertainties*, IEE Proc., Part D, 129 (1982), pp. 242–250.

[9] W. FELLER, *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd ed., Wiley, New York, 1968.

[10] B.A. FRANCIS, *A Course in H_∞ Control Theory*, Springer-Verlag, New York, 1987.

[11] D.T. GILLESPIE, *A theorem for physicists in the theory of random variables*, Amer. J. Phys., 51 (1983), pp. 520–533.

[12] A. GRAHAM, *Kronecker Products and Matrix Calculus with Applications*, Wiley, New York, 1981.

[13] T.T. GUDMUNDSSON, *Implicit Matrix Approximations in Control Theory*, Ph.D. thesis, University of California, ECE Dept., Santa Barbara, September 1992.

[14] A.T. JAMES, *Normal multivariate analysis and the orthogonal group*, Ann. Math. Statist., 25 (1954), pp. 40–75.

[15] C.S. KENNEY AND G. HEWER, *The sensitivity of the algebraic and differential Riccati equations*, SIAM J. Control Opt., 28 (1990), pp. 50–69.

[16] C.S. KENNEY AND A.J. LAUB, *Controllability and stability radii for companion form systems*, Math. Control, Signals, and Sys., 1 (1988), pp. 239–256.

[17] ———, *Condition estimates for matrix functions*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 191–209.

[18] ———, *Small-sample statistical condition estimates for general matrix functions*, SIAM J. Sci. Comput., 15 (1994), pp. 36–61.

[19] C.S. KENNEY, A.J. LAUB, AND M. WETTE, *A stability-enhancing scaling procedure for Schur–Riccati solvers*, Sys. Control Lett., 12 (1989), pp. 241–250.

[20] J. KUCZYŃSKI AND H. WOŹNIAKOWSKI, *Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1094–1122.

[21] D.G. LUENBERGER, *Linear and Nonlinear Programming*, 2nd ed., Addison-Wesley, Reading, MA, 1984.

[22] C.B. MOLER AND C.F. VAN LOAN, *Nineteen dubious ways to compute the exponential of a matrix*, SIAM Rev., 20 (1978), pp. 801–836.

[23] R.J. MUIRHEAD, *Aspects of Multivariate Statistical Theory*, Wiley, New York, 1982.

[24] A.M. OSTROWSKI, *On the spectrum of a one-parametric family of matrices*, J. Reine Angew. Math., band 193, heft 3/4 (1954), pp. 143–160.

[25] J.R. RICE, *A theory of condition*, SIAM J. Numer. Anal., 3 (1966), pp. 287–310.

[26] S. SAKS AND A. ZYGMUND, *Analytic Functions*, Elsevier Publishing Co., Amsterdam, The Netherlands, 1971.

[27] M. SIOTANI, T. HAYAKAWA, AND Y. FUJIKOSHI, *Modern Multivariate Statistical Analysis: A Graduate Course and Handbook*, American Sciences Press, Columbus, OH, 1985.

[28] A.J. STAM, *Limit theorems for uniform distributions on spheres in high-dimensional Euclidean spaces*, J. Appl. Probab., 19 (1982), pp. 221–228.

[29] G.W. STEWART, *The efficient generation of random orthogonal matrices with an application to condition estimators*, SIAM J. Numer. Anal., 17 (1980), pp. 403–409.

[30] R.C. WARD, *Numerical computation of the matrix exponential with accuracy estimate*, SIAM J. Numer. Anal., 14 (1977), pp. 600–610.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 793–810, July 1995

007

DOWNDATING THE SINGULAR VALUE DECOMPOSITION*

MING GU† AND STANLEY C. EISENSTAT‡

Abstract. Let A be a matrix with known singular values and left and/or right singular vectors, and let A' be the matrix obtained by deleting a row from A. We present efficient and stable algorithms for computing the singular values and left and/or right singular vectors of A'. We also show that the problem of computing the singular values of A' is well conditioned when the left singular vectors of A are given, but can be ill conditioned when they are not. Our algorithms reduce the problem to computing the eigendecomposition or singular value decomposition of a matrix that has a simple structure, and solve the reduced problem via finding the roots of a secular equation. Previous algorithms of this type can be unstable and always solve the ill-conditioned problem.

Key words. singular value decomposition, downdating, secular equation

AMS subject classifications. 65F15, 15A18

1. Introduction.

Let

$$A = U\Sigma V^T \tag{1}$$

be the *singular value decomposition* (SVD) of a matrix $A \in \mathbf{R}^{m\times n}$, where $U \in \mathbf{R}^{m\times m}$ and $V \in \mathbf{R}^{n\times n}$ are orthogonal and $\Sigma \in \mathbf{R}^{m\times n}$ is zero except on the main diagonal, which has nonnegative entries in nonincreasing order. The columns of U and V are the *left singular vectors* and the *right singular vectors* of A, respectively, and the diagonal entries of Σ are the *singular values* of A.

In many least squares and signal processing applications (see [5], [21], and [27] and the references therein) we repeatedly update A by appending a row or a column or downdate A by deleting a row or a column. After each update or downdate we must compute the SVD of the resulting matrix. We consider the updating problem in [15] and [17]; here we consider the downdating problem.

Since deleting a column of A is tantamount to deleting a row of A^T, we only consider row deletions. Without loss of generality we further assume that the last row is deleted. Thus we can write

$$A = \begin{pmatrix} A' \\ a^T \end{pmatrix}, \tag{2}$$

where $A' \in \mathbf{R}^{(m-1)\times n}$ is the downdated matrix. Let the SVD of A' be

$$A' = U'\Sigma' V'^T, \tag{3}$$

where $U' \in \mathbf{R}^{(m-1)\times(m-1)}$ and $V' \in \mathbf{R}^{n\times n}$ are orthogonal and $\Sigma' \in \mathbf{R}^{(m-1)\times n}$ is zero except on the main diagonal, which has nonnegative entries in nonincreasing order. We would like to take advantage of our knowledge of the SVD of A when computing the SVD of A'.

* Received by the editors July 2, 1993; accepted for publication by N. J. Higham (in revised form) May 13, 1994. This research was supported in part by U. S. Army Research Office contract DAAL03-91-G-0032.

† Department of Mathematics and Lawrence Berkeley Laboratory, University of California, Berkeley, California 94720 (minggu@math.berkeley.edu).

‡ Department of Computer Science, Yale University, P. O. Box 208285, New Haven, Connecticut 06520-8285 (eisenstat-stan@cs.yale.edu).

We assume that $m > n$; the case $m \leq n$ is similar and is treated in detail in [15] and [16]. We write

$$U = (U_1 \;\; U_2), \quad \Sigma = \begin{pmatrix} D \\ 0 \end{pmatrix} \qquad \text{and} \qquad U' = (U_1' \;\; U_2'), \quad \Sigma' = \begin{pmatrix} D' \\ 0 \end{pmatrix},$$

where $U_1 \in \mathbf{R}^{m \times n}$, $U_2 \in \mathbf{R}^{m \times (m-n)}$, $U_1' \in \mathbf{R}^{(m-1) \times n}$, $U_2' \in \mathbf{R}^{(m-1) \times (m-n-1)}$, and $D, D' \in \mathbf{R}^{n \times n}$ are diagonal matrices. Then (1) and (3) can be rewritten as

$$A = U \Sigma V^T = (U_1 \;\; U_2) \begin{pmatrix} D \\ 0 \end{pmatrix} V^T = U_1 D V^T \tag{4}$$

and

$$A' = U' \Sigma' V'^T = (U_1' \;\; U_2') \begin{pmatrix} D' \\ 0 \end{pmatrix} V'^T = U_1' D' V'^T. \tag{5}$$

There are three downdating problems to consider.

1. Given V, D, and a, compute V' and D'.
2. Given U (or U_1), V, and D, compute U' (or U_1'), V', and D'.
3. Given U (or U_1) and D, compute U' (or U_1') and D'.

We assume that Problem 1 has a solution, i.e., that a is the last row of some matrix A with singular value decomposition (4). Equations (1) and (2) imply that

$$A'^T A' = V' D'^2 V'^T = V(D^2 - zz^T)V^T,$$

where $z = V^T a \in \mathbf{R}^n$. Thus the singular values of A' can found by computing the eigendecomposition $D^2 - zz^T = S \Omega^2 S^T$, where $S \in \mathbf{R}^{n \times n}$ is orthogonal and $\Omega \in \mathbf{R}^{n \times n}$ is nonnegative and diagonal. The diagonal elements of $D' = \Omega$ are the singular values. The right singular vector matrix V' can be computed as VS. We present Algorithm I to solve Problem 1 stably[1] in §§2–3.

Since Problem 1 requires computing the eigendecomposition of $D^2 - zz^T$, small perturbations in V, D, and a can cause large perturbations in V' and D'. We analyze the ill-conditioning of the singular values in §6. Our perturbation results are similar to those of Stewart [26] in the context of downdating the Cholesky/QR factorization.

Problems 2 and 3 always have a solution. We show that there exists a column orthogonal matrix $X \in \mathbf{R}^{(m-1) \times n}$ such that

$$A' = XCV^T,$$

where $C \in \mathbf{R}^{n \times n}$ is given by

$$C = \left(I - \frac{1}{1+\mu} u_1 u_1^T \right) D,$$

with u_1 a vector and $\mu \geq 0$ a scalar. The singular values of A' can found by computing the singular value decomposition $C = Q \Omega W^T$, where Q, $W \in \mathbf{R}^{n \times n}$ are orthogonal

[1] We use the definition of stability in Stewart [25, pp. 75–76]. Let $\mathcal{F}(\mathcal{X})$ be a function of the input data $\mathcal{X}$. We say that an algorithm for computing $\mathcal{F}(\mathcal{X})$ is *stable* if its output is a small perturbation of $\mathcal{F}(\bar{\mathcal{X}})$, where $\bar{\mathcal{X}}$ is a small perturbation of $\mathcal{X}$. This notion of stability is similar to that of *mixed stability* [2], [3] and is used in the context of downdating least squares solutions and Cholesky/QR factorizations [2], [3], [22], [26].

and $\Omega \in \mathbf{R}^{n \times n}$ is nonnegative and diagonal. The diagonal elements of $D' = \Omega$ are the singular values. The left singular vector matrix U_1' can be computed as XQ. The right singular vector matrix V' can be computed as VW. We present Algorithm II to solve Problems 2 and 3 stably in §§4–5.

For Problems 2 and 3 the singular values are well conditioned with respect to perturbations in the input data, whereas the singular vectors can be very sensitive to such perturbations (see §4.1).

Bunch and Nielsen [5] also reduce Problem 1 to computing the eigendecomposition of $D^2 - zz^T$, but their scheme for finding this eigendecomposition is based on results from [6] and [11] and can be unstable [5], [6]. They solve Problem 2 by reducing it to Problem 1, which risks solving a well-conditioned problem using an ill-conditioned process.

Algorithm I solves Problem 1 in $O(n^3)$ time, and Algorithm II solves Problems 2 and 3 in $O(mn^2)$ time when U_1 is given. As with the SVD updating algorithm in [15] and [17], Algorithm I can be accelerated using the fast multipole method of Carrier, Greengard, and Rokhlin [7], [14] to solve Problem 1 in $O(n^2 \log_2^2 \epsilon)$ time, and Algorithm II can be accelerated to solve Problems 2 and 3 in $O(mn \log_2^2 \epsilon)$ time, where ϵ is the machine precision. This is an important advantage for large matrices. Since the techniques are essentially the same as those in [15] and [17], we do not elaborate on this issue.

We take the usual model of arithmetic:[2]

$$fl(\alpha \circ \beta) = (\alpha \circ \beta)\,(1 + \nu),$$

where α and β are floating point numbers; $\circ$ is one of $+$, $-$, $\times$, and $\div$; $fl(\alpha \circ \beta)$ is the floating point result of the operation $\circ$; and $|\nu| \leq \epsilon$. We also require that

$$fl(\sqrt{\alpha}) = \sqrt{\alpha}\,(1 + \nu)$$

for any positive floating point number α. For simplicity we ignore the possibility of overflow and underflow.

2. Solving Problem 1. From (2), (4), and (5) we have

$$A \equiv \begin{pmatrix} A' \\ a^T \end{pmatrix} = U \begin{pmatrix} D \\ 0 \end{pmatrix} V^T \quad \text{and} \quad A' = U' \begin{pmatrix} D' \\ 0 \end{pmatrix} V'^T,$$

so that

$$VD^2V^T = A^TA = A'^TA' + aa^T = V'D'^2V'^T + aa^T.$$

Letting $z = V^Ta$, this equation can be rewritten as

$$V'D'^2V'^T = V\left(D^2 - zz^T\right)V^T. \tag{6}$$

Thus the eigenvalues of $D^2 - zz^T$ are the diagonal elements of D'^2 and must be nonnegative. If $S\Omega^2S^T$ is the eigendecomposition of $D^2 - zz^T$, then $V' = VS$ and $D' = \Omega$.

[2] This model excludes machines like the CRAY and CDC Cyber that do not have a guard digit. Algorithms I and II can easily be modified for such machines.

Algorithm I uses the scheme in §3 to compute a *numerical eigendecomposition* $\tilde{S}\tilde{D}'^2\tilde{S}^T$ satisfying

$$\tilde{S} = \bar{S} + O(\epsilon) \quad \text{and} \quad \tilde{D}' = \bar{D}' + O(\epsilon\|D\|_2),$$

where the eigendecomposition

$$\bar{D}^2 - \bar{z}\bar{z}^T = \bar{S}\bar{D}'^2\bar{S}^T$$

is exact and

$$\bar{D} = D + O(\epsilon\|D\|_2) \quad \text{and} \quad \bar{z} = z + O(\epsilon\|D\|_2).$$

It then computes a right singular vector matrix satisfying

$$\tilde{V}' = V\bar{S} + O(\epsilon).$$

Since V is orthogonal, the error in z can be attributed to an error in a:

$$\bar{a} = V\bar{z} = a + O(\epsilon\|D\|_2).$$

Thus $\bar{D}'$ and $V\bar{S}$ are the exact solution to Problem 1 with slightly perturbed input data V, $\bar{D}$, and $\bar{a}$, so that Algorithm I is stable.

Since small perturbations in D and a can cause large perturbations in D' and S, it follows that $\tilde{D}'$ and $\tilde{S}$ can be very different from D' and S, respectively. We analyze the ill-conditioning of the singular values in §6.

The scheme in §3 takes $O(n^2)$ time, and computing VS takes $O(n^3)$ time. Thus the total time for Algorithm I is $O(n^3)$.

Barlow, Zha, and Yoon [1] compute the eigendecomposition of $D^2 - zz^T$ by using a variant of the LINPACK downdating procedure [10] to "reduce" D to bidiagonal form and then solving the bidiagonal singular value problem. The total time appears to be at least as large as that for Algorithm I.

3. Computing the eigendecomposition of $D^2 - zz^T$. In this section we present an algorithm for computing the eigendecomposition of $D^2 - zz^T$, where $D = \operatorname{diag}(d_1, \ldots, d_k)$, with $d_1 \geq \cdots \geq d_k \geq 0$, and $z = (\zeta_1, \ldots, \zeta_k)^T$. In light of (6) we assume that the eigenvalues of $D^2 - zz^T$ are nonnegative.

We further assume that D and z satisfy

$$d_k > 0, \quad d_i - d_{i+1} \geq \theta\|D\|_2, \quad \text{and} \quad |\zeta_i| \geq \theta\|D\|_2, \tag{7}$$

where θ is a small multiple of ϵ to be specified in §3.4. Any matrix of the form $D^2 - zz^T$ can be stably reduced to one that satisfies these conditions by using the deflation procedure described in §3.5.

3.1. Properties of the eigendecomposition. The following lemma characterizes the eigenvalues and eigenvectors of $D^2 - zz^T$.

LEMMA 3.1 (Bunch and Nielsen [5]). *The eigenvalues of $D^2 - zz^T$ are nonnegative if and only if $z^T D^{-2} z \leq 1$.*

Assume that $z^T D^{-2} z \leq 1$. Then the eigendecomposition of $D^2 - zz^T$ can be written as $S\Omega^2 S^T$, where $S = (s_1, \ldots, s_k)$ and $\Omega = \operatorname{diag}(\omega_1, \ldots, \omega_k)$. The eigenvalues $\{\omega_i^2\}_{i=1}^k$ satisfy the secular equation

$$f_1(\omega) \equiv -1 + \sum_{j=1}^{k} \frac{\zeta_j^2}{d_j^2 - \omega^2} = 0 \tag{8}$$

and the interlacing property

$$d_1 > \omega_1 > d_2 > \cdots > d_k > \omega_k \geq 0. \tag{9}$$

The eigenvectors are given by

$$s_i = \left(\frac{\zeta_1}{d_1^2 - \omega_i^2}, \ldots, \frac{\zeta_k}{d_k^2 - \omega_i^2} \right)^T \Bigg/ \sqrt{\sum_{j=1}^{k} \frac{\zeta_j^2}{(d_j^2 - \omega_i^2)^2}}. \tag{10}$$

Conversely, given D and the eigenvalues of $D^2 - \hat{z}\hat{z}^T$, we can reconstruct $\hat{z}$.

LEMMA 3.2. *Given a diagonal matrix* $D = \operatorname{diag}(d_1, \ldots, d_k)$ *and a set of numbers* $\{\hat{\omega}_i\}_{i=1}^k$ *satisfying the interlacing property*

$$d_1 > \hat{\omega}_1 > d_2 > \cdots > d_k > \hat{\omega}_k \geq 0, \tag{11}$$

there exists a vector $\hat{z} = (\hat{\zeta}_1, \ldots, \hat{\zeta}_k)^T$ *such that the eigenvalues of* $D^2 - \hat{z}\hat{z}^T$ *are* $\{\hat{\omega}_i^2\}_{i=1}^k$. *The components of* $\hat{z}$ *are given by*

$$|\hat{\zeta}_i| = \sqrt{(d_i^2 - \hat{\omega}_k^2) \prod_{j=1}^{i-1} \frac{\hat{\omega}_j^2 - d_i^2}{d_j^2 - d_i^2} \prod_{j=i}^{k-1} \frac{\hat{\omega}_j^2 - d_i^2}{d_{j+1}^2 - d_i^2}}, \qquad 1 \leq i \leq k, \tag{12}$$

where the sign of $\hat{\zeta}_i$ *can be chosen arbitrarily.*

Proof. This is Löwner's construction [20] of $\hat{z}$ given $-D^2$ and the eigenvalues of $(-D^2) + \hat{z}\hat{z}^T$. □

3.2. Computing the eigenvectors. In practice we can only hope to compute an approximation $\hat{\omega}_i$ to ω_i. But problems can arise if we approximate s_i by

$$\hat{s}_i = \left(\frac{\zeta_1}{d_1^2 - \hat{\omega}_i^2}, \ldots, \frac{\zeta_k}{d_k^2 - \hat{\omega}_i^2} \right)^T \Bigg/ \sqrt{\sum_{j=1}^{k} \frac{\zeta_j^2}{(d_j^2 - \hat{\omega}_i^2)^2}}$$

(i.e., replace ω_i by $\hat{\omega}_i$ in (10), as in [5]). For even if $\hat{\omega}_i$ is close to ω_i, the approximate ratio $\zeta_j/(d_j^2 - \hat{\omega}_i^2)$ can still be very different from the exact ratio $\zeta_j/(d_j^2 - \omega_i^2)$, resulting in a unit eigenvector very different from s_i. After all the $\{\hat{\omega}_i\}_{i=1}^k$ are computed and all the corresponding eigenvectors are approximated in this manner, the resulting eigenvector matrix may not be orthogonal.

But Lemma 3.2 allows us to overcome this problem (cf. [18]). After we have computed all the approximations $\{\hat{\omega}_i\}_{i=1}^k$, we find a *new* vector $\hat{z}$ such that $\{\hat{\omega}_i^2\}_{i=1}^k$ are the *exact* eigenvalues of $D^2 - \hat{z}\hat{z}^T$ and then use (10) to compute the eigenvectors of $D^2 - \hat{z}\hat{z}^T$. Note that each difference

$$\hat{\omega}_j^2 - d_i^2 = (\hat{\omega}_j - d_i)(\hat{\omega}_j + d_i) \quad \text{and} \quad d_j^2 - d_i^2 = (d_j - d_i)(d_j + d_i)$$

in (12) can be computed to high relative accuracy, as can each ratio and each product. Thus $|\hat{\zeta}_i|$ can be computed to high relative accuracy. We choose the sign of $\hat{\zeta}_i$ to be the sign of ζ_i. Substituting the *exact* eigenvalues $\{\hat{\omega}_i^2\}_{i=1}^k$ and the computed $\hat{z}$ into (10), each eigenvector of $D^2 - \hat{z}\hat{z}^T$ can also be computed to componentwise high relative accuracy. Consequently, after all the singular vectors of $D^2 - \hat{z}\hat{z}^T$ are computed, the eigenvector matrix will be numerically orthogonal.

To ensure the existence of $\hat{z}$, we need $\{\hat{\omega}_i\}_{i=1}^k$ to satisfy the interlacing property (11). But since $\{\omega_i\}_{i=1}^k$ satisfy the same interlacing property (see (9)), this is only an accuracy requirement on $\{\hat{\omega}_i\}_{i=1}^k$ and is not an additional restriction on $D^2 - zz^T$.

We use the eigendecomposition of $D^2 - \hat{z}\hat{z}^T$ as an approximation to the eigendecomposition of $D^2 - zz^T$. This is stable as long as $\hat{z}$ is close to z.

3.3. Finding the eigenvalues. To guarantee that $\hat{z}$ is close to z, we must ensure that the approximations $\{\hat{\omega}_i\}_{i=1}^k$ to the singular values are sufficiently accurate. The key is the stopping criterion for the root-finder, which requires a slight reformulation of the secular equation (cf. [5], [18]).

Consider the root $\omega_i \in (d_{i+1}, d_i)$, for $1 \le i \le k-1$; the root $\omega_k \in [0, d_k)$ is treated in a similar manner.

First assume that[3] $\omega_i \in \left(d_{i+1}, \frac{d_i+d_{i+1}}{2}\right)$. Let $\delta_j = d_j - d_{i+1}$, and let

$$\psi_i(\xi) \equiv \sum_{j=1}^{i} \frac{\zeta_j^2}{(\delta_j - \xi)(d_j + d_{i+1} + \xi)} \quad \text{and} \quad \phi_i(\xi) \equiv \sum_{j=i+1}^{k} \frac{\zeta_j^2}{(\delta_j - \xi)(d_j + d_{i+1} + \xi)}.$$

Setting $\omega = d_{i+1} + \xi$, we seek the root $\xi_i = \omega_i - d_{i+1} \in (0, \delta_i/2)$ of the reformulated secular equation

$$g_i(\xi) \equiv f_1(\xi + d_{i+1}) = -1 + \psi_i(\xi) + \phi_i(\xi) = 0.$$

Note that we can compute each ratio $\zeta_j^2/((\delta_j - \xi)(d_j + d_{i+1} + \xi))$ in $g_i(\xi)$ to high relative accuracy for any $\xi \in (0, \delta_i/2)$. Indeed, either $\delta_j - \xi$ is a sum of negative terms or $|\xi| \le |\delta_j|/2$, and $d_j + d_{i+1} + \xi$ is a sum of positive terms. Thus, since both $\psi_i(\xi)$ and $\phi_i(\xi)$ are sums of terms of the same sign, we can bound the error in computing $g_i(\xi)$ by

$$\eta k(1 + |\psi_i(\xi)| + |\phi_i(\xi)|),$$

where η is a small multiple of ϵ that is independent of k and ξ.

Next we assume that $\omega_i \in \left[\frac{d_i+d_{i+1}}{2}, d_i\right)$. Let $\delta_j = d_j - d_i$, and let

$$\psi_i(\xi) \equiv \sum_{j=1}^{i} \frac{\zeta_j^2}{(\delta_j - \xi)(d_j + d_i + \xi)} \quad \text{and} \quad \phi_i(\xi) \equiv \sum_{j=i+1}^{k} \frac{\zeta_j^2}{(\delta_j - \xi)(d_j + d_i + \xi)}.$$

Setting $\omega = d_i + \xi$, we seek the root $\xi_i = \omega_i - d_i \in [\delta_{i+1}/2, 0)$ of the equation

$$g_i(\xi) \equiv f_1(\xi + d_i) = -1 + \psi_i(\xi) + \phi_i(\xi) = 0.$$

For any $\xi \in [\delta_{i+1}/2, 0)$, we can compute each ratio $\zeta_j^2/((\delta_j - \xi)(d_j + d_i + \xi))$ to high relative accuracy (either $\delta_j - \xi$ is a sum of positive terms or $|\xi| \le |\delta_j|/2$, and $d_j + d_i + \xi = d_j + (d_i + \xi)$, where $|\xi| \le d_i/2$), and we can bound the error in computing $g_i(\xi)$ as before.

In practice a root-finder cannot make any progress at a point ξ where it is impossible to determine the sign of $g_i(\xi)$ numerically. Thus we propose the stopping criterion

$$|g_i(\xi)| \le \eta k(1 + |\psi_i(\xi)| + |\phi_i(\xi)|), \tag{13}$$

[3] This condition can easily be checked by computing $f_1\left(\frac{d_i+d_{i+1}}{2}\right)$. If $f_1\left(\frac{d_i+d_{i+1}}{2}\right) > 0$, then $\omega_i \in \left(d_{i+1}, \frac{d_i+d_{i+1}}{2}\right)$, otherwise $\omega_i \in \left[\frac{d_i+d_{i+1}}{2}, d_i\right)$.

where, as before, the right-hand side is an upper bound on the round-off error in computing $g_i(\xi)$. Note that for each i there is at least one floating point number that satisfies this stopping criterion numerically, namely, $fl(\xi_i)$.

We have not specified the scheme for finding the root of $g(\xi)$. We can use the bisection method or the rational interpolation strategies in [4], [5], [13], and [19]. What is most important is the stopping criterion and the fact that, with the reformulation of the secular equation given above, we can find a ξ that satisfies it.

3.4. Numerical stability. We now show that the vector $\hat{z}$ defined in (12) is close to z.

THEOREM 3.3. *If* $\theta = 2\eta k^2$ *in* (7) *and each* $\hat{\xi}_i$ *satisfies* (13), *then*

$$|\hat{\zeta}_i - \zeta_i| \leq 4\eta k^2 \|z\|_2, \qquad 1 \leq i \leq k. \tag{14}$$

The proof is nearly identical to that of the analogous result in [18]. As argued there, the factor k^2 in θ and (14) is likely to be $O(k)$ in practice.

3.5. Deflation. We now show that we can stably reduce $D^2 - zz^T$ to a matrix of the same form that further satisfies

$$d_k > 0, \quad d_i - d_{i+1} \geq \theta\|D\|_2, \quad \text{and} \quad |\zeta_i| \geq \theta\|D\|_2,$$

where θ is specified in §3.4. Similar reductions appear in [5] and [9].

Partition D and z as

$$D = \begin{pmatrix} D_1 & \\ & d_k \end{pmatrix} \quad \text{and} \quad z = \begin{pmatrix} z_1 \\ \zeta_k \end{pmatrix}.$$

First assume that $d_k = 0$. Since $D^2 - zz^T$ is nonnegative definite, its diagonal elements must be nonnegative, so that $d_k^2 - \zeta_k^2 \geq 0$. Thus $\zeta_k = 0$ and

$$D^2 - zz^T = \begin{pmatrix} D_1^2 - z_1 z_1^T & \\ & 0 \end{pmatrix}.$$

The eigenvalue 0 can be deflated, and the matrix $D_1^2 - z_1 z_1^T$ has nonnegative eigenvalues and is of the same form but of smaller dimensions. This reduction is exact.

In the following reductions we assume that $d_k > 0$. Recall from Lemma 3.1 that the eigenvalues of $D^2 - zz^T$ are nonnegative if and only if

$$\sum_{i=1}^{k} \frac{\zeta_i^2}{d_i^2} \leq 1. \tag{15}$$

Assume that $|\zeta_i| < \theta\|D\|_2$. We illustrate the reduction for $i = k$. Changing ζ_k to 0 perturbs z by $O(\theta\|D\|_2)$. In the perturbed matrix

$$\begin{pmatrix} D_1^2 - z_1 z_1^T & \\ & d_k^2 \end{pmatrix},$$

the eigenvalue d_k^2 can be deflated, and the matrix $D_1^2 - z_1 z_1^T$ satisfies (15) and is of the same form but of smaller dimensions. This reduction is stable.

Now assume that $d_i - d_{i+1} < \theta\|D\|_2$. We illustrate the reduction for $i = k - 1$. Changing d_k to d_{k-1} perturbs D by $O(\theta\|D\|_2)$. Let G be a Givens rotation in the

$(k-1,k)$ plane such that $(Gz)_k = 0$. Then when we symmetrically apply G to the perturbed matrix, we get

$$G\left(\begin{pmatrix} D_1 & \\ & d_{k-1} \end{pmatrix}^2 - \begin{pmatrix} z_1 \\ \zeta_k \end{pmatrix}\begin{pmatrix} z_1 \\ \zeta_k \end{pmatrix}^T\right) G^T = \begin{pmatrix} D_1^2 - \check{z}_1\check{z}_1^T & \\ & d_{k-1}^2 \end{pmatrix},$$

where $\check{z}_1 = \left(\zeta_1, \ldots, \zeta_{k-2}, \sqrt{\zeta_{k-1}^2 + \zeta_k^2}\right)^T$. The eigenvalue d_{k-1}^2 can be deflated, and the matrix $D_1^2 - \check{z}_1\check{z}_1^T$ satisfies (15) and is of the same form but of smaller dimensions. This reduction is also stable.

4. Solving Problems 2 and 3. In this section we present an algorithm that solves Problems 2 and 3 by reducing them to the problem of finding the singular value decomposition of a simple matrix.

4.1. The algorithm. Partition U_1 and U_2 as

$$U_1 = \begin{pmatrix} U_{11} \\ u_1^T \end{pmatrix} \quad \text{and} \quad U_2 = \begin{pmatrix} U_{12} \\ u_2^T \end{pmatrix},$$

where $U_{11} \in \mathbf{R}^{(m-1)\times n}$, $u_1 \in \mathbf{R}^n$, $U_{12} \in \mathbf{R}^{(m-1)\times(m-n)}$, and $u_2 \in \mathbf{R}^{m-n}$. Then from (2) and (4) we get

$$A' = (\,U_{11} \quad U_{12}\,)\begin{pmatrix} D \\ 0 \end{pmatrix} V^T = U_{11}DV^T \quad \text{and} \quad a^T = u_1^T DV^T. \tag{16}$$

The decomposition of A' in (16) is almost a singular value decomposition — U_{11} is close to being column orthogonal since it is obtained by deleting the last row from U_1. In the following we decompose U_{11} into a product of an $(m-1)\times n$ column orthogonal matrix and a simple $n \times n$ matrix. To this end we will need a scalar $\mu \geq 0$ and a vector $x \in \mathbf{R}^{m-1}$ such that $\|u_1\|^2 + \mu^2 = 1$ and the matrix

$$Y = \begin{pmatrix} U_{11} & x \\ u_1^T & \mu \end{pmatrix} \tag{17}$$

is column orthogonal. We show how to compute Y in §4.2.

Note that if $\mu = 1$, then $u_1 = 0$ and $x = 0$, so that U_{11} is column orthogonal. In general $\mu \neq 1$, but we can orthogonally transform the rows of Y so that $\mu = 1$. The matrix

$$H = \begin{pmatrix} I - \dfrac{1}{1+\mu}u_1u_1^T & u_1 \\ -u_1^T & \mu \end{pmatrix}$$

is orthogonal and $(u_1^T, \mu)H = (0, \ldots, 0, 1)^T$. Since YH is column orthogonal, it follows that

$$YH = \begin{pmatrix} U_{11}\left(I - \dfrac{1}{1+\mu}u_1u_1^T\right) - x\,u_1^T & U_{11}u_1 + \mu\,x \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} X & 0 \\ 0 & 1 \end{pmatrix}, \tag{18}$$

where

$$X = U_{11}\left(I - \frac{1}{1+\mu}u_1u_1^T\right) - x\,u_1^T$$

is column orthogonal.[4] Thus

$$(U_{11} \quad x) = (U_{11} \quad x)\, HH^T = (X \quad 0)\, H^T = X \begin{pmatrix} I - \dfrac{1}{1+\mu} u_1 u_1^T & -u_1 \end{pmatrix},$$

which implies that

$$U_{11} = X\left(I - \frac{1}{1+\mu} u_1 u_1^T\right). \tag{19}$$

Plugging (19) into (16), we get

$$A' = X\left(I - \frac{1}{1+\mu} u_1 u_1^T\right) DV^T \equiv XCV^T. \tag{20}$$

Let $Q\,\Omega\,W^T$ be the SVD of C, where Q, $W \in \mathbf{R}^{n\times n}$ are orthogonal and $\Omega \in \mathbf{R}^{n\times n}$ is nonnegative and diagonal. Substituting into (20), we get

$$A' = XQ\,\Omega\,W^T V^T = (XQ)\,\Omega\,(VW)^T. \tag{21}$$

Comparing with (5), we have $U_1' = XQ$, $D' = \Omega$, and $V' = VW$. We specify U_2' in §4.2.

Algorithm II computes a numerically column orthogonal matrix $\tilde{Y}$ and a *numerical* singular value decomposition $\tilde{Q}\,\tilde{\Omega}\,\tilde{W}^T$ satisfying (see §4.2 and §5)

$$\tilde{Y} = \bar{Y} + O(\epsilon), \quad \tilde{Q} = \bar{Q} + O(\epsilon), \quad \tilde{\Omega} = \bar{\Omega} + O(\epsilon\|D\|_2), \quad \tilde{W} = \bar{W} + O(\epsilon), \tag{22}$$

where

$$\bar{Y} = \begin{pmatrix} \bar{U}_{11} & \bar{x} \\ \bar{u}_1^T & \bar{\mu} \end{pmatrix}$$

is a column orthogonal matrix with

$$\bar{U}_{11} = U_{11} + O(\epsilon), \quad \bar{u}_1 = u_1 + O(\epsilon), \quad \text{and} \quad \bar{\mu} = \mu + O(\epsilon)$$

and

$$\bar{C} \equiv \left(I - \frac{1}{1+\bar{\mu}} \bar{u}_1 \bar{u}_1^T\right) \bar{D} = \bar{Q}\,\bar{\Omega}\,\bar{W}$$

is an exact SVD with

$$\bar{D} = D + O(\epsilon\|D\|_2).$$

Let

$$\bar{X} = \bar{U}_{11}\left(I - \frac{1}{1+\bar{\mu}} \bar{u}_1 \bar{u}_1^T\right) - \bar{x}\,\bar{u}_1^T.$$

Algorithm II then computes numerical approximations to U' and V' satisfying

$$\tilde{U}_1' = \bar{X}\bar{Q} + O(\epsilon), \quad \tilde{U}_2' = \bar{U}_2' + O(\epsilon), \quad \text{and} \quad \tilde{V}' = V\bar{W} + O(\epsilon), \tag{23}$$

[4] Paige [22] has proven similar relations.

where $(\bar{X}\bar{Q}, \bar{U}_2') \in \mathbf{R}^{(m-1)\times(m-1)}$ is orthogonal (see §4.2). Since $\bar{X}\bar{Q}$, $\bar{\Omega}$, and $V\bar{W}$ solve Problems 2 and 3 exactly for slightly perturbed input data $\bar{U}_1$, $\bar{D}$, and V, Algorithm II is stable.

It is well known that the singular values of A' are always well conditioned with respect to perturbations in A', but that the singular vectors of A' can be very sensitive to such perturbations [12], [25]. Since

$$A' = U_{11} D V^T = \bar{U}_{11} \bar{D} V^T + O(\epsilon \|D\|_2) = \bar{X} \bar{C} V^T + O(\epsilon \|D\|_2),$$

this guarantees that $\tilde{D}' = \tilde{\Omega}'$ is close to D'. However, $\tilde{Q}$ and $\tilde{W}$ can be very different from Q and W, respectively, and thus $\tilde{U}_1'$ and $\tilde{V}'$ can be very different from U_1' and V', respectively.

Consider the case where U_1 is given. It takes $O(mn)$ time to compute μ and x (see §4.2), it takes $O(mn)$ time to compute X, it takes $O(n^2)$ time to compute the SVD of C (see §5), and it takes $O(mn^2)$ and $O(n^3)$ time to compute XQ and VW, respectively. Algorithm II computes both XQ and VW for Problem 2 and computes XQ for Problem 3. Thus the total times for solving Problems 2 and 3 are $O((m+n)\,n^2)$ and $O(mn^2)$, respectively.

4.2. Computing Y. In this subsection we show how to compute the column orthogonal matrix Y (see (17)).

First we assume that U_2 is known. Let P be an orthogonal matrix such that $Pu_2 = \|u_2\|_2\, e_1$, where $e_1 = (1, 0, \ldots, 0)^T$, and define $(z_2, X_{12}) = U_{12} P^T$, where $z_2 \in \mathbf{R}^{m-1}$ and $X_{12} \in \mathbf{R}^{(m-1)\times(m-n-1)}$. Since

$$\begin{pmatrix} U_{11} & U_{12} \\ u_1^T & u_2^T \end{pmatrix} \begin{pmatrix} I_n & \\ & P^T \end{pmatrix} = \begin{pmatrix} U_{11} & z_2 & X_{12} \\ u_1^T & \|u_2\|_2 & 0 \end{pmatrix}$$

is orthogonal, the matrix

$$\begin{pmatrix} U_{11} & z_2 \\ u_1^T & \|u_2\|_2 \end{pmatrix}$$

is column orthogonal and $\|u_1\|_2^2 + \|u_2\|_2^2 = 1$. Thus we set $x = z_2$ and $\mu = \|u_2\|_2$. It takes $O(m\,(m-n))$ time to compute x and μ. This computation is stable (see (22)).

From (18) we have

$$\begin{pmatrix} U_{11} & x & X_{12} \\ u_1^T & \mu & 0 \end{pmatrix} \begin{pmatrix} H & \\ & I_{m-n-1} \end{pmatrix} = \begin{pmatrix} X & 0 & X_{12} \\ 0 & 1 & 0 \end{pmatrix},$$

and thus $(X, X_{12}) \in \mathbf{R}^{(m-1)\times(m-1)}$ is orthogonal. We set $U_2' = X_{12}$ (see (5) and (21)). It takes $O(m\,(m-n))$ time to compute X_{12}. This computation is also stable (see (23)).

Next we assume that U_2 is *not* known. Let $u = (x^T, \mu)$ be the result of applying the Gram–Schmidt procedure with reorthogonalization [8, §4] to orthonormalize $e_n = (0, \ldots, 0, 1)^T$ to the columns of U_1. If $u \neq 0$, then

$$Y = (U_1 \quad u) \equiv \begin{pmatrix} U_{11} & x \\ u_1^T & \mu \end{pmatrix}$$

is column orthogonal and $YY^T e_n = e_n$, so that

$$1 = (YY^T e_n)_n = u_1^T u_1 + \mu^2.$$

If $u = 0$, then $U_1 U_1^T e_n = e_n$, so that $1 = (U_1 U_1^T e_n)_n = u_1^T u_1$, and we get a nonzero u by repeating the Gram–Schmidt procedure with a random unit vector in place of e_n (note that in this case $\mu = 0$).[5] The time for computing x and μ is $O(lmn)$, where l is the number of reorthogonalization steps, which is a small constant in practice [8]. These computations are stable (see (22)).

4.3. Another perspective. In this subsection we present another derivation of the decomposition $A' = XCV^T$, which relates Algorithm II to a method for downdating the QR decomposition (cf. [23]).

Consider the augmented matrix

$$\mathcal{A} = \begin{pmatrix} u_1 & D \\ \mu & 0 \end{pmatrix}.$$

From (18), (17), and (16) we have

$$Y \begin{pmatrix} u_1 & D \\ \mu & 0 \end{pmatrix} = \begin{pmatrix} 0 & U_{11} D \\ 1 & u_1^T D \end{pmatrix} = \begin{pmatrix} 0 & A'V \\ 1 & a^T V \end{pmatrix}.$$

On the other hand, from (18) we get

$$Y \begin{pmatrix} u_1 & D \\ \mu & 0 \end{pmatrix} = (YH)\, H^T \begin{pmatrix} u_1 & D \\ \mu & 0 \end{pmatrix} = \begin{pmatrix} X & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & C \\ 1 & u_1^T D \end{pmatrix} = \begin{pmatrix} 0 & XC \\ 1 & u_1^T D \end{pmatrix}.$$

Thus $A'V = XC$, and the result follows.

Park and Van Huffel [24] downdate by using plane rotations to reduce $\mathcal{A}$ to a matrix of the form

$$\begin{pmatrix} 0 & B \\ 1 & w^T \end{pmatrix},$$

where $B = {F'}^T A' V G'$ is bidiagonal and F' and G' are orthogonal, and then solving the bidiagonal singular value problem. The total time appears to be at least as large as that for Algorithm II.

5. Computing the SVD of C. In this section we present an algorithm for computing the singular value decomposition of the matrix $C \in \mathbf{R}^{k \times k}$ given by

$$C = \left(I - \frac{1}{1+\mu} u_1 u_1^T \right) D, \tag{24}$$

where $D = \operatorname{diag}(d_1, \ldots, d_k)$ with $d_1 \geq d_2 \geq \cdots \geq d_k \geq 0$, $u_1 = (\mu_1, \ldots, \mu_k)^T$ with $\|u_1\|_2 \leq 1$, and $\mu = \sqrt{1 - \|u_1\|_2^2}$. For convenience we define $d_{k+1} = 0$ and $\mu_{k+1} = \mu$.

We assume that

$$d_i - d_{i+1} \geq \theta \|D\|_2 \quad \text{and} \quad |\mu_i| \geq \theta, \tag{25}$$

where θ is a small multiple of ϵ to be specified in §5.3. Any matrix of the form (24) can be reduced to one that satisfies these conditions by using the deflation procedure described in §5.4.

[5] The same construction is used in downdating the QR decomposition [8].

5.1. Properties of the SVD. In this subsection we establish some properties of the singular value decomposition of C. The following lemma characterizes the singular values and singular vectors.

LEMMA 5.1. *Let* $Q\,(\Omega, 0)\,W^T$ *be the* SVD *of* C *with*

$$Q = (q_1, \ldots, q_k), \quad \Omega = \operatorname{diag}(\omega_1, \ldots, \omega_k), \quad \textit{and} \quad W = (w_1, \ldots, w_k).$$

Then the singular values $\{\omega_i\}_{i=1}^k$ *satisfy the* secular equation

$$f_2(\omega) \equiv \sum_{j=1}^{k+1} \frac{\mu_j^2}{d_j^2 - \omega^2} = 0 \tag{26}$$

and the interlacing property

$$d_1 > \omega_1 > d_2 > \cdots > d_k > \omega_k > 0. \tag{27}$$

The singular vectors are given by

$$q_i = \left(\frac{\gamma_{i,1}\mu_1}{d_1^2 - \omega_i^2}, \ldots, \frac{\gamma_{i,k}\mu_k}{d_k^2 - \omega_i^2} \right)^T \Bigg/ \sqrt{\sum_{j=1}^{k} \left(\frac{\gamma_{i,j}\mu_j}{d_j^2 - \omega_i^2} \right)^2}, \tag{28}$$

where $\gamma_{i,j} = \omega_i^2 + \mu d_j^2$, *and*

$$w_i = \left(\frac{d_1\mu_1}{d_1^2 - \omega_i^2}, \ldots, \frac{d_k\mu_k}{d_k^2 - \omega_i^2} \right)^T \Bigg/ \sqrt{\sum_{j=1}^{k} \left(\frac{d_j\mu_j}{d_j^2 - \omega_i^2} \right)^2}. \tag{29}$$

Proof. Since $\mu > 0$ and $d_k > 0$ (see (25)), C is nonsingular and $\omega_k > 0$. Since C is square and $C^TC = D(I - u_1u_1^T)D$, the squares of the singular values $\{\omega_i^2\}_{i=1}^k$ and the right singular vectors $\{q_i\}_{i=1}^k$ are the eigenvalues and eigenvectors, respectively, of $D^2 - (Du_1)(Du_1)^T$. Relations (27) and (29) follow immediately from Lemma 3.1 with $z = Du_1$. Moreover, the singular values satisfy the secular equation

$$0 = f_1(\omega) = -1 + \sum_{j=1}^{k} \frac{(d_j\mu_j)^2}{d_j^2 - \omega^2} = -\sum_{j=1}^{k+1} \mu_i^2 + \sum_{j=1}^{k+1} \frac{d_j^2\mu_j^2}{d_j^2 - \omega^2} = \omega^2 \sum_{j=1}^{k+1} \frac{\mu_j^2}{d_j^2 - \omega^2}, \tag{30}$$

which implies that they satisfy (26) as well.

From (29) we see that w_i is a multiple of $(D^2 - \omega_i^2 I)^{-1}Du_1$. Since $\omega_i q_i = Cw_i$, it follows that q_i is a multiple of $C(D^2 - \omega_i^2 I)^{-1}Du_1$. Simplifying,

$$C(D^2 - \omega_i^2 I)^{-1}Du_1 = D(D^2 - \omega_i^2 I)^{-1}Du_1 - \frac{u_1^T D(D^2 - \omega_i^2 I)^{-1}Du_1}{1 + \mu} u_1. \tag{31}$$

Because ω_i satisfies (30), we have

$$u_1^T D(D^2 - \omega_i^2 I)^{-1}Du_1 = \sum_{j=1}^{k} \frac{d_j^2\mu_j^2}{d_j^2 - \omega_i^2} = 1.$$

Plugging this into (31), we have

$$\begin{aligned} C(D^2-\omega_i^2 I)^{-1}Du_1 &= \left(\omega_i^2 I+(D^2-\omega_i^2 I)\right)(D^2-\omega_i^2 I)^{-1}u_1-\frac{1}{1+\mu}u_1 \\ &= \omega_i^2(D^2-\omega_i^2 I)^{-1}u_1+\frac{\mu}{1+\mu}u_1 \\ &= \frac{1}{1+\mu}\left(\omega_i^2 I+\mu D^2\right)(D^2-\omega_i^2 I)^{-1}u_1. \end{aligned}$$

Ignoring the first factor and normalizing, we get (28). □

The following lemma allows one to construct a matrix of the form (24) using D and all the singular values.

LEMMA 5.2. *Given a diagonal matrix* $D=\operatorname{diag}(d_1,\ldots,d_k)$ *and a set of numbers* $\{\hat{\omega}_i\}_{i=1}^k$ *satisfying the interlacing property*

$$d_1>\hat{\omega}_1>d_2>\cdots>d_k>\hat{\omega}_k>d_{k+1}\equiv 0, \tag{32}$$

there exists a vector $\hat{u}_1$ *and a scalar* $\hat{\mu}\geq 0$ *with* $\|\hat{u}_1\|_2^2+\hat{\mu}^2=1$ *such that* $\{\hat{\omega}_i\}_{i=1}^k$ *are the singular values of*

$$\hat{C}=\left(I-\frac{1}{1+\hat{\mu}}\hat{u}_1\hat{u}_1^T\right)D.$$

The vector $\hat{u}_1=(\hat{\mu}_1,\ldots,\hat{\mu}_k)^T$ *and scalar* $\hat{\mu}=\hat{\mu}_{k+1}$ *are given by*

$$|\hat{\mu}_i|=\sqrt{\prod_{j=1}^{i-1}\frac{\hat{\omega}_j^2-d_i^2}{d_j^2-d_i^2}\prod_{j=i}^{k}\frac{\hat{\omega}_j^2-d_i^2}{d_{j+1}^2-d_i^2}},\qquad 1\leq i\leq k+1, \tag{33}$$

where the sign of $\hat{\mu}_i$ *can be chosen arbitrarily for* $1\leq i\leq k$.

Proof. The numbers $\{\hat{\omega}_i\}_{i=1}^k$ satisfy the interlacing property (11). By Lemma 3.2 there exists a vector $\hat{z}=(\hat{\zeta}_1,\ldots,\hat{\zeta}_k)^T$ satisfying (12) such that the eigenvalues of $D^2-\hat{z}\hat{z}^T$ are $\{\hat{\omega}_i^2\}_{i=1}^k$. Defining $\hat{u}_1=D^{-1}\hat{z}$, it follows that $\hat{\mu}_i$ satisfies (33) for $1\leq i\leq k$. The first result of Lemma 3.1 implies that $\hat{u}_1^T\hat{u}_1=\hat{z}^TD^{-2}\hat{z}\leq 1$, so that we can define $\hat{\mu}\equiv\hat{\mu}_{k+1}=\sqrt{1-\|\hat{u}_1\|_2^2}$. It then follows that

$$\hat{C}^T\hat{C}=D^2-D\hat{u}_1\hat{u}_1^TD=D^2-\hat{z}\hat{z}^T,$$

so that $\{\hat{\omega}_i\}_{i=1}^k$ are the singular values of $\hat{C}$. Consequently,

$$\prod_{j=1}^{k}\hat{\omega}_j=\det(\hat{C})=\det\left(I-\frac{1}{1+\hat{\mu}}\hat{u}_1\hat{u}_1^T\right)\det(D)=\hat{\mu}\prod_{j=1}^{k}d_j,$$

and hence

$$\hat{\mu}_{k+1}=\prod_{j=1}^{k}\frac{\hat{\omega}_j}{d_j},$$

which is (33) for $i=k+1$. □

5.2. Computing the singular vectors. In practice we can only hope to compute an approximation $\hat{\omega}_i$ to ω_i. Yet it is well known that equations similar to (28) and (29) can be very sensitive to small errors in ω_i (see §3.2). Lemma 5.2 allows us to overcome this problem. After we have computed all the approximate singular values $\{\hat{\omega}_i\}_{i=1}^k$ of C, we find a *new* matrix $\hat{C}$ whose *exact* singular values are $\{\hat{\omega}_i\}_{i=1}^k$ and then compute the singular vectors of $\hat{C}$ using Lemma 5.1. Note that each difference, each product, and each ratio in (33) can be computed to high relative accuracy.[6] Thus $|\hat{\mu}_i|$ can be computed to high relative accuracy. We choose the sign of $\hat{\mu}_i$ to be the sign of μ_i. Substituting the computed $\hat{u}_1$ and $\hat{\mu}$ and the *exact* singular values $\{\hat{\omega}_i\}_{i=1}^k$ into (28) and (29), each singular vector of $\hat{C}$ can also be computed to componentwise high relative accuracy. Consequently, after all the singular vectors are computed, the singular vector matrices of $\hat{C}$ will be numerically orthogonal.

To ensure the existence of $\hat{C}$, we need $\{\hat{\omega}_i\}_{i=1}^k$ to satisfy the interlacing property (32). But since the exact singular values of C satisfy the same interlacing property (see (27)), this is only an accuracy requirement on the computed singular values and is not an additional restriction on C. We can use the SVD of $\hat{C}$ as an approximation to the SVD of C. This is stable as long as $\hat{u}_1$ and $\hat{\mu}$ are close to u_1 and μ, respectively.

5.3. Stably computing the singular values. To guarantee that $\hat{u}_1$ and $\hat{\mu}$ are close to u_1 and μ, respectively, we must ensure that the approximations $\{\hat{\omega}_i\}_{i=1}^k$ to the singular values are sufficiently accurate. As in §3.3, the key is the stopping criterion for the root-finder, namely,

$$|g_i(\xi)| \le \eta k(|\psi_i(\xi)| + |\phi_i(\xi)|), \tag{34}$$

where the secular equation (26) has been reformulated as $g_i(\xi) \equiv \psi_i(\xi) + \phi_i(\xi) = 0$ in the analogous manner.

THEOREM 5.3. *If* $\theta = 2\eta k^2$ *in* (25) *and each* $\hat{\xi}_i$ *satisfies* (34), *then*

$$|\hat{\mu}_i - \mu_i| \le 4\eta k^2 \|u\|_2, \qquad 1 \le i \le k+1. \tag{35}$$

The proof is again nearly identical to that of the corresponding result in [18]. As argued there, the factor k^2 in θ and (35) is likely to be $O(k)$ in practice.

We have been assuming that $\|u_1\|_2 + \mu^2 = 1$. In practice this is not always true due to round-off errors. However, since a vector with norm near unity is close to an *exact* unit vector to componentwise high relative accuracy, in practice u_1 and μ are given to componentwise high relative accuracy. This implies that each term in the secular equation (26) is still computed to high relative accuracy after the reformulation. Hence the stopping criterion (34) holds, and $\hat{u}_1$ and $\hat{\mu}$ are close to u_1 and μ, respectively.

5.4. Deflation. We now show that we can reduce C to a matrix of the same form that further satisfies

$$d_i - d_{i+1} \ge \theta \|D\|_2 \quad \text{and} \quad |\mu_i| \ge \theta,$$

where θ is specified in §5.3.

[6] Note that $\hat{\mu} = \hat{\mu}_{k+1}$ is *not* computed from $\hat{\mu} = \sqrt{1 - \|\hat{u}_1\|_2^2}$, which might not give high relative accuracy.

Assume that $\mu \equiv \mu_{k+1} < \theta$. Changing μ to θ perturbs μ by $O(\theta)$. The perturbed matrix

$$\left(I - \frac{1}{1+\theta} u_1 u_1^T\right) D$$

has the same form but with $\mu \geq \theta$. This reduction is stable (see §5.3).

Next assume that $|\mu_i| < \theta$ for some $i \leq k$. We illustrate the case $i = 1$. Changing μ_1 to 0 perturbs u_1 by $O(\theta)$. Partition u_1 and D as

$$u_1 = \begin{pmatrix} \mu_1 \\ \breve{u}_1 \end{pmatrix} \quad \text{and} \quad D = \begin{pmatrix} d_1 & \\ & \breve{D} \end{pmatrix}.$$

Then in the perturbed matrix

$$\begin{pmatrix} 1 & \\ & I - \frac{1}{1+\mu} \breve{u}_1 \breve{u}_1^T \end{pmatrix} \begin{pmatrix} d_1 & \\ & \breve{D} \end{pmatrix} \equiv \begin{pmatrix} d_1 & \\ & \breve{C} \end{pmatrix},$$

the singular value d_1 can be deflated, and $\breve{C}$ is another matrix with the same form but smaller dimensions. This reduction is also stable (see §5.3).

Now assume that $d_i - d_{i+1} < \theta \|D\|_2$ for some $i \leq k-1$. We illustrate the reduction for $i = 1$. Changing d_1 to d_2 perturbs D by $O(\theta \|D\|_2)$. Let G be a Givens rotation in the $(1,2)$ plane such that $(Gu)_1 = 0$, and let

$$\breve{u}_1 = \left(\sqrt{\mu_1^2 + \mu_2^2}, \mu_3, \ldots, \mu_k\right)^T \quad \text{and} \quad \breve{D} = \operatorname{diag}(d_2, d_3, \ldots, d_k).$$

Then symmetrically applying G to the perturbed matrix, we get

$$\begin{aligned} & G\left(I - \frac{1}{1+\mu} u_1 u_1^T\right) \begin{pmatrix} d_2 & \\ & \breve{D} \end{pmatrix} G^T \\ &= \left(G - \frac{1}{1+\mu} G u_1 u_1^T\right) G^T \begin{pmatrix} d_2 & \\ & \breve{D} \end{pmatrix} \\ &= \left(I - \frac{1}{1+\mu} \begin{pmatrix} 0 \\ \breve{u}_1 \end{pmatrix} \begin{pmatrix} 0 \\ \breve{u}_1 \end{pmatrix}^T\right) \begin{pmatrix} d_2 & \\ & \breve{D}_1 \end{pmatrix} \\ &= \begin{pmatrix} d_2 & \\ & \left(I - \frac{1}{1+\mu} \breve{u}_1 \breve{u}_1^T\right) \breve{D}_1 \end{pmatrix}. \end{aligned}$$

The singular value d_2 can be deflated, and the remaining matrix has the same form but smaller dimensions. This reduction is stable as well.

Finally assume that $d_k < \theta \|D\|_2$ and $d_{k-1} - d_k \geq \theta \|D\|_2$. Changing d_k to $\theta \|D\|_2$ perturbs D by $O(\theta \|D\|_2)$. Let

$$\breve{D} = \operatorname{diag}(d_1, \ldots, d_{k-2}, d_{k-1}, \theta \|D\|_2).$$

Then the perturbed matrix

$$\left(I - \frac{1}{1+\mu} u_1 u_1^T\right) \breve{D}$$

has the same form but with $d_k \geq \theta \|D\|_2$. If the relation $d_{k-1} - d_k \geq \theta \|D\|_2$ no longer holds, then we can apply the previous reduction to reduce the matrix size. This reduction is again stable.

6. Ill-conditioning of Problem 1. In this section we bound the effect of perturbations in a on the singular values of A'. The effect of perturbations in V and D is similar. We assume that D is nonsingular.

From (20) and the second relation in (16), we have $A' = XCV^T$, where X is column orthogonal and

$$C = D - \frac{1}{1+\mu} u_1 u_1^T D,$$

with $u_1 = D^{-1}V^T a$ and $\mu = \sqrt{1 - \|u_1\|_2^2}$.

Let $\bar{a}$ be a vector slightly perturbed from a with $\|D^{-1}V^T\bar{a}\|_2 \leq 1$, and let $\bar{A}'$ be the downdated matrix for the input data V, D, and $\bar{a}$. As before, we have $\bar{A}' = \bar{X}\bar{C}V^T$, where $\bar{X}$ is column orthogonal and

$$\bar{C} = D - \frac{1}{1+\bar{\mu}} \bar{u}_1 \bar{u}_1^T D,$$

with $\bar{u}_1 = D^{-1}V^T\bar{a}$ and $\bar{\mu} = \sqrt{1 - \|\bar{u}_1\|_2^2}$.

Let ω_i and $\bar{\omega}_i$ be the ith largest singular values of A and $\bar{A}$, respectively. Since the singular values of A' and $\bar{A}'$ are the singular values of C and $\bar{C}$, respectively, we have $|\bar{\omega}_i - \omega_i| \leq \|\bar{C} - C\|_2$ (see [12, p. 428]).

Since

$$\bar{u}_1 - u_1 = D^{-1}V^T(\bar{a} - a),$$

we have

$$\|\bar{u}_1 - u_1\|_2 \leq \|D^{-1}\|_2 \, \|\bar{a} - a\|_2.$$

Similarly,

$$\bar{\mu} - \mu = \frac{(1 - \|\bar{u}_1\|_2^2) - (1 - \|u_1\|_2^2)}{\sqrt{1 - \|\bar{u}_1\|_2^2} + \sqrt{1 - \|u_1\|_2^2}} = -\frac{(\bar{u}_1 + u_1)^T (\bar{u}_1 - u_1)}{\sqrt{1 - \|\bar{u}_1\|_2^2} + \sqrt{1 - \|u_1\|_2^2}},$$

so that

$$|\bar{\mu} - \mu| \leq \frac{2 \, \|\bar{u}_1 - u_1\|_2}{\sqrt{1 - \|u_1\|_2^2}} \leq \frac{2 \, \|D^{-1}\|_2 \|\bar{a} - a\|_2}{\sqrt{1 - \|u_1\|_2^2}}.$$

Since

$$\begin{aligned} \bar{C} - C &= \frac{1}{1+\mu} u_1 a^T V - \frac{1}{1+\bar{\mu}} \bar{u}_1 \bar{a}^T V \\ &= \left(\frac{(\bar{\mu} - \mu) \, u_1}{(1+\mu)(1+\bar{\mu})} - \frac{\bar{u}_1 - u_1}{1+\bar{\mu}} \right) a^T V - \frac{\bar{u}_1}{1+\bar{\mu}} (\bar{a} - a)^T V, \end{aligned}$$

we have

$$\begin{aligned} |\bar{\omega}_i - \omega_i| &\leq \|\bar{C} - C\|_2 \\ &\leq \frac{\|(\bar{\mu} - \mu) \, u_1\|_2}{(1+\mu)(1+\bar{\mu})} \, \|a\|_2 + \frac{\|\bar{u}_1 - u_1\|_2}{1+\bar{\mu}} \, \|a\|_2 + \frac{\|\bar{u}_1\|_2}{1+\bar{\mu}} \, \|\bar{a} - a\|_2 \\ &\leq |\bar{\mu} - \mu| \, \|a\|_2 + \|\bar{u}_1 - u_1\|_2 \, \|a\|_2 + \|\bar{a} - a\|_2 \\ &\leq \frac{4 \max\left\{ \|D^{-1}\|_2 \, \|a\|_2, 1 \right\}}{\sqrt{1 - \|u_1\|_2^2}} \, \|\bar{a} - a\|_2. \end{aligned}$$

When the factor $\|D^{-1}\|_2\|a\|_2$ is very large, or when $\|u_1\|_2$ is near unity, we cannot guarantee that $\bar{\omega}_i$ is close to ω_i. This result parallels that of Stewart [26, p. 205] in the context of downdating the Cholesky/QR factorization.

To better explain the role of $\|u_1\|_2$, Stewart [26] also shows that

$$\omega_n \leq \|D\|_2 \sqrt{1 - \|u_1\|_2^2} \quad \text{and} \quad \|u_1\|_2^2 \geq \frac{(d_i/\omega_i)^2 - 1}{(d_i/\omega_i)^2 + 1}.$$

Thus if $\|u_1\|_2$ is near unity, then ω_n is close to zero and C (and hence A') is close to being singular. And if any d_i is reduced (to ω_i) by a large factor, then $\|u_1\|_2$ is near unity.

REFERENCES

[1] J. L. Barlow, H. Zha, and P. A. Yoon, *Stable chasing algorithms for modifying complete and partial singular value decompositions*, Tech. Report CSE-93-19, Dept. of Computer Science, The Pennsylvania State University, University Park, PA, Sept. 1993.

[2] Å. Björck, H. Park, and L. Eldén, *Accurate downdating of least squares solutions*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 549–568.

[3] A. W. Bojanczyk, R. P. Brent, P. van Dooren, and F. R. de Hoog, *A note on downdating the Cholesky factorization*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 210–221.

[4] C. F. Borges and W. B. Gragg, *A parallel divide and conquer algorithm for the generalized real symmetric definite tridiagonal eigenproblem*, in Numerical Linear Algebra and Scientific Computing, L. Reichel, A. Ruttan, and R. S. Varga, eds., de Gruyter, Berlin, 1993, pp. 10–28.

[5] J. R. Bunch and C. P. Nielsen, *Updating the singular value decomposition*, Numer. Math., 31 (1978), pp. 111–129.

[6] J. R. Bunch, C. P. Nielsen, and D. C. Sorensen, *Rank-one modification of the symmetric eigenproblem*, Numer. Math., 31 (1978), pp. 31–48.

[7] J. Carrier, L. Greengard, and V. Rokhlin, *A fast adaptive multipole algorithm for particle simulations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 669–686.

[8] J. W. Daniel, W. B. Gragg, L. Kaufman, and G. W. Stewart, *Reorthogonalization and stable algorithms for updating the Gram–Schmidt QR factorization*, Math. Comp., 30 (1976), pp. 772–795.

[9] J. J. Dongarra and D. C. Sorensen, *A fully parallel algorithm for the symmetric eigenvalue problem*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. s139–s154.

[10] P. E. Gill, G. H. Golub, W. Murray, and M. A. Saunders, *Methods for modifying matrix factorizations*, Math. Comp., 28 (1974), pp. 505–535.

[11] G. H. Golub, *Some modified matrix eigenvalue problems*, SIAM Rev., 15 (1973), pp. 318–334.

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

[13] W. B. Gragg, J. R. Thornton, and D. D. Warner, *Parallel divide and conquer algorithms for the symmetric tridiagonal eigenproblem and bidiagonal singular value problem*, in Proc. 23rd Annual Pittsburgh Conference, University of Pittsburgh School of Engineering, Vol. 23, Modelling and Simulation, Pittsburgh, 1992.

[14] L. Greengard and V. Rokhlin, *A fast algorithm for particle simulations*, J. Comput. Phys., 73 (1987), pp. 325–348.

[15] M. Gu, *Studies in Numerical Linear Algebra*, Ph.D. thesis, Department of Computer Science, Yale University, New Haven, CT, 1993.

[16] M. Gu and S. C. Eisenstat, *Downdating the singular value decomposition*, Research Report YALEU/DCS/RR-939, Dept. of Computer Science, Yale University, New Haven, CT, May 1993.

[17] ———, *A stable and fast algorithm for updating the singular value decomposition*, Research Report YALEU/DCS/RR-966, Dept. of Computer Science, Yale University, New Haven, CT, June 1993.

[18] ———, *A stable and efficient algorithm for the rank-one modification of the symmetric eigenproblem*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1266–1276.

[19] R.-C. Li, *Solving secular equations stably and efficiently*, Working paper, Dept. of Mathematics, University of California at Berkeley, Oct. 1992.

[20] K. LÖWNER, *Über monotone matrixfunktionen*, Math. Z., 38 (1934), pp. 177–216.
[21] M. MOONEN, P. VAN DOOREN, AND J. VANDEWALLE, *A singular value decomposition updating algorithm for subspace tracking*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1015–1038.
[22] C. C. PAIGE, *Error analysis of some techniques for updating orthogonal decompositions*, Math. Comp., 34 (1980), pp. 465–471.
[23] H. PARK AND L. ELDÉN, *Downdating the rank-revealing URV decomposition*, Tech. Report LiTH-MAT-R-1992-47, Dept. of Mathematics, Linköping University, Linköping, Sweden, Dec. 1992.
[24] H. PARK AND S. VAN HUFFEL, *Two-way bidiagonalization scheme for downdating the singular value decomposition*, Preprint 93-035, Army High Performance Computing Research Center, University of Minnesota, Minneapolis, MN, Apr. 1993.
[25] G. W. STEWART, *Introduction to Matrix Computations*, Academic Press, New York, 1973.
[26] ———, *The effects of rounding error on an algorithm for downdating a Cholesky factorization*, J. Inst. Maths. Applics., 23 (1979), pp. 203–213.
[27] ———, *Determining rank in the presence of error*, Tech. Report CS TR–2972 and UMIACS TR–92–108, Dept. of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD, Oct. 1992.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 811–827, July 1995

008

OPTIMALLY WEIGHTED MUSIC FOR FREQUENCY ESTIMATION*

PETRE STOICA†, ANDERS ERIKSSON†, AND TORSTEN SÖDERSTRÖM†

Abstract. This paper introduces a weighted MUSIC (**mu**ltiple **si**gnal **c**lassification) algorithm for estimating the frequencies of sinusoidal signals from noise-corrupted measurements. The large-sample variance of the weighted MUSIC is determined, and the optimal weighting matrix which minimizes that variance is derived. The optimally weighted MUSIC is shown to provide more accurate frequency estimates than the unweighted MUSIC and ESPRIT (estimation of **s**ignal **p**arameters via **r**otation **i**nvariance **t**echniques).

Key words. signal processing, spectral line analysis, frequency estimation, eigenanalysis-based methods, statistical analysis, centro-symmetric matrices

AMS subject classifications. 93E10, 62F12, 62M15, 15A99

1. Introduction.

MUSIC (multiple signal classification) has received considerable attention in the recent signal processing literature, owing to its high-resolution capabilities and its conceptual simplicity; see [1], [6], [8]–[10], [14], [15], [18]. In this paper we are interested in using MUSIC for frequency estimation. This application of MUSIC is less studied, from the statistical standpoint, than the use of MUSIC in array processing. For thorough statistical analyses of the performance of MUSIC in the array processing application, the reader is referred to [10], [14], [15].

The statistical properties of the frequency estimates obtained with MUSIC have been recently established in [18]. Making use of the results in [18], Eriksson, Stoica, and Söderström [2]. considered the problem of optimizing the statistical accuracy of an eigenanalysis-based frequency estimation method related to MUSIC by using a Markov-based estimation approach. In this paper we tackle the problem of optimally designing a MUSIC frequency estimator from a different perspective. In contrast to [2], we herein consider the standard form of MUSIC, which we extend by the appropriate inclusion of a weighting matrix. We derive an optimal form for the weight and show, by means of simulations, that the optimally weighted MUSIC performs slightly better in the studied cases than both MUSIC and ESPRIT. Note that the unweighted MUSIC is usually less accurate than ESPRIT, even though none of these two methods is uniformly better than the other; see [18]. The work reported in this paper was inspired by the related work in [10] which deals with weighted MUSIC for the array processing problem. Other related works are [11] and [16] which derive optimally weighted ESPRIT algorithms for array processing. It is interesting to note that in the standard array processing case, the optimal weight for MUSIC is equal to the identity matrix (i.e., the unweighted MUSIC is optimal) [15], whereas this is not the case for ESPRIT; see [11], [16]. An optimally weighted ESPRIT algorithm for the frequency estimation problem has not yet been derived, but this could be done by combining the analysis techniques in this paper and in [11], [16].

* Received by the editors May 6, 1992; accepted for publication (in revised form) by G. Cybenko May 17, 1994. This work was supported by the Swedish Research Council for Engineering Sciences contract 91–676.

† Systems and Control Group, Department of Technology, Uppsala University, P. O. Box 27, S–751 03 Uppsala, Sweden (ae@syscon.uu.se). The first author is currently on leave from the Department of Automatic Control, Polytechnic Institute of Bucharest, Splaiul Independentei 313, R-77206 Bucharest, Romania.

2. Statement of the problem. Consider the following complex-valued sinusoidal signal

$$s(t) = \sum_{k=1}^{n} x_k(t); \quad x_k(t) = \alpha_k e^{i(\omega_k t + \phi_k)}, \tag{2.1}$$

where it is assumed that the frequencies $\{\omega_k\}$ are distinct, the phases $\{\phi_k\}$ are independent random variables uniformly distributed on $[0, 2\pi)$, and the amplitudes $\{\alpha_k\}$ are strictly positive. It is further assumed that noisy measurements of $s(t)$ are obtained as

$$z(t) = s(t) + e(t), \tag{2.2}$$

where the noise $e(t)$ is white, Gaussian distributed, and independent of $\{x_k(t)\}$, with

$$\mathbf{E}e(t) = 0; \quad \mathbf{E}e(t)e(s) = 0; \quad \mathbf{E}e(t)e^*(s) = \sigma^2\delta_{t,s} \; \forall t, s. \tag{2.3}$$

Hereafter, $\mathbf{E}$ stands for the expectation operator, the superscript * denotes the complex conjugate, and $\delta_{t,s}$ denotes the Kronecker delta. Concerning other notational conventions to be used in the sequel, the superscripts ^{T} and ^{H} denote the transpose and complex conjugate transpose, respectively. Furthermore, $\mathcal{R}(\cdot)$ and $\mathcal{N}(\cdot)$ denote the range and the null space of the matrix in question.

The problem of interest in this work concerns estimation of the frequencies $\{\omega_k\}$ from the samples $\{z(t)\}_{t=1}^{N+m-1}$ (here m is a positive integer to be chosen by the user, see below, and $N + m - 1$ is the number of available data samples). The number n of sine waves in the observed signal is assumed to be known. For estimation of n, we refer to [4], [8], and [9].

3. Weighted MUSIC. The following additional notation is required to describe the MUSIC algorithm and its weighted version considered in this paper. Let

$$y(t) = \begin{pmatrix} z(t) & z(t+1) & \dots & z(t+m-1) \end{pmatrix}^T, \tag{3.1}$$

where $m > n$. Also, let

$$a(\omega) = \begin{pmatrix} 1 & e^{i\omega} & \dots & e^{i(m-1)\omega} \end{pmatrix}^T, \tag{3.2}$$

$$A = \begin{pmatrix} a(\omega_1) & \cdots & a(\omega_n) \end{pmatrix}, \tag{3.3}$$

$$x(t) = \begin{pmatrix} x_1(t) & \dots & x_n(t) \end{pmatrix}^T, \tag{3.4}$$

$$\varepsilon(t) = \begin{pmatrix} e(t) & \dots & e(t+m-1) \end{pmatrix}^T. \tag{3.5}$$

Note that A is a Vandermonde matrix which has full column rank under the assumptions that $\{\omega_k\}$ are distinct and $m \geq n$ (see, e.g., [5]). Using (3.2)–(3.5), $y(t)$ can be expressed as

$$y(t) = Ax(t) + \varepsilon(t). \tag{3.6}$$

Under the assumptions made, the covariance matrix of $y(t)$ can be readily derived from (3.6) as

$$R \triangleq \mathbf{E}y(t)y^H(t) = APA^H + \sigma^2 I, \tag{3.7}$$

where P is the covariance matrix of the signal vector,

$$P \triangleq \mathbf{E}x(t)x^H(t) = \begin{pmatrix} \alpha_1^2 & & 0 \\ & \ddots & \\ 0 & & \alpha_n^2 \end{pmatrix}. \tag{3.8}$$

Let $\{\lambda_k\}_{k=1}^m$ denote the eigenvalues of R, arranged in nonincreasing order, and let $\{v_k\}_{k=1}^m$ denote the corresponding orthonormal eigenvectors. The so-called signal eigenvalues $\{\lambda_k\}_{k=1}^n$ are assumed to be distinct (this condition is required to conduct the analysis later on, and apparently cannot be relaxed; however, it is a minor restriction that excludes very few combinations of signal parameters, if any). Introduce the following matrices,

$$S = (v_1 \quad \cdots \quad v_n), \quad G = (v_{n+1} \quad \cdots \quad v_m) \tag{3.9}$$

and note the following well-known properties of the range spaces of S and G (see, e.g., [2], [8]–[12], [14], [15], [18]):

$$\mathcal{R}(S) = \mathcal{R}(A), \quad \mathcal{R}(G) = \mathcal{N}(A^H), \tag{3.10}$$

which can be rewritten in the following equivalent forms,

$$SS^H = A(A^HA)^{-1}A^H, \qquad \Pi \triangleq GG^H = I - A(A^HA)^{-1}A^H. \tag{3.11}$$

MUSIC relies on the above properties. The frequencies $\{\omega_k\}_{k=1}^n$ are estimated as the locations of the n smallest minima of the function

$$f(\omega) = a^H(\omega)\hat{\Pi}a(\omega), \quad \omega \in [-\pi, \pi), \tag{3.12}$$

where

$$\hat{\Pi} = \hat{G}\hat{G}^H,$$

and where $\hat{G}$ (and, similarly, $\hat{S}$) is the matrix G (S) made from the orthonormal eigenvectors of the sample covariance matrix

$$\hat{R} = \begin{pmatrix} \hat{r}_0 & \cdots & \hat{r}_{m-1} \\ \vdots & \ddots & \vdots \\ \hat{r}_{m-1}^* & \cdots & \hat{r}_0 \end{pmatrix}, \qquad \hat{r}_k = \frac{1}{N}\sum_{t=1}^{N+m-1-k} z(t)z^*(t+k), \quad (k \geq 0). \tag{3.13}$$

The MUSIC frequency estimator described above is often referred to as the "spectral MUSIC." Another form of MUSIC, called the "root MUSIC," is outlined in Appendix A.1. Root MUSIC usually is preferred to spectral MUSIC, owing to its simpler implementation and higher finite-sample resolution. In Appendix A.1, however, we show that the asymptotic variances of the frequency estimates obtained with these two versions of MUSIC coincide. This means that in the following, all results derived for the spectral MUSIC also apply to the root MUSIC.

Let $\hat{W}$ denote a nonnegative definite weighting matrix. We modify the MUSIC cost function (3.12) by including $\hat{W}$ in the following way,

$$f(\omega) = a^H(\omega)\hat{\Pi}\hat{W}\hat{\Pi}a(\omega). \tag{3.14}$$

Since $\hat{\Pi}$ is idempotent, (3.14) reduces to (3.12) for $\hat{W} = I$ (the unweighted case). The method which estimates the frequencies by minimizing (3.14) is called weighted (spectral) MUSIC.

It should be noted that a seemingly simpler form of weighting the MUSIC cost function is

$$f(\omega) = a^H(\omega)\hat{G}\tilde{W}\hat{G}^H a(\omega). \tag{3.15}$$

However, the weighted form (3.15), if treated directly, leads to a more complicated statistical analysis than (3.14) does. The reason is the nonuniqueness of the so-called noise-eigenvector matrix G (any post-multiplication of G by a unitary matrix gives another valid G). This means that, in a large-sample analysis, $\hat{G}$ cannot be just replaced by G; or if it is, then G should be considered as random (dependent on the realization). See [15] for more details on this aspect. Now, interestingly enough, (3.15) can be obtained from (3.14) by setting $\hat{W} = \hat{G}\tilde{W}\hat{G}^H$ in (3.14) (and vice versa, (3.15) with $\tilde{W} = \hat{G}^H\hat{W}\hat{G}$ reduces to (3.14)). This means that (3.15) does not need to be analysed separately from (3.14).

4. Some preliminary results. The derivation of the large-sample variance of the weighted MUSIC and the minimization of that variance with respect to the weight, require a number of results that are presented in the following. A central role in the subsequent analysis in §§5 and 6 is played by the notion of centro-symmetric matrices. That is why the next definition and properties refer to this type of matrices.

Let $\tilde{I}$ denote the so-called exchange matrix,

$$\tilde{I} = \begin{pmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{pmatrix} \tag{4.1}$$

(the dimension of which will follow from the context).

DEFINITION 1. *A $n \times n$-matrix Q is said to be centro-symmetric (***cs***) if*

$$\tilde{I}Q\tilde{I} = Q^* \quad \textit{or equivalently} \quad Q_{n-i+1,n-j+1} = Q_{i,j}^*,$$

where $Q_{i,j}$ denotes the (i,j)-element of Q.

If Q is Hermitian then the above definition reduces to requiring that

$$Q_{i,j} = Q_{n-j+1,n-i+1},$$

which, in turn, is equivalent to requiring that Q is symmetric about its "northeast-southwest" diagonal (such a matrix is called persymmetric; see, e.g., [5]). The matrix R defined in (3.7) is Hermitian and Toeplitz (hence, persymmetric), which means that it is **cs** as well.

LEMMA 1. *Let Q be* **cs** *and let $Q^\dagger$ denote the Moore–Penrose pseudoinverse of Q. Then $Q^\dagger$ is also* **cs**.

Proof. Let the singular value decomposition of Q be

$$Q = U\Sigma V^H.$$

Then the Moore–Penrose pseudoinverse of Q is (see, e.g., [13], [5])

$$Q^\dagger = V\Sigma^\dagger U^H.$$

It follows at once that

$$\tilde{I}Q^{\dagger}\tilde{I} = (\tilde{I}V)\Sigma^{\dagger}(U^H\tilde{I}) = (\tilde{I}U\Sigma V^H\tilde{I})^{\dagger} = (Q^*)^{\dagger}.$$

However, $(Q^*)^{\dagger} = (Q^{\dagger})^*$, and the lemma is proven. □

LEMMA 2. *The signal-eigenvector matrix S satisfies*

$$\tilde{I}S = S^*D,$$

where

$$D = \begin{pmatrix} e^{i\gamma_1} & & 0 \\ & \ddots & \\ 0 & & e^{i\gamma_n} \end{pmatrix}$$

for some $\{\gamma_i\} \in [0, 2\pi)$.

Proof. Since R is **cs** it follows that

$$Rv_k = \lambda_k v_k \Rightarrow (\tilde{I}R\tilde{I})(\tilde{I}v_k) = \lambda_k(\tilde{I}v_k) \Rightarrow$$
$$R(\tilde{I}v_k)^* = \lambda_k(\tilde{I}v_k)^*, \quad (k = 1, \ldots, n).$$

Since the orthonormal eigenvectors associated with distinct eigenvalues are unique, modulo a multiplication by a unit-amplitude scalar, it readily follows that

$$\tilde{I}v_k = v_k^* e^{i\gamma_k} \text{ for some } \gamma_k \in [0, 2\pi) \quad (k = 1, \ldots, n),$$

which proves the assertion columnwise. □

LEMMA 3. *Let*

$$\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}, \qquad \tilde{\Lambda} = \Lambda - \sigma^2 I. \tag{4.2}$$

The matrix $S\tilde{\Lambda}^{-1}S^H$ is **cs**.

Proof. Since $\lambda_{n+1} = \cdots = \lambda_m = \sigma^2$, it follows that one can write R as

$$R = S\Lambda S^H + \sigma^2 GG^H = S\tilde{\Lambda}S^H + \sigma^2 I, \tag{4.3}$$

which implies at once that $S\tilde{\Lambda}S^H$ is **cs**. Then, by Lemma 1, $S\tilde{\Lambda}^{-1}S^H = (S\tilde{\Lambda}S^H)^{\dagger}$ is **cs** as well.

Alternatively, by Lemma 2 we get

$$\tilde{I}(S\tilde{\Lambda}^{-1}S^H)\tilde{I} = S^*D\tilde{\Lambda}^{-1}D^HS^T = S^*\tilde{\Lambda}^{-1}S^T = (S\tilde{\Lambda}^{-1}S^H)^*$$

which is the sought result. □

LEMMA 4. *The orthogonal projector* Π *defined in* (3.11) *is* **cs**.

Proof. Using Lemma 2 and (4.3), we get

$$R^* = \tilde{I}R\tilde{I} = \tilde{I}(S\Lambda S^H)\tilde{I} + \sigma^2\tilde{I}\Pi\tilde{I} = (S\Lambda S^H)^* + \sigma^2(\tilde{I}\Pi\tilde{I})$$

which immediately gives the stated result. □

LEMMA 5. *Let*

$$d(\omega) = \frac{da(\omega)}{d\omega}. \tag{4.4}$$

Then,

$$\tilde{I}a(\omega) = a^*(\omega)e^{i(m-1)\omega} \quad \textit{for}\ \omega \in [0,\, 2\pi), \tag{4.5}$$

$$d^T(\omega)\Pi^* = d^H(\omega)\Pi\tilde{I}e^{i(m-1)\omega} \quad \text{for}\ \omega = \omega_k \ \ (k = 1, \ldots, n). \tag{4.6}$$

Proof. Equation (4.5) follows by direct calculation. To prove (4.6), first note (for example from (3.11)) that

$$\Pi a(\omega_k) = 0, \quad k = 1, \ldots, n.$$

In the above equation $\{\omega_k\}_{k=1}^n$ in Π and $a(\omega_k)$ are to be seen as n scalar-valued variables. Taking the derivative of both sides of the above equation with respect to ω_k, we get

$$0 = [\Pi a(\omega_k)]' = \Pi' a(\omega_k) + \Pi d(\omega_k),$$

which gives

$$\Pi d(\omega_k) = -\Pi' a(\omega_k). \tag{4.7}$$

From Lemma 4, (4.5), and (4.7) we get

$$\begin{aligned}\Pi^* d^*(\omega_k) &= -(\tilde{I}\Pi\tilde{I})'\tilde{I}a(\omega_k)e^{-i(m-1)\omega_k} = -\tilde{I}\Pi' a(\omega_k)e^{-i(m-1)\omega_k} \\ &= \tilde{I}\Pi d(\omega_k)e^{-i(m-1)\omega_k}\end{aligned}$$

and the proof is finished. □

The next section makes use of the above results to establish some relevant properties of the weighted MUSIC cost function, (3.14), and of its minimizing arguments (which give the frequency estimates).

5. Statistical analysis. First we establish a simple but essential result.

THEOREM 1. *Any function of the form* (3.14), *possibly corresponding to a non-***cs** *weighting matrix* $\hat{W}$, *can be realized using a* **cs** *weighting matrix.*

Proof. Let $\hat{W}$ in (3.14) be an arbitrary Hermitian (nonnegative definite) matrix. For $N \to \infty$, the orthogonal projection matrix $\hat{\Pi}$ in (3.14) is **cs**, by Lemma 4. Next, we prove that $\hat{\Pi}$ is **cs** also for $N < \infty$. First note that since the eigenvalues of $\hat{R}$ are distinct (with probability 1) and since $\hat{R}$ is **cs**, it can be shown by paralleling the proof of Lemma 2, that $\hat{G}$ satisfies

$$\tilde{I}\hat{G} = \hat{G}^*\tilde{D},$$

where $\tilde{D}$ is diagonal with diagonal elements of unit magnitude. This readily implies that $\hat{\Pi}$ is **cs** for $N < \infty$ as well. Using this observation along with Lemma 5, one can write (3.14) as

$$\begin{aligned} f(\omega) &= a^T(\omega)\hat{\Pi}^*\hat{W}^*\hat{\Pi}^* a^*(\omega) = a^H(\omega)\tilde{I}e^{i(m-1)\omega}\hat{\Pi}^*\hat{W}^*\hat{\Pi}^*\tilde{I}a(\omega)e^{-i(m-1)\omega} \\ &= a^H(\omega)(\tilde{I}\hat{\Pi}^*\tilde{I})(\tilde{I}\hat{W}^*\tilde{I})(\tilde{I}\hat{\Pi}^*\tilde{I})a(\omega) = a^H(\omega)\hat{\Pi}(\tilde{I}\hat{W}^*\tilde{I})\hat{\Pi}a(\omega). \end{aligned} \tag{5.1}$$

Combining (3.14) and (5.1) gives

$$f(\omega) = a^H(\omega)\hat{\Pi}V\hat{\Pi}a(\omega),$$

where

$$V = \frac{1}{2}(\hat{W} + \tilde{I}\hat{W}^*\tilde{I}).$$

Since V is **cs**, the assertion follows. □

From here on, we assume that the weighting matrix $\hat{W}$ in (3.14) is **cs**. According to Theorem 1, this is no restriction.

Next, we proceed to derive the large-sample variance of the weighted MUSIC estimates. The following notation will be required:

$$R_p \triangleq \mathbf{E}y(t)y^H(t-p), \tag{5.2}$$

$$Q_p \triangleq \mathbf{E}\varepsilon(t)\varepsilon^H(t-p) = \sigma^2 \begin{pmatrix} 0 & \cdots & \overset{p+1}{1} & \cdots & 0 \\ & & & \ddots & \vdots \\ \vdots & & & & 1 \\ & & & & \vdots \\ 0 & & \cdots & & 0 \end{pmatrix} \triangleq \sigma^2 J_p. \tag{5.3}$$

In addition, it should be noted that the matrix $\hat{W}$ in (3.14) is allowed to depend on sample variables. Usually, $\hat{W}$ is a consistent estimate of some desired but unknown weighting matrix W. The theorem that follows shows that *replacement of W by $\hat{W}$ in* (3.14) *has no effect on the asymptotic variance of the weighted* MUSIC *estimates.* Thus, in the theorem, W denotes the limit of $\hat{W}$ as N tends to infinity.

THEOREM 2. *Let $\{\hat{\omega}_k\}$ denote the estimates of $\{\omega_k\}$ obtained by minimizing the weighted MUSIC cost function* (3.14). *The asymptotic (for $N \gg 1$) variances of $\{\hat{\omega}_k\}$ are given by*

$$\operatorname{var}(\hat{\omega}_k) = \frac{d_k^H \Pi W C_k W \Pi d_k}{(d_k^H \Pi W \Pi d_k)^2}, \quad k = 1, \dots, n, \tag{5.4}$$

where

$$C_k = \Pi \Gamma_k \Pi, \tag{5.5}$$

$$\Gamma_k = \frac{\sigma^4}{N} \sum_{|p| \le m-1} (\beta_k^H J_p \beta_k) J_p^T, \tag{5.6}$$

$$\beta_k = S\tilde{\Lambda}^{-1} S^H a_k, \tag{5.7}$$

and where a_k and d_k are shorthand notations for $a(\omega_k)$ and $d(\omega_k)$, respectively. Furthermore, the matrix C_k introduced above is **cs**.

Proof. See Appendix A.2. □

For $W = I$, formula (5.4) gives a *simplified* version of the variance formula for the unweighted MUSIC derived in [18]. The simplification consists of reducing the two-term formula for Γ_k in [18] to the one-term formula in (5.6). This simplification, for the general case of $W \neq I$, relies heavily on Theorem 1, and the result proved in §4 turns out to have important consequences for both the ensuing analysis and the implementation of the optimally weighted MUSIC which is derived in the next section.

6. Derivation of optimal weight. This section addresses the problem of minimizing $\mathrm{var}(\hat{\omega}_k)$, as given by (5.4), with respect to W. First, however, we need to show that the matrix F_k defined as

$$F_k \triangleq G^H \Gamma_k G \tag{6.1}$$

is positive definite.

LEMMA 6. *The matrix F_k, (6.1), is positive definite for any finite value of m.*

Proof. See Appendix A.3. □

Using the above lemma, the minimization of $\mathrm{var}(\hat{\omega}_k)$ is immediate. Let

$$\psi_k = G^H d_k, \quad H = G^H W G. \tag{6.2}$$

Then, using (6.1) and (6.2) in (5.4) we have

$$\mathrm{var}(\hat{\omega}_k) = \frac{\psi_k^H H F_k H \psi_k}{(\psi_k^H H \psi_k)^2}. \tag{6.3}$$

By the Cauchy–Schwartz inequality,

$$(\psi_k^H H \psi_k)^2 = |\psi_k^H H F_k^{1/2} \cdot F_k^{-1/2} \psi_k|^2 \leq (\psi_k^H H F_k H \psi_k)(\psi_k^H F_k^{-1} \psi_k), \tag{6.4}$$

where use was made of the fact that $F_k^{-1/2}$ exists (c.f. Lemma 6). From (6.3) and (6.4) we immediately get the following result.

THEOREM 3. *The large sample variance (5.4) of $\hat{\omega}_k$ satisfies the inequality*

$$\mathrm{var}(\hat{\omega}_k) \geq \frac{1}{\psi_k^H F_k^{-1} \psi_k}. \tag{6.5}$$

Furthermore, the lower bound in (6.5) is achieved with the weighting matrix

$$W_o = G F_k^{-1} G^H. \tag{6.6}$$

Proof. Straightforward utilization of (6.2)–(6.4). □

Since $C_k = G F_k G^H$ (c.f. (5.5) and (6.1)), a straightforward calculation shows that W_o is the Moore–Penrose pseudoinverse of C_k,

$$W_o = C_k^{\dagger}. \tag{6.7}$$

This observation shows (by Lemma 1 and Theorem 2, last part) that W_o is **cs**, as required. It also makes the connection between Theorem 3 and a similar result derived in [10] for the related problem of array processing with spatially smoothed data. It should be noted, however, that the derivation of (6.7) in [10] is incorrect (see Appendix A.4). In order to obtain a correct derivation, the algebraic structure of the problem under study must be fully exploited, as done above.

The optimal weight (6.6), when inserted into (3.14) in an estimated form, converts that form of weighted MUSIC cost function to a simpler expression of the form of (3.15). To see this, let $\hat{F}_k$ denote an estimate of F_k (obtained, for example, as described in §7), and let

$$\hat{W}_o = \hat{G} \hat{F}_k^{-1} \hat{G}^H. \tag{6.8}$$

Since (6.6) is invariant to post-multiplication of G by a unitary matrix, it follows that $\hat{W}_o$ in (6.8) is a consistent estimate of W_o. By inserting $\hat{W}_o$ in (3.14), in lieu of $\hat{W}$, we obtain the following *optimally weighted* MUSIC *cost function:*

$$f_o(\omega) = a^H(\omega)\hat{G}\hat{F}_k^{-1}\hat{G}^H a(\omega). \tag{6.9}$$

The implementation of the frequency estimator obtained by minimizing (6.9) with respect to ω, is described in the next section.

7. Implementation of the optimally weighted MUSIC. There is one feature of the optimal weight previously derived that should be discussed. The weight (6.6) depends on the specific frequency whose variance is to be minimized, which means that the optimal weight is not the same for all frequencies. This, however, might have been expected as MUSIC estimates the frequencies one by one, and there is no reason why a weighting matrix which is optimal for some ω_k should also be optimal for another $\omega_l \neq \omega_k$. The dependence of W_o on ω_k, $(k = 1, \ldots, n)$, increases the computational cost of the optimal MUSIC estimator. This difficulty aside, the following multistep procedure may be used to implement *the optimally weighted* MUSIC *frequency estimator.*

1. Compute $\hat{R}$ and its eigenelements. Determine unweighted (root) MUSIC estimates of ω_k.

2. For $k = 1, \ldots, n$ perform the following: Let $\hat{\omega}_k$ denote the estimate of ω_k obtained in Step 1. Use $\hat{\omega}_k$ and the eigenelements of $\hat{R}$ to obtain a consistent estimate $\hat{F}_k$ of F_k in (6.1).[1] Then determine an improved estimate $\hat{\omega}_k^o$ of ω_k, by locally minimizing the function (6.9) around $\hat{\omega}_k$ (a Gauss–Newton algorithm, as developed in [3], appears to be a good choice for solving the minimization problem).

Note that estimation of $\hat{\omega}_k^o$ in Step 2 is completely decoupled from the determination of the other frequencies. Thus, if so desired, Step 2 may be performed for only some frequencies in which one has particular interest. Note also that Step 2 might be repeated using $\hat{\omega}_k^o$, in lieu of $\hat{\omega}_k$, to estimate F_k. This additional step may have a beneficial effect on the estimation accuracy whenever the standard MUSIC estimates $\{\hat{\omega}_k\}$ have poor accuracy.

The next section studies, by means of numerical calculations and simulations, the statistical performance achieved by the optimally weighted MUSIC frequency estimator implemented as outlined above, and makes comparisons with the performance corresponding to the unweighted MUSIC (Step 1 above) and ESPRIT [12].

8. Numerical examples. *Example* 1. Consider the case of estimating the frequency of a single sine wave, $n = 1$. Some straightforward calculations show that the large sample variance of the *optimally weighted* MUSIC is given by

$$\mathrm{var}_{\text{opt MUSIC}}(\hat{\omega}) = \frac{2\sigma^4}{\alpha^4 N m^2 (m-1)}, \tag{8.1}$$

which is the same as the large sample variance of ESPRIT (see [18]). The variance of unweighted MUSIC is given by (once more, see [18])

$$\mathrm{var}(\hat{\omega}) = \frac{12\sigma^4(m^2+1)}{5\alpha^4 N m^3 (m^2-1)}$$

[1] The factor σ^4/N in the expression for Γ_k (5.6) may be omitted since it has no effect on the estimates obtained. Also, recall from the discussion preceding Theorem 2 that the frequency estimates, obtained from the minimization of (6.8) with or without the $\hat{}$ on G and F_k, have the same asymptotic accuracy.

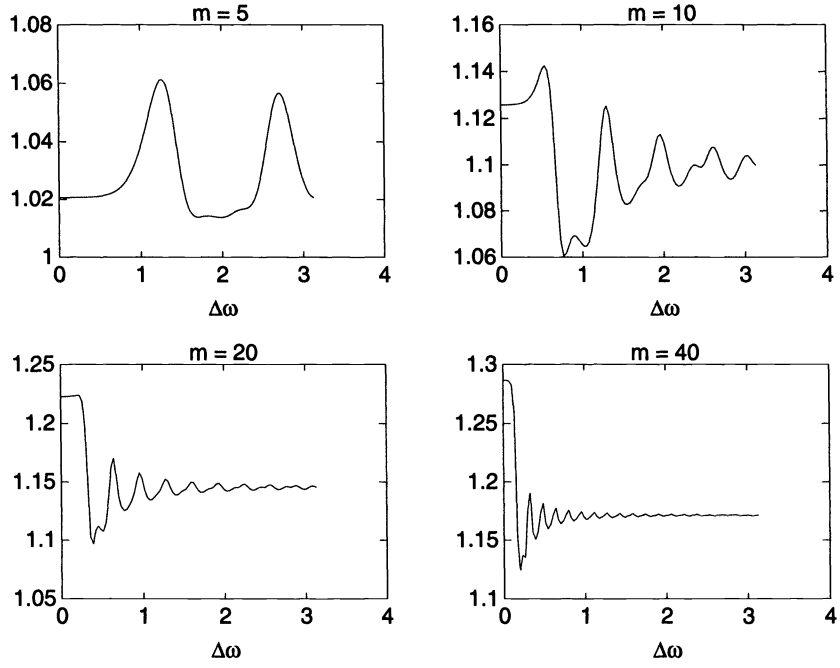


FIG. 8.1. *Comparison between* MUSIC *and optimally weighted* MUSIC *for estimating two frequencies.* $\text{var}_{\text{MUSIC}}(\hat{\omega}_j)/\text{var}_{\text{optMUSIC}}(\hat{\omega}_j)$ *versus frequency separation* $\Delta\omega$, $m = 5$, 10, 20, *and* 40.

which is slightly larger than (8.1).

Example 2. Consider the case of two sine waves, $n = 2$. For this case, the variance of the estimates $\hat{\omega}_1$ and $\hat{\omega}_2$ depends only on the frequency separation $\Delta\omega = |\omega_1 - \omega_2|$ (see, e.g., [18]). In Fig. 8.1, the *optimally weighted* MUSIC is compared to the unweighted MUSIC. The ratio $\text{var}_{\text{MUSIC}}/\text{var}_{\text{optMUSIC}}$ is displayed versus frequency separation, for $m = 5$, 10, 20, and 40. Note that the large sample variance expressions can be written as $f(\Delta\omega, m)/(N\text{SNR}^2)$, and thus the ratio $\text{var}_{\text{MUSIC}}/\text{var}_{\text{optMUSIC}}$ does not depend on the number of data samples, N, and the signal-to-noise ratio (SNR). In Fig. 8.2, a comparison between *optimally weighted* MUSIC and ESPRIT is presented. The ratio $\text{var}_{\text{ESPRIT}}/\text{var}_{\text{optMUSIC}}$ is displayed versus frequency separation, for $m = 5$, 10, 20, and 40. From these diagrams we see that *optimally weighted* MUSIC always is more accurate than unweighted MUSIC, as expected. The performance gain offered by the *optimally weighted* MUSIC over MUSIC is more noticeable in the practically relevant case of (very) small frequency separations. It can also be noted that the large sample variance of the *optimally weighted* MUSIC frequency estimates is less than, or equal to, the large sample variance of ESPRIT in the case of this example (whether or not this is true generally, however, is an open question).

Example 3. For the two cases in Examples 1 and 2, we compared the asymptotic variance expressions with empirical mean-square-errors (MSEs) obtained from Monte Carlo simulations. Table 8.1 displays empirical and theoretical variances for estimating the frequency of a single sine wave, the case considered in Example 1. The number of data points is $N = 100$, the true frequency is $\omega = 1.00$ rad/s and the signal-to-noise ratio is SNR = 0 dB. Two different values of m are used; viz. $m = 5$ and 10. In Table 8.2, the theoretical variance and empirical MSE are displayed for

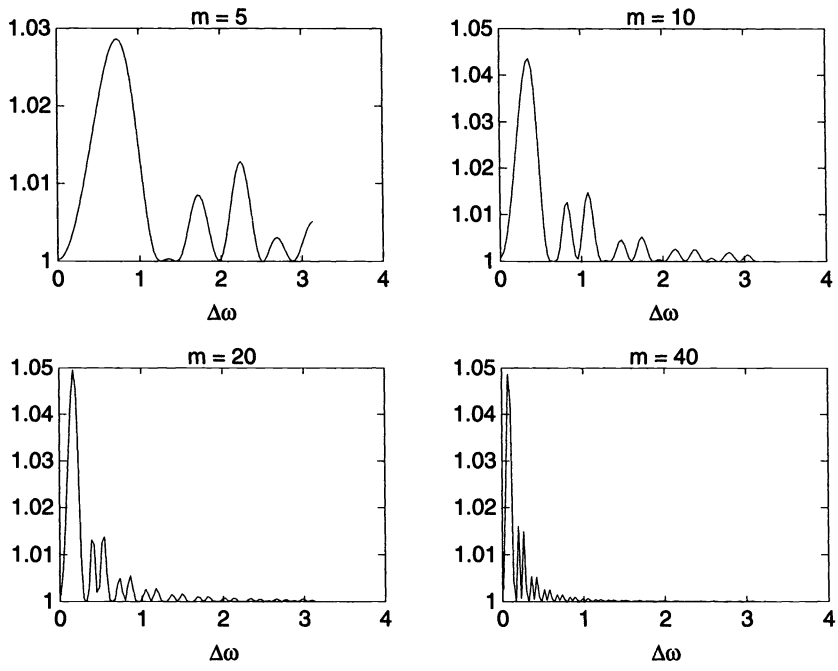


FIG. 8.2. *Comparison between* ESPRIT *and optimally weighted* MUSIC *for estimating two frequencies.* $\text{var}_{\text{ESPRIT}}(\hat{\omega}_j)/\text{var}_{\text{optMUSIC}}(\hat{\omega}_j)$ *versus frequency separation* $\Delta\omega$, $m = 5$, 10, 20, *and* 40.

TABLE 8.1
Theoretical variance and empirical MSE *for estimating a single frequency at* $\omega = 1.00$ rad/s *from* 100 *data samples,* SNR = 0 *dB,* $m = 5$ *and* 10.

$N = 100$, $\omega = 1.00$ rad/s, SNR = 0 dB						
	MUSIC		optimal MUSIC		ESPRIT	
m	MSE	variance	MSE	variance	MSE	variance
5	$2.13 \cdot 10^{-4}$	$2.08 \cdot 10^{-4}$	$2.10 \cdot 10^{-4}$	$2.00 \cdot 10^{-4}$	$2.17 \cdot 10^{-4}$	$2.00 \cdot 10^{-4}$
10	$4.33 \cdot 10^{-5}$	$2.45 \cdot 10^{-5}$	$3.89 \cdot 10^{-5}$	$2.22 \cdot 10^{-5}$	$4.01 \cdot 10^{-5}$	$2.22 \cdot 10^{-5}$

the case of estimating the frequencies of two sine waves, separated by 0.2 rad/s; cf. Example 2. Here we have $\omega_1 = 1.00$ rad/s and $\omega_2 = 1.20$ rad/s, $N = 400$, SNR = 0 dB, and $m = 20$. From Tables 8.1 and 8.2, we see that there is a reasonable agreement between the theoretical and empirical results.

9. Conclusions. We have introduced a weighted MUSIC algorithm for estimating the frequencies of sinusoidal signals. From the large sample variance expression for the weighted MUSIC frequency estimates, the *optimal weighting* matrix has been derived. In a numerical study, we have shown that the variance of the *optimally weighted* MUSIC frequency estimates always is less than the variance of the unweighted MUSIC estimates. Also, the variance of the *optimally weighted* MUSIC frequency estimates has been found in some numerical examples to be less than the variance of the ESPRIT estimates. However, *optimally weighted* MUSIC is computationally more demanding than unweighted MUSIC and ESPRIT, and from an application viewpoint the gain in performance may not motivate the extra computational complexity. On the other

TABLE 8.2
Theoretical variance and empirical MSE *for estimating two frequencies at* $\omega_1 = 1.00$ *and* $\omega_2 = 1.20$ rad/s *from* 400 *data samples,* SNR = 0 *dB,* $m = 20$.

$N = 400$, $\omega_1 = 1.00$ rad/s, $\omega_2 = 1.20$ rad/s, SNR = 0 dB, m = 20						
	MUSIC		optimal MUSIC		ESPRIT	
	MSE	variance	MSE	variance	MSE	variance
ω_1	$2.80 \cdot 10^{-6}$	$2.41 \cdot 10^{-6}$	$2.29 \cdot 10^{-6}$	$1.97 \cdot 10^{-6}$	$2.57 \cdot 10^{-6}$	$2.06 \cdot 10^{-6}$
ω_2	$2.88 \cdot 10^{-6}$	$2.41 \cdot 10^{-6}$	$2.34 \cdot 10^{-6}$	$1.97 \cdot 10^{-6}$	$2.51 \cdot 10^{-6}$	$2.06 \cdot 10^{-6}$

hand, from a theoretical standpoint, the *optimally weighted* MUSIC is believed to have a special place as the first eigenanalysis based frequency estimation method that was shown to asymptotically outperform both (unweighted) MUSIC and ESPRIT. In addition, the statistical and matrix analysis tools developed in this paper to study the weighted MUSIC frequency estimator may also prove useful in other similar performance studies.

Appendix A.

A.1. On root MUSIC and its asymptotic equivalence to spectral MUSIC. For the simplicity of the notation, consider the unweighted case. Introduce the symmetric polynomials

$$B_k(e^{i\omega}) \triangleq |\hat{g}_k^H a(\omega)|^2 \triangleq \sum_{|p| \le m-1} b_{k,p} e^{ip\omega}, \quad b_{k,p} = b_{k,m-p}^*$$

$$\text{for } k = 1, \ldots, m-n, \tag{A.1}$$

where $\hat{g}_k$ denotes the kth column of $\hat{G}$. Next, observe that the MUSIC cost function (3.12) can be written as

$$f(\omega) = \sum_{k=1}^{m-n} B_k(e^{i\omega}) \triangleq \sum_{|p| \le m-1} f_p e^{ip\omega}. \tag{A.2}$$

The coefficients of $f(\omega)$ are readily determined from those of $B_k(e^{i\omega})$:

$$f_p = \sum_{k=1}^{m-n} b_{k,p}, \quad (f_p = f_{m-p}^*). \tag{A.3}$$

Let $B(e^{i\omega})$ denote the spectral factor of $f(\omega) \ge 0$. Then one can write the MUSIC cost function as

$$f(\omega) = |B(e^{i\omega})|^2. \tag{A.4}$$

Note that the coefficients $\{b_{k,p}\}$, defining $B_k(e^{i\omega})$ and $f(\omega)$, are easily obtained from $\{\hat{g}_k\}$. Then $B(e^{i\omega})$ can be obtained from $f(\omega)$ by using a standard spectral factorization algorithm (see, e.g., [13]). The root MUSIC estimates the frequencies as the angular positions of the n roots of $B(e^{i\omega})$ which are closest to the unit circle.

In the above setting, the spectral MUSIC determines the frequencies as the locations of the n largest peaks of the pseudospectrum $1/|B(e^{i\omega})|^2$. It then follows from the general equivalence result in [17] that root MUSIC and spectral MUSIC have the same asymptotic properties.

The previous considerations clearly apply mutatis mutandis to the weighted MUSIC as well.

A.2. Proof of Theorem 2. Since $\hat{\omega}_k$ minimizes $f(\omega)$, we have

$$f'(\hat{\omega}_k) = 0. \tag{A.5}$$

For large values of N, $\hat{\omega}_k$ is close to ω_k. Then a Taylor series expansion of (A.5) around ω_k gives

$$0 \simeq f'(\omega_k) + f''(\omega_k)(\hat{\omega}_k - \omega_k). \tag{A.6}$$

Hereafter, the symbol $\simeq$ is used to denote a first-order approximation. In (A.6),

$$f'(\omega_k) = 2\mathrm{Re}[d_k^H \hat{\Pi} \hat{W} \hat{\Pi} a_k] \simeq 2\mathrm{Re}[d_k^H \Pi W \hat{\Pi} a_k], \tag{A.7}$$

$$f''(\omega_k) = 2\mathrm{Re}[d_k^H \hat{\Pi} \hat{W} \hat{\Pi} d_k] + 2\mathrm{Re}[(d_k^H)' \hat{\Pi} \hat{W} \hat{\Pi} a_k] \simeq 2 d_k^H \Pi W \Pi d_k. \tag{A.8}$$

The approximations in both (A.7) and (A.8) have been obtained by using the fact that $\hat{\Pi} a_k = (\hat{\Pi} - \Pi) a_k$ tends to zero as $N \to \infty$. Inserting (A.7) and (A.8) in (A.6), gives the following expression for the asymptotic estimation error:

$$\hat{\omega}_k - \omega_k = -\frac{\mathrm{Re}[d_k^H \Pi W \hat{\Pi} a_k]}{d_k^H \Pi W \Pi d_k}. \tag{A.9}$$

In order to prove (5.4) it remains to derive the (asymptotic) second-order moment of the numerator in (A.9). Let

$$\mu_k \triangleq -d_k^H \Pi W \hat{\Pi} a_k. \tag{A.10}$$

A simple calculation gives,

$$\begin{aligned} \hat{\Pi} a_k &= \hat{\Pi}\hat{\Pi} a_k \simeq \Pi \hat{\Pi} a_k = \Pi (I - \hat{S}\hat{S}^H) a_k = -\Pi \hat{S} \hat{S}^H a_k \\ &\simeq -\Pi \hat{S} S^H a_k = -G(G^H \hat{S}) S^H a_k. \end{aligned} \tag{A.11}$$

Let $\{\hat{\lambda}_k\}$ denote the eigenvalues of $\hat{R}$, and let

$$\hat{\Lambda} = \begin{pmatrix} \hat{\lambda}_1 & & 0 \\ & \ddots & \\ 0 & & \hat{\lambda}_n \end{pmatrix}, \quad \hat{\Sigma} = \begin{pmatrix} \hat{\lambda}_{n+1} & & 0 \\ & \ddots & \\ 0 & & \hat{\lambda}_m \end{pmatrix}. \tag{A.12}$$

Using this notation, one can write

$$\begin{aligned} G^H \hat{R} S &= G^H (\hat{S} \hat{\Lambda} \hat{S}^H + \hat{G} \hat{\Sigma} \hat{G}^H) S = (G^H \hat{S}) \hat{\Lambda} (\hat{S}^H S) + (G^H \hat{G}) \hat{\Sigma} (\hat{G}^H S) \\ &\simeq (G^H \hat{S}) \Lambda + \sigma^2 G^H \hat{G} \hat{G}^H S = (G^H \hat{S}) \Lambda - \sigma^2 (G^H \hat{S}) \hat{S}^H S \\ &\simeq (G^H \hat{S}) \tilde{\Lambda}. \end{aligned} \tag{A.13}$$

Using (A.13) in (A.11) and the so-obtained result in (A.10), gives the following asymptotic expression for μ_k,

$$\mu_k = d_k^H \Pi W \Pi \hat{R} S \tilde{\Lambda}^{-1} S^H a_k = d_k^H \Pi W \Pi \hat{R} \beta_k. \tag{A.14}$$

Using Theorem 1 and Lemmas 3, 4, and 5, it can be shown that μ_k is real-valued, as follows:

$$\begin{aligned} \mu_k^* &= d_k^T \Pi^* W^* \Pi^* \hat{R}^* (S \tilde{\Lambda}^{-1} S^H)^* a_k^* \\ &= e^{i(m-1)\omega_k} d_k^H \Pi \tilde{I} W^* \tilde{I} \cdot \tilde{I} \Pi^* \tilde{I} \cdot \tilde{I} \hat{R}^* \tilde{I} \cdot \tilde{I} (S \tilde{\Lambda}^{-1} S^H)^* \tilde{I} \cdot a_k e^{-i(m-1)\omega_k} \\ &= d_k^H \Pi W \Pi \hat{R} (S \tilde{\Lambda}^{-1} S^H) a_k = \mu_k. \end{aligned} \tag{A.15}$$

Thus,

$$(\hat{\omega}_k - \omega_k) = \frac{\mu_k}{d_k^H \Pi W \Pi d_k}. \tag{A.16}$$

In order to write μ_k in a more convenient form, introduce

$$\rho_k^H = d_k^H \Pi W \Pi \tag{A.17}$$

and observe that replacement in (A.14) of $\hat{R}$ given by (3.13) with

$$\frac{1}{N}\sum_{t=1}^{N} y(t)y^H(t) \tag{A.18}$$

has no effect on the asymptotic behaviour of μ_k. Thus, from (A.14)

$$\begin{aligned} \mathbf{E}\mu_k^2 &= \mathbf{E}(\rho_k^H \frac{1}{N}\sum_{t=1}^{N} y(t)y^H(t)\beta_k)^2 \\ &= \frac{1}{N^2}\sum_{t=1}^{N}\sum_{s=1}^{N} \mathbf{E}\rho_k^H \varepsilon(t)y^H(t)\beta_k \varepsilon^H(s)\rho_k\beta_k^H y(s), \end{aligned} \tag{A.19}$$

where use was made of the fact that $\rho_k^H A = 0$. Using a well-known formula for the expectation of the product of four Gaussian random variables (see, e.g., [7]) as well as the independence of $\varepsilon(t)$ and $x(s)$, we get

$$\begin{aligned} \mathbf{E}\rho_k^H \varepsilon(t)y^H(t)\beta_k \varepsilon^H(s)\rho_k\beta_k^H y(s) &= \sigma^4|\rho_k^H\beta_k|^2 + (\rho_k^H Q_{t-s}\rho_k)(\beta_k^H R_{s-t}\beta_k) \\ &= (\rho_k^H Q_{t-s}\rho_k)(\beta_k^H R_{s-t}\beta_k) \end{aligned} \tag{A.20}$$

(since $\rho_k^H\beta_k = 0$). Inserting (A.20) in (A.19) and noting the fact that $Q_p = 0$ for $|p| \geq m$, give

$$\mathbf{E}\mu_k^2 = \frac{1}{N^2}\sum_{|p|\leq m-1} (N - |p|)(\rho_k^H Q_p^T \rho_k)(\beta_k^H R_p \beta_k),$$

which asymptotically (for $N \gg 1$) becomes

$$\mathbf{E}\mu_k^2 = \frac{1}{N}\sum_{|p|\leq m-1} (\rho_k^H Q_p^T \rho_k)(\beta_k^H R_p \beta_k). \tag{A.21}$$

Next we show that R_p in the right-hand side of (A.21) can be replaced by Q_p, which leads to a much simplified variance formula. By comparing the two expressions of R in (3.7) and (4.3), we obtain

$$APA^H = S\tilde{\Lambda}S^H. \tag{A.22}$$

Pre- and post-multiplying (A.22) by S^H and S, respectively, gives

$$(S^H A)P(A^H S) = \tilde{\Lambda} \Rightarrow (A^H S)^{-1}P^{-1}(S^H A)^{-1} = \tilde{\Lambda}^{-1} \Rightarrow P^{-1} = A^H S\tilde{\Lambda}^{-1}S^H A$$

or, columnwise,

$$\frac{1}{\alpha_k^2}\ \big(0\ \ldots\ 0\ \overset{k}{1}\ 0\ \ldots\ 0\big)^T = A^H\beta_k,\ \ (k = 1, \ldots, n). \tag{A.23}$$

Next, note that

$$R_p = AP\Delta^p A^H + Q_p, \tag{A.24}$$

where

$$\Delta = \begin{pmatrix} e^{i\omega_1} & & 0 \\ & \ddots & \\ 0 & & e^{i\omega_n} \end{pmatrix}. \tag{A.25}$$

Using (A.23) and (A.24), gives

$$\beta_k^H R_p \beta_k = \beta_k^H (AP\Delta^p A^H + Q_p)\beta_k = \frac{1}{\alpha_k^2} e^{ip\omega_k} + \beta_k^H Q_p \beta_k. \tag{A.26}$$

The first term in (A.26) contributes to (A.21) the following term:

$$\frac{1}{N\alpha_k^2} \sum_{|p|\le m-1} (\rho_k^H Q_p^T \rho_k) e^{ip\omega_k} = \frac{\sigma^2}{N\alpha_k^2} \left[\sum_{|p|\le m-1} e^{-ip\omega_k} (\rho_k^H J_p \rho_k) \right]^*. \tag{A.27}$$

The bracketed sum in (A.27) can be shown to be zero as follows:

$$\begin{aligned} \rho_k^H \left(\sum_{p=-m+1}^{m-1} J_p e^{-ip\omega_k} \right) \rho_k &= \rho_k^H \begin{pmatrix} 1 & \cdots & e^{-i(m-1)\omega_k} \\ \vdots & \ddots & \vdots \\ e^{i(m-1)\omega_k} & \cdots & 1 \end{pmatrix} \rho_k \\ &= \rho_k^H a(\omega_k) a^H(\omega_k) \rho_k = 0. \end{aligned}$$

Thus, (A.21) reduces to

$$\mathbf{E}\mu_k^2 = \frac{\sigma^4}{N} \sum_{|p|\le m-1} (\rho_k^H J_p^T \rho_k)(\beta_k^H J_p \beta_k) \tag{A.28}$$

which, when used together with (A.16), gives (5.4).

It only remains to show that C_k is **cs**. A simple way to see this consists of making use of the expression of C_k that follows from (A.15),

$$C_k = \Pi \mathbf{E} \left[\hat{R}(S\tilde{\Lambda}^{-1}S^H) a_k \cdot a_k^H (S\tilde{\Lambda}^{-1}S^H) \hat{R} \right] \Pi. \tag{A.29}$$

From (A.29) and Lemmas 3, 4, and 5, we get

$$\begin{aligned} \tilde{I}C_k\tilde{I} &= \tilde{I}\Pi\tilde{I} \cdot \mathbf{E} \left[\tilde{I}\hat{R}\tilde{I} \cdot \tilde{I}(S\tilde{\Lambda}^{-1}S^H)\tilde{I} \cdot \tilde{I}a_k \cdot a_k^H \tilde{I} \cdot \tilde{I}(S\tilde{\Lambda}^{-1}S^H)\tilde{I} \cdot \tilde{I}\hat{R}\tilde{I} \right] \tilde{I}\Pi\tilde{I} \\ &= \Pi^* \mathbf{E} \left[\hat{R}^*(S\tilde{\Lambda}^{-1}S^H)^* a_k^* a_k^T (S\tilde{\Lambda}^{-1}S^H)^* \hat{R}^* \right] \Pi^* = C_k^*, \end{aligned}$$

and the proof is completed.

A.3. Proof of Lemma 6. In this appendix, we omit the index k of β_k, etc. (indicating the dependence on the kth frequency), in order to simplify the notation. Let $\{\beta_j\}_0^{m-1}$ denote the elements of β, defined in (5.7), and

$$\gamma_p^* \triangleq \beta^H J_p \beta = (\beta_0^* \quad \cdots \quad \beta_{m-1}^*) \begin{pmatrix} \beta_p \\ \vdots \\ \beta_{m-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \sum_{j=0}^{m-1} \beta_j^* \beta_{j+p} \quad \text{for } p \ge 0, \tag{A.30}$$

where we made the convention that $\beta_j = 0$ for $j \geq m$. Using (A.30), we get

$$\beta^H J_p \beta = (\beta^H J_p^T \beta)^H = (\beta^H J_{-p} \beta)^* = \gamma_p \quad \text{for } p < 0. \tag{A.31}$$

It follows from (5.6) and (A.30), (A.31) that

$$\Gamma = \begin{pmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{m-1} \\ \gamma_1^* & \gamma_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \gamma_1 \\ \gamma_{m-1}^* & \cdots & \gamma_1^* & \gamma_0 \end{pmatrix}$$

is the Toeplitz covariance matrix corresponding to a moving average process of order $(m-1)$ and with the coefficient vector equal to β. Thus, Γ is positive definite for any finite m (see, e.g., [13]), which implies that F is so since G in (6.1) has full column rank.

A.4. On the approach in [10]. Let

$$h = \Pi d.$$

(In this appendix we omit the index k of d_k, etc., for notational convenience). Then, (5.4) can be written as

$$\text{var}(\hat{\omega}) = \frac{h^H W C W h}{(h^H W h)^2}. \tag{A.32}$$

It is claimed in [10] that, for a given C,

$$\text{var}(\hat{\omega}) \geq \frac{1}{h^H C^\dagger h} \tag{A.33}$$

for any positive semidefinite matrix W. However, this is not generally true as the following example shows. Let $W = I$ and $C = uu^H$, for some vector u of dimension m. It is readily verified that the Moore–Penrose pseudoinverse of C is given by

$$C^\dagger = \frac{uu^H}{(u^H u)^2}.$$

Thus, (A.33) becomes

$$|h^H u|^4 \geq (h^H h)^2 (u^H u)^2$$

which, in view of the Cauchy–Schwartz inequality, is invalid for almost any u.

REFERENCES

[1] K. M. BUCKLEY AND X.-L. XU, *Statistical analysis of eigenspace-based source location estimates*, in Proc. Internat. Conf. Acoustics, Speech, and Signal Processing, Toronto, Ontario, April 1991, pp. 3317–3320.

[2] A. ERIKSSON, P. STOICA, AND T. SÖDERSTRÖM, *Markov-based eigenanalysis method for frequency estimation*, IEEE Trans. Signal Processing, SP-42 (1994), pp. 586–594.

[3] ———, *On-line subspace algorithms for tracking moving sources*, IEEE Trans. Signal Processing, SP-42 (1994), pp. 2319–2330.

[4] J. J. FUCHS, *Estimating the number of sinusoids in additive white noise*, IEEE Trans. Acoustics, Speech, and Signal Processing, Vol ASSP-36, (1988), pp. 1846–1853.
[5] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
[6] S. HAYKIN, ED., *Advances in Spectrum Analysis and Array Processing*, Vol. II, Prentice Hall, Englewood Cliffs, NJ, 1991.
[7] P. H. JANSSEN AND P. STOICA, *On the expectation of the product of four matrix-valued Gaussian random variables*, IEEE Trans. Automatic Control, AC-33 (1988), pp. 867–870.
[8] S. M. KAY, *Modern Spectral Estimation*, Prentice Hall, Englewood Cliffs, NJ, 1988.
[9] S. L. MARPLE, *Digital Spectral Analysis, with Applications*, Prentice Hall, Englewood Cliffs, NJ, 1987.
[10] B. D. RAO AND K. V. S. HARI, *Spatial smoothing and* MUSIC: *Further results*, in SVD and Signal Processing, II, R. J. Vaccaro, ed., Elsevier Science Publishers, 1991, pp. 261–276.
[11] ———, *Weighted state space methods/*ESPRIT *and spatial smoothing*, in Proc. Internat. Conf. Acoustics, Speech, and Signal Processing, Toronto, Ontario, April 1991, pp. 3317–3320,
[12] R. ROY AND T. KAILATH, ESPRIT—*estimation of signal parameters via rotational invariance techniques*, IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-37 (1989), pp. 984–995.
[13] T. SÖDERSTRÖM AND P. STOICA, *System Identification*, Prentice Hall International, London, 1989.
[14] P. STOICA AND A. NEHORAI, MUSIC, *maximum likelihood, and Cramér–Rao bound*, IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-37 (1989), pp. 720–741.
[15] ———, MUSIC, *maximum likelihood, and Cramér–Rao bound: further results and comparisons*, IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-38 (1990), pp. 2140–2150.
[16] ———, *Performance comparison of subspace rotation and* MUSIC *methods for direction estimation*, IEEE Trans. Signal Processing, SP-39 (1991), pp. 446–453.
[17] P. STOICA AND T. SÖDERSTRÖM, *On spectral and root forms of sinusoidal frequency estimators*, Signal Processing, 24 (1991), pp. 93–103.
[18] ———, *Statistical analysis of* MUSIC *and subspace rotation estimates of sinusoidal frequencies*, IEEE Trans. Signal Processing, SP-39 (1991), pp. 1836–1847.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 828–842, July 1995

009

SIGN PATTERN ANALYSIS OF CONTROL COEFFICIENTS OF METABOLIC PATHWAYS *

THOMAS LUNDY† AND ASOK K. SEN†

Abstract. Using directed graphs we analyze the sign pattern of the control coefficients of the enzymes in linear metabolic pathways. The following pathways are examined: linear pathways with (a) neither feedback nor feedforward regulation and (b) possible feedback and/or feedforward loops. We establish the different pathway topologies that lead to a sign-nonsingular elasticity matrix. For a given topology with a sign-nonsingular elasticity matrix, we determine the control coefficients that have their signs unambiguously determined and the control coefficients that are sign-indeterminate. The enzymes and metabolites whose control coefficients are sign-indeterminate can be identified directly from the topology of the feedback and feedforward loops in the metabolic pathway.

Key words. sign pattern matrices, control coefficients, metabolic pathways

AMS subject classifications. 15A09, 92C40

1. Introduction.

The sign pattern of a real-valued matrix $A = [a_{ij}]$ is customarily described by a sign-pattern matrix $S = [s_{ij}]$ whose entries are defined by

$$s_{ij} = \begin{cases} + & \text{if} \quad a_{ij} > 0, \\ - & \text{if} \quad a_{ij} < 0, \\ 0 & \text{if} \quad a_{ij} = 0. \end{cases}$$

The symbols $1, -1$, and 0 are also used to designate such a matrix. Two real matrices A and B have the same sign pattern if for all i and $j, a_{ij}b_{ij} > 0$ or $a_{ij} = b_{ij} = 0$. A matrix A is said to be *sign-nonsingular* if every matrix with the same sign pattern as A is nonsingular. For a sign-nonsingular matrix the signs of all the entries in the inverse may be unambiguously determined, or some of the entries in the inverse may be sign-indeterminate.

Sign-nonsingular matrices are encountered in a wide variety of applications ranging from economics [1] to ecology [2]. They are found also in the study of metabolic regulation. The purpose of this paper is to examine the sign-nonsingularity properties of the elasticity matrices of linear metabolic pathways and to determine the sign pattern of the control coefficients of the various enzymes. The sign of a flux (concentration) control coefficient of an enzyme determines if the metabolic flux (concentration) will increase/decrease when the enzyme concentration is increased/decreased. A knowledge of the signs of the control coefficients may be useful in interpreting the results of experiments that involve perturbations of enzyme activity.

We examine linear metabolic pathways with neither feedback nor feedforward regulation, as well as pathways containing feedback and/or feedforward loops. We establish the different topologies of a linear pathway that have a sign-nonsingular elasticity matrix. For a given topology with a sign-nonsingular elasticity matrix, we determine the control coefficients whose signs are unambiguously determined and those that are sign-indeterminate. The results are illustrated with linear pathways containing four enzymes.

---

* Received by the editors March 24, 1992; accepted for publication (in revised form) by J. S. Maybee May 23, 1994.

†Department of Mathematical Sciences, Purdue University School of Science, 402 N. Blackford Street, Indianapolis, Indiana 46202 (ixyl100@indyvax.iupui.edu).

In recent work [3]–[6], directed graphs have been used for calculating the control coefficients of enzymes in metabolic pathways. In this paper we will use signed digraphs to analyze the sign pattern of the control coefficients.

The main results of this paper can be summarized as follows. In the absence of feedback and feedforward regulation, the elasticity matrix of a linear metabolic pathway is sign-nonsingular and the flux control coefficients and the concentration control coefficients are all sign-determined. The elasticity matrix is also sign-nonsingular for a linear pathway containing an arbitrary number of feedback inhibition loops or feedforward activation loops. If the pathway contains feedforward inhibition or feedback activation, the elasticity matrix will not be sign-nonsingular. Finally, if a feedback inhibition loop and a feedforward activation loop are both present, then the elasticity matrix remains sign-nonsingular except when the two loops overlap each other in an overhanging fashion and the enzyme being inhibited is located upstream of the activating metabolite in the pathway. If a linear metabolic pathway has a sign-nonsingular elasticity matrix, then all the enzymes and metabolites whose concentration control coefficients are sign-indeterminate can be identified directly from the topology of the feedback and feedforward loops in the metabolic pathway.

2. Metabolic control analysis: A review. Consider a linear metabolic pathway in which a substrate X_1 is converted to a product P by a series of enzyme-catalyzed reactions.

$$\text{(A)} \qquad X_1 \xrightarrow{E_1} X_2 \xrightarrow{E_2} X_3 \xrightarrow{E_3} \cdots \xrightarrow{E_{n-1}} X_n \xrightarrow{E_n} P.$$

Here $E_1 - E_n$ denote the enzymes and $X_2 - X_n$ represent the intermediate metabolites. The metabolite immediately preceding an enzyme is referred to as its substrate, whereas the metabolite immediately following the enzyme is called its product. In general, there may be additional interactions between metabolites and enzymes, with certain metabolites either inhibiting or activating nonadjacent enzymes. The pathway will then include *feedback* or *feedforward loops.* In any metabolic pathway, the regulatory effect of an enzyme on the metabolic flux or a metabolite concentration can be assessed in terms of its *flux control coefficient* or *concentration control coefficient* [7], [8]. The flux (or concentration) control coefficient of an enzyme is defined as the fractional change in metabolic flux (or concentration of a metabolite) in response to a fractional change in enzyme concentration. Mathematically, the flux control coefficient of an enzyme E_i is given by

$$C_i^J = \frac{e_i}{J}\frac{\partial J}{\partial e_i},$$

where J is the flux and e_i is the concentration of the enzyme E_i. The concentration control coefficient of an enzyme E_i with respect to a metabolite X_j has the definition

$$C_i^{X_j} = \frac{e_i}{x_j}\frac{\partial x_j}{\partial e_i},$$

x_j being the concentration of the metabolite X_j. These control coefficients are governed by a system of linear algebraic equations that can be written in a matrix form as $\tilde{E}\tilde{Z} = I$, where the matrix $\tilde{E}$, usually referred to as an elasticity matrix, is an $n \times n$ matrix containing the so-called elasticity coefficients; the control matrix $\tilde{Z}$ contains the flux control coefficients and concentration control coefficients, and I is the $n \times n$ identity matrix. The elasticity coefficient (ε_{ji}) of an enzyme E_i towards a metabolite

X_j is a measure of the sensitivity of the rate of the reaction catalyzed by the enzyme with respect to a fractional change in concentration of the metabolite. This elasticity coefficient is expressed as

$$\varepsilon_{ji} = \frac{x_j}{v_i}\frac{\partial v_i}{\partial x_j}.$$

Here v_i is the rate of the reaction catalyzed by enzyme E_i. For pathway (A), we have

$$\tilde{E} = \begin{bmatrix} -1 & -1 & -1 & -1 & \cdots & -1 & -1 & -1 & -1 \\ -\varepsilon_{21} & -\varepsilon_{22} & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & -\varepsilon_{32} & -\varepsilon_{33} & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & -\varepsilon_{43} & -\varepsilon_{44} & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -\varepsilon_{n-3,n-3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & -\varepsilon_{n-2,n-3} & -\varepsilon_{n-2,n-2} & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -\varepsilon_{n-1,n-2} & -\varepsilon_{n-1,n-1} & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & -\varepsilon_{n,n-1} & -\varepsilon_{nn} \end{bmatrix}.$$

It should be pointed out that $\varepsilon_{i+1,i}$ is the elasticity coefficient of enzyme E_i due to inhibition by its product and is always negative, whereas ε_{ii} refers to the elasticity coefficient of enzyme E_i with respect to its substrate X_i and is always positive. The control matrix $\tilde{Z}$ is given by

$$\tilde{Z} = \begin{bmatrix} -C_1^J & C_1^{X_2} & C_1^{X_3} & C_1^{X_4} & \cdots & C_1^{X_{n-3}} & C_1^{X_{n-2}} & C_1^{X_{n-1}} & C_1^{X_n} \\ -C_2^J & C_2^{X_2} & C_2^{X_3} & C_2^{X_4} & \cdots & C_2^{X_{n-3}} & C_2^{X_{n-2}} & C_2^{X_{n-1}} & C_2^{X_n} \\ -C_3^J & C_3^{X_2} & C_3^{X_3} & C_3^{X_4} & \cdots & C_3^{X_{n-3}} & C_3^{X_{n-2}} & C_3^{X_{n-1}} & C_3^{X_n} \\ -C_4^J & C_4^{X_2} & C_4^{X_3} & C_4^{X_4} & \cdots & C_4^{X_{n-3}} & C_4^{X_{n-2}} & C_4^{X_{n-1}} & C_4^{X_n} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ -C_{n-3}^J & C_{n-3}^{X_2} & C_{n-3}^{X_3} & C_{n-3}^{X_4} & \cdots & C_{n-3}^{X_{n-3}} & C_{n-3}^{X_{n-2}} & C_{n-3}^{X_{n-1}} & C_{n-3}^{X_n} \\ -C_{n-2}^J & C_{n-2}^{X_2} & C_{n-2}^{X_3} & C_{n-2}^{X_4} & \cdots & C_{n-2}^{X_{n-3}} & C_{n-2}^{X_{n-2}} & C_{n-2}^{X_{n-1}} & C_{n-2}^{X_n} \\ -C_{n-1}^J & C_{n-1}^{X_2} & C_{n-1}^{X_3} & C_{n-1}^{X_4} & \cdots & C_{n-1}^{X_{n-3}} & C_{n-1}^{X_{n-2}} & C_{n-1}^{X_{n-1}} & C_{n-1}^{X_n} \\ -C_n^J & C_n^{X_2} & C_n^{X_3} & C_n^{X_4} & \cdots & C_n^{X_{n-3}} & C_n^{X_{n-2}} & C_n^{X_{n-1}} & C_n^{X_n} \end{bmatrix}.$$

Clearly, if all the elasticity coefficients are known, then the flux control coefficients and concentration control coefficients can be found from the inverse matrix $\tilde{E}^{-1}$. In particular, the entries in the first column of $\tilde{E}^{-1}$ give the negatives of the flux control coefficients of the enzymes $E_1 - E_n$; the entries in columns two through n yield the concentration control coefficients of the various enzymes with respect to the metabolites $X_2, X_3 \ldots X_n$, respectively. The sign of a flux control coefficient of an enzyme indicates the direction in which the flux will change in response to a change in activity of the enzyme. For example, if the flux control coefficient of an enzyme is negative, the flux will decrease (increase) with an increase (decrease) in enzyme activity. Similarly, the sign of a concentration control coefficient of an enzyme with respect to a particular metabolite indicates the direction of change in the concentration of the metabolite due to a change in enzyme activity.

We indicate the elasticity matrix of a general linear pathway, i.e., one with possible feedback and/or feedforward loops, by $\bar{E}$. Our goal is to investigate the sign-nonsingularity of the elasticity matrix and determine, whenever possible, the signs of the flux control and concentration control coefficients. Our approach of a purely qualitative analysis encompasses the quantitative restriction that each entry in the first

row of an elasticity matrix is always equal to -1, since by definition, if an elasticity matrix is sign-nonsingular, it must be nonsingular. On the other hand, if a matrix has the property that every elasticity matrix with the same sign pattern is nonsingular, then it must itself be sign-nonsingular. To put this in perspective, consider an $n \times n$ elasticity matrix $\bar{E}$. Suppose that a matrix $B = [b_{ij}]$ has the same sign pattern as $\bar{E}$ and B is singular. If $F = [f_{ij}]$ is the diagonal matrix defined by

$$f_{jj} = -\frac{1}{b_{1j}},$$

then it follows that $\bar{E}' = BF$ is also singular, $\bar{E}'$ has the same sign pattern as $\bar{E}$, and every entry in the first row of $\bar{E}'$ is equal to -1. A similar reasoning demonstrates that we may also use a purely qualitative approach for the problem of finding which flux control and concentration control coefficients have their signs unambiguously determined. In the following development we represent all matrices as sign pattern matrices.

Consider the n-enzyme linear pathway (A), which has neither feedback nor feedforward regulation. We refer to this pathway as the *unregulated* pathway. The sign pattern of the elasticity matrix of this pathway is given by

$$\hat{E}_n = \begin{bmatrix} - & - & - & - & \cdots & - & - & - & - \\ + & - & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & + & - & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & + & - & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & - & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & + & - & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & + & - & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & + & - \end{bmatrix}.$$

Note that this sign-pattern matrix has a bidiagonal structure with a first row of negative entries. The entries on the main diagonal are all negative, whereas the entries on the subdiagonal are all positive.

3. Sign-nonsingularity. We begin this section with some graph theoretic definitions. A *directed graph,* or *digraph,* D, is an ordered pair (N, E) where N is a finite set and E is a set of ordered pairs of elements of N such that E contains no pair of the form (v, v) for $v \in N$. Each element of N is called a *vertex* and each element of E is called an *arc* of D. For any vertex i we may define an *indegree* and an *outdegree,* denoted respectively by $\mathrm{id}(i)$ and $\mathrm{od}(i)$. The indegree (outdegree) of a vertex is the total number of arcs entering (leaving) the vertex. A *subdigraph* of a digraph D is an ordered pair (N', E'), with $N' \subseteq N$ and $E' \subseteq E$. Suppose that in a digraph D there is an ordered set of distinct vertices $(i_1, i_2, \ldots, i_q)$ with $q > 1$, and each of the arcs $(i_k, i_{k+1}) \in E$ for $k = 1, 2, \ldots, q-1$. This ordered set is called a *path* of D. We commonly indicate a path from vertex i to vertex j by the symbol $p(i \to j)$. A digraph is said to be *strongly connected* or simply *strong* if, given any two distinct vertices i and j, there exists a path $p(i \to j)$. An ordered set of vertices $(i_1, i_2, \ldots, i_q, i_1)$ is called a *cycle* of D if $(i_1, i_2, \ldots, i_q)$ is a path and $(i_q, i_1) \in E$. With any real $n \times n$ matrix $A = [a_{ij}]$ we may associate a digraph $D = (N, E)$ where $N = \{1, 2, \ldots, n\}$ and $(i, j) \in E$ if and only if $a_{ji} \neq 0$. It is well known that a matrix is irreducible if and only if its digraph is strongly connected.

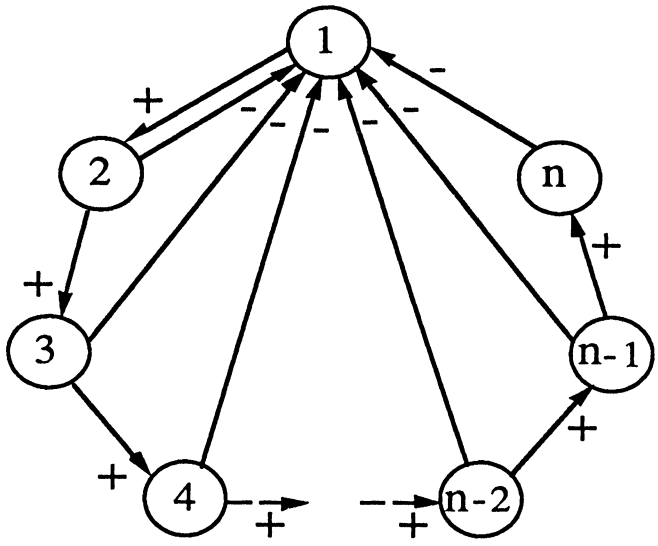


FIG. 1. *The signed digraph* $\hat{S}_n$.

A *signed digraph* $S = (D; \sigma)$ is a digraph D, along with a mapping σ: $E \to \{+, -\}$, i.e., a digraph with a plus or minus sign given to each arc. A path (cycle) in a signed digraph is *negative* (*positive*) if the product of the signs of the arcs constituting the path (cycle) is negative (positive). If all the cycles in a signed digraph S are negative, then S is said to have the *negative cycle property.*

With the elasticity matrix $\bar{E} = [e_{ij}]$ of a linear pathway, we may associate a digraph $\bar{D}(\bar{E})$ and a signed digraph $\bar{S}(\bar{E}) = (\bar{D}(\bar{E}); \sigma)$ that is generated from $\bar{D}(\bar{E})$ by assigning to each arc (i, j) the sign $\sigma(i, j)$, where $\sigma(i, j) = +$ if $e_{ji} > 0$ and $\sigma(i, j) = -$ if $e_{ji} < 0$. Thus, to construct the signed digraph of an $n \times n$ elasticity matrix $\bar{E}, n$ vertices are drawn numbered $1, 2, 3, \ldots, n$. If $e_{ij} \neq 0$, for $i \neq j$, then an arc is directed from vertex j to vertex i; this arc is given the sign of e_{ij}. The diagonal entries of $\bar{E}$ do not contribute to any arcs or cycles in the digraph.

We will denote by $\bar{E}_n$ the sign-pattern elasticity matrix of an n-enzyme linear pathway with possible feedback and feedforward loops. The corresponding signed digraph will be designated by $\bar{S}_n$. For the unregulated n-enzyme pathway (A), the sign-pattern elasticity matrix and its digraph are denoted by $\hat{E}_n$ and $\hat{S}_n$, respectively. The matrix $\hat{E}_n$ is given in §2. The signed digraph $\hat{S}_n$ has the following structure, as shown in Fig. 1.

A characteristic feature of $\hat{S}_n$ is the Hamilton cycle $\bar{H}_n = (1, 2, \ldots, n, 1)$ in which every arc is positive, except the arc $(n, 1)$. Thus $\hat{S}_n$ is strongly connected; accordingly, the matrix $\hat{E}_n$ is irreducible. Since $\bar{S}_n$ contains $\hat{S}_n$ as a subdigraph, we may conclude that the elasticity matrix of any linear pathway is irreducible.

From the signed digraph of a matrix, sign-nonsingularity of the matrix can be characterized by the following theorem due to Bassett, Maybee, and Quirk [9].

THEOREM A. *Let* $A = [a_{ij}]$ *be a* $(0, 1, -1)$ *matrix of order* n *with* $a_{ii} < 0, i = 1, 2, \ldots, n$. *Then* A *is sign-nonsingular if and only if the associated signed digraph has the negative cycle property.*

It should be mentioned that the signed digraph referred to in this theorem has the following definition: it contains n vertices designated $1, 2, \ldots, n$, and if $a_{ij} \neq 0$, then an arc is drawn from vertex i to vertex j carrying the sign of a_{ij}. In contrast, in Fig. 1,

$$\hat{E}_4 = \begin{bmatrix} - & - & - & - \\ + & - & 0 & 0 \\ 0 & + & - & 0 \\ 0 & 0 & + & - \end{bmatrix}$$

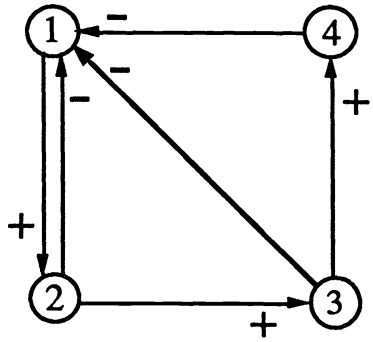


FIG. 2

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

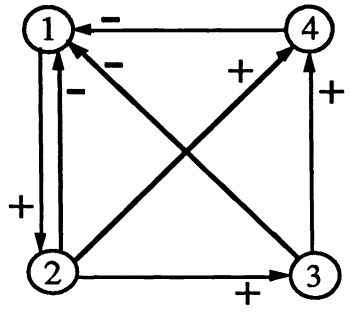


FIG. 3

an arc is drawn from vertex j to vertex i carrying the sign of e_{ij} whenever $e_{ij} \neq 0$. Thus, our procedure for constructing the signed digraph for an elasticity matrix $\bar{E}$ is equivalent to that of Bassett, Maybee, and Quirk [9] for the transpose of $\bar{E}$. Theorem A applies for our convention as well, since there is a one-to-one correspondence, which preserves signs, between the cycles of the two types of signed digraphs. Our convention is more prevalent in the engineering literature and was used earlier in metabolic control analysis [3]–[6].

To illustrate the application of Theorem A for determining the nonsingularity of an elasticity matrix, we present a few examples. First, consider a four-enzyme pathway (B) that contains neither feedback nor feedforward regulation.

(B) $$X_1 \xrightarrow{E_1} X_2 \xrightarrow{E_2} X_3 \xrightarrow{E_3} X_4 \xrightarrow{E_4} P.$$

For this pathway the sign pattern of the elasticity matrix and the associated signed digraph are shown in Fig. 2.

Clearly, since all the cycles in this signed digraph are negative, Theorem A implies that the elasticity matrix for pathway (B) is sign-nonsingular.

Next we examine the effect of feedback inhibition. Consider a four-enzyme pathway in which the enzyme E_2 is inhibited by the metabolite X_4.

(C) $$X_1 \xrightarrow{E_1} X_2 \xrightarrow{E_2} X_3 \xrightarrow{E_3} X_4 \xrightarrow{E_4} P.$$

The sign-pattern elasticity matrix and its signed digraph are depicted in Fig. 3.

Note that the sign-pattern matrix in Fig. 3 is derived from the matrix $\hat{E}_4$ (see Fig. 2) by replacing the zero in the $(4, 2)$ position with a plus sign to account for

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

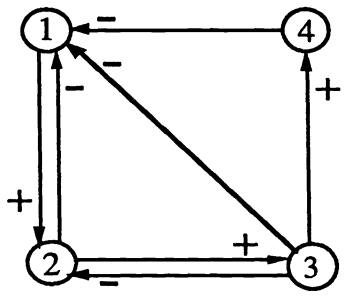


FIG. 4

feedback inhibition. Accordingly, the signed digraph of Fig. 3 results from that of Fig. 2 with the addition of the positive arc from vertex 2 to vertex 4. Since the addition of this arc introduces only a negative cycle $(1, 2, 4, 1)$ into the signed digraph, it follows from Theorem A that the elasticity matrix of pathway (C) is also sign-nonsingular.

Consider now the role of feedforward activation. For this purpose, we examine pathway (D) in which the enzyme E_3 is activated by the metabolite X_2.

(D) $$X_1 \xrightarrow{E_1} X_2 \xrightarrow{E_2} X_3 \xrightarrow{E_3} X_4 \xrightarrow{E_4} P.$$

The sign-pattern elasticity matrix of this pathway and the associated digraph are seen in Fig. 4.

The matrix in Fig. 4 is obtained from $\hat{E}_4$ by replacing the zero in the $(2, 3)$ position with a minus sign, representing the feedforward activation loop. The elasticity matrix of this pathway is sign-nonsingular, since all the cycles in the signed digraph are negative.

More generally, we consider an n-enzyme pathway with (a) neither feedback nor feedforward regulation, (b) an arbitrary number of feedback inhibition loops, and (c) an arbitrary number of feedforward activation loops.

THEOREM 3.1. *If a linear pathway has neither feedback nor feedforward regulation, or contains only feedforward activation loops, then its elasticity matrix is sign-nonsingular and is contained in a maximal sign-nonsingular matrix in upper Hessenberg form. If a linear pathway contains only feedback inhibition loops, then its elasticity matrix is sign-nonsingular and is contained in a maximal sign-nonsingular matrix that is permutation equivalent to a matrix in lower Hessenberg form.*

A square matrix $A = [a_{ij}], i, j = 1, 2, \ldots, n$, is called a lower (upper) Hessenberg matrix if $a_{ij} = 0$ for all pairs (i, j) such that $i + 1 < j (j + 1 < i)$.

Proof. Note that the elasticity matrix $\hat{E}_n$ (see §2) of an unregulated n-enzyme linear pathway is in upper Hessenberg form. For an n-enzyme pathway with feedforward activation, the elasticity matrix $\bar{E}_n$ is obtained from $\hat{E}_n$ by replacing certain zeros in the upper triangle of $\hat{E}_n$ with minus signs. The resulting matrix is obviously contained in a maximal sign-nonsingular matrix in upper Hessenberg form. To establish the result for a linear pathway with feedback inhibition loops, let $P_n = [p_{ij}]$ be the $n \times n$ permutation matrix with $p_{ij} = 1$ if and only if $j = (i + 1) \bmod n$ and let $T_n = [t_{ij}]$ be the $n \times n$ signature matrix defined by $t_{ii} = -1, i \in \{1, 2, \ldots, n - 1\}$ and $t_{nn} = 1$. It is now easy to verify that if an n-enzyme linear pathway, with elasticity matrix $\bar{E}_n$, contains only feedback inhibition loops, then $T_n P_n \bar{E}_n$ is sign-nonsingular and is

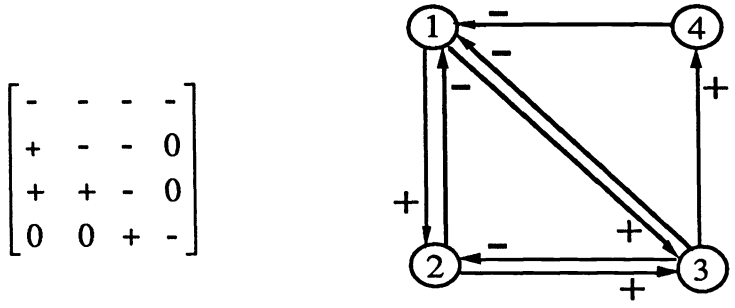


FIG. 5

contained in a maximal sign-nonsingular matrix in lower Hessenberg form. □

Let us now address the effect of *feedback activation* or *feedforward inhibition* on the sign-nonsingularity of the elasticity matrix of a linear pathway. If an n-enzyme linear pathway contains a feedback activation loop, its effect can be taken into account by replacing an appropriate zero in the lower triangular region of the elasticity matrix $\hat{E}_n$ with a plus. Thus, if a linear pathway has a feedback activation loop involving the metabolite X_k and the enzyme E_l (with $k > l$), then its signed digraph $\bar{S}$ contains the positive cycle $\bar{H}_n(k \to l)(l, k)$, where $\bar{H}_n(k \to l)$ refers to the path in $\bar{H}_n$. Accordingly, it follows from Theorem A that if a linear pathway possesses feedback activation, its elasticity matrix cannot be sign-nonsingular. Similarly, if a linear pathway possesses feedforward inhibition involving a metabolite X_k and the enzyme E_l (with $k < l$), then its signed digraph contains the positive cycle $\bar{H}_n(k \to l)(l, k)$. As a consequence, Theorem A implies that its elasticity matrix cannot be sign-nonsingular.

If a linear pathway contains both feedback inhibition and feedforward activation loops, then the sign-nonsingularity of the elasticity matrix depends on the relative positions of these loops. In a four-enzyme pathway there are five possible configurations in which a feedback inhibition loop and a feedforward activation loop can be present. These are: (a) the feedback loop lies completely inside the feedforward loop, (b) the feedforward loop lies completely inside the feedback loop, (c) the same metabolite inhibits one enzyme and activates a different enzyme, (d) the feedback and feedforward loops overlap each other in an *overhanging* fashion with the enzyme being inhibited lying downstream from the activating metabolite, and (e) the two loops overlap each other in an overhanging fashion but the enzyme undergoing feedback inhibition is located upstream from the activating metabolite. It can be shown that of these five possibilities, the first four lead to a sign-nonsingular elasticity matrix, whereas the last configuration does not have a sign-nonsingular elasticity matrix. To put this in perspective, consider the following pathway.

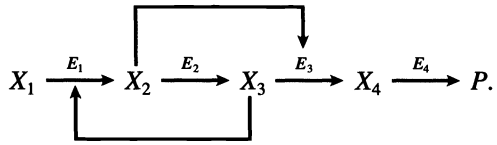
(E) 

Figure 5 shows the sign pattern of the elasticity matrix of this pathway and the associated signed digraph.

In this digraph, notice that the cycle $(1, 3, 2, 1)$ is positive. As a result, by Theorem A, the elasticity matrix of this pathway cannot be sign-nonsingular.

More generally, consider an n-enzyme linear pathway that contains a pair of

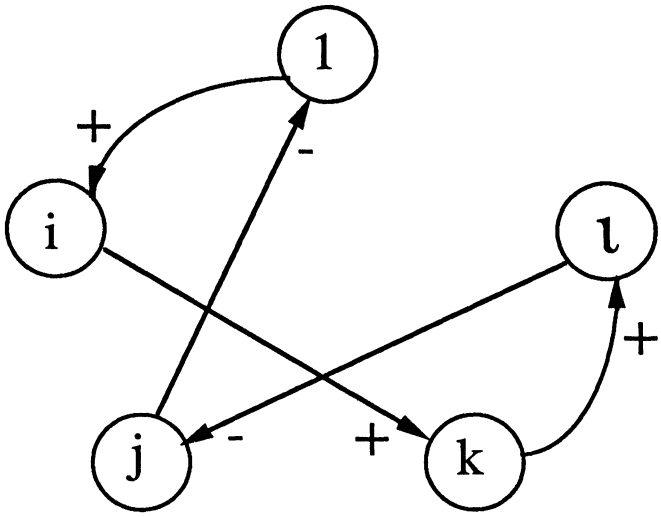


FIG. 6

overhanging loops with the following characteristics:

(i) the metabolite X_k inhibits the enzyme E_i;

(ii) the metabolite X_j activates the enzyme E_l;

(iii) $1 \leq i < j < k \leq l \leq n$.

In other words, the enzyme E_i that undergoes feedback inhibition lies upstream of the activating metabolite X_j. Such a pathway is said to contain a *singular pair of loops.* For example, pathway (E) contains a singular pair of loops, with $i = 1, j = 2, k = l = 3$. If a linear metabolic pathway contains a singular pair of loops, then its elasticity matrix cannot be sign-nonsingular. To see this, consider the n-enzyme linear pathway described above. Let $\bar{S}_n$ be the signed digraph of the elasticity matrix of this pathway. If $i \neq 1$, then $\bar{S}_n$ contains the positive cycle $\bar{H}_n(1 \to i)(i,k)\bar{H}_n(k \to l)(l,j)(j,1)$, provided $k \neq 1$; see Fig. 6. The positive cycles for $i = 1$ and/or $k = l$ are also evident from this figure.

The following theorem describes the connection between the presence of feedback activation, feedforward inhibition, or a singular pair of loops in a linear metabolic pathway, and the sign-nonsingularity of the elasticity matrix.

THEOREM 3.2. *The elasticity matrix of a linear pathway is not sign-nonsingular if and only if the pathway contains* (a) *a feedback activation loop,* (b) *a feedforward inhibition loop, or* (c) *a singular pair of loops.*

The sufficiency of the theorem has been established in the preceding paragraphs. Before we can prove the necessary part of Theorem 3.2, we need a few definitions and results.

Let $S = (D, \sigma) = (N, E; \sigma)$ be a strong signed digraph on n vertices. Suppose S has a vertex i with indegree one, with an arc (j,i) directed from vertex j to vertex i. Suppose further that the following conditions apply: (a) if h is another vertex with $(i,h) \in E$ and $(j,h) \in E$, then $\sigma(j,h) = \sigma(j,i)\sigma(i,h)$, and (b) if $(i,j) \in E$, then $\sigma(i,j) = -\sigma(j,i)$. Under these conditions, we can use the following rules to eliminate the vertex i from S and create a strong digraph S' on $n-1$ vertices: $S' = (D', \sigma') = (N/\{i\}, E'; \sigma')$.

Rule (i). If $(k,l) \in E$ and $k, l \neq i$, then $(k,l) \in E'$ with $\sigma'(k,l) = \sigma(k,l)$.

Rule (ii). If $(i,l) \in E$ with $l \neq i$, then $(j,l) \in E'$ with $\sigma'(j,l) = \sigma(i,l)\sigma(j,i)$.

Elimination of the vertex i using Rule (ii) may be referred to as *contracting* the arc (j,i), which is incident on the vertex i. Similar procedures were used earlier [6] to eliminate vertices from a digraph, for the purpose of simplifying the calculation of the control coefficients. It is well known (see [10]) that S' has the negative cycle property if and only if S has the negative cycle property. Similarly, suppose that the vertex i has outdegree one, with the arc $(i,j) \in E$. Furthermore, suppose that the following conditions hold: (a) if h is another vertex with $(h,i) \in E$ and $(h,j) \in E$, then $\sigma(h,j) = \sigma(h,i)\sigma(i,j)$; and (b), if $(j,i) \in E$, then $\sigma(j,i) = -\sigma(i,j)$. We may now create a new signed digraph S' on $n-1$ vertices, by using Rule (i) and the Rule (iii).

Rule (iii). If $(l,i) \in E$ with $l \neq i$, then $(l,j) \in E'$ with $\sigma'(l,j) = \sigma(l,i)\sigma(i,j)$. It follows that S' has the negative cycle property if and only if S does.

We use these operations in the context of signed digraphs that are associated with linear metabolic pathways. Consider an n-enzyme linear pathway with possible feedback and feedforward loops. In the following development, we refer to this pathway as L. The signed digraph of the elasticity matrix of such a pathway is denoted by $\bar{S}$. We utilize the following mappings, whenever appropriate, to map $\bar{S}$ onto a signed digraph on $(n-1)$ vertices that contains $\hat{S}_{n-1}$ as a subdigraph.

Type 1. An arc of the form $(k,k+1)$ for some $1 \leq k \leq n-1$ is contracted, according to Rules (i) and (ii), with every vertex $k+2 \leq j \leq n$ being relabeled $j-1$.

Type 2. The arc $(n,1)$ is contracted according to Rules (i) and (iii).

Type 3. The arc $(1,2)$ is contracted according to Rules (i) and (iii), with every vertex $2 \leq j \leq n$ being relabeled $j-1$.

We will call these mappings *contraction mappings* of digraphs. If $\bar{S}$ contains a vertex i of indegree or outdegree one, but i cannot be eliminated by a contraction mapping of digraphs, then L must contain (a) feedback activation or (b) feedforward inhibition. To see this, suppose that the vertex n has outdegree one, but cannot be eliminated by a Type 2 contraction mapping. Then the metabolite X_n must activate an enzyme located upstream from it in the pathway. If the vertex 1 has outdegree one, but cannot be eliminated by a Type 3 contraction mapping, then the metabolite X_2 must inhibit an enzyme located downstream from it. Now suppose that the vertex $k+1$ has indegree one, for some k with $1 \leq k \leq n-1$, and it cannot be eliminated by a Type 1 contraction mapping. If the vertex $k+1$ cannot be eliminated because $(k+1,k) \in E$ and $(k+1,k)$ is positive, then the enzyme E_{k+1} is inhibited by the metabolite X_k. If there is some vertex $h \neq k, k+1$ such that $(k,h) \in E, (k+1,h) \in E$, and $\sigma(k,h) \neq \sigma(k+1,h)$, then one of the two enzymes E_k and E_{k+1} must undergo (a) feedback activation or (b) feedforward inhibition.

We now extend the idea of contraction mapping of digraphs to transform an n-enzyme linear pathway into a linear pathway with $(n-1)$ enzymes. Suppose that we may contract $\bar{S}$ to produce a signed digraph S' which contains $\hat{S}_{n-1}$ as a subdigraph. Clearly, S' corresponds to a linear pathway L' with $(n-1)$ enzymes, and possibly containing feedforward and feedback loops. We have

$$L \to \bar{E} \to \bar{S} \xrightarrow{\text{contraction mapping}} S' \to L'.$$

We will call such a mapping a *contraction mapping of pathways,* and identify it as Type 1, Type 2, or Type 3 according to the type of contraction mapping used to map $\bar{S}$ to S'. We have the following result.

LEMMA 3.3. *Suppose that L is a linear pathway with n enzymes, and that L' is a linear pathway with $(n-1)$ enzymes. Furthermore, suppose that L' is obtained from*

L by a contraction mapping. The pathway L' contains (a) *a feedback activation loop,* (b) *a feedforward inhibition loop, or* (c) *a singular pair of loops only if L does.*

Proof. We prove the result by explicitly describing how the topology of L' is obtained from the topology of L, given the type of contraction mapping employed. If the mapping is of Type 2, we may obtain L' directly from L by deleting X_n, E_n and all the feedback loops involving X_n. Therefore, if L' contains (a) a feedback activation loop, (b) a feedforward inhibition loop, or (c) a singular pair of loops, then so must L. If the mapping is of Type 3, we may obtain L' from L by deleting X_1, E_1, and all the feedforward loops involving X_1, and then renumbering all the remaining enzymes and metabolites with their indices reduced by one. Therefore, if L' contains (a) a feedback activation loop, (b) a feedforward inhibition loop, or (c) a singular pair of loops, so must L. Now suppose that the contraction mapping is of Type 1. In this case, we may obtain L' from L as follows: identify the metabolites X_k and X_{k+1}, delete the enzyme E_k and replace all loops containing E_k with a loop containing E_{k+1} and the same metabolite, and renumber all the remaining enzymes and metabolites that are downstream of X_k. Thus, L' contains (a) a feedback activation loop, (b) a feedforward inhibition loop, or (c) a singular pair of loops only if L does. □

We need one more result that concerns signed digraphs with the negative cycle property. It appears in [10], although it is expressed there in terms of sign-nonsingular matrices.

THEOREM B. *Suppose that $S = (N, E; \sigma)$ is a strong signed digraph on n vertices, S possesses the negative cycle property, and there exists $i \in N$ with* $\mathrm{id}(i) = n-1$. *Then there exists $j \in N$ with* $\mathrm{id}(j) = 1$.

We may now proceed to the proof of our main result.

Proof of Theorem 3.2. The result is easily verified for linear pathways with a small number of enzymes by constructing the associated signed digraphs. Now suppose that Theorem 3.2 is not true in general. In particular, let M be a shortest linear pathway, with m enzymes, for which Theorem 3.2 does not hold. We will denote the elasticity matrix of M by $\bar{E}$ and its signed digraph by $\bar{S}$. Note that $\bar{S}$ cannot have the negative cycle property. Suppose $\bar{S}$ contains a vertex with indegree or outdegree one. If we cannot perform a contraction mapping, then M must contain a feedback activation loop or a feedforward inhibition loop. On the other hand if we are able to perform a contraction mapping, we may reduce the pathway M to a linear pathway M' containing $(m-1)$ enzymes. The elasticity matrix of M' cannot be sign-nonsingular. Furthermore, our choice of m as minimal implies that M' must contain (a) a feedback activation loop, (b) a feedforward inhibition loop, or (c) a singular pair of loops. Thus, according to Lemma 3.3, M itself must contain (a) a feedback activation loop, (b) a feedforward inhibition loop, or (c) a singular pair of loops. In other words, Theorem 3.2 must apply for the pathway M. We may therefore conclude that every vertex of $\bar{S}$ must have indegree and outdegree at least two. As a consequence of this, it can be shown that $(1, m) \notin E$ in $\bar{S}$. For if we had $(1, m) \in E$, we must have $(m, j) \in E$ for some j with $2 \leq j \leq m-1$, since $\mathrm{od}(m) \geq 2$. However, this implies that L contains a singular pair of loops, contrary to our choice of L as a counterexample. Now let k be the least index such that $(1, k) \in E$; we must have $2 \leq k \leq m-1$. Obviously, the subdigraph of $\bar{S}$ that is induced by $\{1, 2, \ldots, k\}$ is strongly connected. Theorem A and our choice of m as minimal imply that this induced digraph has the negative cycle property, since it is the signed digraph associated with the linear pathway, say, L'' which is formed from L by deleting every enzyme and metabolite downstream from the enzyme E_k. Therefore, if L'' contains (a) a feedback activation loop, (b) a feedforward inhibition loop, or (c) a singular pair of loops, so does L. However,

Theorem B implies that for some pair of indices l and j, with $l > k > j, (l, j) \in E$. This in turn implies that L contains a singular pair of loops, so that L cannot be a counterexample, and Theorem 3.2 is proved. □

4. Sign-determined and sign-indeterminate control coefficients. Consider an n-enzyme linear metabolic pathway which has a sign-nonsingular elasticity matrix. We show that if the pathway has neither feedback nor feedforward regulation, then the signs of *all* the control coefficients can be determined unambiguously. If the pathway contains feedback inhibition and/or feedforward activation loops, then certain control coefficients will be sign-indeterminate. The enzymes and metabolites whose control coefficients are sign-indeterminate can be identified directly from the topology of the metabolic pathway.

To begin our discussion, we introduce some concepts from Lady and Maybee [11]. Let $A = [a_{ij}]$ be an irreducible $n \times n$ sign-nonsingular matrix. The entry $a_{ij} = 0$ is called an *essential zero* of A if any matrix obtained by setting $a_{ij} \neq 0$ is not sign-nonsingular, irrespective of the sign of a_{ij}. We have the following results from [11].

THEOREM C. *Let A be an irreducible $n \times n$ sign-nonsingular matrix with $a_{ii} < 0, i = 1, 2, \ldots, n$. Furthermore, suppose that A has associated signed digraph $S(A)$ and that $A^{-1} = [\alpha_{ij}]$. Then*

(i) $\alpha_{ii} < 0, i = 1, 2, \ldots, n$.

(ii) *if $a_{ij} \neq 0$,* $\operatorname{sign} \alpha_{ji} = \operatorname{sign} a_{ij}$, *and*

(iii) *if $a_{ij} = 0$, then the sign of α_{ji} is unambiguously determined if and only if every path $p(i \to j)$ in $S(A)$ has the same sign, say, $(-1)^{\delta}$. In this case,* $\operatorname{sign} \alpha_{ji} = (-1)^{\delta+1}$.

THEOREM D. *Let A satisfy the conditions of Theorem* C. *Then the entry $a_{ij} = 0$ is an essential zero of A if and only if there exist paths $p_1(i \to j)$ and $p_2(i \to j)$ in $S(A)$ with opposite signs.*

COROLLARY 4.1. *Suppose A satisfies the conditions of Theorem* C. *Then the entry α_{ji} is sign-indeterminate if and only if a_{ij} is an essential zero.*

We now apply these results to determine the signs of the control coefficients of linear metabolic pathways. First we consider an n-enzyme pathway with neither feedback nor feedforward regulation. The sign-pattern elasticity matrix, the associated signed digraph, and the control matrix for this pathway are given by $\hat{E}_n, \hat{S}_n$, and $\tilde{Z}$, respectively (see §2). Using Theorem C, the signs of the various control coefficients can be ascertained as follows. Since all the diagonal entries in $\hat{E}_n$ are negative, it follows from part (i) of Theorem C that the diagonal entries of the control matrix $\tilde{Z}$ are also negative. Furthermore, by part (ii) of Theorem C, we find that (a) the first column of $\tilde{Z}$ must have all negative entries, and (b) the entries in the first superdiagonal of $\tilde{Z}$ must be positive. To determine the signs of the remaining entries in $\tilde{Z}$, we now consider the zero entries below the first subdiagonal of $\hat{E}_n$. Note that if $i > j + 1$, then every path $p(i \to j)$ in $\hat{S}_n$ is negative, since it must pass through the vertex 1. Therefore, according to part (iii) of Theorem C, all the entries above the first superdiagonal in $\tilde{Z}$ are positive. Finally we consider the entries in the upper triangle of $\hat{E}_n$ other than those in the first row. For $i < j$ with $i > 1$, it is easy to see that $\bar{H}_n(i \to j)$ is the only path in $\hat{S}_n$ from vertex i to vertex j. Since this path is positive, we conclude that every entry in the lower triangle of $\tilde{Z}$ is negative. Collectively, all the entries on and below the main diagonal of $\tilde{Z}$ are negative, whereas the entries above the main diagonal of $\tilde{Z}$ are positive. In other words, for an unregulated linear pathway, the signs of all the control coefficients are unambiguously determined.

Next we examine an n-enzyme linear pathway with possible feedback inhibition

and feedforward activation loops, and which has a sign-nonsingular elasticity matrix. For such a pathway, the control coefficients which are sign-indeterminate can be identified with the aid of Corollary 4.1; they correspond to the essential zeros of the elasticity matrix. The remaining control coefficients have the same fixed signs as those in the unregulated pathway. This follows from part (iii) of Theorem C and the fact that in the digraph $\bar{S}_n$, there exists a path $\bar{H}_n(i \to j)$ from vertex i to vertex j.

Theorem 3.2 tells us exactly which zero entries in an elasticity matrix are essential zeros. An essential zero corresponds to a feedback or a feedforward loop whose addition would create a singular pair of loops in the metabolic pathway. It follows that if a linear pathway has a feedback inhibition loop, then the concentration control coefficients of the enzymes that lie downstream from the feedback loop, with respect to each of the metabolites located inside the loop, are sign-indeterminate. On the other hand, if a linear pathway contains a feedforward loop, then the concentration control coefficients of the enzymes located upstream of the loop, with respect to each of the metabolites that are located inside the loop, are sign-indeterminate.

We illustrate these results with a few examples. Consider first the four-enzyme unregulated pathway (B). All the control coefficients of this pathway have their signs unambiguously determined. They are presented below in a matrix form.

$$\begin{array}{c} \\ E_1 \\ E_2 \\ E_3 \\ E_4. \end{array} \begin{array}{c} \begin{array}{cccc} J & X_2 & X_3 & X_4 \end{array} \\ \begin{bmatrix} - & + & + & + \\ - & - & + & + \\ - & - & - & + \\ - & - & - & - \end{bmatrix} \end{array} .$$

For clarity of interpretation we have written the variables J, X_2, X_3, and X_4 horizontally above this matrix and the enzymes $E_1 - E_4$ vertically on the left side. As an example, the $(2,3)$ entry in this matrix denotes the sign of $C_2^{X_3}$, which is the concentration control coefficient of the enzyme E_2 with respect to the metabolite X_3. Note that the entries in the first column represent the signs of $-C_i^J, i = 1,2,3,4$, where C_i^J is the flux control coefficient of the enzyme E_i. Clearly, the flux control coefficients of all the enzymes are positive.

For the pathway (C), the signs of the various control coefficients are given by the following matrix.

$$\begin{array}{c} \\ E_1 \\ E_2 \\ E_3 \\ E_4 \end{array} \begin{array}{c} \begin{array}{cccc} J & X_2 & X_3 & X_4 \end{array} \\ \begin{bmatrix} - & + & + & + \\ - & - & + & + \\ - & - & - & + \\ - & - & * & - \end{bmatrix} \end{array} .$$

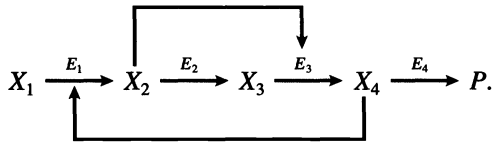
Note that the control coefficients of this pathway have the same signs as those of the unregulated pathway (B) except for $C_4^{X_3}$, which is sign-indeterminate. This is because the enzyme E_4 lies downstream from the feedback loop in the pathway, and the metabolite X_3 is located inside the loop.

Using our results, the signs of the control coefficients of the enzymes in pathway (D) are found to be as shown in the following matrix.

$$\begin{array}{c} \\ E_1 \\ E_2 \\ E_3 \\ E_4 \end{array} \begin{array}{c} \begin{array}{cccc} J & X_2 & X_3 & X_4 \end{array} \\ \begin{bmatrix} - & + & * & + \\ - & - & + & + \\ - & - & - & + \\ - & - & - & - \end{bmatrix} \end{array}$$

Observe that in this pathway the signs of the control coefficients are the same as those in the pathway (B) except for $C_1^{X_3}$ which is sign-indeterminate.

Finally, we examine the control coefficients of a linear pathway containing both feedback and feedforward loops. In particular, consider pathway (F), which has a sign-nonsingular elasticity matrix.

(F) 

All the control coefficients of this pathway except $C_4^{X_2}, C_4^{X_3}$, and $C_1^{X_3}$ have fixed signs; these fixed signs are the same as those in pathway (B). In a matrix form the result is

$$\begin{array}{c} \\ E_1 \\ E_2 \\ E_3 \\ E_4 \end{array} \begin{array}{c} \begin{array}{cccc} J & X_2 & X_3 & X_4 \end{array} \\ \begin{bmatrix} - & + & * & + \\ - & - & + & + \\ - & - & - & + \\ - & * & * & - \end{bmatrix} \end{array}.$$

5. Concluding remarks. We have analyzed the sign pattern of the control coefficients of the enzymes in linear metabolic pathways. It is shown that the elasticity matrices of the following linear pathways are sign-nonsingular: (a) a linear pathway with neither feedback nor feedforward regulation and (b) a linear pathway with only feedback inhibition or feedforward activation loops. If a linear pathway contains feedback activation or feedforward inhibition, its elasticity matrix cannot be sign-nonsingular. When a linear pathway contains a feedback inhibition loop and a feedforward activation loop, then it has a sign-nonsingular elasticity matrix except when the two regulatory loops are located in an overhanging fashion and the enzyme undergoing feedback inhibition lies upstream of the activating metabolite.

For a linear pathway with neither feedback nor feedforward regulation, we have shown that the signs of all the control coefficients are unambiguously determined. If a linear pathway contains a feedback inhibition loop, then all the control coefficients are sign-determined except for the concentration control coefficients of the enzymes which are located downstream of the feedback loop, with respect to each of the metabolites which are inside the loop. On the other hand, if a linear pathway has a feedforward activation loop, the concentration control coefficients of the enzymes lying upstream of the feedforward loop with respect to each of the metabolites located inside the loop are sign-indeterminate; the remaining control coefficients are sign-determined. Finally, if a linear pathway has a feedback inhibition loop and a feedforward activation loop, the concentration control coefficients that are sign-indeterminate can be identified by examining the feedback inhibition loop and the feedforward activation loop separately. The situation for a linear pathway with multiple feedback inhibition and feedforward activation loops can be analyzed in a similar fashion.

REFERENCES

[1] J. QUIRK AND R. RUPPERT, *Qualitative economics and the stability of equilibrium*, Rev. Econom. Stud., 32 (1965), pp. 311–326.

[2] C. JEFFRIES, *Qualitative stability and digraphs in model ecosystems*, Ecology, 55 (1974), pp. 1415–1419.

[3] A. K. SEN, *Metabolic control analysis: An application of signal flow graphs*, Biochem. J., 269 (1990), pp. 141–147.
[4] A. K. SEN, *Topological analysis of metabolic control*, Math. Biosci., 102 (1990), pp. 191–223.
[5] ———, *Quantitative analysis of metabolic regulation: A graph-theoretic approach using spanning trees*, Biochem. J., 275 (1991), pp. 253–258.
[6] ———, *A graph-theoretic analysis of metabolic regulation in linear pathways with multiple feedback loops and branched pathways*, Biochim. Biophys. Acta, 1059 (1991), pp. 293–311.
[7] H. KACSER AND J. A. BURNS, *The control of flux*, Symp. Soc. Exp. Biol., 27 (1973), pp. 65–104.
[8] R. HEINRICH AND T. RAPOPORT, *A linear steady-state treatment of enzymatic chains*, Eur. J. Biochem., 42 (1974), pp. 97–105.
[9] L. BASSETT, J. MAYBEE, AND J. QUIRK, *Qualitative economics and the scope of the correspondence principle*, Econometrica, 36 (1968), pp. 544–563.
[10] C. THOMASSEN, *Sign-nonsingular matrices and even cycles in directed graphs*, Linear Algebra Appl., 74 (1986), pp. 27–41.
[11] G. LADY AND J. MAYBEE, *Qualitatively invertible matrices*, J. Math. Soc. Sciences, 6 (1983), pp. 397–407.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 843–862, July 1995

010

THE CONVERGENCE OF GENERALIZED LANCZOS METHODS FOR LARGE UNSYMMETRIC EIGENPROBLEMS*

ZHONGXIAO JIA†

Abstract. In this paper, we investigate the convergence theory of generalized Lanczos methods for solving the eigenproblems of large unsymmetric matrices. Bounds for the distances between normalized eigenvectors and the Krylov subspace $\mathcal{K}_m(v_1, A)$ spanned by $v_1, Av_1, \ldots, A^{m-1}v_1$ are established, and a priori theoretical error bounds for eigenelements are presented when matrices are defective. Using them we show that the methods will still favor the outer part eigenvalues and the associated eigenvectors of A usually though they may converge quite slowly in the case of A being defective. Meanwhile, we analyze the relationships between the speed of convergence and the spectrum of A. However, a detailed analysis exposes that the approximate eigenvectors, Ritz vectors, obtained by generalized Lanczos methods for any unsymmetric matrix cannot be guaranteed to converge in theory even if approximate eigenvalues, Ritz values, do. Therefore, generalized Lanczos algorithms including Arnoldi's algorithm and IOMs with correction are provided with necessary theoretical background.

Key words. generalized Lanczos methods, Arnoldi's method, IOMs, orthogonal projection, orthonormal basis, defective, nonderogatory, Chebyshev polynomial, derivative

AMS subject classifications. 65F15, 15A18, 41A10

1. Introduction.

In practice one often wants to compute a few, say r, eigenvalues with largest (smallest) real parts and possibly the corresponding eigenvectors of large unsymmetric matrices, e.g., [21], [12]. For this kind of problem, one of the most useful techniques is Arnoldi's method [1] developed by Saad [17], where he defines one class of generalized Lanczos methods. They are orthogonal projection methods on a Krylov subspace, and reduce to Arnoldi's method when the basis of Krylov subspace is taken to be orthonormal [17], [19]. It is necessary to note that orthogonal projection methods have no restriction to choices of the basis of Krylov subspace, instead they only require that residuals of approximate eigenelements, called Ritz elements of the matrix A in the Krylov subspace, be orthogonal to this subspace other than requiring the basis of Krylov subspace to be orthonormal; see [17], [19]. Therefore, from the definition of generalized Lanczos methods, it can be seen that different choices of the basis of Krylov subspace will give rise to different generalized Lanczos algorithms [17], and Arnoldi's method is only a special representative most often used in practice when the basis of Krylov subspace is taken to be orthonormal. Another typical kind of method, incomplete orthogonalization methods (IOMs) *with correction* proposed by Saad [17], is also among generalized Lanczos methods since the approximate eigenelements obtained by them are just Ritz elements of the matrix A in the Krylov subspace and their residuals are orthogonal to this subspace. Obviously, the basis of Krylov subspace generated by IOMs with correction is not orthonormal usually.

The idea for generalized Lanczos methods is the following: Given an initial vector v_1, we realize an orthogonal projection process onto the Krylov subspace $\mathcal{K}_m(v_1, A)$ spanned by $v_1, Av_1, \ldots, A^{m-1}v_1$, where v_1 is the initial vector, $m \leq N$ [17] and N is the order of A. Let π_m be the orthogonal projector on $\mathcal{K}_m(v_1, A)$. We then

* Received by the editors April 2, 1993; accepted for publication (in revised form) by A. Greenbaum May 30, 1994.

† Present address. Fakultät für Mathematik, Universität Bielefeld, Postfach 100131, 33501 Bielefeld, Germany (jia@mathematik.uni-bielefeld.de). Permanent address. Department of Applied Mathematics, Dalian University of Technology, Dalian 116023, P.R. China. This work was supported by the Graduiertenkellog Mathematik at the University of Bielefeld.

compute the eigenelements $\lambda^{(m)}, \varphi^{(m)}$ of the restriction of the linear operator $\pi_m A$ to $\mathcal{K}_m(v_1, A)$, and then take them as approximations to some eigenelements λ, φ of A. Here $\lambda^{(m)}, \varphi^{(m)}$ are referred to as Ritz elements of A in $\mathcal{K}_m(v_1, A)$. In practice, Arnoldi's method is the simplest algorithm that achieves generalized Lanczos methods and generates an orthonormal basis of $\mathcal{K}_m(v_1, A)$. The matrix representation of the restriction of the linear operator $\pi_m A$ to $\mathcal{K}_m(v_1, A)$ is an upper Hessenberg matrix in this basis. In contrast, IOMs with correction generate a nonorthonormal basis and, in such a basis, the matrix representation of the restriction of $\pi_m A$ to $\mathcal{K}_m(v_1, A)$ is a banded upper Hessenberg matrix except for the last column being full. For more details, see [17].

Practical computations have shown that the methods usually favor the outer part eigenvalues and the associated eigenvectors of A as m increases, e.g., [6], [9], [14]–[17], [20], [21], [23]. However, the convergence theory of the methods is still incomplete. Saad [17] made a convergence analysis of the methods for the case that all eigenvalues of A are real simple, and he established bounds for the distances between normalized eigenvectors and $\mathcal{K}_m(v_1, A)$. His results show that these distances will usually converge to zero first for the eigenvectors associated with outer part eigenvalues. In a later paper [19], Saad extended the convergence theory to the case where complex eigenvalues are present, and he gave such bounds for the most right outer one eigenelement λ_1, φ_1 when the matrix A is diagonalizable. In his Master's thesis [7], the author further analyzed the convergence theory of generalized Lanczos methods, and established bounds for such distances for more than one eigenelement when A is diagonalizable and complex eigenvalues are present. It is shown that these distances will usually converge to zero first for the eigenvectors corresponding to eigenvalues with largest (smallest) real parts. In the literature, however, there have been no results concerning the implications of these distances on the behaviors of eigenvectors, which are very important in understanding how generalized Lanczos methods converge. Concerning eigenvalues, Saad [19] gave an error bound for them, where he assumes that the matrices derived by Arnoldi's method are diagonalizable. But they can be defective even if A is diagonalizable, which will be seen later. Thus, a further analysis is obviously necessary.

Since A is unsymmetric, it can be defective, which can indeed arise in applications; see, e.g., [3]. For this kind of matrices, how do the methods converge? This paper investigates this difficult problem. In the context, we establish bounds for the distances between normalized eigenvectors φ and $\mathcal{K}_m(v_1, A)$, and present a priori theoretical error bounds for eigenelements. The results show that generalized Lanczos methods still favor the outer part eigenvalues and usually the associated eigenvectors of A, though they may converge relatively slowly in the case of A being defective. However, a detailed analysis exposes that the approximate eigenvectors, Ritz vectors, obtained by generalized Lanczos methods cannot be guaranteed to converge in theory for any unsymmetric A even if approximate eigenvalues, Ritz values, do. This can happen when the approximate eigenproblems derived by generalized Lanczos methods are too ill conditioned. Therefore, our theory can provide necessary background for all generalized Lanczos methods, e.g., Arnoldi's method and IOMs with correction.

The paper is organized as follows: In §2 we review preliminaries; in §3 we analyze the convergence of generalized Lanczos methods in detail and prove our results; finally, in §4 we conclude the paper and point out the extension and applicability of our results as well as future work.

2. Preliminaries. Assume A to be an $N \times N$ real defective and nonderogatory matrix. Let A have M distinct eigenvalues labeled in the decreasing order of their real parts

$$\text{Re}(\lambda_1) \geq \text{Re}(\lambda_2) \geq \cdots \geq \text{Re}(\lambda_M), \tag{1}$$

where the multiplicities of λ_i are $d_i, i = 1, 2, \ldots, M$ (If A has the eigenvalues with the same real parts. but different imaginary parts, we first label those with the larger imaginary parts; for a complex pair of eigenvalues, we first label the one with positive real part). We will mainly be concerned with a few, say r, eigenvalues with largest (smallest) real parts and possibly the associated eigenvectors.

We denote by C^N the complex space of dimension N and by $\theta(u, \mathcal{K}_m(v_1, A))$ the acute angle between a nonzero vector u and the Krylov subspace $\mathcal{K}_m(v_1, A)$, defined by

$$\theta(u, \mathcal{K}_m(v_1, A)) = \arcsin \frac{\|(I - \pi_m)u\|}{\|u\|}, \tag{2}$$

where π_m is the orthogonal projector on $\mathcal{K}_m(v_1, A)$. In the whole context we denote by superscript H the conjugate transpose of a matrix or vector.

According to the assumptions, the Jordan form of A has nontrivial Jordan blocks. Since A is nonderogatory, there exists a nonsingular matrix S for which $A = SJS^{-1}$, where

$$J = \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_M \end{pmatrix}, \quad J_i = \begin{pmatrix} \lambda_i & 1 & & & \\ & \lambda_i & \ddots & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ & & & & \lambda_i \end{pmatrix}, \quad i = 1, 2, \ldots, M.$$

Let Q_m denote the set of all polynomials of degree not exceeding m. Then for any $p_m \in Q_m$, we have $p_m(A) = Sp_m(J)S^{-1}$, where

$$p_m(J) = \begin{pmatrix} p_m(J_1) & & & \\ & p_m(J_2) & & \\ & & \ddots & \\ & & & p_m(J_M) \end{pmatrix},$$

$$p_m(J_i) = \begin{pmatrix} p_m(\lambda_i) & p_m'(\lambda_i) & \frac{1}{2!}p_m''(\lambda_i) & \cdots & \frac{1}{(d_i-1)!}p_m^{(d_i-1)}(\lambda_i) \\ & p_m(\lambda_i) & p_m'(\lambda_i) & \cdots & \frac{1}{(d_i-2)!}p_m^{(d_i-2)}(\lambda_i) \\ & & \ddots & & \vdots \\ & & & p_m(\lambda_i) & p_m'(\lambda_i) \\ & & & & p_m(\lambda_i) \end{pmatrix},$$

$$i = 1, 2, \ldots, M.$$

3. Convergence analysis. Now let us study the convergence theory of generalized Lanczos methods. First, it is necessary to remind the reader that the theory to be

developed in this paper will, in fact, work for this whole class of methods, e.g., IOMs with correction, not only for Arnoldi's method; namely, we have no restriction to choices of the basis of $\mathcal{K}_m(v_1, A)$ except for Theorem 3.7 which can be easily modified for a nonorthonormal basis. Second, we point out that the same theory in this section applies as well to the left outer part eigenvalues and the associated eigenvectors of A with essentially the same results once A is replaced by $-A$. The section is divided into two parts. In §3.1 we establish bounds for the distances $\|(I-\pi_m)\varphi\|$, and using them in §3.2 we derive a priori theoretical error bounds for eigenelements and expose some important features of generalized Lanczos methods.

3.1. Inequalities on $\|(I-\pi_m)\varphi\|$. First, we need the following useful theorem, which is proved in [17], [19].

THEOREM 3.1. *Let $\gamma_m = \|\pi_m A(I-\pi_m)\|$, and λ, φ an eigenelement of A. Then*

$$\begin{aligned} \|(A_m - \lambda I)\pi_m\varphi\| &\leq \gamma_m\|(I-\pi_m)\varphi\|, \\ \|(A_m - \lambda I)\varphi\| &\leq \sqrt{\gamma_m^2 + |\lambda|^2}\|(I-\pi_m)\varphi\|, \end{aligned} \tag{3}$$

where $A_m = \pi_m A \pi_m$.

Note that $\gamma_m \leq \|A\|$, so the coefficients of the right-hand sides of (3) are of the same orders as $\|(I-\pi_m)\varphi\|$.

Assume the normalized vectors $v_{lj}, j = 1, 2, \ldots, d_l$ to be a chain of principal vectors of A associated with λ_l, where $v_{l1} = \varphi_l$'s are the corresponding normalized eigenvectors, $l = 1, 2, \ldots, M$. Let $S = (S_1, S_2, \ldots, S_M)$ and $S_l = (v_{l1}, v_{l2}, \ldots, v_{ld_l})$, $l = 1, 2, \ldots, M$. Then the initial vector v_1 can be expressed as

$$v_1 = \alpha_{i1}\varphi_i + \sum_{j=2}^{d_i} \alpha_{ij} v_{ij} + \sum_{l \neq i} \sum_{j=1}^{d_l} \alpha_{lj} v_{lj}. \tag{4}$$

Define $b_l = (\alpha_{l1}, \alpha_{l2}, \ldots, \alpha_{ld_l})^H$, $l = 1, 2, \ldots, M$, and $b_i' = (0, \alpha_{i2}, \ldots, \alpha_{id_i})^H$, and let $b^{(i)} = (b_1^H, \ldots, b_{i-1}^H, b_i'^H, b_{i+1}^H, \ldots, b_M^H)^H$.

Having the above notation, we can prove the following proposition.

PROPOSITION 3.2. *Let $S = (S_1, S_2, \ldots, S_M)$ and the initial vector v_1 as above, and define $\sigma(p) = (\sum_{j=0}^{d_i - 1} |\frac{1}{j!} p^{(j)}(\lambda_i)|^2)^{1/2}$. Then*

$$\|(I-\pi_m)\varphi_i\| \leq \xi_i \min_{p \in Q_{m-1}, \sigma(p)=1} \max_{j \neq i} \{\|p(J_j)\|, \|p(J_i')\|\} \tag{5}$$

$$= \xi_i \epsilon_i^{(m)}, \tag{6}$$

where J_i' is the one replacing the first row of J_i with zero elements,

$$\xi_i = \frac{\sum_{j=2}^{d_i} |\alpha_{ij}| + \sum_{l \neq i} \sum_{j=1}^{d_l} |\alpha_{lj}|}{|\sum_{j=0}^{d_i - 1} \frac{1}{j!} p_o^{(j)}(\lambda_i) \alpha_{ij+1}|} \max_{j \neq i} \{\|S_j\|, \|S_i'\|\} \tag{7}$$

with $S_i' = (0, v_{i2}, \ldots, v_{id_i})$, and $p_o \in Q_{m-1}$ is the optimal polynomial achieving the minimum in $\epsilon_i^{(m)}$.

Proof. It follows from the definition that

$$\|(I-\pi_m)\varphi_i\| = \sin\theta(\varphi_i, \mathcal{K}_m(v_1, A)) = \min_{u \in \mathcal{K}_m(v_1, A)} \|u - \varphi_i\|.$$

Due to $u \in \mathcal{K}_m(v_1, A)$, there exists a polynomial $q \in Q_{m-1}$ such that

$$u = q(A)v_1.$$

In terms of (4), we obtain

$$\begin{aligned} u &= q(A)\left(\alpha_{i1}\varphi_i + \sum_{j=2}^{d_i}\alpha_{ij}v_{ij} + \sum_{l\neq i}\sum_{j=1}^{d_l}\alpha_{lj}v_{lj}\right) \\ &= \left(\sum_{j=0}^{d_i-1}\frac{1}{j!}q^{(j)}(\lambda_i)\alpha_{ij+1}\right)\varphi_i + \sum_{k=2}^{d_i}\left(\sum_{j=0}^{d_i-k}\frac{1}{j!}q^{(j)}(\lambda_i)\alpha_{ij+k}\right)v_{ik} \\ &\quad + \sum_{l\neq i}\left[\sum_{k=1}^{d_l}\left(\sum_{j=0}^{d_l-k}\frac{1}{j!}q^{(j)}(\lambda_l)\alpha_{lj+k}\right)v_{lk}\right]. \end{aligned}$$

Now let $S^{(i)} = (S_1, \ldots, S_{i-1}, S_i', S_{i+1}, \ldots, S_M)$ and $J^{(i)}$ be the one replacing the ith Jordan block J_i of A by J_i'. Then noticing that the first row of $q(J_i')$ is zero for any $q \in Q_{m-1}$, we can get from the above relation

$$u = \left(\sum_{j=0}^{d_i-1}\frac{1}{j!}q^{(j)}(\lambda_i)\alpha_{ij+1}\right)\varphi_i + S^{(i)}q(J^{(i)})b^{(i)}.$$

Thus we have

$$\frac{u}{\sum_{j=0}^{d_i-1}\frac{1}{j!}q^{(j)}(\lambda_i)\alpha_{ij+1}} - \varphi_i = \frac{(\sum_{j=0}^{d_i-1} \mid \frac{1}{j!}q^{(j)}(\lambda_i)\mid^2)^{\frac{1}{2}}}{\sum_{j=0}^{d_i-1}\frac{1}{j!}q^{(j)}(\lambda_i)\alpha_{ij+1}} \frac{S^{(i)}q(J^{(i)})b^{(i)}}{(\sum_{j=0}^{d_i-1} \mid \frac{1}{j!}q^{(j)}(\lambda_i)\mid^2)^{\frac{1}{2}}}.$$

Hence taking $p(z) = q(z)/(\sum_{j=0}^{d_i-1} \mid \frac{1}{j!}q^{(j)}(\lambda_i)\mid^2)^{\frac{1}{2}}$ gives $\sigma(p) = 1$.

Because

$$\frac{u}{\sum_{j=0}^{d_i-1}\frac{1}{j!}q^{(j)}(\lambda_i)\alpha_{ij+1}} \in \mathcal{K}_m(v_1, A),$$

and $p \in Q_{m-1}$ such that $\sigma(p) = 1$ is arbitrary, we have

$$\|(I - \pi_m)\varphi_i\| = \min_{u\in\mathcal{K}_m(v_1,A)}\|u - \varphi_i\| \leq \min_{\sigma(p)=1}\frac{\|S^{(i)}p(J^{(i)})b^{(i)}\|}{\mid\sum_{j=0}^{d_i-1}\frac{1}{j!}p^{(j)}(\lambda_i)\alpha_{ij+1}\mid}.$$

For the numerator in the right-hand side of the above relation, we can derive

$$\begin{aligned} \|S^{(i)}p(J^{(i)})b^{(i)}\| &= \left\|\sum_{j\neq i}S_jp(J_j)b_j + S_i'p(J_i')b_i'\right\| \\ &\leq \sum_{j\neq i}\|S_j\|\|p(J_j)\|\|b_j\| + \|S_i'\|\|p(J_i')\|\|b_i'\| \\ &\leq \left(\|b_i'\| + \sum_{j\neq i}\|b_j\|\right)\max_{j\neq i}\{\|S_j\|, \|S_i'\|\}\cdot\max_{j\neq i}\{\|p(J_j)\|, \|p(J_i')\|\} \\ &\leq \left(\sum_{j=2}^{d_i}|\alpha_{ij}| + \sum_{l\neq i}\sum_{j=1}^{d_l}|\alpha_{lj}|\right)\max_{j\neq i}\{\|S_j\|, \|S_i'\|\}\cdot\max_{j\neq i}\{\|p(J_j)\|, \|p(J_i')\|\} \\ &= \xi_i'\max_{j\neq i}\{\|p(J_j)\|, \|p(J_i')\|\}. \end{aligned}$$

Combining the above relations, we can get

$$\begin{aligned}\|(I-\pi_m)\varphi_i\| &\le \frac{\xi_i'}{|\sum_{j=0}^{d_i-1}\frac{1}{j!}p_o^{(j)}(\lambda_i)\alpha_{ij+1}|}\min_{p\in Q_{m-1},\sigma(p)=1}\max_{j\neq i}\{\|p(J_j)\|,\|p(J_i')\|\}\\ &= \xi_i\min_{p\in Q_{m-1},\sigma(p)=1}\max_{j\neq i}\{\|p(J_j)\|,\|p(J_i')\|\}\\ &= \xi_i\epsilon_i^{(m)},\end{aligned}$$

which completes the proof. □

Remark 1. If A has only simple eigenvalues, then all $d_j = 1, S_j = \varphi_j$, $j = 1,2,\ldots,N$. We thus have from $\sigma(p_o) = |\, p_o(\lambda_i)\,| = 1$ and $\sigma(p) = |\, p(\lambda_i)\,| = 1$

$$\xi_i = \sum_{l\neq i}\frac{|\,\alpha_{l1}\,|}{|\,\alpha_{i1}\,|}, \qquad \epsilon_i^{(m)} = \min_{p\in Q_{m-1},p(\lambda_i)=1}\max_{j\neq i}|\,p(\lambda_j)\,|\,. \tag{8}$$

So this proposition completely reduces to Proposition 2.1 in [17].

Remark 2. In view of the definition, since $\sigma(p_o) = 1$, we have for the denominator in ξ_i

$$\begin{aligned}\left|\sum_{j=0}^{d_i-1}\frac{1}{j!}p_o^{(j)}(\lambda_i)\alpha_{ij+1}\right| &\le \sum_{j=0}^{d_i-1}\left|\frac{1}{j!}p_o^{(j)}(\lambda_i)\right||\alpha_{ij+1}|\\ &\le \left(\sum_{j=0}^{d_i-1}\left|\frac{1}{j!}p_o^{(j)}(\lambda_i)\right|^2\right)^{1/2}\left(\sum_{j=1}^{d_i}|\,\alpha_{ij}\,|^2\right)^{1/2}\\ &= \left(\sum_{j=1}^{d_i}|\,\alpha_{ij}\,|^2\right)^{1/2}\\ &\le \sum_{j=1}^{d_i}|\,\alpha_{ij}\,|\,.\end{aligned}$$

It follows from this and (7) that

$$\xi_i \ge \frac{\sum_{j=2}^{d_i}|\,\alpha_{ij}\,| + \sum_{l\neq i}\sum_{j=1}^{d_l}|\,\alpha_{lj}\,|}{\sum_{j=1}^{d_i}|\,\alpha_{ij}\,|}\max_{j\neq i}\{\|S_j\|,\|S_i'\|\}.$$

So, it can be seen that if the initial vector v_1 is nearly deficient in the directions of $\varphi_i, v_{i2},\ldots,v_{id_i}$, then ξ_i is very large. However, if at least one component of v_1 in these directions is not small and the principal vectors matrix S not very ill conditioned, we can then expect that ξ_i is of reasonable size since $\max_{j\neq i}\{\|S_j\|, \|S_i'\|\} \le \max_{j\neq i}\{\sqrt{d_j}, \sqrt{d_i-1}\}$ and $\sum_{l\neq i}\sum_{j=1}^{d_l}|\,\alpha_{lj}\,|$ is not large under such assumptions. Therefore, the problem of estimating the right-hand side of (6) reduces to that of estimating $\epsilon_i^{(m)}$. Note that it is impossible to give an upper bound for ξ_i.

Remark 3. For brevity, we used $\max_{j\neq i}\{\|S_j\|,\|S_i'\|\}$ instead of the correct but somewhat lengthy $\max\{\|S_i'\|, \max_{j\neq i}\|S_j\|\}$.

Now let us estimate $\epsilon_i^{(m)}$. From Proposition 3.2 and the previous preliminaries, we want to find a polynomial $p \in Q_{m-1}$ that satisfies $\sigma(p) = 1$, such that its kth

derivatives $p^{(k)}(\lambda_j), k < d_j, j \neq i$, and $p^{(k)}(\lambda_i), k < d_i - 1$, are as small as possible on the spectrum of A.

Before proceeding, we need some lemmas.

LEMMA 3.3. *Let $T_m(z)$ be the first kind Chebyshev polynomial of degree m in the complex plane* [13]. *Then the following hold.*

1. *If $z \in [-1, 1]$ and $0 \leq j \leq m$, then*

$$| T_m^{(j)}(z) | \leq | T_m^{(j)}(\pm 1) | = \frac{m^2(m^2-1)\cdots(m^2-(j-1)^2)}{1 \cdot 3 \cdot 5 \cdots (2j-1)} = m^{2j} C(m,j), \tag{9}$$

where for

$$j \geq 1,\ C(m,j) = \frac{(1-\frac{1}{m^2})(1-\frac{2^2}{m^2})\cdots(1-\frac{(j-1)^2}{m^2})}{1 \cdot 3 \cdot 5 \cdots (2j-1)}, \quad \text{and} \quad C(m,0) = 1.$$

2. *If $z \neq \pm 1$ and $0 \leq j \leq m$, and for $\mathrm{Re}(z) \geq 0$, $B_m(z,j)$ is defined by*

$$T_m^{(j)}(z) = m^j \frac{(z+\sqrt{z^2-1})^m}{(z^2-1)^j} B_m(z,j), \tag{10}$$

then $B_m(z,j)$ is uniformly bounded in m, and is of order $O(z^j)$ if $z \notin [-1,1]$. If $z > 1$ is fixed and $0 \leq j \leq m$, $T_m^{(j)}(z)$ is increasing in j.

3. *Assume $E(0,1,a)$ to be an ellipse with the center at the origin, the foci at ± 1 and the main semiaxis a. Then for $0 \leq j \leq m$*

$$\max_{z \in E(0,1,a)} | T_m^{(j)}(z) | = m^j \frac{(a+\sqrt{a^2-1})^m}{(a^2-1)^j} B_m(a,j). \tag{11}$$

Proof. Part 1. See Rivlin [13, p. 33].

Part 2. In terms of one of the definitions of $T_m(z)$, we have for $\mathrm{Re}(z) \geq 0$

$$\begin{aligned} T_m(z) &= \frac{1}{2}\left[(z+\sqrt{z^2-1})^m + (z-\sqrt{z^2-1})^m\right] \\ &= \frac{1}{2}(z+\sqrt{z^2-1})^m Q_m(z), \end{aligned} \tag{12}$$

where $Q_m(z) = 1 + (z+\sqrt{z^2-1})^{-2m}$ (If $\mathrm{Re}(z) < 0$, $T_m(z) = \frac{1}{2}(z-\sqrt{z^2-1})^m Q_m(z)$, where $Q_m(z) = 1 + (z-\sqrt{z^2-1})^{-2m}$).

Therefore, we can obtain

$$\lim_{m \to \infty} Q_m(z) = 1 \quad \text{for } z \notin [-1,1],$$

$$| Q_m(z) | \leq 2 \quad \text{for } z \in [-1,1].$$

It is clear that both $(z+\sqrt{z^2-1})^m$ and $Q_m(z)$ are analytic in the complex plane excluding the points ± 1 though $T_m(z)$ is analytic in the whole complex plane.

Now we can get from (12)

$$\begin{aligned} T_m'(z) &= \frac{1}{2} m \frac{(z+\sqrt{z^2-1})^m}{\sqrt{z^2-1}} Q_m(z) + \frac{1}{2}(z+\sqrt{z^2-1})^m Q_m'(z) \\ &= m \frac{(z+\sqrt{z^2-1})^m}{z^2-1}\left(\frac{1}{2}\sqrt{z^2-1} Q_m(z) + \frac{1}{2m}(z^2-1) Q_m'(z)\right) \\ &= m \frac{(z+\sqrt{z^2-1})^m}{(z^2-1)} B_m(z,1). \end{aligned}$$

Obviously, from the above analysis, for $z \notin [-1,1]$ we can see that $B_m(z,0) = \frac{1}{2}Q_m(z)$ contains the term $(z+\sqrt{z^2-1})^{-2m}$ and is of order $O(1)$. It follows that $B_m(z,1)$ also contains this factor. Therefore, $B_m(z,1)$ is uniformly bounded in m, and $B_m(z,1) = O(z)O(1) + O(z^2)O(z^{-1}) = O(z)$ for $z \notin [-1,1]$. So the assertion holds for $j=1$.

Now suppose for $k=j$ that the assertion is valid and $B_m(z,j)$ contains the term $(z+\sqrt{z^2-1})^{-2m}$. We then get by induction

$$\begin{aligned} T_m^{(j+1)}(z) &= m^j \left(\frac{m(z+\sqrt{z^2-1})^m(\sqrt{z^2-1})^{2j-1} - 2j(z+\sqrt{z^2-1})^m z(z^2-1)^{j-1}}{(z^2-1)^{2j}} \right) B_m(z,j) \\ &\quad + m^j \frac{(z+\sqrt{z^2-1})^m}{(z^2-1)^j} B_m'(z,j) \\ &= m^{j+1} \frac{(z+\sqrt{z^2-1})^m}{(z^2-1)^{j+1}} \left(\sqrt{z^2-1} B_m(z,j) - \frac{2j}{m} z B_m(z,j) + \frac{1}{m}(z^2-1) B_m'(z,j) \right) \\ &= m^{j+1} \frac{(z+\sqrt{z^2-1})^m}{(z^2-1)^{j+1}} B_m(z,j+1). \end{aligned}$$

Since $B_m(z,j)$ contains the term $(z+\sqrt{z^2-1})^{-2m}$, it is easily seen that $B_m(z,j+1)$ is uniformly bounded in m, and for $z \notin [-1,1]$,

$$B_m(z,j+1) = O(z)O(z^j) + O(z)O(z^j) + O(z^2)O(z^{j-1}) = O(z^{j+1}).$$

According to a result of [13, p. 51],

$$T_m^{(j)}(z) = \sum_{l=0}^{m-j} A_{lj} T_l(z), 0 \le j \le m, \tag{13}$$

where $A_{lj} \ge 0, 0 \le j \le m$, we can get

$$T_m^{(j+1)}(z) - T_m^{(j)}(z) = \sum_{l=0}^{m-j} A_{lj}(T_l'(z) - T_l(z)), \ 1 \le j+1 \le m.$$

It is then known that the left-hand side of the last relation is nonnegative since it is easy to show that $T_l'(z) - T_l(z) > 0$ for $z > 1$. So, the assertion is proved.

Part 3. Assume ∂E to be the boundary of $E(0,1,a)$. Then, by the maximum modulus principle and (13), we can get

$$\begin{aligned} \max_{z \in E(0,1,a)} | T_m^{(j)}(z) | &= \max_{z \in \partial E} | T_m^{(j)}(z) | \\ &= \max_{z \in \partial E} \left| \sum_{l=0}^{m-j} A_{lj} T_l(z) \right|. \end{aligned}$$

Since $A_{lj} \ge 0$ and $T_l(z)$, $0 \le l \le m-j$, achieves the maximum at the point a [13], it follows immediately that

$$\begin{aligned} \max_{z \in E(0,1,a)} | T_m^{(j)}(z) | &= T_m^{(j)}(a) \\ &= m^j \frac{(a+\sqrt{a^2-1})^m}{(a^2-1)^j} B_m(a,j). \end{aligned}$$

Thus, assertion (11) holds. □

Remark. By a continuity argument, if z lies in a neighborhood of one, then for fixed j we have, by comparing Parts 1 and 2 of Lemma 3.3,

$$\frac{B_m(z,j)}{(z^2-1)^j} = O(m^j) \text{ as } m \text{ increases.} \tag{14}$$

It implies that the left-hand side of the above relation cannot be uniformly bounded in m in the ellipse $E(0,1,a)$.

Having this lemma, we can establish the following result.

LEMMA 3.4. *Let*

$$p_m(z) = q_k(z)T_{m-k}(z),$$

where $q_k(z)$ is some fixed polynomial of degree k, and $T_{m-k}(z)$ is the first kind Chebyshev polynomial of degree $m-k$ in the complex plane. Then we have the following statement.

1. *If $z \notin [-1,1]$ and $0 \le j \le m-k$,*

$$p_m^{(j)}(z) = T_{m-k}^{(j)}(z)C_1(m,z,j), \tag{15}$$

where $C_1(m,z,j)$ is uniformly bounded in m for fixed z,j and of order $O(z^k)$.

2. *Assume $E(0,1,a)$ to be the ellipse described before. Then for $0 \le j \le m-k$*

$$\max_{z\in E(0,1,a)} | p_m^{(j)}(z) | = T_{m-k}^{(j)}(a)C_2(m,a,j), \tag{16}$$

where $C_2(m,a,j)$ is uniformly bounded in m for fixed j and of order $O(a^k)$.

Proof. Part 1. Since all roots of $T_{m-k}(z)$ are in $[-1,1]$, for $1 \le j \le m-k-1$ the roots of $T_{m-k}^j(z)$ lie in $[-1,1]$ by the Rolle rule. Also, noticing that $T_{m-k}^{(m-k)}(z)$ is a constant, thus for $z \notin [-1,1]$, $T_{m-k}^{(j)}(z) \neq 0$ if $0 \le j \le m-k$. Therefore, by the binomial expansion of derivatives, for $0 \le j \le m-k$, we can get

$$\begin{aligned} p_m^{(j)}(z) &= \sum_{n=0}^{j} \binom{j}{n} q_k^{(j-n)}(z) T_{m-k}^{(n)}(z) \\ &= \frac{\sum_{n=0}^{j} \binom{j}{n} q_k^{(j-n)}(z) T_{m-k}^{(n)}(z)}{T_{m-k}^{(j)}(z)} T_{m-k}^{(j)}(z) \\ &= C_1(m,z,j) T_{m-k}^{(j)}(z). \end{aligned}$$

From Part 2 of Lemma 3.3, for $0 \le n \le j$,

$$\begin{aligned} \frac{T_{m-k}^{(n)}(z)}{T_{m-k}^{(j)}(z)} &= \frac{(m-k)^n}{(m-k)^j}(z^2-1)^{j-n}\frac{B_{m-k}(z,n)}{B_{m-k}(z,j)} \\ &= \frac{1}{(m-k)^{j-n}} O(z^{2(j-n)}) O(z^{n-j}) \\ &= \frac{1}{(m-k)^{j-n}} O(z^{j-n}). \end{aligned}$$

Noticing that $q_k^{(j-n)}(z) = O(z^{k-(j-n)})$ for $z \notin [-1,1]$, therefore, by the definition $C_1(m,z,j)$ is uniformly bounded in m for fixed z,j and of order $O(z^k)$ for $z \notin [-1,1]$.

Part 2. According to the maximum modulus principle, we get

$$\begin{aligned}\max_{z\in E(0,1,a)} \mid p_m^{(j)}(z) \mid &= \max_{z\in E(0,1,a)} \left| \sum_{n=0}^{j} \binom{j}{n} q_k^{(j-n)}(z) T_{m-k}^{(n)}(z) \right| \\ &= \max_{z\in E(0,1,a)} \frac{\mid \sum_{n=0}^{j} \binom{j}{n} q_k^{(j-n)}(z) T_{m-k}^{(n)}(z) \mid}{\max_{z\in E(0,1,a)} \mid T_{m-k}^{(j)}(z) \mid} \max_{z\in E(0,1,a)} \mid T_{m-k}^{(j)}(z) \mid \\ &= C_2(m,a,j) \max_{z\in E(0,1,a)} \mid T_{m-k}^{(j)}(z) \mid .\end{aligned}$$

From Part 3 of Lemma 3.3, for $z \in E(0,1,a)$ and $0 \le n \le j$ we have

$$\begin{aligned}\frac{\mid T_{m-k}^{(n)}(z) \mid}{\max_{z\in E(0,1,a)} \mid T_{m-k}^{(j)}(z) \mid} &\le \frac{\max_{z\in E(0,1,a)} \mid T_{m-k}^{(n)}(z) \mid}{\max_{z\in E(0,1,a)} \mid T_{m-k}^{(j)}(z) \mid} \\ &= \frac{1}{(m-k)^{j-n}} (a^2-1)^{j-n} \frac{B_{m-k}(a,n)}{B_{m-k}(a,j)} \\ &= \frac{1}{(m-k)^{j-n}} O(a^{2(j-n)}) O(a^{n-j}) \\ &= \frac{1}{(m-k)^{j-n}} O(a^{j-n}).\end{aligned}$$

Thus, $C_2(m,a,j)$ is uniformly bounded in m.

It follows from this and

$$\max_{z\in E(0,1,a)} \mid q_k^{(j-n)}(z) \mid = O(a^{k-(j-n)})$$

that $C_2(m,a,j)$ is uniformly bounded in m for fixed j and of order $O(a^k)$. □

Now let us estimate $\epsilon_i^{(m)}$.

Assume $\bar{\lambda}_{i-1} = \lambda_i$ if λ_i is complex, and $\mathrm{Re}(\lambda_i) > \mathrm{Re}(\lambda_{i+1}) \ge \cdots \ge \mathrm{Re}(\lambda_M)$. From the above assumptions, there is an ellipse $E_i(c,e,a)$ with real center c, foci $c+e, c-e$, major semiaxis a to contain the set $\{\lambda_{i+1}, \ldots, \lambda_M\}$ but $\{\lambda_1, \ldots, \lambda_i\}$. Furthermore, due to the conjugation of the eigenvalues as A is real, $E_i(c,e,a)$ can be symmetric with respect to the real axis, that is, a, e must be either real or purely imaginary. Let us call these ellipses first kind and second kind ellipses, respectively, and write $E_i(c,e,a)$ simply as E_i.

We will consider two cases, respectively.

Case 1. λ_i is simple. In this case $d_i = 1$ and $\sigma(p) = \mid p(\lambda_i) \mid = 1$.

THEOREM 3.5. *Assume λ_i to be simple and $E_i = E_i(c,e,a)$ as described above. Let*

$$\tilde{d}_i = \max_{j>i} d_j, \quad i' = \sum_{j=1}^{i-1} d_j, \quad \alpha = a/e, b = \alpha^2 - 1, \tau = \alpha + \sqrt{\alpha^2 - 1}, \gamma_i = (\lambda_i - c)/e.$$

Then

$$\epsilon_i^{(m)} \le \tilde{d}_i c_1 m_i^{\tilde{d}_i - 1} b^{-\tilde{d}_i + 1} B_{m_i}(\alpha, \tilde{d}_i - 1) \tau^{m_i} / \mid T_{m_i}(\gamma_i) \mid, \tag{17}$$

where c_1 is a function uniformly bounded in m and of order $O(\alpha^{i'})$, and $T_{m_i}(z)$ is the first kind Chebyshev polynomial of degree $m_i = m - i' - 1$.

Proof. Write

$$J_j(z) = \begin{pmatrix} z & 1 & & & \\ & z & 1 & & \\ & & \ddots & \ddots & \\ & & & z & 1 \\ & & & & z \end{pmatrix}, \quad \dim(J_j(z)) = d_j, j = 1, 2, \ldots, M.$$

Let $\tilde{Q}_{m-1}$ be the set of polynomials of the form $p(z) = q_{i'}(z)r(z)$, where

$$q_{i'}(z) = \begin{cases} \prod_{j=1}^{i-1} \frac{(z-\lambda_j)^{d_j}}{(\lambda_i-\lambda_j)^{d_j}} & \text{if } i \neq 1, \\ 1 & \text{if } i = 1, \end{cases}$$

and $r(z)$ is a polynomial of degree m_i and satisfies the condition $r(\lambda_i) = 1$. It is clear that $p \in Q_{m-1}$, $\sigma(p) = \mid p(\lambda_i) \mid = 1$ and $p(J_j) = 0$ for $j < i$. It then follows from Proposition 3.2 that

$$\epsilon_i^{(m)} \leq \min_{p \in \tilde{Q}_{m-1}} \max_{j>i} \|p(J_j)\| \leq \min_{p \in \tilde{Q}_{m-1}} \max_{z \in E_i, j>i} \|p(J_j(z))\|.$$

For the above minimax problem, we do not know how to find the optimal polynomial explicitly. Thus we seek an approximate optimal one. For the following minimax problem:

$$\min_{r \in Q_{m_i}} \max_{z \in E_i} \mid r(z) \mid,$$

it is proved in [4] that the scaled and transformed Chebyshev polynomial

$$r(z) = T_{m_i}\left(\frac{z-c}{e}\right) / T_{m_i}(\gamma_i) \tag{18}$$

is often optimal though it is not always the case; if it is not optimal, it is still a very good approximate one. So we choose

$$\bar{p}(z) = q_{i'}(z) T_{m_i}\left(\frac{z-c}{e}\right) / T_{m_i}(\gamma_i).$$

Therefore, we have by [5, p. 541]

$$\begin{aligned} \epsilon_i^{(m)} &\leq \max_{z \in E_i, j>i} \|\bar{p}(J_j(z))\| \leq \max_{z \in E_i, j>i} \| \mid \bar{p}(J_j(z)) \mid \| \leq \max_{j>i} \| \max_{z \in E_i} \mid \bar{p}(J_j(z)) \mid \| \\ &\leq \max_{j>i} d_j \max_{0 \leq r \leq d_j - 1, z \in E_i} \left| \frac{\bar{p}^{(r)}(z)}{r!} \right| \leq \tilde{d}_i \max_{0 \leq r \leq \tilde{d}_i - 1, z \in E_i} \left| \frac{\bar{p}^{(r)}(z)}{r!} \right|. \end{aligned}$$

Since the transform $z' = (z-c)/e$ maps E_i into the ellipse $E(0, 1, \alpha)$ and by Part 2 of Lemma 3.3 $T_{m_i}^{(r)}(\alpha)$ is increasing in r for $0 \leq r \leq m_i$, we can obtain from Part 2 of Lemma 3.4 by manipulation

$$\epsilon_i^{(m)} \leq \tilde{d}_i c_1 m_i^{\tilde{d}_i - 1} b^{-\tilde{d}_i + 1} B_{m_i}(\alpha, \tilde{d}_i - 1) \tau^{m_i} / \mid T_{m_i}(\gamma_i) \mid,$$

where $c_1 = \max_{0 \leq r \leq \tilde{d}_i - 1} \mid e \mid^{-r} C_2(m, \alpha, r)$ is uniformly bounded in m and of order $O(\alpha^{i'})$ because of Part 2 of Lemma 3.4, which is just (17). □

Remark 1. Without loss of generality, assume that the real part of γ_i is nonnegative, and let $\tau_i = \mid \gamma_i + \sqrt{\gamma_i^2 - 1} \mid$. Then, from (17) and $\mathrm{Re}(\lambda_i) > \mathrm{Re}(\lambda_{i+1}) \geq \cdots \geq \mathrm{Re}(\lambda_M)$, we can see that $\epsilon_i^{(m)}$ converges to zero as m increases because

$$\tau^{m_i} / \mid T_{m_i}(\gamma_i) \mid \leq 2(\tau/\tau_i)^{m_i}$$

converges to zero as m tends to infinity. How rapidly the methods converge depends strongly on $m_i^{\tilde{d}_i - 1} b^{-\tilde{d}_i + 1} B_{m_i}(\alpha, \tilde{d}_i - 1)$ and $\tau^{m_i} / \mid T_{m_i}(\gamma_i) \mid$, which asymptotically depends on the following factors:

$$\kappa_{1i} = \frac{\mid (\lambda_i - c) + \sqrt{(\lambda_i - c)^2 - e^2} \mid}{a + \sqrt{a^2 - e^2}} \quad \text{if } a \text{ and } e \text{ real,} \tag{19}$$

$$\kappa_{2i} = \frac{\mid (\lambda_i - c) + \sqrt{(\lambda_i - c)^2 + \mid e \mid^2} \mid}{\mid a \mid + \sqrt{\mid a \mid^2 - \mid e \mid^2}} \quad \text{if } a \text{ and } e \text{ purely imaginary.} \tag{20}$$

We refer to κ_{1i}, κ_{2i} as the crucial factors of convergence associated with first and second kind ellipses, respectively. Note that in view of Lemma 3.3 and (14)

$$m_i^{\tilde{d}_i - 1} b^{-\tilde{d}_i + 1} B_{m_i}(\alpha, \tilde{d}_i - 1)$$

is of order

$$m_i^{\tilde{d}_i - 1} O(\alpha^{-\tilde{d}_i + 1})$$

if α is not close to one and of order

$$O(m_i^{2(\tilde{d}_i - 1)})$$

if α is near one.

Remark 2. For the first kind ellipse, the convergence is likely to be better if the eigenvalues are close to the real line since κ_{1i} will be larger in this case; if the ellipse E_i has nearly a circular shape, the convergence is likely to be slower since κ_{1i} is small. For the second kind ellipse, the convergence is likely to be slower if the eigenvalues are almost purely imaginary since κ_{2i} will be smaller at this moment.

Remark 3. $\|(I - \pi_m)\varphi_i\|$ will usually converge to zero first for the eigenvectors associated with eigenvalues with largest real parts. However, the presence of nonlinear elementary divisors may decrease the speed of convergence considerably.

Case 2. λ_i is multiple.

THEOREM 3.6. *Let* $\tilde{d}_i, i', \alpha, \gamma_i, \tau$ *be in Theorem* 3.5 *and* $\tau_i = \mid \gamma_i + \sqrt{\gamma_i^2 - 1} \mid$. *Then*

$$\epsilon_i^{(m)} \leq \max\left\{ \tilde{d}_i c_2 m_i^{\tilde{d}_i - d_i} \left(\frac{\tau}{\tau_i}\right)^{m_i}, (d_i - 1)\frac{1}{m_i} O(1) \right\}, \tag{21}$$

where c_2 *is of order* $O(m_i^{\tilde{d}_i - 1})$ *if* α *is very near one and of order* $O(1)$ *if* α *is well isolated from one, and* $m_i = m - i' - 1$.

Proof. Let $J_j(z)$ be in the proof of Theorem 3.5, and $\tilde{Q}_{m-1}$ the set of polynomials of the form

$$p(z) = \pi(z) r(z) / \left(\sum_{k=0}^{d_i - 1} \left| \frac{1}{k!} (\pi(\lambda_i) r(\lambda_i))^{(k)} \right|^2 \right)^{1/2},$$

where

$$\pi(z) = \begin{cases} \prod_{j=1}^{i-1}(z-\lambda_j)^{d_j} & \text{if } i \neq 1, \\ 1 & \text{if } i = 1, \end{cases}$$

and $r(z)$ is a polynomial of m_i. Obviously $\sigma(p) = 1$ and $p(J_j) = 0$ for $j < i$. Therefore, from Proposition 3.2 and [5, p. 541], we can get

$$\begin{aligned} \epsilon_i^{(m)} &\leq \min_{\tilde{Q}_{m-1}} \max_{j>i} \{\|p(J_j)\|, \|p(J_i')\|\} \\ &\leq \min_{p\in\tilde{Q}_{m-1}} \max_{z\in E_i, j>i} \{\|p(J_j(z))\|, \|p(J_i')\|\} \\ &\leq \min_{p\in\tilde{Q}_{m-1}} \max_{j>i} \left\{ \| \max_{z\in E_i} | p(J_j(z)) | \|, \|p(J_i')\| \right\} \\ &\leq \max \left\{ \tilde{d}_i \max_{0\leq k\leq \tilde{d}_i-1, z\in E_i} \left| \frac{p^{(k)}(z)}{k!} \right|, (d_i-1) \max_{0\leq k\leq d_i-2} \left| \frac{p^{(k)}(\lambda_i)}{k!} \right| \right\}. \end{aligned}$$

Since it is impossible to find the polynomial attaining the minimum in the above relation, we can only find a good approximate one.

Take

$$\bar{r}(z) = T_{m_i}\left(\frac{z-c}{e}\right), \quad \bar{p}(z) = \pi(z)\bar{r}(z) / \left(\sum_{k=0}^{d_i-1} \left| \frac{1}{k!} (\pi(\lambda_i)\bar{r}(\lambda_i))^{(k)} \right|^2 \right)^{1/2}.$$

It is then clear that $\sigma(\bar{p}) = 1$.

Since $(z-c)/e \in E(0,1,\alpha)$ and $T_{m_i}^{(k)}(\alpha)$ is increasing in k for $0 \leq k \leq m_i$, by Lemma 3.4 we can get (21) by complicated calculations, where

$$c_2 = (d_i-1)! \frac{(\alpha^2-1)^{-\tilde{d}_i+1} B_{m_i}(\alpha, \tilde{d}_i - 1) \max_{0\leq r\leq \tilde{d}_i-1} | e |^{-r} C_2(m,\alpha,r)}{| (\gamma_i^2-1)^{-d_i+1} B_{m_i}(\gamma_i, d_i-1) e^{-d_i+1} C_1(m,\gamma_i,d_i-1) |}.$$

It follows from (14) and Part 2 of Lemma 3.3 as well as Lemma 3.4 that c_2 is of order

$$O(m_i^{\tilde{d}_i-1})$$

if α is very near one and of order of $O(1)$ if α is well isolated from one. □

Remark 1. It is easy to see from the remarks following Theorem 3.5 that, if $d_i = 1$, assertion (21) essentially reduces to (17).

Remark 2. We see from (21) that if m is large, then $\epsilon_i^{(m)}$ may be of order $(d_i-1)\frac{1}{m_i}O(1)$. Thus, the right-hand side of (21) will eventually converge to zero geometrically.

As pointed out previously, if A is an $N \times N$ diagonalizable matrix, then Proposition 3.2 reduces to Proposition 2.1 [17]. Saad [19] gives such an example:

Assume $m = N-1$, the eigenvalues $\lambda_k = e^{i2(k-1)\pi/N}$. Then $\epsilon_1^{(m)} = \frac{1}{m}$.

It can be seen from this example that $\epsilon_1^{(m)}$ converges to zero geometrically. Thus, our estimates can be realistic.

3.2. A priori theoretical error bounds for eigenelements. Previously, we have established inequalities on $\|(I-\pi_m)\varphi_i\|$, which mean that there exists a vector, i.e., $\pi_m\varphi_i$, in $\mathcal{K}_m(v_1,A)$ to approximate φ_i as m increases. We now study the implications of these inequalities on the behaviors of eigenelements, which are very important in understanding how generalized Lanczos methods converge.

Define the matrix $V_m=(v_1,v_2,\ldots,v_m)$, and assume that its columns constitute a basis of $\mathcal{K}_m(v_1,A)$. Concerning error bounds for eigenvalues, for simplicity, assume V_m to be orthonormal. Let us set $V_m^H\varphi_i=\tilde{y}_i^{(m)}$ and define $H_m=V_m^HAV_m$. Then noticing that $\|\tilde{y}_i^{(m)}\|=\|V_m\tilde{y}_i^{(m)}\|=\|\pi_m\varphi_i\|$, the first inequality of (3) translates itself into

$$\frac{\|(H_m-\lambda_iI)\tilde{y}_i^{(m)}\|}{\|\tilde{y}_i^{(m)}\|}\le\gamma_m\frac{\|(I-\pi_m)\varphi_i\|}{\|\pi_m\varphi_i\|}. \tag{22}$$

Remark. Note that

$$\frac{\|(I-\pi_m)\varphi_i\|}{\|\pi_m\varphi_i\|}=\tan\theta(\varphi_i,\mathcal{K}_m(v_1,A)),$$

and H_m is just the matrix representation of A_m in the basis $\{v_i\}_1^m$ of $\mathcal{K}_m(v_1,A)$ and equal to the upper Hessenberg matrix generated by Arnoldi's method starting with v_1.

With (22), we can establish the following result.

THEOREM 3.7. *Let $S_m^{-1}H_mS_m=J^{(m)}$ be the Jordan form of H_m and cond$(S_m)=\|S_m\|\|S_m^{-1}\|$. Assume $\|(I-\pi_m)\varphi_i\|$ to be small enough. Then there exists an eigenvalue $\lambda_i^{(m)}$ of H_m with the* index *l_i such that we have*

$$|\lambda_i^{(m)}-\lambda_i|\le 2(\gamma_m\mathrm{cond}(S_m))^{1/l_i}\left(\frac{\|(I-\pi_m)\varphi_i\|}{\|\pi_m\varphi_i\|}\right)^{1/l_i}. \tag{23}$$

Proof. We need only consider the case that $\lambda_i^{(m)}$ is not an eigenvalue of A. Let $(H_m-\lambda_iI)\tilde{y}_i^{(m)}=r_i^{(m)}$. Then obviously $\|r_i^{(m)}\|\le\gamma_m\|(I-\pi_m)\varphi_i\|$. By the definition of $r_i^{(m)}$, we have

$$\begin{aligned}\|\pi_m\varphi_i\|=\|\tilde{y}_i^{(m)}\|&=\|(H_m-\lambda_iI)^{-1}r_i^{(m)}\|\\&=\|S_m(J^{(m)}-\lambda_iI)^{-1}S_m^{-1}r_i^{(m)}\|\\&\le\|S_m\|\|S_m^{-1}\|\|(J^{(m)}-\lambda_iI)^{-1}\|\|r_i^{(m)}\|\\&\le\mathrm{cond}(S_m)\gamma_m\|(J^{(m)}-\lambda_iI)^{-1}\|\|(I-\pi_m)\varphi_i\|.\end{aligned}$$

From [2], the above relation means that there exists a $\lambda_i^{(m)}$ with the index l_i such that

$$\frac{|\lambda_i^{(m)}-\lambda_i|^{l_i}}{1+|\lambda_i^{(m)}-\lambda_i|^{l_i-1}}\le\mathrm{cond}(S_m)\gamma_m\frac{\|(I-\pi_m)\varphi_i\|}{\|\pi_m\varphi_i\|},$$

which shows that (23) holds. □

From the analysis of §3.1, (23) shows that $\lambda_i^{(m)}\to\lambda_i$ as m increases. However, note that if $l_i\ne1$, then the approximate eigenproblem

$$H_my_i^{(m)}=\lambda_i^{(m)}y_i^{(m)} \tag{24}$$

is ill conditioned. In this case, though $\lambda_i^{(m)} \to \lambda_i$ as m increases in theory if $\text{cond}(S_m)$ is uniformly bounded in m, in numerical computations, it is very difficult for us to determine $\lambda_i^{(m)}$.

Theorem 3.7 can be simplified when some $l_i = 1$, as described below.

THEOREM 3.8. *Assume that some $l_i = 1$ and the associated $\|(I - \pi_m)\varphi_i\|$ are small enough. Let $P_i^{(m)}$ be the spectral projectors associated with $\lambda_i^{(m)}$. Then*

$$|\lambda_i^{(m)} - \lambda_i| \le \|P_i^{(m)}\|\gamma_m \frac{\|(I-\pi_m)\varphi_i\|}{\|\pi_m\varphi_i\|} + O\left(\left(\frac{\|(I-\pi_m)\varphi_i\|}{\|\pi_m\varphi_i\|}\right)^2\right). \tag{25}$$

Proof. In terms of the first inequality of (3), it can be easily shown from [24, p. 69] and [5, p. 344] that (25) holds. □

Remark. Equation (25) indicates that $\lambda_i^{(m)} \to \lambda_i$ as m increases if only $\|P_i^{(m)}\|$ is uniformly bounded in m.

We now give error bounds for eigenvectors. First, we need a lemma.

LEMMA 3.9. *Let $x_1, x_2, \ldots, x_k$ be k vectors and $\alpha_1, \alpha_2, \ldots, \alpha_k$ k scalars, and define the matrix $X = (x_1, x_2, \ldots, x_k)$. Then*

$$\begin{aligned} &\|\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k\| \\ &\ge \frac{1}{\inf_D \text{diag.}\, \text{cond}(XD)} \min_{1\le j\le k} |\alpha_j| \, \|x_1 + x_2 + \cdots + x_k\|, \end{aligned} \tag{26}$$

where D's are $k \times k$ nonsingular diagonal matrices.

Proof. Let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)^H$ and $e = (1, 1, \ldots, 1)^H$. Then for any $k \times k$ nonsingular diagonal matrix D with the diagonal entries $\delta_i, i = 1, 2, \ldots, k$, we have

$$\begin{aligned} \|\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k\|^2 &= \alpha^H X^H X \alpha \\ &= (D^{-1}\alpha)^H (XD)^H (XD)(D^{-1}\alpha) \\ &\ge \sigma_{\min}^2(XD)\|D^{-1}\alpha\|^2 \\ &= \sigma_{\min}^2(XD) \sum_{i=1}^k \left|\frac{\alpha_i}{\delta_i}\right|^2 \\ &\ge \sigma_{\min}^2(XD) \min_{1\le j\le k} |\alpha_j|^2 \sum_{i=1}^k \left|\frac{1}{\delta_i}\right|^2, \end{aligned}$$

where $\sigma_{\min}(XD)$ is the smallest singular value of XD. On the other hand, similarly, we have

$$\begin{aligned} \|x_1 + x_2 + \cdots + x_k\|^2 = e^H X^H X e &= (D^{-1}e)^H (XD)^H (XD)(D^{-1}e) \\ &\le \sigma_{\max}^2(XD)\|D^{-1}e\|^2 \\ &= \sigma_{\max}^2(XD) \sum_{i=1}^k \left|\frac{1}{\delta_i}\right|^2, \end{aligned}$$

where $\sigma_{\max}(XD)$ is the largest singular value of XD.

Combining the above results, we can get

$$\alpha^H X^H X \alpha \ge \frac{\sigma_{\min}^2(XD) \min_{1\le j\le k} |\alpha_j|^2 \sum_{i=1}^k |\frac{1}{\delta_i}|^2}{\sigma_{\max}^2(XD) \sum_{i=1}^k |\frac{1}{\delta_i}|^2} \sigma_{\max}^2(XD) \sum_{i=1}^k \left|\frac{1}{\delta_i}\right|^2$$

$$\geq \frac{\sigma_{\min}^2(XD)\min_{1\leq j\leq k} \mid \alpha_j \mid^2}{\sigma_{\max}^2(XD)} e^H X^H X e$$
$$= \frac{\min_{1\leq j\leq k} \mid \alpha_j \mid^2}{\mathrm{cond}^2(XD)} \|x_1 + x_2 + \cdots + x_k\|^2.$$

Since D is any nonsingular diagonal matrix, it follows that (26) holds. □

Remark. If the columns of X are orthogonal (not orthonormal), $\inf_D \text{ diag.} \mathrm{cond}(XD) = 1$.

THEOREM 3.10. *Assume that A_m is diagonalizable and has s distinct eigenvalues $\lambda_j^{(m)}$ in $\mathcal{K}_m(v_1, A)$. Let $P_j^{(m)}$ denote the spectral projectors associated with $\lambda_j^{(m)}$, and $d_{i,m} = \min_{j\neq i} \mid \lambda_i - \lambda_j^{(m)} \mid$ and $\gamma_m = \|\pi_m A(I - \pi_m)\|$, and define the matrix*

$$\Phi_i^{(m)} = (P_1^{(m)}\varphi_i, \ldots, P_{i-1}^{(m)}\varphi_i, P_{i+1}^{(m)}\varphi_i, \ldots, P_s^{(m)}\varphi_i).$$

Then

(27)
$$\|(I - P_i^{(m)})\varphi_i\| \leq \left(1 + \frac{\inf_D \text{ diag.} \mathrm{cond}(\Phi_i^{(m)}D)(1 + \|P_i^{(m)}\|)\gamma_m}{d_{i,m}}\right) \|(I - \pi_m)\varphi_i\|.$$

Let $P_i^{(m)}\varphi_i / \|P_i^{(m)}\varphi_i\| = \varphi_i^{(m)}$. Then

(28)
$$\sin\theta(\varphi_i, \varphi_i^{(m)}) \leq \left(1 + \frac{\inf_D \text{ diag.} \mathrm{cond}(\Phi_i^{(m)}D)(1 + \|P_i^{(m)}\|)\gamma_m}{d_{i,m}}\right) \sin\theta(\varphi_i, \mathcal{K}_m(v_1, A)).$$

Proof. We first prove the inequality

$$\|(\pi_m - P_i^{(m)})\varphi_i\| \leq \frac{\inf_D \text{ diag.} \mathrm{cond}(\Phi_i^{(m)}D)(1 + \|P_i^{(m)}\|)\gamma_m}{d_{i,m}} \|(I - \pi_m)\varphi_i\|.$$

Let $\lambda_1^{(m)}, \lambda_2^{(m)}, \ldots, \lambda_s^{(m)}$ be the distinct eigenvalues of the linear operator $\pi_m A \pi_m$ in $\mathcal{K}_m(v_1, A)$, and $P_j^{(m)}, j = 1, 2, \ldots, s$ the associated spectral projectors. Then it is well known that

$$P_j^{(m)} P_i^{(m)} = \delta_{ij} P_j^{(m)},$$
(29)
$$\sum_{j=1}^{s} P_j^{(m)} = \pi_m,$$

where δ_{ij} is the Kronecker delta. Hence

$$(\pi_m A - \lambda_i I)\pi_m \varphi_i = (\pi_m A - \lambda_i I) \sum_{j=1}^{s} P_j^{(m)} \varphi_i$$
$$= \sum_{j=1}^{s} (\lambda_j^{(m)} - \lambda_i) P_j^{(m)} \varphi_i.$$

Premultiplying the two hand sides of the above relation by $I - P_i^{(m)}$, we get

$$(I - P_i^{(m)})(\pi_m A - \lambda_i I)\pi_m \varphi_i = \sum_{j \neq i} (\lambda_j^{(m)} - \lambda_i) P_j^{(m)} \varphi_i.$$

From Lemma 3.9, we have

$$\|(I - P_i^{(m)})(\pi_m A - \lambda_i I)\pi_m \varphi_i\| \geq \frac{d_{i,m}}{\inf_D \text{ diag. } \text{cond}(\Phi_i^{(m)} D)} \left\| \sum_{j \neq i} P_j^{(m)} \varphi_i \right\|.$$

On the other hand, in terms of Theorem 3.1

$$\begin{aligned} \|(I - P_i^{(m)})(\pi_m A - \lambda_i I)\pi_m \varphi_i\| &\leq \|I - P_i^{(m)}\| \|(\pi_m A - \lambda_i I)\pi_m \varphi_i\| \\ &\leq (1 + \|P_i^{(m)}\|) \|\pi_m (A - \lambda_i I)\pi_m \varphi_i\| \\ &= (1 + \|P_i^{(m)}\|) \|(A_m - \lambda_i I)\pi_m \varphi_i\| \\ &\leq (1 + \|P_i^{(m)}\|)\gamma_m \|(I - \pi_m)\varphi_i\|. \end{aligned}$$

From (29), we obtain

$$\|(\pi_m - P_i^{(m)})\varphi_i\| = \left\| \sum_{j \neq i} P_j^{(m)} \varphi_i \right\|.$$

Therefore,

$$\begin{aligned} \|(\pi_m - P_i^{(m)})\varphi_i\| &\leq \frac{\inf_D \text{ diag. } \text{cond}(\Phi_i^{(m)} D)}{d_{i,m}} \|(I - P_i^{(m)})(\pi_m A - \lambda_i I)\pi_m \varphi_i\| \\ &\leq \frac{\inf_D \text{ diag. } \text{cond}(\Phi_i^{(m)} D)(1 + \|P_i^{(m)}\|)\gamma_m}{d_{i,m}} \|(I - \pi_m)\varphi_i\|. \end{aligned}$$

To get result (27), let us decompose

$$(I - P_i^{(m)})\varphi_i = (I - \pi_m)\varphi_i + (\pi_m - P_i^{(m)})\varphi_i.$$

Hence

$$\begin{aligned} \|(I - P_i^{(m)})\varphi_i\| &\leq \|(I - \pi_m)\varphi_i\| + \frac{\inf_D \text{ diag. } \text{cond}(\Phi_i^{(m)} D)(1 + \|P_i^{(m)}\|)\gamma_m}{d_{i,m}} \|(I - \pi_m)\varphi_i\| \\ &= \left(1 + \frac{\inf_D \text{ diag. } \text{cond}(\Phi_i^{(m)} D)(1 + \|P_i^{(m)}\|)\gamma_m}{d_{i,m}} \right) \|(I - \pi_m)\varphi_i\|, \end{aligned}$$

which is just (27).

Let $P_i^{(m)} \varphi_i / \|P_i^{(m)} \varphi_i\| = \varphi_i^{(m)}$. Then from the above relation, we have

$$\begin{aligned} \sin \theta(\varphi_i, \varphi_i^{(m)}) &= \min_\alpha \|\varphi_i - \alpha \varphi_i^{(m)}\| \leq \|\varphi_i - P_i^{(m)} \varphi_i\| = \|(I - P_i^{(m)})\varphi_i\| \\ &\leq \left(1 + \frac{\inf_D \text{ diag. } \text{cond}(\Phi_i^{(m)} D)(1 + \|P_i^{(m)}\|)\gamma_m}{d_{i,m}} \right) \sin \theta(\varphi_i, \mathcal{K}_m(v_1, A)). \end{aligned}$$

Thus, assertion (28) is proved. □

According to the results of §3.1 and (27), (28), it is clear that $\varphi_i^{(m)} \to \varphi_i$ as m increases once both $\inf_{D \text{ diag.}} \operatorname{cond}(\Phi_i^{(m)} D)$ and $\|P_i^{(m)}\|$ are uniformly bounded in m. Moreover, if $\lambda_i^{(m)}$ is ill conditioned, then $\varphi_i^{(m)}$ may approximate φ_i quite slowly. However, we cannot guarantee in theory that $\inf_{D \text{ diag.}} \operatorname{cond}(\Phi_i^{(m)} D)$ is uniformly bounded in m, even assuming that $\|P_i^{(m)}\|$ is uniformly bounded in m. Therefore, the right-hand sides of (27) and (28) may not converge to zero unless $\|(I - \pi_m)\varphi_i\| = 0$, which implies that $\varphi_i^{(m)}$ *may not* converge to φ_i even if $\lambda_i^{(m)} \to \lambda_i$ as m increases.

By observing the results of §§3.1 and 3.2, we get the following conclusions. First, in the case of A being defective, generalized Lanczos methods will still favor the outer part eigenvalues and the associated eigenvectors of A usually. Second, they may converge slowly even if $\|(I - \pi_m)\varphi_i\|$, i.e., $\sin\theta(\varphi_i, \mathcal{K}_m(v_1, A))$, tends to zero rapidly. Third, Ritz vectors cannot be guaranteed to converge in theory even if Ritz values do. We should note that this is a case for any unsymmetric A no matter whether its eigenproblem is well conditioned or not. However, this possible nonconvergence of Ritz vectors cannot happen to a symmetric A because then both $\|P_i^{(m)}\| = 1$ and $\inf_{D \text{ diag.}} \operatorname{cond}(\Phi_i^{(m)} D) = 1$.

In fact, there can occur such a phenomenon: Even if the eigenproblem

$$A\varphi = \lambda\varphi \tag{30}$$

is ill conditioned, we may get a well-conditioned approximate eigenproblem (24); on the other hand, we may get an ill-conditioned approximate eigenproblem (24) though (30) is well conditioned.

Example. Take

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \qquad v_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

Then the computed V_2 and H_2 using Arnoldi's method [17] are

$$V_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad H_2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

A has three simple eigenvalues 0, 1, and 2, and not defective, and its eigenproblem is well conditioned. But H_2 is defective, ill conditioned, and has an eigenvalue 1 with multiplicity 2 corresponding to the eigenspace of dimension 1.

4. Conclusion. We have established a convergence theory of generalized Lanczos methods for solving large unsymmetric eigenproblems when matrices are defective and nonderogatory. So now the methods can be used to compute the outer part eigenvalues and the associated eigenvectors of any nonderogatory matrix in theory. However, our analysis has exposed that the approximate eigenvectors, Ritz vectors, obtained by generalized Lanczos methods cannot be guaranteed to converge in theory even if approximate eigenvalues, Ritz values, do. Having looked at Theorem 2.1 [19], we find that many of the results in this paper work for the biorthogonalization Lanczos method without any essential modification.

Concerning numerical aspects, however, there are still many problems to be studied further. At present, the development of generalized Lanczos algorithms is only in an initial stage, and the existing algorithms are only the Arnoldi algorithm and its variants with accelerations as well as IOMs with correction. There should exist other potential efficient algorithms among generalized Lanczos methods to be developed and pursued. For example, how to develop those algorithms, which belong to generalized Lanczos methods but have nonorthogonal basis, is interesting and possibly promising. Another very important issue is how to seek new strategies in order to ensure the convergence of approximate eigenvectors in theory when Ritz values do. The author has recently found a new strategy that can guarantee the convergence of refined approximate eigenvectors if Ritz values do. Numerical experiments there show that the refined iterative algorithms based on Arnoldi's process are considerably more efficient than their counterparts, i.e., the iterative Arnoldi algorithm and the Arnoldi–Chebyshev algorithm. For details, refer to [10], [11].

Finally, we point out that the tools used in the paper could be exploited to develop a convergence theory of generalized Lanczos methods for a defective and derogatory matrix. Besides, Lemma 3.3 can be used to make a convergence analysis of a large class of Krylov subspace-type methods for the solution of large unsymmetric linear systems, e.g., [18], [22], when the matrix is defective. In [8], convergence results were obtained for them.

Note added in proof on §3.2. Essentially, all the theorems in this section hold for general projection methods rather than only generalized Lanczos methods since Theorem 3.1 holds for a general subspace [19] and we do not necessarily limit ourselves to the Krylov subspace $K_m(v_1, A)$ in all the proofs.

Acknowledgments. This paper is part of [11]. The author thanks Professor Ludwig Elsner for reading the manuscript with great care and many helpful discussions, and he is very indebted to the anonymous referees and the editor, Dr. Anne Greenbaum, for many valuable suggestions and comments that enabled him to considerably improve the presentation of this paper.

REFERENCES

[1] W. E. ARNOLDI, *The principle of minimized iteration in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 11–29.

[2] F. CHATELIN, *Ill-conditioned eigenproblems*, Large Scale Eigenvalue Problem, J. Cullum and R. A. Willoughby, eds., 1986, pp. 267–282.

[3] F. CHATELIN AND S. GODET-THOBIE, *Stability analysis in aeronautical industries*, High Performance Computing, M. Durand and F. El. Dabagli, eds., 1991, pp. 415–422.

[4] B. FISCHER AND R. W. FREUND, *Chebyshev polynomials are not always optimal*, J. Approx. Theory, 65 (1991), pp. 261–272.

[5] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd Edition, The John Hopkins University Press, Baltimore, 1989.

[6] D. HO, F. CHATELIN, AND M. BENNANI, *Arnoldi–Tchebychev procedure for large nonsymmetric matrices*, Math. Mod. Numer. Anal., 24 (1990), pp. 53–65.

[7] Z. JIA, *The Studies on the Convergence of the Generalized Lanczos Method for Large Unsymmetric Eigenproblems*, Master's thesis (Chinese), Dalian University of Technology, 1987 (published in part in J. Dalian Univ. of Technology, 30 (1990), pp. 1–7).

[8] ———, *The convergence of Krylov subspace methods for large unsymmetric linear systems*, University of Bielefeld, Germany, 1994, preprint.

[9] ———, *Arnoldi type algorithms for large unsymmetric multiple eigenproblems*, University of Bielefeld, Germany, 1995, preprint.

[10] ———, *Refined iterative algorithms based on Arnoldi's process for large unsymmetric eigenproblems*, Linear Algebra Appl., submitted.

[11] Z. JIA, *Some Numerical Methods for Large Unsymmetric Eigenproblems*, Ph.D. thesis, University of Bielefeld, Germany, 1994.
[12] W. KERNER, *Large-scale complex eigenvalue problems*, J. Comput. Phys., 53 (1989), pp. 1–85.
[13] T. R. RIVLIN, *The Chebyshev Polynomials*, J. Wiley and Sons Inc., New York, 1974.
[14] A. RUHE, *Rational Krylov algorithms for nonsymmetric eigenvalue problems*, Talk at IMA, University of Minnesota, 1992.
[15] ———, *The rational Krylov algorithm for generalized eigenvalue problems*, Talk at the Shanghai International Numerical Algebra and its Applications, 1992.
[16] ———, *Rational Krylov algorithms for nonsymmetric eigenvalue problems* II: *Matrix pairs*, Linear Algebra Appl., 197/198 (1994), pp. 283–296.
[17] Y. SAAD, *Variations on Arnoldi's method for computing eigenelements of large nonsymmetric matrices*, Linear Algebra Appl., 34 (1980), pp. 269–295.
[18] ———, *Krylov subspace methods for solving large unsymmetric linear systems*, Math. Comput., 37 (1981), pp. 105–126.
[19] ———, *Projection methods for solving large sparse eigenvalues problems*, Matrix Pencils, Proceedings Pitea Havsbad, B. K. Kågström and A. Ruhe, eds., Lecture Notes in Math., Springer–Verlag, Berlin, 973, 1983, pp. 121–144.
[20] ———, *Least squares polynomials in the complex plane with applications to solving sparse nonsymmetric matrix problems*, Tech. Report RR-276, Dept. of Computer Science, Yale University, New Haven, CT, 1983.
[21] ———, *Chebyshev acceleration techniques for solving nonsymmetric eigenvalue problems*, Math. Comput., 42 (1984), pp. 567–588.
[22] Y. SAAD AND M. H. SCHULTZ, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[23] D. C. SORENSEN, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 357–385.
[24] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 863–885, July 1995

011

VECTOR ORTHOGONAL POLYNOMIALS AND LEAST SQUARES APPROXIMATION *

ADHEMAR BULTHEEL† AND MARC VAN BAREL†

Abstract. We describe an algorithm for complex discrete least squares approximation, which turns out to be very efficient when function values are prescribed in points on the real axis or on the unit circle. In the case of polynomial approximation, this reduces to algorithms proposed by Rutishauser, Gragg, Harrod, Reichel, Ammar, and others. The underlying reason for efficiency is the existence of a recurrence relation for orthogonal polynomials, which are used to represent the solution. We show how these ideas can be generalized to least squares approximation problems of a more general nature.

Key words. orthogonal vector polynomials, discrete least squares

AMS subject classifications. 41A20, 65F25, 65D05, 65D15, 30E10

1. Introduction.

Let $\{z_k\}_{k=0}^m$ be a set of complex nodes and $\{w_k^2\}_{k=0}^m$ a set of positive weights (let us assume that $w_k > 0$).

We shall first solve the problem of finding the least squares polynomial approximant in the space with positive semidefinite inner product

$$\langle f, g\rangle = \sum_{k=0}^{m} \overline{f(z_k)} w_k^2 g(z_k). \tag{1}$$

Note that this is a positive definite inner product for the space of vectors $(f(z_0), \ldots, f(z_m))$ representing the function values at the given nodes. The polynomial $p \in \mathbb{P}_n$ of degree $n \leq m$ which minimizes

$$\|f - p\|, \quad \text{with} \quad \|v\| = \langle v, v\rangle^{1/2}$$

(note that this is a seminorm) can be found as follows. Find a basis $\{\varphi_0, \ldots, \varphi_n\}$ for $\mathbb{P}_n$, which is orthonormal with respect to $\langle \cdot, \cdot \rangle$. The solution p is the generalized Fourier expansion of f with respect to this basis, truncated after the term of degree n. An algorithm that solves the problem will compute implicitly or explicitly the orthonormal basis and the Fourier coefficients. As we see in the following sections, we can reduce the complexity of such an algorithm by an order of magnitude when a "short recurrence" exists for the orthogonal polynomials. We consider the case where all the z_i are on the real line, in which case a three-term recurrence relation exists, and the case where all the z_i are on the complex unit circle, in which case a Szegő type recurrence relation exists.

The above-mentioned discrete least squares problem is closely related to many other problems in numerical analysis. For example, consider the quadrature formula

$$\int_a^b w(x) f(x) dx \approx \sum_{k=0}^{m} w_k^2 f(z_k),$$

* Received by the editors July 20, 1993; accepted for publication (in revised form) by M. Gutknecht May 25, 1994. This research was supported by the Human Capital and Mobility project ROLLS of the European Community contract ERBCHRXCT930416.

† Department of Computer Science, Katholicke Universiteit, Leuven, Celestijnenlaan 200A, B-3001 Leuven, Belgium (adhemar.bultheel@cs.kuleuven.ac.be).

where $w(x)$ is a positive weight for the real interval $[a, b]$. We get a Gaussian quadrature formula, exact for all polynomials of degree $2m+1$ by a special choice of the nodes and weights. The nodes z_k are the zeros of the $(m+1)$st orthogonal polynomial with respect to $\langle f, g\rangle = \int_a^b f(x)g(x)w(x)dx$. These are also the eigenvalues of the truncated Jacobi matrix which is associated with this orthogonal system. The weights w_i^2 are proportional to q_{i0}^2 where q_{i0} is the first component of the corresponding eigenvector.

Another link can be made with inverse spectral problems. These come in several forms. One variant is precisely the inverse of the previous quadrature problem: find the Jacobi matrix, when its eigenvalues and the first entries of the normalized eigenvectors are given.

We shall call the computation of the quadrature formula or the eigenvalue decomposition of the Jacobi matrix direct problems, while the inverse spectral problem, and the least squares problem will be called inverse problems.

For a survey of inverse spectral problems, we refer to Boley and Golub [5]. One of the methods mentioned there is the Rutishauser–Gragg–Harrod algorithm. This algorithm can be traced back to Rutishauser [14] and was adapted by Gragg–Harrod [11] with a technique of Kahan–Pal–Walker for chasing a nonzero element in the matrix.

For a discrete least squares interpretation of these algorithms we refer to Reichel [12]. When the z_i are not on the real line, but on the unit circle, similar ideas lead to algorithms discussed by Ammar and He [4] and Ammar, Gragg, and Reichel [2] for the inverse eigenvalue problem and to Reichel, Ammar, and Gragg [13] for a least squares interpretation.

We first survey the general theory in the context of discrete least squares approximation where the z_k are arbitrary complex numbers in §§2, 3, and 4. In §5, we explain how the complexity can be reduced with an order of magnitude when short recurrences exist.

The next step (§6) is to generalize these results to the problem of minimizing

$$\min \sum_{k=0}^{m} |w_{0k}p_0(z_k) + \cdots + w_{\alpha k}p_\alpha(z_k)|^2, \tag{2}$$

where the $\{w_{0k}, \ldots, w_{\alpha k}\}_{k=0}^m$ are given complex numbers and the polynomials p_i of degree at most d_i, $i = 0, \ldots, \alpha$ must be found, with the constraint that at least one of them is monic of strict degree.

When $\alpha = 1$, this generalization is related with rational approximation, in contrast with the previously described problem, which is related to polynomial approximation. We refer to the generalized problem as the matrix case, while the simpler polynomial case is referred to as the scalar case.

For the matrix case, we may distinguish between two levels of complication. When all the degrees d_i are equal, it turns out (§7) that the solution method can be described in terms of square matrix orthogonal polynomials of size $\alpha+1$, and the previous theory of scalar orthogonal polynomials is readily generalized.

When not all the degrees d_i are equal, we are in the most general case that we consider here (§8). The solution can now be described in terms of vector orthogonal polynomials, which allows the scalar orthogonal polynomials of the first case and the matrix orthogonal polynomials of the second case to combine, both of which show up during the solution of the problem.

The breakdown of the algorithm will only occur in the case of exact interpolation. This is discussed in §9.

To avoid an unduly complicated notation, we mainly restrict ourselves in this paper to the case $\alpha = 1$, but the generalization to general $\alpha \geq 1$ should be obvious.

2. Polynomial least squares approximation. Discrete least squares approximation by polynomials is a classical problem in numerical analysis where orthogonal polynomials play a central role.

Given an inner product $\langle \cdot, \cdot \rangle$ defined on $\mathbb{P}_m \times \mathbb{P}_m$, the polynomial $p \in \mathbb{P}_n$ of degree at most $n \leq m$, which minimizes the error

$$\|f - p\|, \qquad p \in \mathbb{P}_n,$$

is given by

$$p = \sum_{k=0}^{n} \varphi_k a_k, \qquad a_k = \langle f, \varphi_k \rangle$$

when the $\{\varphi_k\}_0^n$ form an orthonormal set of polynomials:

$$\varphi_k \in \mathbb{P}_k - \mathbb{P}_{k-1}, \qquad \mathbb{P}_{-1} = \varnothing, \qquad \langle \varphi_k, \varphi_l \rangle = \delta_{kl}.$$

The inner product we consider here is of the discrete form (1) where the z_i are distinct complex numbers.

Note that when $m = n$, the least squares solution is the interpolating polynomial, so that interpolation can be seen as a special case.

To illustrate where the orthogonal polynomials show up in this context, we start with an arbitrary polynomial basis $\{\psi_k\}$, $\psi_k \in \mathbb{P}_k - \mathbb{P}_{k-1}$. Setting

$$p = \sum_{k=0}^{n} \psi_k a_k^{\Psi}, \qquad a_k^{\Psi} \in \mathbb{C},$$

the least squares problem can be formulated as finding the weighted least squares solution of the system of linear equations

$$\sum_{k=0}^{n} \psi_k(z_i) a_k^{\Psi} = f(z_i), \qquad i = 0, \ldots, m,$$

which is the same as the least squares solution of

$$W \Psi_n A_n^{\Psi} = WF,$$

where $W = \operatorname{diag}(w_0, \ldots, w_m)$ and

$$\Psi_n = \begin{bmatrix} \psi_0(z_0) & \ldots & \psi_n(z_0) \\ \vdots & & \vdots \\ \psi_0(z_m) & \ldots & \psi_n(z_m) \end{bmatrix}, \quad A_n^{\Psi} = \begin{bmatrix} a_0^{\Psi} \\ \vdots \\ a_n^{\Psi} \end{bmatrix}, \quad F = \begin{bmatrix} f(z_0) \\ \vdots \\ f(z_m) \end{bmatrix}.$$

Note that when $\psi_k(z) = z^k$, the power basis, then Ψ_n is a rectangular Vandermonde matrix.

The normal equations for this system are

$$(\Psi_n^H W^2 \Psi_n) A_n^{\Psi} = \Psi_n^H W^2 F.$$

When the ψ_k are chosen to be the orthonormal polynomials φ_k, then $\Psi_n^H W^2 \Psi_n = I_{n+1}$ and the previous system gives the solution $A_n^\Psi = \Psi_n^H W^2 F$ immediately.

When the least squares problem is solved by QR factorization, i.e., when Q is an $m \times m$ unitary matrix such that $Q^H W \Psi_n = [R^T \; 0^T]^T$ is upper triangular, we must solve the triangular system given by the first $n+1$ rows of

$$\begin{bmatrix} R \\ 0 \end{bmatrix} A_n^\Psi = Q^H W F + \begin{bmatrix} 0 \\ X \end{bmatrix},$$

where X is related to the residual vector r by

$$\begin{bmatrix} 0 \\ X \end{bmatrix} = Q^H r, \qquad r = W \Psi_n A_n^\Psi - WF.$$

Note that the least squares error is $\|X\| = \|r\|$. Again, when the ψ_k are replaced by the orthonormal polynomials φ_k, we get the trivial system ($m \geq n$)

$$\begin{bmatrix} I_{n+1} \\ 0 \end{bmatrix} A_n^\Phi = Q^H W F + \begin{bmatrix} 0 \\ X \end{bmatrix}.$$

Note that a unitary matrix Q is always related to the orthonormal polynomials φ_k by

$$Q = W\Phi,$$

where

$$\Phi = \Phi_m = \begin{bmatrix} \varphi_0(z_0) & \cdots & \varphi_m(z_0) \\ \vdots & & \vdots \\ \varphi_0(z_m) & \cdots & \varphi_m(z_m) \end{bmatrix}$$

since

$$Q^H Q = \Phi^H W^2 \Phi = I_{m+1}.$$

3. The Hessenberg matrix. From the previous discussion, it follows that the central problem is to construct the orthonormal basis $\{\varphi_k\}$. In general, the polynomial $z\varphi_{k-1}(z)$ can be expressed as a linear combination of the polynomials $\varphi_0, \ldots, \varphi_k$, leading to a relation of the form

$$z\varphi_{k-1}(z) = \eta_{kk}\varphi_k(z) + \cdots + \eta_{0k}\varphi_0(z), \qquad k = 1, \ldots, m+1.$$

We can express the previous relations as

$$z[\varphi_0(z), \ldots, \varphi_m(z)] = [\varphi_0(z), \ldots, \varphi_m(z)]H + e_{m+1}^T \varphi_{m+1}(z)\eta_{m+1,m+1}, \tag{3}$$

where H is an upper Hessenberg matrix

$$H = \begin{bmatrix} \eta_{01} & \cdots & \eta_{0m} & \eta_{0,m+1} \\ \eta_{11} & \cdots & \eta_{1m} & \eta_{1,m+1} \\ & \ddots & \vdots & \vdots \\ & & \eta_{mm} & \eta_{m,m+1} \end{bmatrix}$$

and $e_{m+1}^T = [0 \; 0 \; \ldots \; 0 \; 1]$.

Note that a discrete inner product of the proposed form will cause a breakdown in the generation of the polynomials at stage $m+1$. Indeed, we should identify a function with the $(m+1)$-vector of its function values in z_k, $k = 0, \ldots, m$. Thus when we say the "polynomial p," we actually mean the vector $(p(z_0), \ldots, p(z_m))$. Thus our "function space" is a space of $(m+1)$-vectors, which is inherently $(m+1)$-dimensional, and thus the $(m+1)$st orthogonal polynomial will be orthogonal to the whole space, hence it must be zero. Thus, if φ_k are these orthogonal polynomials, then $[\varphi_{m+1}(z_0), \ldots, \varphi_{m+1}(z_m)]^T$ will be the zero vector. This is equivalent to saying that φ_{m+1} is proportional to $(z - z_0) \ldots (z - z_m)$.

Even when we use terms as "functions" and "polynomials," the problem considered is in fact a vectorial problem, which can be best formulated in terms of matrices, which we do below.

Setting $\Phi = \Phi_m$ as before, we rewrite the relation (3) as

$$Z\Phi = \Phi H$$

with $Z = \operatorname{diag}(z_0, \ldots, z_m)$.

Multiplying with the diagonal matrix W and using $WZ = ZW$, we are led to

$$H = (W\Phi)^H Z (W\Phi) = Q^H Z Q,$$

which means that the diagonal matrix Z and the Hessenberg matrix H are unitarily similar.

The constant polynomial φ_0 is normalized when it is equal to η_{00}^{-1} with η_{00} given by

$$Q^H \mathbf{w}_1 = [\eta_{00}, 0, \ldots, 0]^T,$$

where $\mathbf{w}_1 = [w_0, \ldots, w_m]^T$. Indeed, using $Q = W\Phi$ and supposing $\|\varphi_0\| = 1$, we see that all the entries in $Q^H \mathbf{w}_1$ are zero by orthogonality, except for the first one, which is $1/\varphi_0$.

This condition is not sufficient to characterize Q completely. We can fix it uniquely by making the φ_k have positive leading coefficients. This will be obtained when all the η_{kk}, $k = 0, 1, \ldots, m$ are positive. Since we assume that all the weights w_i^2 are positive, the η_{kk} are nonzero and therefore this normalization can always be realized.

We thus obtained a one-to-one relation between the data $\{z_i, w_i\}_0^m$, the unitary matrix Q and the elements η_{ij}, $i = 0, \ldots, m$, $j = 0, \ldots, m+1$ of an extended (with η_{00}) Hessenberg matrix. This also fixes the orthonormal polynomials.

Since Z and H are unitarily similar, they have the same spectrum and the construction of H from Z by unitary similarity transformations is in fact an inverse spectral problem: given the spectrum Z and the first components of the eigenvectors, find the set of orthonormal eigenvectors (the columns of Q^H), such that $Q^H Z Q$ is the eigenvalue decomposition of some upper Hessenberg matrix with the normalization described above.

In the direct problem, one computes the eigenvalues $\{z_k\}_0^m$ and the eigenvectors Q from the Hessenberg matrix, e.g., with the QR algorithm. For the inverse problem, the Hessenberg matrix is reconstructed from the spectral data by an algorithm that could be called an inverse QR algorithm. This is the Rutishauser–Gragg–Harrod algorithm for the case of the real line [11], [12] and the unitary inverse QR algorithm described in [2] for the case of the unit circle. For the least squares problem, we add the function values $f(z_k)$ and when these are properly transformed by the similarity transformations of the inverse QR algorithm, this will result in the generalized

Fourier coefficients of the approximant and some information about the corresponding residual. Indeed, the solution of the approximation problem is given by

$$p = [\varphi_0, \ldots, \varphi_n] A_n^{\Phi}, \qquad A_n^{\Phi} = \Phi_n^H W^2 F.$$

Note that the normal equations are never explicitly formed.

The whole scheme can be collected in one table giving the relations

$$Q^H \left[\begin{array}{c|c||c} \mathbf{w}_0 & \mathbf{w}_1 & Z \end{array} \right] \left[\begin{array}{cc} I_2 & \\ & Q \end{array} \right] = \left[\begin{array}{c|c||ccc} & \eta_{00} & \eta_{01} & \cdots & \eta_{0m} & \eta_{0,m+1} \\ A_n^{\Phi} & 0 & \eta_{11} & \cdots & \eta_{1m} & \eta_{1,m+1} \\ -- & \vdots & & \ddots & \vdots & \vdots \\ X & 0 & & & \eta_{mm} & \eta_{m,m+1} \end{array} \right]$$

with $\mathbf{w}_0 = WF$ and $\mathbf{w}_1 = [w_0, \ldots, w_m]^T$ as before. The approximation error is $\|X\|$. For further reference we refer to the matrix of the right-hand side as the extended Hessenberg matrix.

4. Updating. Suppose that A_n^{Φ} was computed by the last scheme for some data set $\{z_i, f_i, w_i\}_0^m$. We then end up with a scheme of the following form ($n = 3, m = 5$):

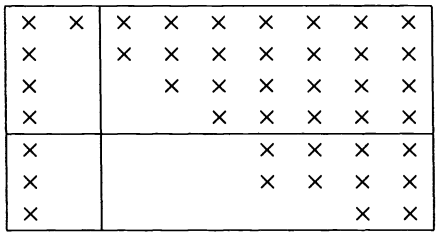
$$\left[\begin{array}{cc|cccccc} \times & \times & \times & \times & \times & \times & \times & \times \\ \times & & \times & \times & \times & \times & \times & \times \\ \times & & & \times & \times & \times & \times & \times \\ \times & & & & \times & \times & \times & \times \\ \hline \times & & & & & \times & \times & \times \\ \times & & & & & & \times & \times \end{array} \right].$$

A new data triple $(z_{m+1}, f_{m+1}, w_{m+1})$ can be added, for example, in the top line. The three crosses in the top line of the left scheme below represent $w_{m+1}f_{m+1}$, w_{m+1} and z_{m+1}, respectively. The other crosses correspond to the ones we had in the previous scheme.

$$\left[\begin{array}{cc|c|cccccc} \times & \times & \times & & & & & & \\ \hline \times & \times & & \times & \times & \times & \times & \times & \times \\ \times & & & \times & \times & \times & \times & \times & \times \\ \times & & & & \times & \times & \times & \times & \times \\ \times & & & & & \times & \times & \times & \times \\ \hline \times & & & & & & \times & \times & \times \\ \times & & & & & & & \times & \times \end{array} \right] \Rightarrow \left[\begin{array}{cc|ccccccc} \times & \times & \times & \times & \times & \times & \times & \times & \times \\ \times & & \times & \times & \times & \times & \times & \times & \times \\ \times & & & \times & \times & \times & \times & \times & \times \\ \times & & & & \times & \times & \times & \times & \times \\ \hline \times & & & & & \times & \times & \times & \times \\ \times & & & & & & \times & \times & \times \\ \times & & & & & & & \times & \times \end{array} \right]$$

This left scheme must be transformed by unitary similarity transformations into the right scheme, which has the same form as the original one but with one extra row and one extra column. This result is obtained by eliminating the (2,2) element by an elementary rotation/reflection in the plane of the first two rows. The corresponding transformation on the columns will influence columns 3 and 4 and will introduce a nonzero element at position (3,3), which should not be there. This is eliminated by a rotation/reflection in the plane of rows 2 and 3, etc. We call this procedure chasing the elements down the diagonal. In the first column of the result, we find above the horizontal line the updated coefficients A_n^{Φ}. When we do not change n, it is sufficient to perform only the operations that influence these coefficients. Thus we could have

stopped after we obtained the form



This can be done with $O(n^2)$ operations per new data point. In the special case of data on the real line or on the unit circle, this reduces to $O(n)$ operations as we see in the next section.

5. Recurrence relations. The algorithm described above simplifies considerably when the orthogonal polynomials satisfy a particular recurrence relation. A classical situation occurs when the $z_i \in \mathbb{R}$, $i = 0, 1, \ldots, m$. Since also the weights w_i are real, the Q and H matrix will be real, which means that we can drop the complex conjugation from our notation. However, in view of the generalization to follow, where we have complex numbers instead of the w_i, we keep for the moment the bar, although it has no effect, being applied to real numbers. Thus we observe that for $z_i \in \mathbb{R}$, the Hessenberg matrix H satisfies

$$H^H = (Q^H Z Q)^H = Q^H Z Q = H.$$

This means that H is Hermitian and therefore tridiagonal. The matrix H reduces to the classical Jacobi matrix

$$H = \begin{bmatrix} a_0 & \bar{b}_1 & & \\ b_1 & a_1 & \ddots & \\ & \ddots & \ddots & \bar{b}_m \\ & & b_m & a_m \end{bmatrix}$$

containing the coefficients of the three term recurrence relation

$$\varphi_{-1} = 0, \quad z\varphi_k(z) = \bar{b}_k \varphi_{k-1}(z) + a_k \varphi_k(z) + b_{k+1}\varphi_{k+1}(z), \quad k = 0, 1, \ldots, m-1.$$

A similar situation occurs when the z_i are purely imaginary, in which case the matrix H is skew Hermitian. We do not discuss this case separately.

The algorithm we described before now needs to perform rotations (or reflections) on vectors of length 3 or 4, which reduces the complexity of the algorithm by an order. This is the basis of the Rutishauser–Gragg–Harrod algorithm [14], [11]. See also [5], [12], [6].

In this context, it was observed only lately [9], [10], [2], [13], [3] that the situation where the $z_i \in \mathbb{T}$ (the unit circle) also leads to a simplification. It follows from

$$H^H H = Q^H Z^H Z Q = Q^H Q = I_{m+1}$$

that H is then a unitary Hessenberg matrix. The related orthogonal polynomials are orthogonal with respect to a discrete measure supported on the unit circle. The three-term recurrence relation is replaced by a recurrence of Szegö-type

$$z\varphi_{k-1}(z) = \varphi_k(z)\sigma_k + \varphi_{k-1}^*(z)\gamma_k$$

with

$$\varphi_k^*(z) = z^k\overline{\varphi_k(1/\bar{z})} \in \mathbb{P}_k \quad \text{and} \quad \sigma_k^2 = 1 - |\gamma_k|^2, \quad \sigma_k > 0,$$

where the γ_k are the so-called reflection coefficients or Schur parameters. Just like in the case of a tridiagonal matrix, the Hessenberg matrix is built up from the recurrence coefficients γ_k, σ_k. However, the connection is much more complicated. For example, for $m = 3$, H has the form

$$H = \begin{bmatrix} -\gamma_1 & -\sigma_1\gamma_2 & -\sigma_1\sigma_2\gamma_3 & -\sigma_1\sigma_2\sigma_3\gamma_4 \\ \sigma_1 & -\bar{\gamma}_1\gamma_2 & -\bar{\gamma}_1\sigma_2\gamma_3 & -\bar{\gamma}_1\sigma_2\sigma_3\gamma_4 \\ & \sigma_2 & -\bar{\gamma}_2\gamma_3 & -\bar{\gamma}_2\sigma_3\gamma_4 \\ & & -\sigma_3 & \bar{\gamma}_3\gamma_4 \end{bmatrix}.$$

The Schur parameters can be recovered from the Hessenberg matrix by

$$\sigma_j = \eta_{jj}, \quad j = 1, \ldots, m, \quad \eta_{00} = 1/\varphi_0 = \sigma_0,$$

$$\gamma_j = -\eta_{0j}/(\sigma_1\sigma_2 \ldots \sigma_{j-1}), \qquad j = 1, \ldots, m+1.$$

The complexity reduction in the algorithm is obtained from the important observation that any unitary Hessenberg matrix H can be written as a product of elementary unitary factors

$$H = G_1 G_2 \ldots G_m G'_{m+1}$$

with

$$G_k = I_{k-1} \oplus \begin{bmatrix} -\gamma_k & \sigma_k \\ \sigma_k & \bar{\gamma}_k \end{bmatrix} \oplus I_{m-k}, \qquad k = 1, \ldots, m$$

and

$$G'_{m+1} = \operatorname{diag}(1, \ldots, 1, -\gamma_{m+1}).$$

This result can be found, e.g., in [9], [2].

Now an elementary similarity transformation on rows/columns k and $k+1$ of H, represented in this factored form, will only affect the factors G_k and part of the factors G_{k-1} and G_{k+1}. Again, these operations require computations on short vectors of length 3, making the algorithm very efficient again. For the details consult [9], [2], [13]. For example, the interpolation problem ($n = m$) is solved in $O(m^2)$ operations instead of $O(m^3)$.

6. Vector approximants. The previous situation of polynomial approximation can be generalized as follows.

Given $\{z_i; f_{0i}, \ldots, f_{\alpha i}; w_{0i}, \ldots, w_{\alpha i}\}_{i=0}^m$, find polynomials $p_k \in \mathbb{P}_{d_k}$, $k = 0, \ldots, \alpha$, such that

$$\sum_{i=0}^{m} |w_{0i} f_{0i} p_0(z_i) + \cdots + w_{\alpha i} f_{\alpha i} p_\alpha(z_i)|^2$$

is minimized. Now it does not really matter whether the w_{ji} are positive or not, since the products $w_{ji} f_{ji}$ will now play the role of the weights and the f_{ji} are arbitrary

complex numbers. Thus, to simplify the notation, we could as well write w_{ji} instead of $w_{ji}f_{ji}$ since these numbers will always appear as products. Thus the problem is to minimize

$$\sum_{i=0}^{m} |w_{0i}p_0(z_i) + \cdots + w_{\alpha i}p_\alpha(z_i)|^2.$$

Setting $\mathbf{d} = (d_0, \ldots, d_\alpha)$, $\mathbb{P}_{\mathbf{d}} = [\mathbb{P}_{d_0}, \ldots, \mathbb{P}_{d_\alpha}]^T$,

$$w_i = [w_{0i}, \ldots, w_{\alpha i}], \qquad p(z) = [p_0(z), \ldots, p_\alpha(z)]^T \in \mathbb{P}_{\mathbf{d}},$$

we can write this as

$$\min \sum_{i=0}^{m} |w_i p(z_i)|^2, \qquad p \in \mathbb{P}_{\mathbf{d}}.$$

Of course, this problem has the trivial solution $p = 0$, unless we require at least one of the $p_i(z)$ to be of strict degree d_i, e.g., by making it monic. This, or any other normalization condition could be imposed for that matter.

In this paper we require that p_α is monic of degree d_α, and rephrase this as $p_\alpha \in \mathbb{P}^M_{d_\alpha}$.

To explain the general idea, we restrict ourselves to $\alpha = 1$, the case of a general α being a straightforward generalization, which would only increase the notational burden. Thus we consider the problem

$$\min \sum_{i=0}^{m} |w_{0i}p_0(z_i) + w_{1i}p_1(z_i)|^2, \qquad p_0 \in \mathbb{P}_{d_0}, p_1 \in \mathbb{P}^M_{d_1}.$$

Note that when $w_{0i} = w_i > 0$, $w_{1i} = -w_i f_i$, and $p_1 \equiv 1 \in \mathbb{P}^M_0$, (i.e., $d_1 = 0$), then we get the polynomial approximation problem discussed before.

When we set $w_{0i} = w_i f_{0i}$ and $w_{1i} = -w_i f_{1i}$ with $w_i > 0$, the problem becomes

$$\min \sum_{i=0}^{m} w_i^2 |f_{0i}p_0(z_i) - f_{1i}p_1(z_i)|^2,$$

which is a linearized version of the rational least squares problem of determining the rational approximant p_0/p_1 for the data f_{1i}/f_{0i}, or equivalently the rational approximant p_1/p_0 for the data f_{0i}/f_{1i}. Note that in the linearized form, it is as easy to prescribe pole information ($f_{0i} = 0$) as it is to fix a finite function value ($f_{0i} \neq 0$).

The solution of the general case is partly parallel to the polynomial case $d_1 = 0$ discussed before, and partly parallel to another simple case, namely, $d_0 = d_1 = n$, which we discuss first in §7. In the subsequent §8, we consider the general case where $d_0 \neq d_1$.

7. Equal degrees. We consider the case $\alpha = 1$, $d_0 = d_1 = n$. This means that $\mathbb{P}_{\mathbf{d}}$ is here equal to $\mathbb{P}^{2\times 1}_n$.

7.1. The optimization problem. We must find

$$\min \sum_{i=0}^{m} |w_i p(z_i)|^2, \quad p_0 \in \mathbb{P}_n, \quad p_1 \in \mathbb{P}^M_n,$$

where $w_i = [w_{0i} \;\; w_{1i}]$ and $p(z) = [p_0(z) \;\; p_1(z)]^T \in \mathbb{P}_n^{2\times 1}$. This problem was considered in [15], [17]. We propose a solution of the form

$$p(z) = \sum_{k=0}^{n} \varphi_k(z) a_k,$$

where

$$\varphi_k(z) \in \mathbb{P}_k^{2\times 2} - \mathbb{P}_{k-1}^{2\times 2}, \quad a_k \in \mathbb{C}^{2\times 1}, \quad k = 0, 1, \ldots, n.$$

Proposing $p(z)$ to be of this form assumes that the leading coefficients of the block polynomials φ_k are nonsingular. Otherwise this would not represent all possible couples of polynomials $(p_0, p_1) \in \mathbb{P}_n^{2\times 1}$. We call this the regular case and assume for the moment that we are in this comfortable situation. In the singular case, a breakdown may occur during the algorithm, and we shall deal with that separately.

Note that the singular case did not show up in the previous scalar polynomial case, unless at the very end when $n = m + 1$, since the weights were assumed to be positive. We see below that in this block polynomial situation, the weights are not positive and could even be singular.

When we denote

$$W = \operatorname{diag}(w_0, \ldots, w_m) \in \mathbb{C}^{(m+1)\times(2m+2)},$$
$$A_n^{\Phi} = [a_0^T, \ldots, a_n^T]^T \in \mathbb{C}^{(2n+2)\times 1},$$
$$\Phi_n = \begin{bmatrix} \varphi_0(z_0) & \cdots & \varphi_n(z_0) \\ \vdots & & \vdots \\ \varphi_0(z_m) & \cdots & \varphi_n(z_m) \end{bmatrix} \in \mathbb{C}^{(2m+2)\times(2n+2)},$$

the optimization problem is to find the least squares solution of the homogeneous linear system

$$W \Phi_n A_n^{\Phi} = 0$$

with the constraint that p_1 should be monic of degree n.

For simplicity reasons, suppose that $m + 1 = 2(m' + 1)$ is even. If it were not, we would have to make a modification in our formulations for the index m'. The algorithm however, does not depend on m being odd or even, as we shall see later.

By making the block polynomials φ_k orthogonal so that

$$\sum_{i=0}^{m} \varphi_k(z_i)^H w_i^H w_i \varphi_l(z_i) = \delta_{kl} I_2, \quad k, l = 0, 1, \ldots, m', \tag{4}$$

we can construct a unitary matrix $Q \in \mathbb{C}^{(m+1)\times(m+1)}$ by setting

$$Q = W\Phi,$$

where $\Phi = \Phi_{m'}$ is a $(2m + 2) \times (m + 1)$ matrix, so that Q is a square matrix of size $m + 1$.

We also assume that the number of data points $m + 1$ is at least equal to the number of unknowns $2n + 1$ (recall that one coefficient is fixed by the monic normalization).

The unitarity of the matrix Q means that

$$Q^H Q = \Phi^H W^H W \Phi = I_{m+1}$$

and the optimization problem reduces to

$$\begin{aligned}\min \sum_{i=0}^{m} p(z_i)^H w_i^H w_i p(z_i) &= \min (A_{m'}^{\Phi})^H \Phi^H W^H W \Phi (A_{m'}^{\Phi}) \\ &= \min (A_{m'}^{\Phi})^H (A_{m'}^{\Phi}) \\ &= \min \sum_{k=0}^{m'} a_k^H a_k \\ &= \min \sum_{k=0}^{m'} (|a_{1k}|^2 + |a_{2k}|^2), \qquad a_k = [a_{1k} \;\; a_{2k}]^T\end{aligned}$$

with the constraint that $p(z) \in \mathbb{P}_n^{2\times 1}$; thus $a_{n+1} = \cdots = a_{m'} = 0$, and $p_1 \in \mathbb{P}_n^M$. Since the leading term of p_1 is only influenced by $\varphi_n a_n$, we are free to choose $a_0, \ldots, a_{n-1}$, so that we can set them equal to zero, to minimize the error. Thus it remains to find

$$\min(|a_{1n}|^2 + |a_{2n}|^2)$$

such that

$$\varphi_n(z) \begin{bmatrix} a_{1n} \\ a_{2n} \end{bmatrix} = \begin{bmatrix} p_0(z) \\ p_1(z) \end{bmatrix} \in \begin{bmatrix} \mathbb{P}_n \\ \mathbb{P}_n^M \end{bmatrix}.$$

To monitor the degree of p_1, we require that the polynomials φ_k have an upper triangular leading coefficient

$$\varphi_k(z) = \begin{bmatrix} \alpha_k & \gamma_k \\ 0 & \beta_k \end{bmatrix} z^k + \cdots$$

with $\alpha_k, \beta_k > 0$. Note that this is always possible in the regular case. The condition $p_1 \in \mathbb{P}_n^M$ then sets $a_{2n} = 1/\beta_n$ and a_{1n} is arbitrary, hence to be set equal to zero if we want to minimize the error.

As a conclusion, we have solved the approximation problem by computing the nth block polynomial φ_n, orthonormal in the sense of (4), and with leading coefficient upper triangular. The solution is

$$p(z) = \begin{bmatrix} p_0(z) \\ p_1(z) \end{bmatrix} = \varphi_n(z) \begin{bmatrix} 0 \\ a_{2n} \end{bmatrix}, \qquad a_{2n} = 1/\beta_n.$$

7.2. The algorithm. As in the scalar polynomial case, expressing $z\varphi_k(z)$ in terms of $\varphi_0, \ldots, \varphi_{k+1}$ for $z \in \{z_0, \ldots, z_m\}$ leads to the matrix relation

$$\mathbf{Z}\Phi = \Phi H,$$

where as before $\Phi = \Phi_{m'}$, $Z = \operatorname{diag}(z_0, \ldots, z_m)$, $\mathbf{Z} = Z \otimes I_2 = \operatorname{diag}(z_0 I_2, \ldots, z_m I_2)$, and H is a block upper Hessenberg matrix with 2×2 blocks. If the leading coefficient

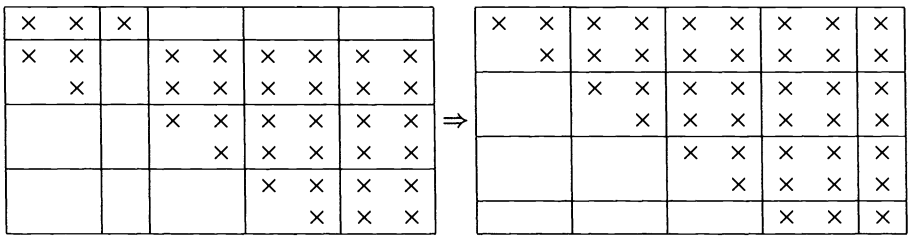
of φ_k is upper triangular, then the subdiagonal blocks of H are upper triangular. The computational scheme is compressed in the formula

$$Q^H[\mathbf{w}|Z]\begin{bmatrix} I_2 & \\ & Q \end{bmatrix} = \left[\begin{array}{c|ccc} \eta_{00} & \eta_{01} & \cdots & \eta_{0m'} & \eta_{0,m'+1} \\ & \eta_{11} & & \eta_{1m'} & \eta_{1,m'+1} \\ & & \ddots & \vdots & \vdots \\ & & & \eta_{m'm'} & \eta_{m',m'+1} \end{array}\right],$$

where $\mathbf{w} = [w_0^T, \ldots, w_m^T]^T$ and where all η_{ij} are 2×2 blocks and the η_{ii} are upper triangular with positive diagonal elements. Thus

$$\varphi_0 = \eta_{00}^{-1}; \quad z\varphi_{k-1}(z) = \varphi_0(z)\eta_{0k} + \cdots + \varphi_k(z)\eta_{kk}, \quad k = 1, \ldots, m'.$$

The updating after adding the data (z_{m+1}, w_{m+1}), where $w_{m+1} = (w_{0,m+1}, w_{1,m+1})$, makes the transformation with unitary similarity transformations from the left to the right scheme below. The three crosses in the top row of the left scheme represent the new data.



The successive elementary transformations eliminate the crosses on the subdiagonal, chasing them down the matrix. This example also illustrates what happens at the end when m is an even number: the polynomial $\varphi_{m'} \in \mathbb{P}_{m'}^{2\times 1}$ instead of $\varphi_{m'} \in \mathbb{P}_{m'}^{2\times 2}$. Again, when finishing this updating after φ_n has been computed, it will require only $O(n^2)$ operations per data point introduced. In the special case of data on the real line or the unit circle, this reduces to $O(n)$ operations. For the details, we refer to [15], [17].

By the same arguments as in the scalar case, it is still true that H is Hermitian, hence block tridiagonal, when all the z_i are real. Taking into account that the subdiagonal blocks are upper triangular, we obtain in this case that H is pentadiagonal and the extended Hessenberg matrix has the form

$$\left[\begin{array}{c|ccc} B_0 & A_0 & B_1^H & \\ & B_1 & A_1 & \ddots \\ & & \ddots & \ddots & B_{m'}^H \\ & & & B_{m'} & A_{m'} \end{array}\right]$$

with the 2×2 blocks B_k upper triangular and the A_k Hermitian. This leads to the following block three-term recurrence

$$\varphi_0 = B_0^{-1}, \quad z\varphi_k(z) = \varphi_{k-1}B_k^H + \varphi_k(z)A_k + \varphi_{k+1}(z)B_{k+1}, \quad 0 \le k < m'.$$

This case was considered in [15].

Similarly, the case where all z_i lie on the unit circle $\mathbb{T}$, leads to a 2×2 block generalization of the corresponding polynomial case. For example, the extended unitary block Hessenberg matrix takes the form ($m' = 3$)

$$[H_0|H] = \left[\begin{array}{c|cccc} \sigma_0 & -\gamma_1 & -\Sigma_1\gamma_2 & -\Sigma_1\Sigma_2\gamma_3 & \Sigma_1\Sigma_2\Sigma_3 \\ & \sigma_1 & -\Gamma_1\gamma_2 & -\Gamma_1\Sigma_2\gamma_3 & \Gamma_1\Sigma_2\Sigma_3 \\ & & \sigma_2 & -\Gamma_2\gamma_3 & \Gamma_2\Sigma_3 \\ & & & \sigma_3 & \Gamma_3 \end{array}\right], \quad \gamma_i, \sigma_i, \Sigma_i, \Gamma_i \in \mathbb{C}^{2\times 2}.$$

The matrices

$$U_k = \begin{bmatrix} -\gamma_k & \Sigma_k \\ \sigma_k & \Gamma_k \end{bmatrix}$$

are unitary: $U_k^H U_k = I_4$. Note that by allowing some asymmetry in the U_k we do not need a $-\gamma_4$ in the last column as we had in the scalar case. We have for $k = 1, \ldots, m'$, the block Szegö recurrence relations

$$\begin{aligned} \varphi_k(z)\sigma_k &= z\varphi_{k-1}(z) + \varphi'_{k-1}(z)\gamma_k, \\ \varphi'_k(z)\Sigma_k^H &= z\varphi_{k-1}(z)\gamma_k^H + \varphi'_{k-1}(z), \end{aligned}$$

which start with $\varphi_0 = \varphi'_0 = \sigma_0^{-1}$.

The block Hessenberg matrix can again be factored as

$$H = G_1 G_2 \ldots G_{m'}$$

with

$$G_k = I_{2(k-1)} \oplus U_k \oplus I_{m-2k-1}, \qquad k = 1, \ldots, m'.$$

The proof of this can be found in [17]. This makes it possible to perform the elementary unitary similarity transformations of the updating procedure only on vectors of maximal length 5, very much like in the case of real points z_i. Thus also here, the complexity of the algorithm reduces to $O(m^2)$ for interpolation. More details can be found in [17]. For the case of the real line, the algorithm was also discussed in [1], solving an open problem in [5, p. 615]. The previous procedure now solves the problem also for the case of the unit circle.

7.3. Summary. The case $\alpha = 1$, $d_0 = d_1 = n$ and also the case $\alpha \geq 1$, $d_0 = d_1 = \cdots = d_\alpha = n$ for that matter, generalizes the polynomial approximation problem by constructing orthonormal polynomials φ_k which are $(\alpha + 1) \times (\alpha + 1)$ polynomial matrices and these are generated by a block three-term recurrence relation when all $z_i \in \mathbb{R}$ and by a block Szegö recurrence relation when all $z_i \in \mathbb{T}$.

The computational algorithm is basically the same, since it reduces the extended matrix

$$[\mathbf{w}|Z] \in \mathbb{C}^{(m+1)\times(\alpha+m+2)}$$

by a sequence of elementary unitary similarity transformations to an upper trapezoidal matrix

$$Q^H[\mathbf{w}|Z] \begin{bmatrix} I_{\alpha+1} & \\ & Q \end{bmatrix} = [H_0|H]$$

with H block upper Hessenberg with $(\alpha+1)\times(\alpha+1)$ blocks and

$$H_0 = Q^H \mathbf{w} = [\eta_{00}^T, 0, \ldots, 0]^T,$$

where $\eta_{00} \in \mathbb{C}^{(\alpha+1)\times(\alpha+1)}$ is upper triangular with positive diagonal elements, as well as all the subdiagonal blocks of H. For $n = m'$, where $(\alpha+1)(m'+1) - 1 = m+1$ (which implies that $\sigma_{m'}$ is of size $\alpha \times (\alpha+1)$) we solve an interpolation problem. It requires $O(m^2)$ operations when $z_i \in \mathbb{R}$ or $\in \mathbb{T}$, instead of $O(m^3)$ when the z_i are arbitrary in $\mathbb{C}$.

8. Arbitrary degrees. In this section we consider the case $\alpha = 1$ with $d_0 \neq d_1$. For more details we refer to [16].

8.1. The problem. We suppose without loss of generality that $d_0 = \delta$ and $d_1 = n + \delta$, $n, \delta \geq 0$. We must find once more

$$\min \sum_{i=0}^{m} |w_i p(z_i)|^2, \quad p_0 \in \mathbb{P}_\delta, \quad p_1 \in \mathbb{P}_{n+\delta}^M$$

with $w_i = [w_{0i} \;\; w_{1i}]$ and $[p_0(z) \;\; p_1(z)]^T \in \mathbb{P}_{\mathbf{d}}$, $\mathbf{d} = (d_0, d_1)$.

The polynomial approximation problem is recovered by setting $\delta = 0$. The case $d_0 = d_1 = \delta$ is recovered by setting $n = 0$.

The simplest approach to the general problem is by starting with the algorithm. In the subsequent subsections, we propose a computational scheme involving unitary similarity transformations, next we give an interpretation in terms of orthogonal polynomials and finally we solve the approximation problem.

8.2. The algorithm. Comparing the cases $\delta = 0$ and $n = 0$, we see that the algorithm applies a sequence of elementary unitary similarity transformations on an extended matrix

$$[\mathbf{w}|Z], \quad \mathbf{w} = [w_0^T, \ldots, w_m^T]^T, \quad Z = \operatorname{diag}(z_0, \ldots, z_m)$$

to bring it in the form of an extended (block) upper Hessenberg

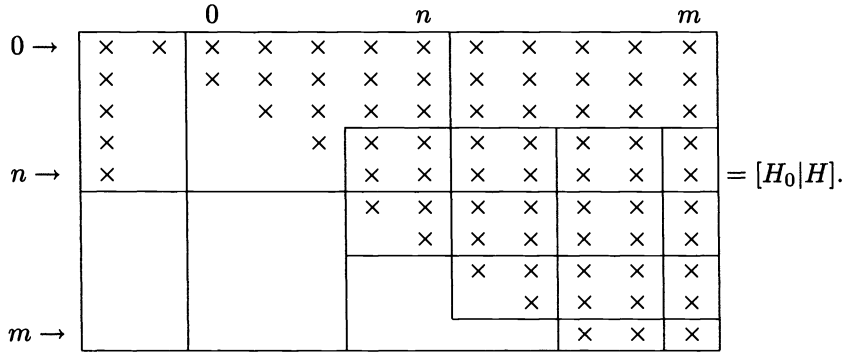
$$Q^H [\mathbf{w}|Z] \begin{bmatrix} I_2 & \\ & Q \end{bmatrix} = [H_0|H].$$

When $n = 0$, the transformations were aimed at chasing down the elements of $[\mathbf{w}|Z]$ below the main diagonal, making $[H_0|H]$ upper triangular. Therefore H turned out to be block upper Hessenberg.

When $\delta = 0$, the transformations had much the same objective, but now, there was no attempt to eliminate elements from the first column of $\mathbf{w}$, only elements from the second column were pushed to the SE part of the matrix. The matrix then turned out to be upper Hessenberg in a scalar sense.

The general case can be treated by an algorithm that combines both of these objectives. We start as in the polynomial case ($n = 0$), chasing only elements from the second column of $\mathbf{w}$. However, once we reach row $n+1$, we start eliminating elements in the first column, too.

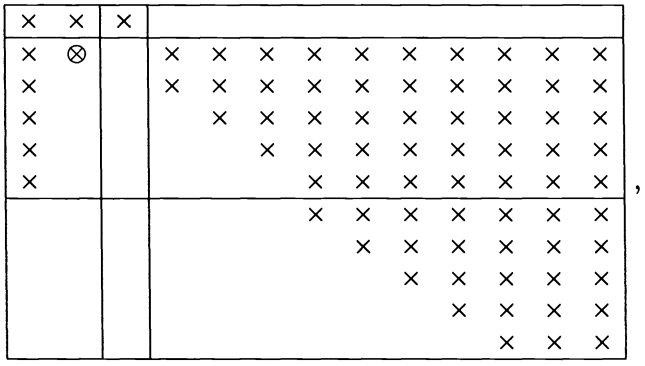
Applying this procedure shows that the extended Hessenberg $[H_0|H]$ has the form



$= [H_0|H].$

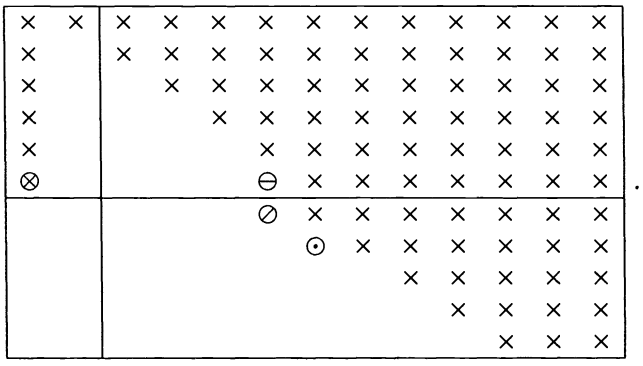
This means that the NW part of H, of size $(n+1) \times (n+1)$, will be scalar upper Hessenberg as in the case $n = 0$, while the SE part of size $(m-n+2) \times (m-n+2)$ has the block upper Hessenberg form of the case $\delta = 0$.

The updating procedure works as follows. Starting with (the new data are found in the first row)



,

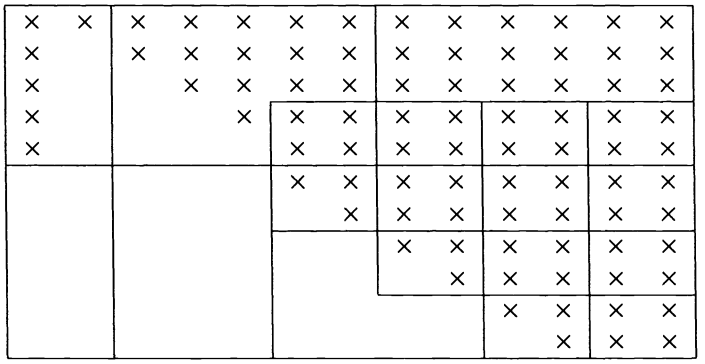
the element $\otimes$ is chased down the diagonal by elementary unitary similarity transformations operating on two successive rows/columns until we reach the following scheme (where $\odot = 0$ and $\oslash$ and $\ominus$ are the last elements introduced which are in general nonzero)



.

Now the element $\otimes$ in row $n+1$ is eliminated by a rotation/reflection in the plane of this row and the previous one. The corresponding transformation on the columns

will introduce a nonzero element at position $\odot$. Then $\oslash$ and $\odot$ are chased down the diagonal in the usual way until we reach the final situation



8.3. Orthogonal vector polynomials. The unitary matrix Q involved in the previous transformation was for the case $\delta = 0$ of the form $Q = W\Phi_m$, where W was a scalar diagonal matrix of the weights and Φ_m was the matrix with ij-element given by $\varphi_j(z_i)$, with φ_j the jth orthonormal polynomial.

When $n = 0$, then $Q = W\Phi_{m'}$, where W is the block diagonal with blocks being the 2×1 "weights" w_i and $\Phi_{m'}$ is the block matrix with 2×2 blocks, where the ij-block is given by $\varphi_j(z_i)$, with φ_j the jth block orthonormal polynomial.

For the general case, we have a mixture of both. For the NW part of the H matrix, we have the scalar situation and for the SE part we have the block situation.

To unify both situations, we turn to vector polynomials π_k of size 2×1. For the block part, we see a block polynomial φ_j as a collection of two columns and set

$$\varphi_j(z) = [\pi_{2j-1}(z)|\pi_{2j}(z)].$$

For the scalar part, we embed the scalar polynomial φ_j in a vector polynomial π_j by setting

$$\pi_j(z) = \begin{bmatrix} 0 \\ \varphi_j(z) \end{bmatrix}.$$

In both cases, the orthogonality of the φ_j translates into the orthogonality relation

$$\sum_{i=0}^{m} \pi_k(z_i)^H w_i^H w_i \pi_l(z_i) = \delta_{kl}$$

for the vector polynomials π_k. Let us apply this to the situation of the previous algorithm. For simplicity, we suppose that all $z_i \in \mathbb{R}$. For $z_i \in \mathbb{T}$, the situation is similar.

For column number $j = 0, 1, \ldots, n-1$, we are in the situation of scalar orthogonal polynomials: $Q_{ij} = w_{1i}\varphi_j(z_i) = w_i\pi_j(z_i)$. Setting

$$[H_0|H] = \left[\begin{array}{cc|cc|cc} \times & b_0 & a_0 & \bar{b}_1 & & \\ \vdots & & b_1 & a_1 & \ddots & \\ \times & & & \ddots & \ddots & \ddots \\ \hline & & & & \ddots & \ddots \end{array}\right],$$

we have for $j = 0, \ldots, n-2$, the three-term recurrence relation

$$z\varphi_j(z) = \varphi_{j-1}(z)\bar{b}_j + \varphi_j(z)a_j + \varphi_{j+1}(z)b_{j+1}, \qquad \varphi_{-1} = 0, \quad \varphi_0 = b_0^{-1}.$$

By embedding, this becomes

$$z\pi_j(z) = \pi_{j-1}(z)\bar{b}_j + \pi_j(z)a_j + \pi_{j+1}(z)b_{j+1}, \qquad \pi_0 = [0 \;\; \varphi_0]^T.$$

Thus, setting

$$\Pi_j = [\pi_j(z_0)^T, \ldots, \pi_j(z_m)^T]^T,$$

we have for the columns Q_j of Q the equality

$$Q_j = W\Pi_j, \qquad j = 0, 1, \ldots, n-1.$$

For the trailing part of Q, i.e., for columns $(n+2j-1, n+2j)$, $j = 0, 1, \ldots$, we are in the block polynomial case. The block polynomials $\varphi_j(z)$ group two vector polynomials

$$\varphi_j(z) = [\pi_{n+2j-1}(z)|\pi_{n+2j}(z)],$$

which correspond to two columns of Q, namely,

$$\mathbf{Q}_j = [Q_{n+2j-1}|Q_{n+2j}].$$

Observe that we have the following relation between $\mathbf{Q}_j$ and the block orthogonal polynomials

$$\mathbf{Q}_{ij} = w_i\varphi_j(z_i) = \begin{bmatrix} Q_{2i,n+2j-1} & Q_{2i,n+2j} \\ Q_{2i+1,n+2j-1} & Q_{2i+1,n+2j} \end{bmatrix},$$

where this time $w_i = [w_{0i} \;\; w_{1i}]$. As above, denote the vector of function values for π_j by Π_j. The block column of function values for φ_j is denoted by $\boldsymbol{\Phi}_j$. Then clearly

$$\mathbf{Q}_j = W\boldsymbol{\Phi}_j, \qquad \boldsymbol{\Phi}_j = [\Pi_{n+2j-1}|\Pi_{n+2j}].$$

Denoting in the extended Hessenberg matrix

$$[H_0|H] = \left[\begin{array}{cc|ccc} \times & b_0 & \ddots & \ddots & \\ \times & & \ddots & \ddots & B_0^T \\ 0 & & & B_0 & A_0 \quad B_1^T \\ 0 & & & & B_1 \quad A_1 \quad \ddots \\ \vdots & & & & \ddots \quad \ddots \end{array}\right], \qquad B_0 = \begin{bmatrix} 0 & b_{n-1} \\ 0 & 0 \end{bmatrix},$$

we have the block recurrence

$$z\varphi_j(z) = \varphi_{j-1}(z)B_j^T + \varphi_j(z)A_j + \varphi_{j+1}(z)B_{j+1}, \qquad j = 0, 1, \ldots.$$

The missing link between the scalar and the block part is the initial condition for this block recurrence. This is related to columns $n-2, n-1$ and n of Q. Because columns $n-2$ and $n-1$ are generated by the scalar recurrence, we know that these columns are $Q_j = W\Pi_j$, $j = n-2, n-1$, where the Π_j are related to the embedded scalar

polynomials. A problem appears in column Q_n where the three-term recurrence of the leading (scalar) part migrates to the block three-term recurrence of the trailing (block) part, i.e., from a three-term to a five-term scalar recurrence. We look at this column in greater detail. Because

$$\begin{bmatrix} d_0 \\ \vdots \\ d_n \\ \\ 0 \\ \vdots \\ 0 \end{bmatrix} = Q^H \begin{bmatrix} w_{00} \\ \vdots \\ \\ \vdots \\ \\ \vdots \\ w_{0m} \end{bmatrix} = Q^H W E, \qquad E = \left[\begin{array}{c} 1 \\ 0 \\ \hline 1 \\ 0 \\ \hline \vdots \\ \hline 1 \\ 0 \end{array}\right],$$

we have

$$Q_0 d_0 + \cdots + Q_n d_n = WE;$$

thus

$$\begin{aligned} Q_n &= \frac{1}{d_n}(WE - [Q_0|\ldots|Q_{n-1}]A_{n-1}^{\Phi}), \quad A_{n-1}^{\Phi} = [d_0, \ldots, d_{n-1}]^T \\ &= \frac{1}{d_n}(WE - W[\Pi_0|\ldots|\Pi_{n-1}]A_{n-1}^{\Phi}) \\ &= W\frac{1}{d_n}(E - [\Pi_0|\ldots|\Pi_{n-1}]A_{n-1}^{\Phi}) \\ &= W\frac{1}{d_n}(E - P_{n-1}), \quad P_{n-1} = [\Pi_0|\ldots|\Pi_{n-1}]A_{n-1}^{\Phi}. \end{aligned}$$

Setting $Q_n = W\Pi_n$, $\Pi_n = [\pi_n(z_0)^T, \ldots, \pi_n(z_m)^T]^T$, we find that

$$\pi_n(z) = \frac{1}{d_n}\begin{bmatrix} 1 \\ p_{n-1}(z) \end{bmatrix}, \tag{5}$$

where

$$p_{n-1}(z) = \varphi_0(z)d_0 + \cdots + \varphi_{n-1}(z)d_{n-1}$$

is the polynomial least squares approximant of degree $n-1$ for the data (z_i, w_i), $i = 0, \ldots, m$.

8.4. Solution of the general problem. Now we are ready to solve the general problem. We start with the degree structure of the polynomials $\pi_j(z)$. Suppose the jth column of Q is Q_j, which we write as

$$Q_j = W\Pi_j, \quad \Pi_j = [\pi_j(z_0)^T, \ldots, \pi_j(z_m)^T]^T$$

with $W = \operatorname{diag}(w_0, \ldots, w_m)$ and $\pi_j(z) = [\psi_j(z) \;\; \phi_j(z)]^T$. Then it follows from the previous analysis that the ϕ_j are the scalar orthogonal polynomials φ_j, and hence the degree of $\phi_j(z)$ is j, for $j = 0, 1, \ldots, n-1$. Moreover, the ψ_j are zero for the same indices (their degree is $-\infty$). For $j = n$, we just found that ψ_n is $1/d_n$, thus of degree 0 and ϕ_n is of degree at most $n-1$, since the latter is proportional to the polynomial

least squares approximant of that degree. With the block recurrence relation, we now easily find that the degree structure of the block polynomials

$$\varphi_j = [\pi_{n+2j-1}|\pi_{n+2j}] = \begin{bmatrix} \psi_{n+2j-1} & \psi_{n+2j} \\ \phi_{n+2j-1} & \phi_{n+2j} \end{bmatrix} \quad \text{is} \quad \begin{bmatrix} j-1 & j \\ n+j-1 & n+j-1 \end{bmatrix}$$

for $j = 1, 2, \ldots$, while φ_0 has degree structure

$$\begin{bmatrix} -\infty & 0 \\ n-1 & n-1 \end{bmatrix}.$$

It can be checked that in the regular case, that is when all the subdiagonal elements $b_0, \ldots, b_{n-1}$ as well as d_n are nonzero and when also all the subdiagonal blocks $B_1, \ldots, B'_m$ are regular (upper triangular), then the degrees of $\phi_k = \varphi_k$ are precisely k for $k = 0, 1, \ldots, n-1$ and in the block polynomials φ_j, the entries ψ_{n+2j} and ϕ_{n+2j-1} have the precise degrees that are indicated, i.e., j and $n+j-1$, respectively. Thus, if we propose a solution to our approximation problem of the form (suppose $m \geq n+2\delta$)

$$p(z) = \sum_{j=0}^{n+2\delta+1} \pi_j(z) a_j, \qquad a_j \in \mathbb{C},$$

then $p(z) = [p_0(z) \;\; p_1(z)]^T$ will automatically satisfy the degree restrictions $d_0 \leq \delta$ and $d_1 \leq n+\delta$. We must find

$$\min (A_n^{\Pi})^H \boldsymbol{\Pi}_{n'}^H W^H W \boldsymbol{\Pi}_{n'} (A_n^{\Pi}), \qquad n' = n + 2\delta + 1,$$

where

$$A_n^{\Pi} = [a_0, \ldots, a_{n'}]^T \quad \text{and} \quad \boldsymbol{\Pi}_{n'} = [\Pi_0 | \ldots | \Pi_{n'}].$$

Since $W\boldsymbol{\Pi}_{n'}$ form the first $n'+1$ columns of the unitary matrix Q, this reduces to

$$\min (A_{n'}^{\Pi})^H (A_{n'}^{\Pi}) = \min \sum_{j=0}^{n'} |a_j|^2.$$

If we require as before that $p_1(z)$ is monic of degree $n+\delta$, then $a_{n'} = 1/\beta_{n'}$ where β_j is the leading coefficient in ϕ_j. The remaining a_j are arbitrary. Hence, to minimize the error, we should make them all zero. Thus our solution is given by

$$p(z) = \pi_{n'}(z) a_{n'}, \quad n' = 2n + \delta + 1, \quad a_{n'} = 1/\beta_{n'}.$$

9. The singular case. Let us start by considering the singular case for $d_0 = d_1 = n$. We then generate a singular subdiagonal block η_{kk} of the Hessenberg matrix. The algorithm performing the unitary similarity transformations will not be harmed by this situation. However, the sequence of block orthogonal polynomials will break down. From the relation

$$z\varphi_{k-1}(z) = \varphi_0(z)\eta_{0k} + \cdots + \varphi_k(z)\eta_{kk},$$

it follows that if η_{kk} is singular, then this cannot be solved for $\varphi_k(z)$. In the regular case, all the η_{jj} are regular and then the leading coefficient of φ_k is $\eta_{00}^{-1} \ldots \eta_{kk}^{-1}$. Thus, if all the η_{jj} are regular upper triangular, then also the leading coefficient of φ_k will

be regular upper triangular. As we said in the Introduction, the singular situation will always occur, even in the scalar polynomial case with positive weights, but only at the very end where $k = m+1$. That is exactly the stage where we reach the situation where the least squares solution becomes the solution of an interpolation problem. We show below that this is precisely what also happens when some premature breakdown occurs.

Suppose that the *scalar* entries of the extended block Hessenberg matrix are $[H_0|H] = [h_{ij}]_{i,j=0,1,\ldots}$. (We use h_{ij} to distinguish them from the block entries η_{ij}.) Suppose that the element h_{kk} is the first element on its diagonal that becomes zero and thus produces some singular subdiagonal block in H. Then it is no problem to construct the successive scalar columns of the matrix $\Phi = \Phi_{m'}$ until the recurrence relation hits the zero entry h_{kk}. If we denote for $j = 0, 1, \ldots, k-1$, the jth column of Φ as Π_j, then we know from what we have seen, that Π_j represents the vector of function values at the nodes $z_0, \ldots, z_m$ of some vector polynomial $\pi_j(z) \in \mathbb{P}^{2\times 1}$. The problem in the singular case is that $\pi_k(z)$ cannot be solved from

$$z\pi_{k-2} = \pi_k h_{kk} + \pi_{k-1}h_{k-1,k} + \cdots + \pi_0 h_{0k}$$

because $h_{kk} = 0$. However, from

$$Q\left[\begin{array}{c|c} H_0 & H \end{array}\right] = \left[\begin{array}{cc|c} \mathbf{w}_0 & \mathbf{w}_1 & Z \end{array}\right]\left[\begin{array}{c|c} I_2 & 0 \\ 0 & Q \end{array}\right],$$

it follows that

$$\mathbf{w}_0 = Q_0 h_{00}; \qquad \mathbf{w}_1 = Q_0 h_{01} + Q_1 h_{11}$$

and for $k \geq 2$

$$ZQ_{k-2} = Q_0 h_{0k} + \cdots + Q_k h_{kk},$$

where Q_j, $j = 0, 1, \ldots$ denotes the jth column of Q. We shall discuss the case $h_{kk} = 0$ separately for $k = 0$, $k = 1$ and $k \geq 2$ separately.

If $h_{00} = 0$, then $\mathbf{w}_0 = 0$. This is a very unlikely situation because then there is only a trivial solution $(p_0, p_1) = (1, 0)$ which fits exactly.

Next consider $h_{11} = 0$; then define π_1' as

$$\pi_1' = \left[\begin{array}{c} 0 \\ 1 \end{array}\right] - \pi_0 h_{01}.$$

Then

$$\begin{aligned} W\Pi_1' &= (\mathbf{w}_1 - W\Pi_0 h_{01}) \\ &= (\mathbf{w}_1 - Q_0 h_{01}) \\ &= Q_1 h_{11} = 0. \end{aligned}$$

This means that we get an exact approximation since $w_i \pi_1'(z_i) = 0$, $i = 0, \ldots, m$.

For the general case $h_{kk} = 0$, $k \geq 2$, we have that

$$ZQ_{k-2} - Q_0 h_{0k} - \cdots - Q_{k-1}h_{k-1,k} = Q_k h_{kk} = 0.$$

Since $Q_j = W\Pi_j$ for $j = 0, \ldots, k-1$, we also have

$$\begin{aligned} 0 &= ZW\Pi_{k-2} - W\Pi_0 h_{0k} - \cdots - W\Pi_{k-1}h_{k-1,k} \\ &= W\Big(\mathbf{Z}\Pi_{k-2} - \Pi_0 h_{0k} - \cdots - \Pi_{k-1}h_{k-1,k}\Big), \end{aligned} \tag{6}$$

where $\mathbf{Z} = Z \otimes I_2$. Define the polynomial

$$\pi'_k(z) = z\pi_{k-2}(z) - \pi_0(z)h_{0k} - \cdots - \pi_{k-1}(z)h_{k-1,k}$$

then, $W\Pi'_k = W[\pi'_k(z_0)^T, \ldots, \pi'_k(z_m)^T]^T$ will be zero since it is equal to the expression (6), which is zero. This means that

$$w_i\pi'_k(z_i) = 0, \quad i = 0, \ldots, m.$$

The latter relations just tell us that this π'_k is an exact solution of the approximation problem, i.e., it interpolates.

In the general situation where $d_0 \neq d_1$, we must distinguish between the scalar and the block part. For the scalar part we can now also have a breakdown in the sequence of orthogonal polynomials since the weights are not positive anymore, but arbitrary complex numbers.

Using the notation

$$[H_0|H] = \left[\begin{array}{cc|ccc|c} \times & h_{00} & h_{01} & \cdots & h_{0,n+1} & \\ \vdots & & \ddots & \ddots & \vdots & \\ \times & & & h_{nn} & h_{n,n+1} & \ddots \\ \hline 0 & & & & \ddots & \ddots \end{array}\right]$$

for the NW part of the extended Hessenberg matrix, the situation is as sketched above: whenever some h_{kk} is zero, we will have an interpolating polynomial solution. It then holds that

$$\pi'_k(z) = z\pi_{k-1}(z) - \pi_0(z)h_{0k} - \cdots - \pi_{k-1}(z)h_{k-1,k}$$

and because $W\Pi'_k = W[\pi'^T_k(z_0), \ldots, \pi'^T_k(z_m)]^T$ is zero, we get

$$w_i\pi'_k(z_i) = 0, \quad i = 0, \ldots, m,$$

identifying $\pi'_k(z)$ as a (polynomial) interpolant.

For the SE part, i.e., for the block polynomial part, a zero on the subsubdiagonal (i.e., when we get a singular subdiagonal block in the Hessenberg matrix), will imply interpolation as we explained above for the block case.

The remaining problem is the case where the bottom element in the first column of the transformed extended Hessenberg matrix becomes zero. That is the element that has been previously denoted by d_n. Indeed, if this is zero, then our derivation, which gave (5)

$$\pi_n(z) = \frac{1}{d_n}\begin{bmatrix} 1 \\ p_{n-1}(z) \end{bmatrix}$$

does not hold anymore. But again, here we will have interpolation, i.e., a least squares error equal to zero. It follows from the derivation in the previous section that when $d_n = 0$,

$$W(E - P_{n-1}) = d_n Q_n = 0.$$

Thus

$$w_i\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix} - \sum_{k=0}^{n-1} \pi_k(z_i)d_k\right) = w_i\begin{bmatrix} 1 \\ p_{n-1}(z_i) \end{bmatrix} = 0,$$

where $p_{n-1}(z) = \sum_{k=0}^{n-1} \varphi_k(z) a_k^{\Phi}$. This is the same as

$$w_{0i} - w_{1i} p_{n-1}(z_i) = 0, \qquad i = 0, \ldots, m,$$

which means that $(1, p_{n-1}(z))/d'$ with $d' \neq 0$ to normalize $p_{n-1}(z)$ as a monic polynomial fits the data exactly.

10. Conclusion. We have shown that the inverse QR algorithm for solving discrete polynomial approximation problems for knots on the real line or on the unit circle can be generalized to more general approximation problems of the form (2).

In the previous section, we only considered the problem of updating, i.e., how to adapt the approximant when one knot is added to the set of data. There also exists a possibility to consider downdating, i.e., when one knot is removed from the set of interpolation points. For the polynomial approximation problem, this was discussed in [6] for real data and in [3] for data on the unit circle. The procedure can be based on a direct QR algorithm which will "diagonalize" the Hessenberg matrix in one row and column (e.g., the last one). This means that the only nonzero element in the last row and the last column of the transformed Hessenberg matrix is z_m on the diagonal. The unitary similarity transformations on the rest of the extended Hessenberg matrix brings out the corresponding weight in its first columns and the leading $m \times m$ part gives the solution for the downdated problem. Of course, just as the updating procedure can be generalized, the downdating procedure can also be adapted to our general situation. A combination of downdating and updating provides a tool for least squares approximation with a sliding window, i.e., where a window slides over the data, letting new data enter and simultaneously forgetting about the oldest data.

The inverse QR algorithm that we described in the previous sections is, in principle, applicable in the situation of arbitrary complex data. However its complexity can be reduced by an order of magnitude if the knots are real or located on the unit circle. The secret of this complexity reduction is the exploitation of a recurrence relation for the corresponding orthogonal polynomials and the parametrization of the Hessenberg matrix involved in terms of the recurrence coefficients.

The polynomial discrete least squares approximation problem discussed in the papers where the algorithm was first conceived gave rise to the construction of a sequence of polynomials orthogonal with respect to a discrete inner product. In the more general problem, these generalize to orthogonal block polynomials when all the degrees of the approximating polynomials are equal, or, in the more general case of arbitrary degrees, both scalar and block orthogonal polynomials appear that can be uniformly treated as vector orthogonal polynomials.

The algorithm has been reported to have excellent numerical stability properties [11], [12] and is preferred over the so-called Stieltjes procedure [12]. See also [7], [8]. Moreover it is well suited for implementation in a pipeline fashion on a parallel architecture [17], [15].

REFERENCES

[1] G. Ammar and W. Gragg, *$O(n^2)$ reduction algorithms for the construction of a band matrix from spectral data*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 426–431.

[2] G. Ammar, W. Gragg, and L. Reichel, *Constructing a unitary Hessenberg matrix from spectral data*, in Numerical Linear Algebra, Digital Signal Processing and Parallel Algorithms, G. Golub and P. Van Dooren, eds., Vol. 70, NATO-ASI Series, F: Computer and Systems Sciences, Springer-Verlag, Berlin, 1991, pp. 385–395.

[3] G. AMMAR, W. GRAGG, AND L. REICHEL, *Downdating of Szegö polynomials and data-fitting applications*, Linear Algebra Appl., 172 (1992), pp. 315–336.

[4] G. AMMAR AND C. HE, *On an inverse eigenvalue problem for unitary Hessenberg matrices*, Linear Algebra Appl., 1994, to appear.

[5] D. BOLEY AND G. GOLUB, *A survey of matrix inverse eigenvalue problems*, in Inverse Problems, Vol. 3, Physics Trust Publications, Bristol, England, 1987, pp. 595–622.

[6] S. ELHAY, G. GOLUB, AND J. KAUTSKY, *Updating and downdating of orthogonal polynomials with data fitting applications*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 327–353.

[7] W. GAUTSCHI, *On generating orthogonal polynomials*, SIAM J. Sci. Statist. Comput., 3 (1982), pp. 289–317.

[8] ———, *Computational problems and applications of orthogonal polynomials*, in Orthogonal Polynomials and Their Applications, C. Brezinski, L. Gori, and A. Ronveaux, eds., Vol. 9, IMACS Annals on Computing and Applied Mathematics, 1991, pp. 61–71.

[9] W. GRAGG, *The QR algorithm for unitary Hessenberg matrices*, J. Comput. Appl. Math., 16 (1986), pp. 1–8.

[10] W. GRAGG AND L. REICHEL, *A divide and conquer method for unitary orthogonal eigenproblems*, Numer. Math., 57 (1990), pp. 695–718.

[11] W. B. GRAGG AND W. J. HARROD, *The numerically stable reconstruction of Jacobi matrices from spectral data*, Numer. Math., 44 (1984), pp. 317–335.

[12] L. REICHEL, *Fast QR decomposition of Vandermonde-like matrices and polynomial least squares approximation*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 552–564.

[13] L. REICHEL, G. AMMAR, AND W. GRAGG, *Discrete least squares approximation by trigonometric polynomials*, Math. Comp., 57 (1991), pp. 273–289.

[14] H. RUTISHAUSER, *On Jacobi rotation patterns*, in Proceedings of Symposia in Applied Mathematics, Vol. 15, Experimental Arithmetic, High Speed Computing and Mathematics, Amer. Math. Soc., Providence, 1963, pp. 219–239.

[15] M. VAN BAREL AND A. BULTHEEL, *A parallel algorithm for discrete least squares rational approximation*, Numer. Math., 63 (1992), pp. 99–121.

[16] ———, *Discrete least squares approximation with polynomial vectors*, Tech. Report TW190, Department of Computer Science, Katholieke Universiteit Leuven, May 1993.

[17] ———, *Discrete linearized least squares rational approximation on the unit circle*, J. Comput. Appl. Math., 50 (1994), pp. 545–563.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 886–897, July 1995

012

ON A MATRIX GENERALIZATION OF AFFINE-SCALING VECTOR FIELDS*

LEONID FAYBUSOVICH†

Abstract. We construct a generalization of affine-scaling vector fields for matrix linear programming problems. We discuss various properties of these vector fields and suggest a generalization of a path-following algorithm that is due to C.Gonzaga [*SIAM Rev.*, 34 (1992), pp. 493–513].

Key words. linear programming, interior point methods, matrix problems

1. Introduction. Affine-scaling vector fields and the related geometric picture play a prominent role in the development of interior-point methods for solving various optimization problems [2]. For example, in so-called path-following algorithms the idea is to construct a finite-step approximation to the trajectory of this vector field. In this way the best known theoretical complexity estimates for polynomial-time algorithms were obtained. See, e.g., [11] and references therein. Affine-scaling vector fields arise naturally in connection with logarithmic barrier functions [2]. They possess an important property of being invariant under scaling transformations.

In the present paper we consider the problem of maximization of a linear function on a convex set determined by linear matrix inequalities:

$$\mathrm{Tr}(CK) \to \max, \tag{1.1}$$

$$\mathrm{Tr}(A_i K) = b_i, i = 1, \ldots, m, \tag{1.2}$$

$$K \in S(n), K \geq 0. \tag{1.3}$$

Here $C, A_i, i = 1, \ldots, m$ are symmetric matrices; $\mathrm{Tr}(A)$ stands for the trace of a matrix A and $K \geq 0$ means $\langle x, Kx \rangle = x^T K x \geq 0$ for any $x \in R^n$; $S(n)$ is the set of symmetric real n by n matrices; $b_i \in R, i = 1, \ldots m$. The optimization problem

$$f_\beta(K) = \beta \mathrm{Tr}(CK) + \ln(\det K) \to \max, \tag{1.4}$$

$$\mathrm{Tr}(A_i K) = b_i, i = 1, \ldots, m, \tag{1.5}$$

$$K \in S(n), K \geq 0, \tag{1.6}$$

where $\beta > 0$ is a scalar parameter, enables us to introduce a barrier function $\ln \det K$ and to choose a matrix generalization of affine-scaling vector fields. Namely, if $K(\beta)$ is a solution to (1.4)–(1.6), then it turns out that $K(\beta)$ as a function of β satisfies a system of differential equations that we call the generalized affine-scaling vector fields. The emerging geometric picture includes a Riemannian metric determined by the Hessian of the barrier function $\ln \det K$ and a group of linear isometries playing the role of scaling transformations for the matrix case. We have chosen a path-following

* Received by the editors July 9, 1993; accepted for publication (in revised form) by R. W. Cottle May 31, 1994.

† Department of Mathematics, University of Notre Dame, Mail Distribution Center, Notre Dame, Indiana, 46556-5683 (leonid.faybusovich@nd.edu).

algorithm that is due to Gonzaga [11] to illustrate the main point of this paper: the analogy between the scalar and the matrix case goes as far as the choice of the step size and resulting complexity estimates. The reader is invited to generalize major constructions and results of [11] for the matrix case using the technique developed in the present paper. A more challenging project is to construct the polynomial-time, path-following algorithms for the case of the entropy-type barrier function [8], [9]. The main problem here is that the corresponding group of isometries acts nonlinearly.

It should be pointed out that a different approach to the construction of path-following algorithms for (1.1)–(1.3) was suggested in [13].

The problem (1.1)–(1.3) has numerous applications in control theory [5]. The primal-dual and potential reduction algorithms for this problem were considered in [4], [1], and [13]. For a review of noninterior-point methods for solving (1.1)–(1.3), see [10].

2. Generalized affine-scaling vector fields. We start with some elementary properties of the function $\phi : K \to \ln \det K$.

LEMMA 2.1. *Let $K, \xi \in S(n), K$ be positive definite and $\xi \neq 0$. Then*

$$\mathrm{Tr}(K^{-1}\xi K^{-1}\xi) > 0.$$

Proof. Indeed,

$$\mathrm{Tr}(K^{-1}\xi K^{-1}\xi) = \mathrm{Tr}(K^{1/2}(K^{-1}\xi K^{-1}\xi)K^{-1/2})$$

$= \mathrm{Tr}([K^{-1/2}\xi K^{-1/2}]^2) \geq 0$. Moreover, the equality holds if and only if $K^{-1/2}\xi K^{-1/2} = 0$ or $\xi = 0$. □

PROPOSITION 2.2. *The function ϕ is strictly concave on the convex set of positive definite symmetric matrices.*

Proof. Indeed,

$$D\phi(K)(\xi) = \mathrm{Tr}(K^{-1}\xi), \tag{2.1}$$

$$D^2\phi(K)(\xi, \eta) = -\mathrm{Tr}(K^{-1}\xi K^{-1}\eta), \tag{2.2}$$

$\xi, \eta \in S(n), K > 0$. Here we use notations $D\phi(K), D^2\phi(K)$ for the first and second Frechet derivatives of the function ϕ at a point K. The result follows from Lemma 2.1. □

Suppose that the convex set P determined by constraints (1.2), (1.3) is compact. Let $T = \mathrm{span}(A_1, \ldots, A_m)$ (the set of all linear combinations of matrices $A_1, \ldots, A_m$) and $T^\perp = \{\xi \in S(n) : \mathrm{Tr}(\xi A_i) = 0, i = 1, \ldots, m\}$. It is clear that $S(n) = T \oplus T^\perp$. Let $\pi : S(n) \to T^\perp$ be the projection of $S(n)$ onto $T^\perp$ along T. Consider the map $\psi : \mathrm{int}(P) \to T^\perp$

$$\psi(K) = \pi(K^{-1}). \tag{2.3}$$

Here $\mathrm{int}(P)$ is the set of positive definite matrices in P. We denote $P \setminus \mathrm{int}(P)$ by ∂P. Observe that $\psi(K) = \pi(\nabla\phi(K))$ (see (2.1).)

LEMMA 2.3. *Suppose that K is a positive definite symmetric matrix and $A_1, \ldots, A_m$ are linearly independent symmetric matrices. Then the m by m matrix*

$$\Gamma(K) = (\mathrm{Tr}(A_i K A_j K)) \tag{2.4}$$

is positive definite.

Proof. It is sufficient to verify that $\xi^T\Gamma(K)\xi > 0$ for any nonzero $\xi \in R^m$. However, $\xi^T\Gamma(K)\xi = \mathrm{Tr}(\eta K \eta K)$, $\eta = \xi_1 A_1 + \cdots + \xi_m A_m$. The result follows by Lemma 2.1. □

THEOREM 2.4. *Suppose that the convex set P determined by* (1.2), (1.3) *is compact,* $\mathrm{int}(P) \neq \emptyset$ *and $A_1, \ldots, A_m$ are linearly independent. Then the map ψ is a diffeomorphism of* $\mathrm{int}(P)$ *onto $T^\perp$.*

Proof. Given $C \in T^\perp$, consider the following extremal problem:

$$f_1(K) = \mathrm{Tr}(CK) + \ln(\det K) \to \max, K \in P. \tag{2.5}$$

Since $\ln(\det K) \to -\infty$ when $K \to \partial P, P$ is compact and f is strictly concave, (2.5) has a unique solution $K(C) \in \mathrm{int}(P)$ for any $C \in T^\perp$. It is clear that $\nabla f_1(K(C)) \in T$. Hence $C + K(C)^{-1} \in T$ or $\pi(C) = -\pi(K(C)^{-1})$. This means that ψ is surjective. If $\psi(K_1) = \psi(K_2) = -C$, then both K_1 and K_2 are solutions to (2.5). We conclude that $K_1 = K_2$ because such a solution is unique. Hence, ψ is injective. It remains to prove that ψ^{-1} is smooth. We have for $\xi \in S(n), K \in \mathrm{int}(P)$:

$$D\psi(K)\xi = -\pi(K^{-1}\xi K^{-1}). \tag{2.6}$$

Suppose that $\mathrm{Tr}(A_i\xi) = 0, i = 1, \ldots, m$, and $D\psi(K)\xi = 0$. This yields

$$K^{-1}\xi K^{-1} = \sum_{i=1}^{m} \mu_i A_i$$

for some real μ_i and consequently

$$\sum_{i=1}^{m} \mu_i \mathrm{Tr}(A_s K A_i K) = 0, s = 1, \ldots, m.$$

Hence by Lemma 2.3 $\mu_i = 0, s = 1, \ldots, m$. In other words, $D\psi(K)$ is an injective and hence bijective map from $T^\perp$ to $T^\perp$. Thus ψ^{-1} is smooth by the implicit function theorem. □

REMARK 2.5. The map ψ is a generalization of a version of the Legendre transform considered in [2].

Let $S^+(n)$ be the set of positive definite symmetric matrices. It is clear that $S^+(n)$ is an open subset in $S(n)$. Consider a Riemannian metric g on $S^+(n)$ defined as follows:

$$g(K; \xi, \eta) = \mathrm{Tr}(K^{-1}\xi K^{-1}\eta). \tag{2.7}$$

Here $K \in S^+(n), \xi, \eta \in S(n)$. Consider the map $GL(n, R) \times S^+(n) \to S^+(n) : (S, K) \to SKS^T$. In this way we define a transitive action of $GL(n, R)$ on $S^+(n)$.

PROPOSITION 2.6. *The group $GL(n, R)$ acts on $S^+(n)$ by isometries.*

Proof. It is sufficient to verify that

$$g(SKS^T; S\xi S^T, S\eta S^T) = g(K; \xi, \eta) \tag{2.8}$$

for any $S \in GL(n, R), K \in S^+(n), \xi, \eta \in S(n)$. But

$$\mathrm{Tr}((SKS^T)^{-1}S\xi S^T(SKS^T)^{-1}(S\eta S^T)) = \mathrm{Tr}(K^{-1}\xi K^{-1}\eta),$$

which is equivalent to (2.8). □

REMARK 2.7. The action described above will play the role of scaling transformations for the matrix case.

It is clear that $\mathrm{int}(P)$ is a submanifold of $S^+(n)$ and hence a Riemannian submanifold. Suppose that f is a smooth function on $\mathrm{int}(P)$. Our next goal is to describe the gradient $\nabla_g f$ of f relative to the metric g.

PROPOSITION 2.8. *Let $A_1, \ldots, A_m$ be linearly independent. Then*

$$\nabla_g f(K) = K\nabla f(K) - \sum_{i=1}^{m} \mu_i(K, f)A_i))K, \tag{2.9}$$

$K \in \mathrm{int}(P)$. Here $Df(K)\xi = \mathrm{Tr}(\nabla f(K)\xi); \xi, \nabla f(K) \in S(n)$ and

$$\begin{bmatrix} \mu_1(K,f) \\ \vdots \\ \mu_m(K,f) \end{bmatrix} = \Gamma(K)^{-1} \begin{bmatrix} \mathrm{Tr}(A_1 K\nabla f(K)K) \\ \vdots \\ \mathrm{Tr}(A_m K\nabla f(K)K) \end{bmatrix}, \tag{2.10}$$

where $\Gamma(K)$ is defined in (2.4).

Proof. It is sufficient to verify that $\mathrm{Tr}(A_i\nabla_g f(K)) = 0, i = 1, \ldots, m$, and that $g(K; \nabla_g f(K), \xi) = \mathrm{Tr}(\nabla f(K)\xi)$ for any $\xi \in T^{\perp}$. We have

$$\mathrm{Tr}(A_i\nabla_g f(K)) = \mathrm{Tr}(A_i K\nabla f(K)K) - \sum_{j=1}^{m} \mu_j \mathrm{Tr}(A_i K A_j K) = 0$$

by (2.10). Furthermore,

$$g(K; \nabla_g f(K), \xi) = \mathrm{Tr}((\nabla f(K) - \sum_{i=1}^{m} \mu_i A_i)\xi) = \mathrm{Tr}(\nabla f(K)\xi),$$

since $\mathrm{Tr}(A_i\xi) = 0, i = 1, \ldots, m$. □

COROLLARY 2.9. *Let $f(K) = \mathrm{Tr}(KC), C \in S(n)$. Then*

$$D\psi(K)\nabla_g f(K) = -\pi(C). \tag{2.11}$$

In other words, all vector fields $\nabla_g f$ arising from linear functions f correspond to constant vector fields under the diffeomorphism ψ of $\mathrm{int}(P)$ onto $T^{\perp}$.

Proof. It follows from (2.6), (2.9), and that $\pi(A_i) = 0, i = 1, \ldots, m$. □

COROLLARY 2.10. *It holds that*

$$D(\psi)(K)\nabla_g f_\beta(K) = -\beta\pi(C) - \psi(K). \tag{2.12}$$

Proof. The proof is exactly the same as in the previous corollary. One should use the fact that $\nabla f_\beta(K) = \beta C + K^{-1}$. □

REMARK 2.11. Observe that, if $C, A_i, i = 1, \ldots, m$, are diagonal matrices, the vector field (2.9) has an invariant manifold consisting of positive definite diagonal matrices in P. The restriction of (2.9) to this manifold coincides with the standard affine-scaling vector field [2].

COROLLARY 2.12. *Let $f(K) = \mathrm{Tr}(CK), C \in S(n)$. Then* (2.9) *has no stationary points in* $\mathrm{int}(P)$, *provided* $\pi(C) \neq 0$.

Proof. This immediately follows by (2.11). □

Set

$$V_C(K) = \nabla_g f(K), \tag{2.13}$$

provided $f(K) = \mathrm{Tr}(CK)$.

COROLLARY 2.13. *For any C, C' vector fields $V_C, V_{C'}$ pairwise commute. In other words, the Lie bracket $[V_C, V_{C'}] = 0$.*

Proof. Since two constant vector fields pairwise commute, hence this follows from (2.11). □

Suppose that $\Gamma(K)^{-1}$ (see (2.4)) can be smoothly extended to ∂P. It is possible then using (2.10) to extend smoothly the vector fields $\nabla_g f$ to P. Suppose that this is the case. For any vector subspace $M \in R^n$ denote by $P(M)$ the set $\{K \in P : M \subset \mathrm{Ker}K\}$.

PROPOSITION 2.14. *The set $P(M)$ is an invariant submanifold for V_C.*

Proof. If $Kx = 0$, then by (2.9) $V_C(K)x = 0$. Hence $K(0) \in P(M)$ implies $K(t) \in P(M)$ for any t. □

EXAMPLE 2.15. Consider the following linear programming problem:

$$\mathrm{Tr}(KC) \to \max, \mathrm{Tr}(K) = 1, K \geq 0.$$

In this case $\Gamma(K) = \mathrm{Tr}(K^2)$. Hence

$$V_C(K) = KCK - \frac{\mathrm{Tr}(KCK)}{\mathrm{Tr}(K^2)} K^2.$$

It is clear that the maximal value of the cost function coincides with the maximal eigenvalue of C. Our results will show that the maximal eigenvalue of C can be obtained by a finite-step procedure with any given accuracy. For realistic eigenvalue algorithms using interior-point technique, see [14], [3], [12].

PROPOSITION 2.16. *Under the assumptions of Theorem* (2.4) *suppose that $K(\beta)$ is a solution to the problem* (1.4)–(1.6). *Then*

$$\frac{dK(\beta)}{d\beta} = V_C(K(\beta)), \tag{2.14}$$

$\lim K(\beta) = K_0, \beta \to 0$, *where K_0 is a solution to the problem* $\ln\det K \to \max, K \in P$.

REMARK 2.17. K_0 is a natural analogue of the analytic center of a polyhedron [15].

Proof. Since $K(\beta) \in \mathrm{int}(P)$, we should have: $(C + K(\beta)^{-1}/\beta) \in T$ or

$$\pi(C) = -\frac{\pi(K(\beta)^{-1})}{\beta}. \tag{2.15}$$

Multiplying by β and then differentiating yields

$$D\psi(K(\beta))\left(\frac{dK(\beta)}{d\beta}\right) = -\pi(C).$$

Comparing this with (2.11), (2.13), we arrive at (2.14). □

PROPOSITION 2.18. *Let K^* be a solution to the problem* (1.1)–(1.3). *Then*

$$\mathrm{Tr}(CK(\beta)) \leq \mathrm{Tr}(CK^*) \leq \mathrm{Tr}(CK(\beta)) + n/\beta.$$

In particular, $\lim \mathrm{Tr}(CK(\beta)) = \mathrm{Tr}(CK^*), \beta \to +\infty$.

Proof. By (2.15), we have

$$C = \sum_{i=1}^{m} w_i A_i - \frac{K(\beta)^{-1}}{\beta} \tag{2.16}$$

for some real w_i. For any $K \in P$ we have

$$\mathrm{Tr}(CK) = \sum_{i=1}^{m} w_i \mathrm{Tr}(A_i K) - \frac{\mathrm{Tr}(K(\beta)^{-1}K)}{\beta}$$

$$= \sum_{i=1}^{m} w_i b_i - \frac{\mathrm{Tr}(K(\beta)^{-1}K)}{\beta} \leq \sum_{i=1}^{m} w_i b_i.$$

Hence $\mathrm{Tr}(CK^*) \leq \sum_{i=1}^{m} w_i b_i$. On the other hand, by (2.16),

$$\mathrm{Tr}(CK(\beta)) = \sum_{i=1}^{m} w_i b_i - \frac{\mathrm{Tr}(K(\beta)K(\beta)^{-1})}{\beta}$$

or $\sum_{i=1}^{m} w_i b_i = \mathrm{Tr}(CK(\beta)) + n/\beta$. □

The best way to understand Proposition 2.18 is to consider the dual of the problem (1.1)–(1.3) (for a detailed relationship between the primal and the dual problem see [1]):

$$\sum_{i=1}^{m} b_i \nu_i \to \min, \tag{2.17}$$

$$\sum_{i=1}^{m} \nu_i A_i + Z = C, Z \leq 0. \tag{2.18}$$

PROPOSITION 2.19. *If* K *satisfies* (1.2), (1.3) *and* $(Z, \nu_1, \ldots, \nu_m)$ *satisfies* (2.18), *then*

$$\Delta(Z, K) = \sum_{i=1}^{m} b_i \nu_i - \mathrm{Tr}(CK) = -\mathrm{Tr}(ZK) \geq 0. \tag{2.19}$$

Proof. Indeed, $\Delta(Z, K) = \mathrm{Tr}((\sum_{i=1}^{m} \nu_i A_i - C)K) = -\mathrm{Tr}(ZK) = -\mathrm{Tr}(K^{1/2} Z K^{1/2})$ ≥ 0, since $Z \leq 0$. □

COROLLARY 2.20. *Suppose that* K^* *is a solution to* (1.1)–(1.3). *Under assumptions of Proposition* 2.19 *we have*

$$\mathrm{Tr}(CK^*) \leq \mathrm{Tr}(CK) + \Delta(Z, K).$$

Proposition 2.18 simply means that $Z = -\frac{K(\beta)^{-1}}{\beta}$ satisfies (2.18) with an appropriate choice of ν_i.

3. Gonzaga's path-following algorithm. Proposition 2.18 shows that the trajectory of the generalized affine-scaling vector field that starts at the analytic center (the central trajectory), converges to the optimal solution of the problem (1.1)–(1.3) (at least in the value of the cost function). Suppose we know that a point $K_0 \in \text{int}(P)$ near the analytic center. Since any point on the central trajectory is the solution to the corresponding problem (1.4)–(1.6), it is natural to try to move along the gradients of functions (1.4). One should try to choose a step size in such a way that the current iteration would remain close enough to the central trajectory. This is the main idea of short-step path-following algorithms. Each time the gradient is calculated relative to the metric g introduced earlier (Newton–Raphson's method). We denote the corresponding vector field by W_β. The most remarkable part of Gonzaga's algorithm and many other path-following algorithms is the chosen measure of proximity of a point $K \in \text{int}(P)$ to the point $K(\beta)$ on the central trajectory—the length of the gradient $W_\beta(K)$ in the metric g.

Since $W_\beta(K) = 0$ if and only if $K = K(\beta)$ (see Proposition 3.1 below), this choice of proximity certainly makes some sense. But the ultimate justification for this choice is due to concrete calculations (see below).

Consider the vector fields $W_\beta(C, A)(K), \beta > 0$, defined as follows:

$$W_\beta(C, A)(K) = \nabla_g f_\beta(K), \tag{3.1}$$

see (1.4), (2.9). Let $K \in \text{int}(P)$ and $\gamma(t)$ be the integral curve of the vector field (3.1) such that $\gamma(0) = K$.

PROPOSITION 3.1. *The only stationary point of W_β in $\text{int}(P)$ is $K(\beta)$. For any $K \in \text{int}(P)$, the corresponding solution $\gamma(t)$ is defined for all $t \in R$ and has the following property:*

$$\gamma(t) \to K(\beta), t \to +\infty. \tag{3.2}$$

Moreover,

$$\gamma(t) = \psi^{-1}(e^{-t}(\psi(K) + \beta\pi(C)) - \beta\pi(C)). \tag{3.3}$$

Proof. Since by (2.15) $\psi(K(\beta)) = -\beta\pi(C)$, (3.2) follows by (3.3). To prove (3.3) set $\xi(t) = \psi(\gamma(t))$. We have by (2.12)

$$\dot{\xi}(t) = D\psi(\gamma(t))W_\beta(\gamma(t)) = -\beta\pi(C) - \xi(t).$$

Hence, (3.3) follows. □

The next two propositions are very simple but important for the construction of the path-following algorithm.

PROPOSITION 3.2. *Let $S \in GL(n, R), K \in \text{int}(P)$. Then*

$$W_\beta(C, A)(SKS^T) = SW_\beta(\tilde{C}, \tilde{A})(K)S^T, \tag{3.4}$$

where $\tilde{C} = S^T C S, \tilde{A}_i = S^T A_i S$.

Proof. This immediately follows by (2.4),(2.9),(2.10). □

PROPOSITION 3.3. *It holds that*

$$W_\beta(C, A)(E) = \pi(\beta C + E) =$$

$$(3.5)\quad \arg\min\left\{\|X\|_F : X = \beta C + E - \sum_{i=1}^{m}\mu_i A_i \text{ for some } (\mu_1,\dots,\mu_m)\in R^m\right\}.$$

Here E is the n by n identity matrix and $\|X\|_F = \mathrm{Tr}(XX^T)^{1/2}$.

Proof. Indeed, the expression for $W_\beta(C,A)(E)$ follows from the fact that $g(E;\xi,\eta) = \mathrm{Tr}(\xi\eta)$ and standard properties of orthogonal projections onto vector subspaces. □

Given $\xi \in S(n)$, define

$$(3.6)\qquad l_K(\xi) = g(K;\xi,\xi)^{1/2}.$$

COROLLARY 3.4. *It holds that*

$$(3.7)\quad l_K(W_\beta(C,A)(K)) = \min\left\{\|\beta\tilde{C} + E - \sum_{i=1}^{m}\mu_i\tilde{A}_i\|_F : (\mu_1,\dots,\mu_m)\in R^m\right\},$$

$\tilde{C} = K^{1/2}CK^{1/2}, \tilde{A}_i = K^{1/2}A_iK^{1/2}$.

Proof. By (3.4) with $S = K^{1/2}$ we have

$$W_\beta(C,A)(K) = K^{1/2}W_\beta(\tilde{C},\tilde{A})(E)K^{1/2} = \xi.$$

Thus

$$l_K(W_\beta(C,A)(K_)^2 = g(K^{1/2}EK^{1/2};\xi,\xi) = g(E;W_\beta(\tilde{C},\tilde{A})(E),W_\beta(\tilde{C},\tilde{A})(E)).$$

Here we used (2.8). The result now follows by (3.5). □

It is convenient to introduce the notation

$$(3.8)\qquad \delta(K;\beta) = l_K(W_\beta(C,A)(K)).$$

The next proposition is well known.

PROPOSITION 3.5. *Let A be an n by n matrix and $x_1,\dots,x_n$ be an orthonormal basis in R^n. Then*

$$\mathrm{Tr}(A) = \sum_{i=1}^{n}\langle x_i, Ax_i\rangle.$$

Proof. Let O be an orthogonal matrix such that $x_i = Oe_i, i = 1,\dots n$, where $e_1,\dots,e_n$ is the standard basis in R^n. We have $\langle x_i, Ax_i\rangle = \langle e_i, O^{-1}AOe_i\rangle$. Hence,

$$\sum_{i=1}^{n}\langle x_i, Ax_i\rangle = \mathrm{Tr}(O^{-1}AO) = \mathrm{Tr}(A). \qquad \square$$

COROLLARY 3.6. *Let $x\in R^n, \|x\| = \sqrt{x^Tx} = 1$ and K be a nonnegative definite symmetric matrix. Then*

$$(3.9)\qquad \langle x, Kx\rangle \le \mathrm{Tr}(K).$$

Proof. There exist an orthonormal basis $x_1,\dots,x_n$ in R^n such that $x_1 = x$. The result follows by Proposition 3.5. □

PROPOSITION 3.7. *Suppose that* $K \in S^+(n), X \in S(n)$ *are such that*

$$g(K; X-K, X-K) \leq 1, \tag{3.10}$$

(*see* (2.7)). *Then* $X \geq 0$. *If* $g(K; X-K, X-K) < 1$, *then* $X \in S^+(n)$.

Proof. Suppose that X satisfies (3.10) but is not nonnegative definite. Then there exists $\lambda > 0, x \in R^n, x \neq 0$ such that $Xx = -\lambda x$. One can choose x in such a way that $\|K^{1/2}x\| = 1$. According to (2.7), $g(K; X-K, X-K) = \mathrm{Tr}(K^{-1}(X-K)K^{-1}(X-K)) = \mathrm{Tr}([K^{-1}(X-K)]^2) = \mathrm{Tr}(K^{1/2}(K^{-1}(X-K))^2K^{-1/2}) = \mathrm{Tr}(Y^2)$, where $Y = K^{-1/2}(X-K)K^{-1/2}$. Let $y = K^{1/2}x$. By Proposition 3.5 we have: $\langle Yy, Yy\rangle = \langle y, Y^2y\rangle \leq \mathrm{Tr}(Y^2)$. On the other hand, $\langle Yy, Yy\rangle = \langle K^{-1/2}(X-K)x, K^{-1/2}(X-K)x\rangle = \langle K^{-1/2}(-\lambda - K)x, K^{-1/2}(-\lambda - K)x\rangle = \lambda^2\langle x, K^{-1}x\rangle + \langle y, y\rangle + 2\lambda\langle x, x\rangle = 1 + \lambda^2\langle x, K^{-1}x\rangle + 2\lambda\|x\|^2 > 1$, a contradiction. Hence X is nonnegative definite. The same reasoning shows that $g(K; X-K, X-K) < 1$ implies $X \in S^+(n)$. □

Let $K \in \mathrm{int}(P)$. Suppose that $\delta(K, \beta) < 1$ (see (3.8)). Then by Proposition 3.7

$$K_1 = K + W_\beta(C, A)(K) \in \mathrm{int}(P).$$

PROPOSITION 3.8. *If* $\delta(K; \beta) < 1$, *then*

$$\delta(K_1; \beta) \leq \delta^2(K; \beta). \tag{3.11}$$

Proof. Consider, first, the case where $K = E$. We have

$$\delta(K, \beta) = \left\| \beta C + E - \sum_{i=1}^m \nu_i A_i \right\|_F$$

for some $\nu_i \in R$. By (3.7)

$$\delta(K_1; \beta) = \min\left\{ \left\| \beta\tilde{C} + E - \sum_{i=1}^m \mu_i \tilde{A}_i \right\|_F : (\mu_1 \dots \mu_m) \in R^m \right\} \leq \left\| \beta\tilde{C} + E - \sum_{i=1}^m \nu_i \tilde{A}_i \right\|_F$$

Here $\tilde{C} = K_1^{1/2}CK_1^{1/2}, \tilde{A}_i = K_1^{1/2}A_iK_1^{1/2}$. Now $K_1 - E = \beta C + E - \sum_{i=1}^m \nu_i A_i$. Using this to substitute for $\sum_{i=1}^m \nu_i \tilde{A}_i$ in the previous inequality, we obtain

$$\delta(K_1; \beta) \leq \|\beta\tilde{C} + E + K_1^2 - K_1 - \beta\tilde{C} - K_1\|_F = \|(K_1 - E)^2\|_F \leq \|K_1 - E\|_F^2 = \delta^2(K; \beta).$$

In general, by Proposition 3.2,

$$K_1 = K + W_\beta(C, A)(K) = K^{1/2}(E + W_\beta(\bar{C}, \bar{A})(E))K^{1/2},$$

where

$$\bar{C} = K^{1/2}CK^{1/2}, \qquad \bar{A} = K^{1/2}AK^{1/2}.$$

Set $E + W_\beta(\bar{C}, \bar{A})(E) = \tilde{K}_1$. We have by Proposition 2.6,

$$\delta(K_1; \beta) = l_{K_1}(W_\beta(C, A)(K_1)) = l_{\tilde{K}_1}(W_\beta(\bar{C}, \bar{A})(\tilde{K}_1)).$$

Since we have already proved our statement for $K = E$,

$$\delta(K_1; \beta) \leq \|W_\beta(\bar{C}, \bar{A})(E)\|_F^2.$$

Using again Proposition 2.6 and Proposition 3.2, we obtain

$$\begin{aligned}\|W_\beta(\bar{C},\bar{A})(E)\|_F &= l_K(K^{1/2}W_\beta(\bar{C},\bar{A})(E)K^{1/2}) \\ &= l_K(W_\beta(C,A)(K)) = \delta(K;\beta). \qquad \square\end{aligned}$$

PROPOSITION 3.9. *Let $K \in \text{int}(P)$ be such that $\delta(K;\beta) \leq 1$. If $W_\beta(K) = K(\beta C + K^{-1} - \Lambda)K, \Lambda \in \text{span}(A_1,\ldots,A_m)$, then $Z = C - \Lambda/\beta$ is a feasible solution to the dual problem* (2.17), (2.18) *with the duality gap $\Delta(Z,K)$ such that*

$$\Delta(Z,K) \leq \frac{n + \delta(K;\beta)\sqrt{n}}{\beta}. \tag{3.12}$$

Proof. Since $\delta(K;\beta) \leq 1, K - W_\beta(K) \in P$ by Proposition 3.7. Hence, $K - K(\beta C + K^{-1} - \Lambda)K \geq 0$ or $\Lambda - \beta C \geq 0$. Consequently, $Z \leq 0$. Since $Z + \Lambda/\beta = C, \Lambda \in \text{span}(A_1,\ldots,A_m)$, it is clear that Z satisfies (2.18).

Now by (2.19)

$$\begin{aligned}\Delta(Z,K) &= -\text{Tr}((C - \Lambda/\beta)K) \\ &= -\frac{\text{Tr}([\beta C - \Lambda + K^{-1}]K - E)}{\beta} = \frac{n - \text{Tr}(K^{-1/2}W_\beta(K)K^{-1/2})}{\beta} \\ &\leq \frac{n + \sqrt{n}[\text{Tr}(K^{-1/2}W_\beta(K)K^{-1}W_\beta(K)K^{-1/2})]^{1/2}}{\beta} \\ &= \frac{n + \delta(K;\beta)\sqrt{n}}{\beta}.\end{aligned}$$

Here we used that $|\text{Tr}(X)| \leq [\text{Tr}(E)]^{1/2}[\text{Tr}(X^2)]^{1/2}$. $\square$

THEOREM 3.10. *Let $K_0 \in \text{int}(P), \beta_0 > 0$ be such that $\delta(K_0;\beta_0) < 0.5$. Set*

$$\beta_{i+1} = \beta_i(1+\mu), \mu = \frac{1}{\sqrt{n}10},$$

$$K_{i+1} = K_i + W_{\beta_{i+1}}(K_i), \qquad i = 0,1\ldots.$$

Then $\delta(K_i,\beta_{i+1}) < 0.7$ and given $\epsilon > 0$

$$\text{Tr}(CK^*) - \text{Tr}(CK_i) < \epsilon$$

for

$$i > \frac{\ln(2n/\epsilon\beta_0)}{\ln(\mu+1)}. \tag{3.13}$$

Here K^* is an optimal solution to the problem (1.1)–(1.3).

Proof. Suppose that $K \in \text{int}(P), \delta(K;\beta) < 0.5$. By (3.7)

$$\delta(K;\beta) = \|\tilde{\pi}(\beta\tilde{C} + E)\|_F,$$

where $\tilde{C} = K^{1/2}CK^{1/2}, \tilde{A}_i = K^{1/2}A_iK^{1/2}$, and $\tilde{\pi} : S(n) \to \text{span}(\tilde{A}_1,\ldots,\tilde{A}_m)^\perp$ is the orthogonal projection. Hence,

$$\beta\|\tilde{\pi}(\tilde{C})\|_F \leq \delta(K;\beta) + \|E\|_F < 0.5 + \sqrt{n} \leq 1.5\sqrt{n}. \tag{3.14}$$

If $\beta' = (1+\mu)\beta$, then

$$\delta(K;\beta') \leq \|\tilde{\pi}(\beta\tilde{C}+E)\|_F + \mu\beta\|\tilde{\pi}(\tilde{C})\|_F < 0.5 + 1.5\sqrt{n}\mu,$$

where we used (3.14). If $\mu = 0.1/\sqrt{n}$, we conclude that $\delta(K,\beta') \leq 0.65 < 0.7$. In particular, by Proposition 3.8 $\delta(K+W_{\beta'}(K);\beta') \leq \delta^2(K,\beta)' < (0.7)^2 < 0.5$. Finally, by Proposition 3.9

$$\mathrm{Tr}(CK^*) - \mathrm{Tr}(CK) \leq \frac{n+\sqrt{n}\delta(K;\beta)}{\beta} \leq \frac{2n}{\beta}.$$

This completes the proof of the theorem since for $\beta = \beta_i = (1+\mu)^i\beta_0$ and i satisfying (3.13), we obviously have

$$\frac{2n}{\beta} < \epsilon. \qquad \square$$

Theorem 3.10 leads in a standard way to polynomial complexity estimates [11].

4. Concluding remarks. In this paper we introduced and studied generalized affine-scaling vector fields. We showed how to use these vector fields for the construction of a polynomial-time path-following algorithm for solving the semidefinite linear programming problem. We mention here that the generalization of Dikin's algorithm [6] is pretty straightforward. In this respect see also [7]. In general, it seems that a substantial part of the geometric structure underlying the interior-point algorithms can be carried over to the matrix case. The reader should be warned, however, that the boundary behavior of the trajectories is much more complicated in the matrix case. We hope to address this question later on. It is worthwhile to consider large-step path-following algorithms for the problem considered. Since the complexity of performing primal and dual iterations may be drastically different, the large-step path-following algorithms may be the right way to go for various classes of optimization problems described by linear matrix inequalities.

Acknowledgments. I would like to express my gratitude to R. Polyak and A. Tits for useful discussions. S. Boyd kindly provided his preprints before publication. A. Nemirovsky sent me a bibliography related to linear matrix inequalities. Thanks are also due to anonymous referees for very valuable remarks.

REFERENCES

[1] F. Alizadeh, *Optimization over the positive-definite cone:interior-point methods and combinatorial applications*, in Advances in Optimization and Parallel Computing, P. Pardalos, ed., North Holland, Amsterdam, 1992.

[2] D. A. Bayer and J. C. Lagarias, *The nonlinear geometry of linear programming.* I, Trans. Amer. Math. Soc., 314 (1989), pp. 499–526.

[3] S. Boyd, *Method of centers for minimizing generalized eigenvalues*, Linear Algebra Appl., 188, 189 (1993), pp. 63–111.

[4] S. Boyd and L. Vandenberghe, *Primal-dual potential reduction method for problems involving matrix inequalities*, Math. Programming, to appear.

[5] S. Boyd, L. Ghaoui, E. Feron and V. Balakrishnan, *Linear matrix inequalities in system and control theory*, Monograph, to appear.

[6] I. Dikin, *Iterative solution of problems of linear and quadratic programming*, Soviet. Math. Dokl., 8(1967), pp. 674–675.

[7] M. Chu and J. Wright, *A revisit of the educational testing problem*, preprint.

[8] L. Faybusovich, Hamiltonian structure of dynamical systems which solve linear programming problems, Physica, D53(1991), pp. 217–232.

[9] L. FAYBUSOVICH, *Dynamical systems which solve linear programming problems*, 1992 IEEE Conference on Decision and Control, pp. 1626–1631.

[10] R. FLETCHER, *Semi-definite constraints in optimization,* SIAM J. Control Optim., 23 (1985), pp. 493–513.

[11] C. GONZAGA,*Path-following methods for linear programming,* SIAM Rev., 34 (1992), pp. 167–224.

[12] F. JARRE, *Interior-point method for minimizing the maximum eigenvalue of a linear combination of matrices*, SIAM J. Control Optim., 31 (1993), pp. 1360–1376.

[13] YU. E. NESTEROV AND A. S. NEMIROVSKY, *Interior Point Polynomial Algorithms in Convex Programming*, Society for Industrial and Applied Mathematics, Philadelphia,1994.

[14] M. L. OVERTON AND R. S. WOMERSLEY, *Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices*, Math. Programming, 62 (1993), pp. 321–357.

[15] G. SONNEVEND, Applications of analytic centers, in Numerical Linear Algebra, Digital Signal Processing and Parallel Algorithms, Springer-Verlag, New York, Berlin, 1991.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 898–921, July 1995

013

TWO-DIMENSIONAL MINIMAL CUBATURE FORMULAS AND MATRIX EQUATIONS*

HANS JOACHIM SCHMID†

Abstract. For strictly positive, linear, and centrally symmetric functionals in two dimensions the existence of cubature formulas attaining the known lower bounds is equivalent to the solvability of certain matrix equations under some constraints. Any solution generates a real ideal the common roots of which are the nodes of the cubature formula. These results are applied to construct an infinite number of minimal positive cubature formulas of an arbitrary degree of exactness for one special, but classical, integral.

Key words. minimal cubature formulas, matrix equation, real ideal

AMS subject classification. 65D32

1. Introduction. The linear space of continuous functions in two variables defined on $\Omega \subset \mathbb{R}^2$ is denoted by $\mathcal{C}(\Omega)$. We consider functionals of the form

$$I : \mathcal{C}(\Omega) \to \mathbb{R} : f \mapsto I(f) = \int_\Omega f(x,y) d\mu(x,y), \qquad I(1) = 1,$$

where μ is a positive measure and Ω is a closed region such that I is strictly positive and centrally symmetric. Hence the following properties of I hold:

(linearity) $$I(\lambda_1 f_1 + \lambda_2 f_2) = \lambda_1 I(f_1) + \lambda_2 I(f_2),\ \lambda_1, \lambda_2 \in \mathbb{R}, f_1, f_2 \in \mathcal{C}(\Omega),$$

(strict positivity) $$I(f) > 0, \text{ whenever } f \geq 0 \text{ on } \Omega,\ 0 \neq f \in \mathcal{C}(\Omega),$$

(central symmetry) $$I(x^i y^j) = 0, \text{ if } i + j \text{ odd},\ i, j \in \mathbb{N}_0.$$

We want to approximate $I(f)$ by a convex combination of point-evaluations of f such that the approximation is exact for all $f \in \mathbb{P}_m$ where

$$\mathbb{P}_m = \text{span}\,\{1, x, y, \ldots, x^m, x^{m-1}y, \ldots, xy^{m-1}, y^m\}.$$

Note that $\dim \mathbb{P}_m = (m+1)(m+2)/2$. More precisely, we have the following definition.

DEFINITION 1.1. *The functional*

$$K(f) = K(m,N)(f) = \sum_{i=1}^{N} C_i f(x_i, y_i),\ C_i > 0,\ (x_i, y_i) \in \Omega,$$

is called cubature formula of degree m, if

(i) $I(f) = K(m,N)(f)$ *for all* $f \in \mathbb{P}_m$,

(ii) $I(f^\star) \neq K(m,N)(f^\star)$ *for at least one* $f^\star \in \mathbb{P}_{m+1}$.

The points (x_i, y_i) are called nodes, N is the number of nodes, and m the degree of exactness. In order to specify m and N we write $K(m,N)$.

DEFINITION 1.2. *Let $m \in \mathbb{N}$ be arbitrary but fixed. A cubature formula $K(m,N)$ is minimal, if the number of nodes N is minimal. Minimal formulas will be denoted by $K(m,\star)$.*

* Received by the editors July 22, 1993; accepted for publication (in revised form) by M. Gutknecht June 3, 1994.

† Mathematisches Institut, Universität Erlangen-Nürnberg, Bismarckstrasse 1.5, D-91054 Erlangen, Germany (schmid@mi.uni-erlangen.de).

We denote by $\mathbb{P} = \mathbb{R}[x, y]$ the ring of polynomials in two variables with real coefficients. We disregard $C(\Omega)$ in order to treat the problem by considering a strictly positive, linear, and centrally symmetric functional on $\mathbb{P}$,

$$I : \mathbb{P}(\Omega) \to \mathbb{R} : P \mapsto I(P) = \int_{\Omega} P(x, y) d\mu(x, y). \tag{1}$$

This functional will be approximated by

$$K(m, N) : \mathbb{P}_m \to \mathbb{R} : P \mapsto K(m, N)(P) = \sum_{i=1}^{N} C_i P(x_i, y_i), \tag{2}$$

such that I and $K(m, N)$ coincide on $\mathbb{P}_m$. If in this setting μ is a positive Radon measure, then, in fact, all strictly positive linear functionals on $C(\Omega)$ are represented by (1) (see [3]). It is of theoretical and practical importance that $K(m, N)$ is strictly positive on $\mathbb{P}_m$, i.e., $C_i > 0$ and $(x_i, y_i) \in \Omega$.

DEFINITION 1.3. *A cubature formula $K(m, N)$ is called interpolatory, if $N \leq \dim \mathbb{P}_m$ and if there are linearly independent polynomials $U_1, U_2, \ldots, U_N \in \mathbb{P}_m$ such that*

$$\det \begin{pmatrix} U_1(x_1, y_1) & U_1(x_2, y_2) & \ldots & U_1(x_N, y_N) \\ U_2(x_1, y_1) & U_2(x_2, y_2) & \ldots & U_2(x_N, y_N) \\ \vdots & \vdots & & \vdots \\ U_N(x_1, y_1) & U_N(x_2, y_2) & \ldots & U_N(x_N, y_N) \end{pmatrix} \neq 0. \tag{3}$$

Due to Tschakalov's Theorem [32] there exist formulas $K(m, N)$ with $N \leq \dim \mathbb{P}_m$. If such a formula is not interpolatory, by applying Steinitz's Austauschsatz an interpolatory cubature formula $K(m, N')$ can be constructed such that $N' < N$ (see [4]). Since we are interested in minimizing the number of nodes, it is appropriate to study such formulas.

If $K(m, N)$ is interpolatory we can determine a basis of $\mathbb{P}_m$ of the form

$$U_1, U_2, \ldots, U_N, Q_1, Q_2, \ldots, Q_t, \; t = \dim \mathbb{P}_m - N.$$

For $i = 1, 2, \ldots, t$ the linear system

$$\begin{pmatrix} U_1(x_1, y_1) & U_2(x_1, y_1) & \ldots & U_N(x_1, y_1) \\ U_1(x_2, y_2) & U_2(x_2, y_2) & \ldots & U_N(x_2, y_2) \\ \vdots & \vdots & & \vdots \\ U_1(x_N, y_N) & U_2(x_N, y_N) & \ldots & U_N(x_N, y_N) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{pmatrix} = \begin{pmatrix} Q_i(x_1, y_1) \\ Q_i(x_2, y_2) \\ \vdots \\ Q_i(x_N, y_N) \end{pmatrix} \tag{4}$$

can be solved such that

$$Q_i(x_j, y_j) = \sum_{l=1}^{N} a_l U_l(x_j, y_j), \qquad j = 1, 2, \ldots, N$$

holds. Hence there are $t = \dim \mathbb{P}_m - N$ linearly independent polynomials

$$R_i = Q_i - \sum_{i=1}^{N} a_i U_i, \; i = 1, 2, \ldots, t$$

vanishing at the nodes. Thus we have constructed a basis $U_1, U_2, \ldots, U_N, R_1, R_2, \ldots, R_t$ of $\mathbb{P}_m$ where all R_i vanish at the nodes of the formula, while the U_i do not vanish at all nodes. The R_i characterize the formula. If they are known, the polynomials U_i can be constructed and the linear system

$$\begin{pmatrix} U_1(x_1,y_1) & U_1(x_2,y_2) & \ldots & U_1(x_N,y_N) \\ U_2(x_1,y_1) & U_2(x_2,y_2) & \ldots & U_2(x_N,y_N) \\ \vdots & \vdots & & \vdots \\ U_N(x_1,y_1) & U_N(x_2,y_2) & \ldots & U_N(x_N,y_N) \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \\ \vdots \\ C_N \end{pmatrix} = \begin{pmatrix} I(U_1) \\ I(U_2) \\ \vdots \\ I(U_N) \end{pmatrix}$$

can be solved to determine the C_i. Due to the degree of exactness, each R_i satisfies $I(R_iQ) = 0$ whenever $R_iQ \in \mathbb{P}_m$.

DEFINITION 1.4. *A polynomial $R \in \mathbb{P}_m$ is called m-orthogonal (with respect to I), if $I(RQ) = 0$ whenever $RQ \in \mathbb{P}_m$.*

There are several books dealing extensively, but from different points of view, with cubature problems; see, e.g., [5], [9], [16], [30], [31]. In numerous papers cubature problems are studied. Some problems related to the approach presented here are attacked by completely different methods in [23]. For an overview we refer to these sources.

We follow the approach taken in [27]. The results will be enlarged and presented in a strict matrix notation. This makes the problem more transparent and allows a better understanding of the nonlinearity involved in the determination of cubature formulas.

2. Some special matrices. To treat cubature formulas in a compact matrix notation, several special matrices and their elementary properties are needed.

All matrices and vectors are real if not declared otherwise. Matrices are denoted by capital Latin or Greek letters. Indices are used in the following way: A_k is a $k+1 \times k+1$ or $k \times k+1$ matrix. Vectors are denoted by small Latin letters, indices are used as above: v_k is in $\mathbb{R}^{k+1}$.

The following matrices will be used quite often in the sequel,

$$F_k = \begin{pmatrix} 0 & \\ \vdots & E_{k-1} \\ 0 & \end{pmatrix}, \qquad L_k = \begin{pmatrix} & 0 \\ E_{k-1} & \vdots \\ & 0 \end{pmatrix} \in \mathbb{R}^{k \times k+1},$$

where $E_{k-1} \in \mathbb{R}^{k \times k}$ is the identity. We have chosen the letter F (L) since cancellation of the first (last) row of $A \in \mathbb{R}^{k+1 \times l}$ can be expressed by F_kA (L_kA), similarly, cancellation of the first (last) column of A by AF_k^t (AL_k^t).

Furthermore, we need the matrices

$$J_{k-1} = \begin{pmatrix} 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & \ldots & 1 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 1 & \ldots & 0 & 0 \\ 1 & 0 & \ldots & 0 & 0 \end{pmatrix}, \qquad T_{k-1} = \begin{pmatrix} 0 & 1 & \ldots & 0 & 0 \\ -1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & \ldots & -1 & 0 \end{pmatrix} \in \mathbb{R}^{k \times k},$$

and the diagonal matrix

$$D_k = \operatorname{diag}\{1, 2, 2, \ldots, 2, 2, 1\} \in \mathbb{R}^{k+1 \times k+1}.$$

LEMMA 2.1. *The following properties are evident.*

$$F_k F_k^t = L_k L_k^t = E_{k-1}, \tag{5}$$

$$F_k L_{k+1} = L_k F_{k+1}, \tag{6}$$

$$L_k^t F_k - F_k^t L_k = T_k \text{ and } F_k L_k^t - L_k F_k^t = T_{k-1}, \tag{7}$$

$$J_k J_k = E_k, \tag{8}$$

$$L_k J_k = J_{k-1} F_k, \tag{9}$$

$$J_k T_k J_k = -T_k. \tag{10}$$

$$D_k = L_k^t L_k + F_k^t F_k. \tag{11}$$

The matrices F_k, L_k can be used to characterize Hankel matrices.

LEMMA 2.2. *Let $A \in \mathbb{R}^{k \times l}$, $l = k, k+1, k+2$ be given. Then A is a Hankel matrix if and only if*

$$F_{k-1} A L_{l-1}^t = L_{k-1} A F_{l-1}^t. \tag{12}$$

If $A_k \in \mathbb{R}^{k \times k+1}$ is a Hankel matrix, then

$$L_k A_k^t = A_k L_k^t, \; F_k A_k^t = A_k F_k^t. \tag{13}$$

3. Orthogonal polynomials. In this section properties of orthogonal polynomials are discussed. An extensive treatment of this topic in n dimensions is given by M. A. Kowalski [11], [12]. The way of treating the recursion formulas in two dimensions has been proposed by G. Renner [22]; we put this into a strict matrix form. For further progress in multidimensional recursion formulas for strictly positive linear functionals, see Y. Xu [34].

Let us denote by

$$P_i^k = P_i^k(x,y) = x^{k-i} y^i + Q_i, \; Q_i \in \mathbb{P}_{k-1}, \qquad i = 0, 1, \ldots, k,$$

the set of orthogonal polynomials of degree k, normalized to a highest monomial term such that $I(P_i^k Q) = 0$, $i = 0, 1, \ldots, k$, for all $Q \in \mathbb{P}_{k-1}$. Let

$$p_k = (P_0^k, P_1^k, \ldots, P_k^k)^t \in (\mathbb{P}_k)^{k+1}, \qquad k = 1, 2, \ldots.$$

The orthogonality relation can be stated as

$$I(p_k p_i^t) = 0, \; i = 0, 1, \ldots, k-1.$$

We will denote by $\dot{\mathbb{P}}_k$ the linear space spanned by $P_0^k, P_1^k, \ldots, P_k^k$. Introducing $M_{ij}^k = M_{ji}^k = I(P_i^k P_j^k)$, $i, j = 0, 1, \ldots, k, k = 1, 2 \ldots$, for the moment matrix we obtain

$$M_k = I(p_k p_k^t) = \begin{pmatrix} M_{00}^k & M_{01}^k & \ldots & M_{0k}^k \\ M_{10}^k & M_{11}^k & \ldots & M_{1k}^k \\ \vdots & \vdots & & \vdots \\ M_{k0}^k & M_{k1}^k & \ldots & M_{kk}^k \end{pmatrix} \in \mathbb{R}^{k+1 \times k+1}, \qquad M_0 = 1.$$

Since I is strictly positive, M_k is positive definite, in particular $M_{ii}^k > 0$, $i = 0, 1, \ldots, k$. The matrices $M_0, M_1, \ldots$ and their inverses play a central role in the sequel and we assume that they are known for the functional I under consideration.

Since I is centrally symmetric an orthogonal polynomial of degree k is even (odd), if k is even (odd); i.e., its monomial terms are even (odd). This implies that the polynomials xP_i^k, yP_i^k are odd (even) and $(2k-1)$-orthogonal. Hence we can write

$$L_{k+1}p_{k+1} = xp_k - A_k^t p_{k-1}, \quad F_{k+1}p_{k+1} = yp_k - B_k^t p_{k-1}, \quad i = 0, 1, \ldots, k,$$

where $A_k, B_k \in \mathbb{R}^{k\times k+1}$ must be chosen such that the desired orthogonality relations hold. By multiplying this equation by p_{k+1}^t and by applying the functional I we obtain the relations

$$L_{k+1}M_{k+1} = I(xp_k p_{k+1}^t),\ F_{k+1}M_{k+1} = I(yp_k p_{k+1}^t)$$

and

$$M_{k+1}L_{k+1}^t = I(xp_{k+1}p_k^t),\ M_{k+1}F_{k+1}^t = I(yp_{k+1}p_k^t).$$

Multiplying the equation by p_{k-1}^t and applying I again, gives

$$I(xp_k p_{k-1}^t) = A_k^t M_{k-1},\ I(yp_k p_{k-1}^t) = B_k^t M_{k-1}.$$

Hence, $A_k^t = M_k L_k^t M_{k-1}^{-1}$, $B_k^t = M_k F_k^t M_{k-1}^{-1}$.

THEOREM 3.1. *The orthogonal polynomials with respect to I satisfy the following recursion formula. Starting with $p_0 = 1$, $p_1 = (x, y)^t$, we get for $k = 2, 3, \ldots$*

$$L_{k+1}p_{k+1} = xp_k - M_k L_k^t M_{k-1}^{-1} p_{k-1}, \qquad F_{k+1}p_{k+1} = yp_k - M_k F_k^t M_{k-1}^{-1} p_{k-1}. \tag{14}$$

Using (14) we can compute $I(xyp_k p_k^t)$ in two different ways. From this we obtain

$$L_{k+1}M_{k+1}F_{k+1}^t - F_{k+1}M_{k+1}L_{k+1}^t = -M_k(L_k^t M_{k-1}^{-1}F_k - F_k^t M_{k-1}^{-1}L_k)M_k. \tag{15}$$

The matrix

$$M_k^\star = L_{k+1}M_{k+1}F_{k+1}^t - F_{k+1}M_{k+1}L_{k+1}^t \tag{16}$$

is fundamental for cubature problems. J. Radon [18] introduced it to study the existence of cubature formulas of degree 5 with seven nodes. I. P. Mysovskikh (see [16]) solved Radon's problem by considering an associated similar matrix. H. M. Möller [14] discovered that the rank of $M_{k-1}^\star$ is involved in the lower bound for formulas of degree $2k-1$ and derived an improved lower bound (Theorem 4.2). We will see in the sequel that this matrix appears in the matrix equations characterizing cubature formulas as well. In the following we prefer writing a polynomial $P \in \mathring{\mathbb{P}}_k$ as

$$P = \sum_{i=0}^{k} \alpha_i P_i^k = a_k^t p_k.$$

If the recursion formula for p_k will be applied, P will be rewritten as

$$P = b_k^t(F_k^t F_k + L_k^t L_k)p_k = b_k^t D_k p_k,\ b_k = D_k^{-1}a_k.$$

4. Lower bounds. For strictly positive linear functionals we obtain the following theorem.

THEOREM 4.1. *A formula $K(m,N)$ satisfies $N \geq \dim \mathbb{P}_{[m/2]}$.*

Proof. Assuming $N' < \dim \mathbb{P}_{[m/2]}$ we can find a polynomial $Q \in \mathbb{P}_{[m/2]}$ vanishing at the nodes (x_i, y_i), $i = 1, 2, \ldots, N'$. Since $Q^2 \in \mathbb{P}_m$ we obtain $I(Q^2) = K(m, N')(Q^2) = 0$ in contradiction to the strict positivity. □

Necessary and sufficient conditions for functionals such that this bound will be attained have been studied for odd m by J. Radon [18] and I. P. Mysovskikh [17]. In particular, if the bound will be attained, then the polynomial vector p_k vanishes at the nodes of the formula. For m even, the following necessary condition holds.

COROLLARY 4.1. *If the bound in Theorem 4.1 is attained by $K(2k-2, \star)$, then there is a matrix $0 \neq \Gamma_k^t \in \mathbb{R}^{k \times k+1}$ such that the polynomials in*

$$r_k = p_k + \Gamma_k p_{k-1} \in (\mathbb{P}_k)^{k+1}.$$

vanish at all nodes of the formula.

Proof. If $K(2k-2, N)$, $N = \dim \mathbb{P}_{k-1}$ exists then no polynomial in $\mathbb{P}_{k-1}$ vanishes at the nodes of the formula due to Theorem 4.1. Denoting a basis of $\mathbb{P}_{k-1}$ by $U_1, U_2, \ldots, U_N$, we can determine, by use of (4), $t = \dim \mathbb{P}_{2k-2} - N$ linearly independent polynomials in $\mathbb{P}_{2k-2}$ vanishing at all nodes. In particular, we can find a polynomial vector $p_k + \Gamma_k p_{k-1}$, $\Gamma_k^t \in \mathbb{R}^{k \times k+1}$, such that all polynomial entries vanish at the nodes of the formula. □

This and Mysovskikh's characterization allows the construction of a big class of functionals such that the lower bound of Theorem 4.1 will be attained for an arbitrary degree of exactness (see [29]). However, these functionals admitting a direct generalization of Gaussian quadrature, are neither classical nor centrally symmetric.

For the functionals in question and m small, the bound will be attained for m even, while it is too pessimistic for m odd. So the lower bounds for centrally symmetric functionals differ in the odd and even case. An improved lower bound for $m = 2k-1$ has been given by H. M. Möller [14]. We will derive this bound below. For linearly independent polynomials $Q_i^k \in \mathbb{P}_k$, $i = 1, 2, \ldots, s$, we consider

$$\mathcal{W} = \operatorname{span} \{Q_i^k, xQ_i^k, yQ_i^k \; i = 1, 2, \ldots, s\}.$$

Obviously $\dim \mathcal{W} \leq s + k + 2$. The exact dimension can be determined, if all linear dependencies of the form

$$\sum_{i=1}^{s} c_i Q_i^k = \sum_{i=1}^{s} a_i x Q_i^k - b_i y Q_i^k, \; a_i, b_i, c_i \in \mathbb{R}.$$

are known. To get a lower bound for $\dim \mathcal{W}$, Möller investigated equations of the form

$$x(F_k p_k)^t a_{k-1} - y(L_k p_k)^t b_{k-1} = p_k^t c_k, \; a_{k-1}, b_{k-1} \in \mathbb{R}^k, c_k \in \mathbb{R}^{k+1}.$$

Since the P_i^k are even (odd) if k is even (odd), the polynomials on the left-hand side are odd (even), while the polynomial on the right-hand side is even (odd). So we can assume c to be equal to 0, and the equation reduces to

$$x p_k^t F_k^t a_{k-1} = y p_k^t L_k^t b_{k-1}.$$

Multiplying this by p_{k+1} and p_{k-1}, respectively, and applying the functional I, we obtain

$$\begin{aligned} I(xp_{k+1}p_k^t)F_k^t a_{k-1} &= I(yp_{k+1}p_k^t)L_k^t b_{k-1}, \\ M_{k+1}L_{k+1}^t F_k^t a_{k-1} &= M_{k+1}F_{k+1}^t L_k^t b_{k-1}, \\ L_k F_{k+1} L_{k+1}^t F_k^t a_{k-1} &= b_{k-1}, \\ F_k L_{k+1} L_{k+1}^t F_k^t a_{k-1} &= b_{k-1}, \\ a_{k-1} &= b_{k-1} \end{aligned}$$

and

$$\begin{aligned} I(xp_{k-1}p_k^t)F_k^t a_{k-1} &= I(yp_{k-1}p_k^t)L_k^t a_{k-1}, \\ L_k M_k F_k^t a_{k-1} &= F_k M_k L_k^t a_{k-1}. \end{aligned}$$

Thus the number of linear dependencies in the given basis of W can be estimated by the number of linearly independent solutions of

$$(L_k M_k F_k^t - F_k M_k L_k^t)a_{k-1} = M_{k-1}^\star a_{k-1} = 0.$$

Note that $M_{k-1}^\star$ is skew. To determine the rank of $M_{k-1}^\star$, we follow G. Renner [22]. If there is an $a_{k-1} \in \mathbb{R}^k$, $a_{k-1} \neq 0$, such that $M_{k-1}^\star a_{k-1} = 0$ then by (15)

$$(L_{k-1}^t M_{k-2}^{-1} F_{k-1} - F_{k-1}^t M_{k-2}^{-1} L_{k-1}) M_{k-1} a_{k-1} = 0. \tag{17}$$

Setting $b_{k-1} = M_{k-1}a_{k-1}$, we find

$$M_{k-2}^{-1} F_{k-1} b_{k-1} = c_{k-2} \neq 0, \qquad M_{k-2}^{-1} L_{k-1} b_{k-1} = d_{k-2} \neq 0.$$

The latter equation holds since, due to (17), the equation $M_{k-2}^{-1}F_{k-1}b_{k-1} = 0$ implies $M_{k-2}^{-1}L_{k-1}b_{k-1} = 0$; i.e., $b_{k-1} = 0$. Thus, (17) implies $L_{k-1}^t c_{k-2} = F_{k-1}^t d_{k-2}$; i.e.,

$$\begin{pmatrix} c_{k-2} \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ d_{k-2} \end{pmatrix} = \begin{pmatrix} 0 \\ b_{k-3} \\ 0 \end{pmatrix}, \quad b_{k-3} \in \mathbb{R}^{k-2}.$$

Hence, we obtain

$$F_{k-1} b_{k-1} = M_{k-2} c_{k-2} = M_{k-2} F_{k-2}^t b_{k-3} \tag{18}$$

and

$$L_{k-1} b_{k-1} = M_{k-2} d_{k-2} = M_{k-2} L_{k-2}^t b_{k-3}. \tag{19}$$

Since $L_{k-2}F_{k-1}b_{k-1} = F_{k-2}L_{k-1}b_{k-1}$,

$$(L_{k-2}M_{k-2}F_{k-2}^t - F_{k-2}M_{k-2}L_{k-2}^t)b_{k-3} = 0,$$

so $M_{k-3}^\star b_{k-3} = 0$. From (18) and (19) we get

$$F_{k-1}^t F_{k-1} b_{k-1} = F_{k-1}^t M_{k-2} F_{k-2}^t b_{k-3}, \qquad L_{k-1}^t L_{k-1} b_{k-1} = L_{k-1}^t M_{k-2} L_{k-2}^t b_{k-3},$$

hence

$$D_{k-1} M_{k-1} a_{k-1} = (F_{k-1}^t M_{k-2} F_{k-2}^t + L_{k-1}^t M_{k-2} L_{k-2}^t) b_{k-3}.$$

For any linearly independent solution a_{k-1}, there will be a linearly independent solution b_{k-3}. Hence Renner's inductive argument shows that it is sufficient to study

$$\text{rank } M_0^\star = \text{rank } (L_1 M_1 F_1^t - F_1 M_1 L_1^t) = 0$$

and

$$\text{rank } M_1^\star = \text{rank } (L_1^t M_0^{-1} F_1 - F_1^t M_0^{-1} L_1) = \text{rank } (L_1^t F_1 - F_1^t L_1 = T_1) = 2.$$

We have thus proved the following lemma.

LEMMA 4.1.

$$\text{rank } M_{k-1}^\star = \text{rank } (L_k M_k F_k^t - F_k M_k L_k^t) = \begin{cases} k-1, & \text{if } k \text{ is odd,} \\ k, & \text{if } k \text{ is even.} \end{cases}$$

THEOREM 4.2 ([14]). *A formula of type $K(2k-1, N)$ satisfies $N \geq \dim \mathbb{P}_{k-1} + [k/2]$. If this bound is attained for odd k, then $a_{k-1}^t(F_k + L_k)p_k$ belongs to the ideal associated with the cubature formula, where a_{k-1} is determined by $M_{k-1}^\star a_{k-1} = 0$.*

Proof. By Theorem 4.1 we know that no polynomial from $\mathbb{P}_{k-1}$ vanishes at all nodes of the formula. If, in addition, no polynomial in $\mathbb{P}_k$ vanishes at all nodes, then $N \geq \dim \mathbb{P}_{k+1} > \dim \mathbb{P}_k + [k/2]$. Let us denote by Q_i^k a maximal linearly independent set of polynomials in $\mathbb{P}_k$ vanishing at all nodes. Hence all polynomials in $W = \text{span } \{Q_i^k, xQ_i^k, yQ_i^k\}$ vanish at the nodes of the formula. As we have shown above, there might be only one linear dependency, if k is odd; thus we get from

$$3s - (k - 2[k/2]) \leq \dim W \leq s + k + 2, \tag{20}$$

$s \leq k - [k/2] + 1$. So there are $k + 1 - k + [k/2] - 1 = [k/2]$ linearly independent polynomials in $\mathbb{P}_k$ that do not vanish at all nodes. If k is odd there is an a_{k-1} such that $M_{k-1}^\star a_{k-1} = 0$. As we have seen, the polynomials $a_{k-1}^t F_k p_k$ and $a_{k-1}^t L_k p_k$ belong to the ideal. □

COROLLARY 4.2. *If the bound in Theorem 4.2 is attained by $K(2k-1, \star)$, then there are polynomials xQ_i^k, yQ_i^k, which vanish at all nodes of the formula, forming a set of $k+2$ linearly independent polynomials of degree $k+1$. This set can be assumed to be of the form*

$$p_{k+1} + \Gamma p_{k-1},\ \Gamma^t \in \mathbb{R}^{k \times k+2}.$$

Furthermore, if k is odd, two polynomials of degree k vanishing at all nodes are known.

Proof. If the bound in Theorem 4.2 is attained there are $s = k+1-[k/2]$ linearly independent polynomials Q_i, $i = 1, 2, \ldots, s$ in $\mathbb{P}_k$ vanishing at the nodes. Thus the estimate in (20) holds. Hence xQ_i, yQ_i, $i = 1, 2, \ldots, s$ form $k+2$ linearly independent polynomials of degree $k+1$ vanishing at the nodes. By the $(2k-1)$-orthogonality, we can transform these polynomials to $p_{k+1} + \Gamma p_{k-1}$, $\Gamma^t \in \mathbb{R}^{k \times k+2}$, with suitably chosen Γ. If k is odd, there are $k+3$ polynomials xQ_i, yQ_i, hence there is a linear dependency of the form $xQ_1^\star = yQ_2^\star$. Thus $Q_1^\star, Q_2^\star$ vanish at the nodes as well, so, two polynomials of degree k vanishing at the nodes are known. □

The corollaries of this section motivate the following

DEFINITION 4.1. *A set of polynomials $\mathcal{S}$ is called fundamental of degree i whenever $i+1$ linearly independent polynomials of the form $x^{i-j}y^j + S_j$, $S_j \in \mathbb{P}_{i-1}$, $j = 0, 1, \ldots, i$, belong to* span $\mathcal{S}$.

COROLLARY 4.3. *If $K(2k-2, \star)$ attains the bound in Theorem 4.1 there exists a fundamental system of degree k vanishing at the nodes of the formula.*

If $K(2k-1, \star)$ attains the bound in Theorem 4.2 there exists a fundamental system of degree $k+1$ vanishing at the nodes of the formula.

5. Cubature formulas and real ideals. In this section we briefly discuss the connection between cubature formulas and ideals. Ideas in algebraic geometry in connection with cubature formulas were introduced by J. Radon [18] and refined in several papers of I. P. Mysovkikh (see [16]). H. M. Möller [13] presented an approach to construct cubature formulas that was completely based on polynomial ideals. His ideas can be applied by restricting the interest to real ideals which turn out to characterize interpolatory cubature formulas [27]. We will summarize the main results of the theory.

A set $\mathcal{A}$ of polynomials in $\mathbb{P}$ is called an ideal if $R_1Q_1 + R_2Q_2 \in \mathcal{A}$ whenever $Q_1, Q_2 \in \mathcal{A}$, $R_1, R_2 \in \mathbb{P}$. The polynomials $Q_1, Q_2, \ldots, Q_s$ form a basis of $\mathcal{A}$ if each $Q \in \mathcal{A}$ can be written as

$$Q = \sum_{i=1}^{s} R_i Q_i, \ R_i \in \mathbb{P}.$$

Ideals generated by $Q_1, Q_2, \ldots, Q_s$ are denoted by $(Q_1, Q_2, \ldots, Q_s)$, the zero-ideal by (0). For a given polynomial vector $r_n \in (\mathbb{P}_n)^{n+1}$, we use (r_n) for the ideal generated by the polynomial entries.

Let var $(\mathcal{A})$ be the real zero-set of an ideal,

$$\text{var } (\mathcal{A}) = \{(x, y) \in \mathbb{R}^2 \ : \ Q(x, y) = 0 \text{ for all } Q \in \mathcal{A}\}.$$

If $\mathcal{N} \subseteq \mathbb{R}^2$ is a set of points, we denote by $\mathcal{A}(\mathcal{N})$ the ideal of all polynomials vanishing at $\mathcal{N}$; i.e.,

$$\mathcal{A}(\mathcal{N}) = \{Q \in \mathbb{P} \ : \ Q(x, y) = 0 \text{ for all } (x, y) \in \mathcal{N}\}.$$

DEFINITION 5.1. *Let $K(m, N)$ be given, where $\mathcal{N}$ denotes the set of nodes. Then $\mathcal{A}_K = \mathcal{A}(\mathcal{N})$ is the ideal associated to $K(m, N)$.*

We have proved the following in §1.

LEMMA 5.1. *Let $K(m, N)$ be an interpolatory cubature formula with a set of nodes $\mathcal{N}$, then there are $s = \dim \mathbb{P}_m - N$ linearly independent polynomials $R_1, R_2, \ldots, R_s$ of degree $\leq m$ which belong to $\mathcal{A}(\mathcal{N})$.*

Introducing real ideals, it can be shown that the polynomials $R_1, R_2, \ldots, R_s$ generate the ideal $\mathcal{A}(\mathcal{N})$.

An ideal $\mathcal{A}$ is called real if all polynomials vanishing at var $(\mathcal{A})$ belong to $\mathcal{A}$, i.e., $\mathcal{A}(\text{var } (\mathcal{A})) = \mathcal{A}$.

THEOREM 5.1. *An ideal $\mathcal{A}$ is real if and only if for all $M \in \mathbb{N}$ and all $R_i \in \mathbb{P}$ $i = 1, 2, \ldots, M$*

$$\sum_{i=1}^{M} R_i^2 \in \mathcal{A} \text{ implies } R_i \in \mathcal{A}, \ i = 1, 2, \ldots, M.$$

For proofs and further details see [7], [8], [24], [25]. We will generalize a theorem of I. P. Mysovskikh [17]; see [27].

THEOREM 5.2. *Let $R_1, R_2, \ldots, R_t$ be linearly independent polynomials in $\mathbb{P}_k$ which are fundamental of degree k, which span the linear space $\mathcal{R}$, and which generate the ideal $\mathcal{A}$. Let $\mathcal{U}$ be an arbitrary but fixed complement of $\mathcal{R}$ in $\mathbb{P}_k$. Then*

$$|\text{var } (\mathcal{A})| \leq \dim \mathbb{P}_k - t = \dim \mathcal{U}.$$

Furthermore, if $\mathcal{R} = \mathcal{A} \cap \mathbb{P}_k$, *then* $\mathcal{A}$ *is real if and only if* $|\text{var}\ (\mathcal{A})| = \dim \mathbb{P}_k - t$.

Based on Theorems 5.1 and 5.2 the following characterization can be obtained; see [27].

THEOREM 5.3. *Let* $r_{s-1} \in (\mathbb{P}_{m+1})^s$ *be fundamental of degree* k, $[m/2]+1 \leq k \leq m+1$, *and* m*-orthogonal. Let* $\mathcal{A} = (r_{s-1}^t)$, $\mathcal{R} = \text{span}\ \{r_{s-1}^t\}$, *and let* $\mathcal{U}$ *be an arbitrary but fixed complement of* $\mathcal{R}$ *in* $\mathbb{P}_{m+1}$.

Then the following conditions are equivalent.

(i) $\mathcal{A}$ *is an ideal associated with a cubature formula* $K(m, N)$, $N = \dim \mathbb{P}_{m+1} - s = \dim \mathcal{U}$.

(ii) $\mathcal{A} \cap \mathcal{U} = (0)$ *and for all* $0 \neq U \in \mathcal{U}$ *the condition* $I(U^2 - R^\star) > 0$ *holds, whenever* $R^\star \in \mathcal{A}$ *can be chosen such that* $U^2 - R^\star \in \mathbb{P}_m$.

If these conditions are satisfied, then $\mathcal{A}$ *is real.*

Simple cases will be obtained, if the fundamentality of the polynomial set is as low as possible. This will occur if minimal or near minimal formulas will be studied. This will be the content of the next two sections. The following lemma, see [27], will turn out to be very useful in order to apply Theorem 5.3.

LEMMA 5.2. *Let*

$$r_k = p_k + \hat{r}_k \in (\mathbb{P}_k)^{k+1}\ ,\ \hat{r}_k \in (\mathbb{P}_{k-1})^{k+1}$$

be a fundamental system of degree k *generating the ideal* $\mathcal{A} = (r_k^t)$. *If*

$$xF_k r_k - yL_k r_k \in \text{span}\ \{r_k^t\},$$

then every $Q \in (r_k) \cap \mathbb{P}_k$ *of the form* $Q = q_k^t r_k$, $q_k \in (\mathbb{P}_n)^{k+1}$, $n > 1$, *can be transformed to* $Q = \hat{q}_k^t r_k$, $\hat{q}_k \in (\mathbb{P}_{n-1})^{k+1}$.

6. Formulas of even degree. Here we study the question of whether minimal formulas of type $K(2k-2, \dim \mathbb{P}_{k-1})$ exist or not. Due to Corollary 4.1, the nodes of such formulas are the common real zeros of

$$r_k = p_k + \Gamma_k^t M_{k-1}^{-1} p_{k-1} \in (\mathbb{P}_k)^{k+1},\ \Gamma_k \in \mathbb{R}^{k \times k+1}. \tag{21}$$

The following theorem can be proved applying Theorem 5.3. It goes back to [15] and [26]. Further studies of even degree formulas have been made by G. G. Rasputin; cf. [19], [20].

THEOREM 6.1. *A cubature formula for* I *of type* $K(2k-2, \dim \mathbb{P}_{k-1})$ *exists if and only if there is a* $\Gamma_k \in \mathbb{R}^{k \times k+1}$ *such that for* r_k *of the form* (21) *the equation*

$$yL_k r_k - xF_k r_k = C_k r_k,\ C_k \in \mathbb{R}^{k \times k+1} \tag{22}$$

holds. If this equation is satisfied, then $\mathcal{A}_K = (r_k^t)$.

Proof. $\Rightarrow$. If $K(2k-2, \dim \mathbb{P}_{k-1})$ exists, we can assume a fundamental system of the form (21). If (22) does not hold we can find a nontrivial polynomial in $\mathbb{P}_{k-1}$ vanishing at the nodes of the formula. This is in contradiction to Theorem 5.3.

$\Leftarrow$. If r_k of the form (21) is given satisfying (22) every $Q \in \mathbb{P}_k \cap (r_k^t)$ can be written as

$$Q = a_k^t r_k,\ a_k \in \mathbb{R}^{k+1}$$

due to Lemma 5.2. Hence $(r_k) \cap \mathbb{P}_{k-1} = (0)$ and Theorem 5.3, (ii) is satisfied for $\mathcal{A}_K = (r_k^t)$. Note that the second condition holds since every $U^2 \in \mathbb{P}_{2k-1}$ for $U \in \mathbb{P}_{k-1}$. The theorem follows from Theorem 5.3. □

Applying the recursion formula we can rewrite (22) in a more constructive matrix equation.

$$\begin{aligned} C_k r_k &= C_k p_k + C_k \Gamma_k^t M_{k-1}^{-1} p_{k-1} \\ &= L_k(y p_k + y \Gamma_k^t M_{k-1}^{-1} p_{k-1}) - F_k(x p_k + x \Gamma_k^t M_{k-1}^{-1} p_{k-1}) \\ &= L_k(F_{k+1} p_{k+1} + M_k F_k^t M_{k-1}^{-1} p_{k-1}) \\ &\quad + L_k \Gamma_k^t M_{k-1}^{-1}(F_k p_k + M_{k-1}^t F_{k-1}^t M_{k-2}^{-1} p_{k-2}) \\ &\quad - F_k(L_{k+1} p_{k+1} + M_k L_k^t M_{k-1}^{-1} p_{k-1}) \\ &\quad - F_k \Gamma_k^t M_{k-1}^{-1}(L_k p_k + M_{k-1}^t L_{k-1}^t M_{k-2}^{-1} p_{k-2}). \end{aligned}$$

This can be written as

$$\begin{aligned} C_k p_k + C_k \Gamma_k^t M_{k-1}^{-1} p_{k-1} = {} & (L_k F_{k+1} - F_k L_{k+1}) p_{k+1} \\ & + (L_k \Gamma_k^t M_{k-1}^{-1} F_k - F_k \Gamma_k^t M_{k-1}^{-1} L_k) p_k \\ & + (L_k M_k F_k^t M_{k-1}^{-1} - F_k M_k L_k^t M_{k-1}^{-1}) p_{k-1} \\ & + (L_k \Gamma_k^t F_{k-1}^t - F_k \Gamma_k^t L_{k-1}^t) M_{k-2}^{-1} p_{k-2}. \end{aligned}$$

Due to (6) the factor of p_{k+1} vanishes. By multiplying the remaining equation by $p_k^t, p_{k-1}^t, p_{k-2}^t$, respectively, and applying the functional I, we obtain

$$C_k = L_k \Gamma_k^t M_{k-1}^{-1} F_k - F_k \Gamma_k^t M_{k-1}^{-1} L_k, \tag{23}$$

$$C_k \Gamma_k^t = L_k M_k F_k^t - F_k M_k L_k^t, \tag{24}$$

$$0 = L_k \Gamma_k^t F_{k-1}^t - F_k \Gamma_k^t L_{k-1}^t. \tag{25}$$

From (25) it follows by (12) that $\Gamma_k^t \in \mathbb{R}^{k+1 \times k}$ is a Hankel matrix. Applying (13) and (15) we obtain from (24) and (25)

$$\begin{aligned} C_k &= \Gamma_k(L_k^t M_{k-1}^{-1} F_k - F_k^t M_{k-1}^{-1} L_k) = -\Gamma_k M_k^{-1} M_k^\star M_k^{-1}, \\ C_k \Gamma_k^t &= M_{k-1}^\star. \end{aligned}$$

Thus

$$-M_{k-1}^\star = \Gamma_k M_k^{-1} M_k^\star M_k^{-1} \Gamma_k^t.$$

COROLLARY 6.1. *A cubature formula for I of type $K(2k-2, \dim \mathbb{P}_{k-1})$ exists if and only if there is a Hankel matrix $\Gamma_k^t \in \mathbb{R}^{k+1 \times k}$ such that the matrix equation*

$$-M_{k-1}^\star = \Gamma_k M_k^{-1} M_k^\star M_k^{-1} \Gamma_k^t \tag{26}$$

holds. If this equation is satisfied, then $\mathcal{A}_K = (r_k^t)$ where r_k is of the form (21).

7. Formulas of degree $2k-1$ and of type $\mathcal{F}_{k+1}$. In this section we study formulas $K(2k-1, N)$. By Corollaries 4.2 and 4.3 such formulas attaining the bound in Theorem 4.2 belong to a real ideal $\mathcal{A}_K$ containing a fundamental set of degree $k+1$.

DEFINITION 7.1. *A formula $K(2k-1, N)$ is called of type $\mathcal{F}_{k+1}$ if the ideal $\mathcal{A}_K$ associated with the formula contains a fundamental set of degree $k+1$.*

Hence minimal formulas attaining Möller's bound are of type $\mathcal{F}_{k+1}$. If a formula $K(2k-1, N)$ of type $\mathcal{F}_{k+1}$ exists, then due to the $(2k-1)$-orthogonality the polynomials in

$$r_{k+1} = p_{k+1} + \Gamma M_{k-1}^{-1} p_{k-1} \in (\mathbb{P}_{k+1})^{k+2}, \ \Gamma^t \in \mathbb{R}^{k \times k+2},$$

with a suitable chosen Γ belong to $\mathcal{A}_K$. Hence the polynomials in

$$yL_{k+1}r_{k+1} - xF_{k+1}r_{k+1} = S_k p_k \in (\mathbb{P}_k)^{k+1},\ S_k \in \mathbb{R}^{k+1\times k+1}$$

belong to $\mathcal{A}_K$ as well. Applying the recursion formula we obtain

$$\begin{aligned}
S_k p_k &= L_{k+1}(yp_{k+1} + \Gamma M_{k-1}^{-1} y p_{k-1}) - F_{k+1}(xp_{k+1} + \Gamma M_{k-1}^{-1} x p_{k-1}) \\
&= L_{k+1}(F_{k+2}p_{k+2} + M_{k+1}F_{k+1}^t M_k^{-1} p_k) \\
&\quad + L_{k+1}\Gamma M_{k-1}^{-1}(F_k p_k + M_{k-1}F_{k-1}^t M_{k-2}^{-1})p_{k-2} \\
&\quad - F_{k+1}(L_{k+2}p_{k+2} + M_{k+1}L_{k+1}^t M_k^{-1} p_k) \\
&\quad - F_{k+1}\Gamma M_{k-1}^{-1}(L_k p_k + M_{k-1}L_{k-1}^t M_{k-2}^{-1})p_{k-2} \\
&= L_{k+1}M_{k+1}F_{k+1}^t M_k^{-1} p_k + L_{k+1}\Gamma M_{k-1}^{-1} F_k p_k \\
&\quad - F_{k+1}M_{k+1}L_{k+1}^t M_k^{-1} p_k - F_{k+1}\Gamma M_{k-1}^{-1} L_k p_k \\
&\quad + L_{k+1}\Gamma F_{k-1}^t M_{k-2}^{-1} p_{k-2} - F_{k+1}\Gamma L_{k-1}^t M_{k-2}^{-1} p_{k-2}.
\end{aligned}$$

The $(2k-1)$-orthogonality implies

$$L_{k+1}\Gamma F_{k-1}^t = F_{k+1}\Gamma L_{k-1}^t.$$

Thus by (12), $\Gamma^t \in \mathbb{R}^{k\times k+2}$ is a Hankel matrix which is written as

$$\Gamma = \begin{pmatrix}
\gamma_0 & \gamma_1 & \cdots & \gamma_{k-2} & \gamma_{k-1} \\
\gamma_1 & \gamma_2 & \cdots & \gamma_{k-1} & \gamma_k \\
\vdots & \vdots & & \vdots & \vdots \\
\gamma_k & \gamma_{k+1} & \cdots & \gamma_{2k-2} & \gamma_{2k-1} \\
\gamma_{k+1} & \gamma_{k+2} & \cdots & \gamma_{2k-1} & \gamma_{2k}
\end{pmatrix}. \tag{27}$$

Let us define

$$\Gamma_k = \begin{pmatrix}
\gamma_0 & \gamma_1 & \cdots & \gamma_{k-1} & \gamma_k \\
\gamma_1 & \gamma_2 & \cdots & \gamma_k & \gamma_{k+1} \\
\vdots & \vdots & & \vdots & \vdots \\
\gamma_{k-1} & \gamma_k & \cdots & \gamma_{2k-2} & \gamma_{2k-1} \\
\gamma_k & \gamma_{k+1} & \cdots & \gamma_{2k-1} & \gamma_{2k}
\end{pmatrix} \in \mathbb{R}^{k+1\times k+1}. \tag{28}$$

Then

$$F_{k+1}\Gamma = \Gamma_k F_k^t,\ \ L_{k+1}\Gamma = \Gamma_k L_k^t \tag{29}$$

holds, and applying (13), (15), and (16), we can write

$$\begin{aligned}
S_k p_k &= (L_{k+1}M_{k+1}F_{k+1}^t - F_{k+1}M_{k+1}L_{k+1}^t)M_k^{-1} p_k \\
&\quad + \Gamma_k(L_k^t M_{k-1}^{-1} F_k - F_k^t M_{k-1}^{-1} L_k)p_k \\
&= M_k^\star M_k^{-1} p_k - \Gamma_k M_k^{-1} M_k^\star M_k^{-1} p_k \\
&= (M_k - \Gamma_k)M_k^{-1} M_k^\star M_k^{-1} p_k.
\end{aligned}$$

Hence we can assume

$$r_{k+1} = p_{k+1} + \Gamma M_{k-1}^{-1} p_{k-1}, \tag{30}$$

where Γ is a Hankel matrix, which can be computed from Γ_k. Furthermore, we have constructed some polynomials of degree k belonging to $\mathcal{A}_K$.

The following theorem characterizes all polynomials of degree k in $\mathcal{A}_K$. It was given by G. Renner [21] for product integrals. We present it in our more general setup.

THEOREM 7.1. *Let $K(2k-1, N)$ be an interpolatory cubature formula of type $\mathcal{F}_{k+1}$, let r_{k+1} be the fundamental set of degree $k+1$, given by* (30), *and let $\mathcal{A}_K$ be the associated ideal. Then the following conditions are equivalent.*

(i) $a_k^t p_k \in \mathcal{A}_K$,

(ii) $xa_k^t p_k, ya_k^t p_k \in \operatorname{span}\{r_{k+1}^t\}$,

(iii) $(M_k - \Gamma_k)a_k = 0$, *where Γ_k is of the form* (28) *and satisfies* (29).

Proof. (i) $\Rightarrow$ (ii). If $a_k^t p_k \in \mathcal{A}_K$ then $xa_k^t p_k,\ ya_k^t p_k \in \mathcal{A}_K$, hence

$$xa_k^t p_k - a_k^t L_{k+1} r_{k+1},\ ya_k^t p_k - a_k^t F_{k+1} r_{k+1}$$

are in $\mathcal{A}_K$. Applying the recursion formula, we see that these polynomials are in $\mathbb{P}_{k-1}$. By Theorem 5.3 $\mathcal{A}_K \cap \mathbb{P}_{k-1} = (0)$ must be satisfied. Hence these polynomials must vanish; i.e.,

$$xa_k^t p_k = a_k^t L_{k+1} r_{k+1},\ ya_k^t p_k = a_k^t F_{k+1} r_{k+1}. \tag{31}$$

(ii) $\Rightarrow$ (iii). Multiplying (31) by p_{k-1}^t and applying the functional I we obtain

$$a_k^t(M_k L_k^t - L_{k+1}\Gamma) = 0,\ \ a_k^t(M_k F_k^t - F_{k+1}\Gamma) = 0.$$

In view of (29) this can be written as

$$a_k^t(M_k - \Gamma_k)L_k^t = 0,\ \ a_k^t(M_k - \Gamma_k)F_k^t = 0.$$

Thus

$$(M_k - \Gamma_k)a_k = 0.$$

(iii) $\Rightarrow$ (i). Applying the recursion formula we see that

$$\begin{aligned}
&L_k^t p_{k-1} p_{k+1}^t L_{k+1}^t + F_k^t p_{k-1} p_{k+1}^t F_{k+1}^t - D_k p_k p_k^t \\
&= L_k^t p_{k-1}(xp_k^t - p_{k-1}^t M_{k-1}^{-1} L_k M_k) + F_k^t p_{k-1}(yp_k^t - p_{k-1}^t M_{k-1}^{-1} F_k M_k) - D_k p_k p_k^t \\
&= L_k^t x p_{k-1} p_k^t - L_k^t p_{k-1} p_{k-1}^t M_{k-1}^{-1} L_k M_k + F_k^t y p_{k-1} p_k^t \\
&\quad - F_k^t p_{k-1} p_{k-1}^t M_{k-1}^{-1} F_k M_k - D_k p_k p_k^t \\
&= L_k^t(L_k p_k + M_{k-1} L_{k-1}^t M_{k-2}^{-1} p_{k-2}) p_k^t + F_k^t(F_k p_k + M_{k-1} F_{k-1}^t M_{k-2}^{-1} p_{k-2}) p_k^t \\
&\quad - L_k^t p_{k-1} p_{k-1}^t M_{k-1}^{-1} L_k M_k - F_k^t p_{k-1} p_{k-1}^t M_{k-1}^{-1} F_k M_k - D_k p_k p_k^t \\
&= L_k^t M_{k-1} L_{k-1}^t M_{k-2}^{-1} p_{k-2} p_k^t + F_k^t M_{k-1} F_{k-1}^t M_{k-2}^{-1} p_{k-2} p_k^t \\
&\quad - L_k^t p_{k-1} p_{k-1}^t M_{k-1}^{-1} L_k M_k - F_k^t p_{k-1} p_{k-1}^t M_{k-1}^{-1} F_k M_k.
\end{aligned}$$

Hence all matrix entries are polynomials in $\mathbb{P}_{2k-1}$. Let $P = a_k^t p_k \in \dot{\mathbb{P}}_k$ be given. We can write P as $P = b_k^t D_k p_k$, where $b_k = D_k^{-1} a_k$. Hence

$$P^2 = b_k^t D_k p_k p_k^t D_k b_k \in \mathbb{P}_{2k}.$$

Subtracting

$$R^\star = b_k^t(L_k^t p_{k-1} r_{k+1}^t L_{k+1}^t + F_k^t p_{k-1} r_{k+1}^t F_{k+1}^t) D_k b_k \in \mathbb{P}_{2k} \cap \mathcal{A}_K,$$

we obtain $P^2 - R^\star \in \mathbb{P}_{2k-1}$. Applying the functional I, we get

$$b_k^t D_k M_k D_k b_k - b_k^t (L_k^t I(p_{k-1} r_{k+1}^t) L_{k+1}^t + F_k^t I(p_{k-1} r_{k+1}^t) F_{k+1}^t) D_k b_k.$$

Since

$$I(p_{k-1} r_{k+1}^t L_{k+1}^t) = \Gamma^t L_{k+1}^t = L_k \Gamma_k, \quad I(p_{k-1} r_{k+1}^t F_{k+1}^t) = \Gamma^t F_{k+1}^t = F_k \Gamma_k,$$

we obtain

$$I(P^2 - R^\star) = b_k^t D_k M_k D_k b_k - b_k^t D_k \Gamma_k D_k b_k = a_k^t (M_k - \Gamma_k) a_k.$$

So, if $(M_k - \Gamma_k) a_k = 0$, then $I(P^2 - R^\star) = 0$. Since $\mathcal{A}_K$ generates a $K(2k-1, N)$-formula we find $I(P^2 - R^\star) = K(2k-1, N)(P^2 - R^\star) = 0$. Thus P^2 vanishes at the zero-set of $\mathcal{A}_K$; by Theorem 5.3 this ideal is real; finally, P^2 and P are in $\mathcal{A}_K$. □

THEOREM 7.2. *An interpolatory cubature formula $K(2k-1, N)$ of type $\mathcal{F}_{k+1}$ with $N = \dim \mathbb{P}_{k-1} + s$ exists if and only if there is a Hankel matrix $\Gamma_k \in \mathbb{R}^{k+1 \times k+1}$ such that*

(i) $(M_k - \Gamma_k) M_k^{-1} M_k^\star M_k^{-1} (M_k - \Gamma_k) = 0$,

(ii) rank $(M_k - \Gamma_k) = s \geq [k/2]$, *and,*

(iii) $M_k - \Gamma_k$ *is positive semidefinite.*

If these conditions hold, then $\mathcal{A}_K = (r_{k+1}^t, (Cp_k)^t)$ is the real ideal generating $K(2k-1, N)$, where $r_{k+1} = p_{k+1} + \Gamma M_{k-1}^{-1} p_k$; Γ and Γ_k are connected by (29), and $C \in \mathbb{R}^{k+1-s \times k+1}$ is of rank $k+1-s$, *and $(M_k - \Gamma_k) C^t = 0$.*

Proof. $\Rightarrow$: Let $K(2k-1, N)$, $N = \dim \mathbb{P}_{k-1} + s$ of type $\mathcal{F}_{k+1}$ be given. Then there is a Hankel matrix $\Gamma \in \mathbb{R}^{k+2 \times k}$ of the form (28) such that by using (29) the fundamental system in $\mathcal{A}_K$ can be written as (30). Applying Theorem 7.1 for the polynomials

$$S_k p_k = (M_k - \Gamma_k) M_k^{-1} M_k^\star M_k^{-1} p_k \in \mathcal{A}_K,$$

we obtain

$$0 = (M_k - \Gamma_k)((M_k - \Gamma_k) M_k^{-1} M_k^\star M_k^{-1})^t;$$

i.e., condition (i).

Since there are $k+1-s$ linearly independent polynomials in $\mathcal{A}_K \cap \dot{\mathbb{P}}_k$, we obtain condition (ii). By the proof of Theorem 4.2 there are at most $k+1-[k/2]$ linearly independent polynomials in $\mathcal{A}_K \cap \dot{\mathbb{P}}_{k-1}$ hence $s \geq [k/2]$.

Combining the last part of the proof of Theorem 7.1 and Theorem 5.3(ii), we obtain for all $P = a_k^t p_k \in \dot{\mathbb{P}}_k$

$$a_k^t (M_k - \Gamma_k) a_k \geq 0,$$

i.e., $M_k - \Gamma_k$ is positive semidefinite.

$\Leftarrow$: Let $r_{k+1} = p_{k+1} + \Gamma M_{k-1}^{-1} p_{k-1}$. By condition (ii) there is a matrix $C \in \mathbb{R}^{k+1-s \times k+1}$ of rank $k+1-s$ such that $(M_k - \Gamma_k) C^t = 0$. Considering the polynomial vectors

$$q_1 = xCp_k - CL_{k+1} r_{k+1}, \qquad q_2 = yCp_k - CF_{k+1} r_{k+1},$$

we find by applying (14) and (29)

$$q_1 = C(M_k L_k^t - L_{k+1}\Gamma) M_{k+1}^{-1} p_{k-1} = C(M_k - \Gamma_k) L_k^t M_{k-1}^{-1} p_{k-1},$$
$$q_2 = C(M_k F_k^t - F_{k+1}\Gamma) M_{k+1}^{-1} p_{k-1} = C(M_k - \Gamma_k) F_k^t M_{k-1}^{-1} p_{k-1}.$$

Since $C(M_k - \Gamma_k) = 0$, it follows that $q_1 = q_2 = 0$, hence $xCp_k, yCp_k \in (r_{k+1}^t)^{k+1-s}$.

For $\mathcal{A} = (r_{k+1}^t, (Cp_k)^t)$ we verify the conditions of Theorem 5.3. By (i) we obtain

$$(M_k - \Gamma_k)M_k^{-1}M_k^\star M_k^{-1}(M_k - \Gamma_k) = (M_k - \Gamma_k)S_k^t = 0.$$

There is a matrix $T \in \mathbb{R}^{k+1\times k+1-s}$ such that $S_k = TC$. We know that $yL_{k+1}r_{k+1} - xF_{k+1}r_{k+1} = S_kp_k$, hence $yL_{k+1}r_{k+1} - xF_{k+1}r_{k+1} \in \operatorname{span}\{(Cp_k)^t\}$.

Any $Q \in \mathcal{A} \cap \mathbb{P}_k$ can be written as

$$Q = u_{k-s}^t Cp_k + v_{k+1}^t r_{k+1},\ u_{k-s} \in (\mathbb{P}_{n_1})^{k+1-s},\ v_{k+1} \in (\mathbb{P}_{n_2})^{k+2}.$$

Since $xCp_k,\ yCp_k \in (r_{k+1}^t)^{+1-s}$ we find $\hat{u}_{k-s}^t Cp_k \in (r_{k+1}^t)$, where $\hat{u}_{k-s}^t$ is the polynomial u_{k-s}^t without the constant term. Thus

$$Q = \tilde{u}_{k-s}^t Cp_k + w_{k+1}^t r_{k+1},\ \tilde{u}_{k-s} \in \mathbb{R}^{k+1-s},\ w_{k+1} \in (\mathbb{P}_n)^{k+2}.$$

Applying Lemma 5.2 we finally obtain $Q = c_{k-s}^t Cp_k,\ c_{k-s} \in \mathbb{R}^{k+1-s}$. Thus $\mathcal{A} \cap \mathbb{P}_{k-1} = (0)$. If we denote by $\mathcal{U}$ the linear space spanned by $\mathbb{P}_{k-1}$ and by the s linearly independent polynomials $U_1, U_2, \ldots, U_s \in \mathbb{P}_k$ which do not belong to $\mathcal{A}$ we get $\mathcal{A} \cap \mathcal{U} = (0)$.

If $U \in \mathbb{P}_{k-1}$, then $I(U^2) > 0$. Let $U = a_k^t p_k \in \mathcal{U}$ be given, i.e., $(M_k - \Gamma_k)a_k \neq 0$. Choosing $R^\star$ such that $U^2 - R^\star \in \mathbb{P}_{2k+1}$ we obtain by the postive semidefiniteness

$$I(U^2 - R^\star) = a_k^t(M_k - \Gamma_k)a_k > 0.$$

Equality will be attained if and only if $(M_k - \Gamma_k)a_k = 0$. Thus we get $I(U^2 - R^\star) > 0$ for all $U \in \mathcal{U}$ and can apply Theorem 5.3 to complete the proof. □

COROLLARY 7.1. *Any formula $K(2k-1, N)$ of type $\mathcal{F}_{k+1}$ satisfies*

$$\dim \mathbb{P}_{k-1} + [k/2] \leq N \leq \dim \mathbb{P}_{k-1} + [k/2] + 1.$$

Proof. Since $(M_k - \Gamma_k)p_k$ are polynomials that do not belong to the ideal $\mathcal{A}_K$, and since the polynomials in $(M_k - \Gamma_k)M_k^{-1}M_k^\star M_k^{-1}p_k$ belong to $\mathcal{A}_K$, we get

$$\operatorname{rank}(M_k - \Gamma_k) + \operatorname{rank}(M_k - \Gamma_k)M_k^{-1}M_k^\star M_k^{-1} \leq k + 1.$$

Considering Lemma 4.1, we obtain

$$\operatorname{rank}(M_k - \Gamma_k) - 1 \leq \operatorname{rank}(M_k - \Gamma_k)M_k^{-1}M_k^\star M_k^{-1}$$

for even k, while

$$\operatorname{rank}(M_k - \Gamma_k) \leq \operatorname{rank}(M_k - \Gamma_k)M_k^{-1}M_k^\star M_k^{-1}$$

holds for odd k. Thus $s \leq \operatorname{rank}(M_k - \Gamma_k) \leq [k/2] + 1$. □

8. Applications. We have characterized the existence of minimal cubature formulas attaining the known lower bounds by the solvability of certain matrix equations under some constraints. However, to obtain formulas, one must solve these equations; to improve the lower bound, one must show that no real solutions exist or the constraints are violated; both are hard.

Some progress has been made, nevertheless. For integrals over the circle with weight function $(1 - x^2 - y^2)^\alpha,\ \alpha > -1$, G. Godzina [10] proved for $k = 5$ that the

matrix equation of Theorem 7.2 has no solution for the α in question. This implies that the lower bound of Theorem 4.2 is not sharp for these integrals. Further results in this direction have been found by using a completely different approach (see [33] and [2]).

In the following we will apply our approach for a special classical integral with a moment matrix which is a multiple of the identity. So both matrix equations take their simplest form and can be solved.

This integral is a special two-dimensional, centrally symmetric product-integral generated by the integral l_α defined on $\mathbb{R}[x]$, the ring of real polynomials,

$$l_\alpha : \mathbb{R}[x] \to \mathbb{R} : \pi \mapsto l_\alpha(\pi) = \int_{-1}^{1} \pi(t)\, w(t) dt,$$

where

$$w(t) = \frac{\Gamma(2\alpha+2)}{\Gamma(\alpha+1)\Gamma(\alpha+1)2^{2\alpha+1}}(1-t^2)^\alpha, \quad \alpha > -1,$$

such that $l_\alpha(1) = 1$. The orthogonal polynomials with respect to l_α are the ultraspherical polynomials. They will be denoted by $\pi_i^\alpha = \pi_i$, $i = 0, 1, \ldots$. The π_is are normalized such that their leading coefficient is 1. The following recursion formula holds:

$$\pi_0(t) = 1,\ \pi_1(t) = t,\ \pi_{i+1}(t) = t\pi_i(t) - \Lambda_i^\alpha \pi_{i-1}(t),\ i = 1, 2, \ldots,$$

where

$$\Lambda_1^\alpha = \frac{1}{2\alpha+3},\ \Lambda_i^\alpha = \Lambda_i = \frac{i(2\alpha+i)}{(2\alpha+2i+1)(2\alpha+2i-1)},\ i = 2, 3, \ldots.$$

The moments are of the form

$$l_\alpha(\pi_0\pi_0) = 1,\ l_\alpha(\pi_i^2) = \Lambda_1\Lambda_2 \ldots \Lambda_i,\ i = 1, 2, \ldots.$$

The product form of l_α is centrally symmetric,

$$I_\alpha : \mathbb{P}^2 \to \mathbb{R} : P \mapsto I_\alpha(P) = \int_{-1}^{1}\int_{-1}^{1} P(x,y) w(x) w(y) dx dy,\ \alpha > -1.$$

The orthogonal polynomials with respect to I_α can be written as

$${}^\alpha P_i^k = P_i^k(x,y) = \pi_i(x)\pi_{k-i}(y),\ i = 0, 1, \ldots, k, k = 0, 1, \ldots,$$

generating the moments ${}^\alpha M_{ij}^k = I_\alpha(P_i^k P_j^k) = l_\alpha(\pi_i^2) l_\alpha(\pi_{k-i}^2)\delta_{ij}$.

We can apply Theorem 3.1, where the M_ks are diagonal matrices of the form

$$M_k = \operatorname{diag}\{\Lambda_1\Lambda_2 \ldots \Lambda_k, \Lambda_1\Lambda_1\Lambda_2 \ldots \Lambda_{k-1}, \Lambda_1\Lambda_2\Lambda_1\Lambda_2 \ldots \Lambda_{k-2}, \ldots, \Lambda_1\Lambda_2 \ldots \Lambda_k\}.$$

Note that for $\alpha^2 = 1/4$ we obtain the moment matrices

$$(32) \quad M_k = 1/2(1/4)^{k-1} D_k \quad \text{if } \alpha = -1/2, \qquad M_k = (1/4)^k E_k \quad \text{if } \alpha = 1/2.$$

8.1. Minimal formulas of even degree for $I_{1/2}$. For $I_{1/2}$, (26) is of a very simple form since $M_k = (1/4)^k E_k$ and $M_k^\star = (1/4)^{k+1} T_k$. Choosing a suitable factor for Γ_k, Corollary 6.1 can be restated as follows.

Any Hankel matrix $\Gamma_k \in \mathbb{R}^{k \times k+1}$ solving

$$-T_{k-1} = \Gamma_k T_k \Gamma_k^t \tag{33}$$

corresponds to a minimal cubature formula of type $K(2k-2, \dim \mathbb{P}_{k-1})$ for $I_{1/2}$. The associated ideal (r_k^t) is generated by

$$r_k = p_k + 1/2 \Gamma_k^t p_{k-1}.$$

Inserting relation (7), (33) can be written as

$$L_k F_k^t - F_k L_k^t = \Gamma_k (L_k^t F_k - F_k^t L_k) \Gamma_k^t.$$

Using (12), this is equivalent to

$$L_k F_k^t - F_k L_k^t = L_k \Gamma_k^t \Gamma_k F_k^t - F_k \Gamma_k^t \Gamma_k L_k^t;$$

i.e.,

$$L_k (E_k - \Gamma_k^t \Gamma_k) F_k^t = F_k (E_k - \Gamma_k^t \Gamma_k) L_k^t.$$

Due to Lemma 2.2 this means that a Hankel matrix Γ_k solves (33) if and only if $E_k - \Gamma_k^t \Gamma_k$ is a Hankel matrix. Some special solutions of (33) can be computed directly, $\Gamma_k = \pm J_{k-1} F_k$ and $\Gamma_k = \pm J_{k-1} L_k$. The corresponding cubature formulas (with all nodes inside the domain of integration) are due to [15]. For $k \geq 6$ all solutions can be derived in a closed form.

LEMMA 8.1 ([28]). *For $k \geq 6$ all minimal formulas of type $K(2k-2, \dim \mathbb{P}_{k-1})$ are generated by the real ideal (r_k^t), where*

$$r_k = p_k + 1/2 \Gamma_k^t p_{k-1},$$

and where Γ_k is equal to $\dot{\Gamma}_k$ and $J_{k-1} \dot{\Gamma}_k J_k$, respectively,

$$\dot{\Gamma}_k = \begin{pmatrix} \gamma_0 & \sigma\gamma_0 & \sigma^2\gamma_0 & \cdots & \sigma^{k-1}\gamma_0 & 1/\sigma \\ \sigma\gamma_0 & \sigma^2\gamma_0 & \sigma^3\gamma_0 & \cdots & 1/\sigma & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \sigma^{k-1}\gamma_0 & 1/\sigma & 0 & \cdots & 0 & 0 \end{pmatrix}, \quad \gamma_0 = \frac{1-\sigma^2}{\sigma^{k+1}}, \quad 0 \neq \sigma \in \mathbb{R}.$$

8.2. Minimal formulas of odd degree for $I_{1/2}$. Since $M_k = (1/4)^k E_k$ and $M_k^\star = (1/4)^{k+1} T_k$ the matrix equation for $K(2k-1, \star)$ is reduced to a simple form as well by choosing a suitable factor for Γ_k in order to replace $M_k - \Gamma_k$ by $E_k - \Gamma_k$. If k is odd, then by Theorem 4.2 two polynomials are known that belong to the associated ideal $\mathcal{A}_K$. If $a_{k-1}^t = (1, 0, 1, \ldots, 0, 1)$ we find $T_{k-1} a_{k-1} = 0$, hence the polynomial

$$a_{k-1}^t (L_k + F_k) p_k = (1, 1, \ldots, 1) p_k \tag{34}$$

belongs to $\mathcal{A}_K$. Thus for odd k we find, due to Theorem 7.2, the necessary condition

$$(E_k - \Gamma_k) \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = 0.$$

These linear conditions on the entries of Γ_k are easy to handle. They induce

$$\Gamma_k = \begin{pmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_k \\ \gamma_1 & \gamma_2 & \gamma_3 & \cdots & \gamma_0 \\ \gamma_2 & \gamma_3 & \gamma_4 & \cdots & \gamma_1 \\ \vdots & \vdots & & \vdots & \vdots \\ \gamma_k & \gamma_0 & \gamma_1 & \cdots & \gamma_{k-1} \end{pmatrix} \tag{35}$$

and

$$\gamma_k = 1 - \sum_{i=0}^{k-1} \gamma_i. \tag{36}$$

Matrices of type (35) are called (-1)-circulant. They have been studied extensively by P. J. Davis in [6]; we follow this book.

Let us introduce the Fourier matrix $G_k \in \mathbb{C}^{k+1\times k+1}$ and its conjugate transpose G_k^H,

$$G_k^H = \frac{1}{\sqrt{k+1}} \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \\ \omega & \omega^2 & \cdots & \omega^{k-1} & \omega^k \\ \vdots & \vdots & & \vdots & \vdots \\ \omega^{k-1} & \omega^k & \cdots & \omega^3 & \omega^2 \\ \omega^k & \omega^{k-1} & \cdots & \omega^2 & \omega \end{pmatrix} \in \mathbb{R}^{k+1\times k+1}, \quad \omega = e^{\frac{2\pi i}{k+1}}.$$

Furthermore, we need the orthogonal matrix

$$\Pi_k = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \in \mathbb{R}^{k+1\times k+1}.$$

LEMMA 8.2 ([6]). *Let $A_k \in \mathbb{R}^{k+1\times k+1}$ be given. The following conditions are equivalent:*

(i) *A_k is (-1)-circulant,*
(ii) $\Pi_k A_k = A_k \Pi_k^t$,
(iii) $A_k = G_k^H J_k \Pi_k \Delta_k G_k$, *where* $\Delta_k = \operatorname{diag}\{\lambda_0, \lambda_1, \ldots, \lambda_k\}$.

LEMMA 8.3 ([6]). *Let $A_k = G_k^H J_k \Pi_k \Delta_k G_k$ be given. Then*

(i) *the eigenvalues of A_k are identical to those of*

$$J_k \Pi_k \Delta_k = \begin{pmatrix} \lambda_0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & \lambda_k \\ 0 & 0 & \cdots & \lambda_{k-1} & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & \lambda_1 & \cdots & 0 & 0 \end{pmatrix},$$

(ii) $(J_k \Pi_k \Delta_k)^2 = \operatorname{diag}\{\lambda_0^2, \lambda_k \lambda_1, \lambda_{k-1}\lambda_2, \ldots, \lambda_1 \lambda_k\}$.

If k is even, let us assume that Γ_k is (-1)-circulant, too. So we do not derive all formulas in this case. To apply Theorem 7.2 to find formulas with dim $\mathbb{P}_{k-1} + [k/2]$ nodes, we need

$$\operatorname{rank}(E_k - \Gamma_k) = [k/2].$$

This means that Γ_k must have the eigenvalue 1 of multiplicity $k + 1 - [k/2]$. Due to Lemma 8.3 the eigenvalues of $J_k \Pi_k \Delta_k$ are of the form

$$\lambda_0,\ \pm\sqrt{\lambda_{k+1-i}\lambda_i},\ i = 1, 2, \ldots, [k/2],$$

and, if k is odd, in addition $\lambda_{[k/2]+1}$. If 1 is an eigenvalue of multiplicity $k+1-[k/2]$, then $\lambda_0^2 = \lambda_{k+1-i}\lambda_i = 1$, $i = 1, 2, \ldots, [k/2]$, and, $\lambda_{[k/2]+1}^2 = 1$, if k is odd. With respect to Lemma 8.3 it follows that $(J_k \Pi_k \Delta_k)^2 = E_k$. This implies

$$\Gamma_k \Gamma_k = G_k^H J_k \Pi_k \Delta_k G_k G_k^H J_k \Pi_k \Delta_k G_k = G_k^H E_k G_k = E_k.$$

Thus the rank condition on $E_k - \Gamma_k$ implies that Γ_k is an orthogonal matrix. Hence the eigenvalues of $E_k - \Gamma_k$ are 0 or 2. From this we conclude that the matrix $E_k - \Gamma_k$ is positive semidefinite. The matrix equation

$$(E_k - \Gamma_k) T_k (E_k - \Gamma_k) = 0$$

can be written as

$$(E_k - \Gamma_k)(\Pi_k - \Pi_k^t + W_k)(E_k - \Gamma_k) = 0,$$

where

$$W_k = \begin{pmatrix} 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 0 \\ -1 & 0 & \ldots & 0 & 0 \end{pmatrix} \in \mathbb{R}^{k+1 \times k+1}.$$

This can be simplified to

$$(E_k - \Gamma_k) W_k (E_k - \Gamma_k) = 0, \tag{37}$$

since for an orthogonal and (-1)-circulant matrix Γ_k we have

$$(E_k - \Gamma_k)(\Pi_k - \Pi_k^t)(E_k - \Gamma_k) = 0.$$

Hence, for $I_{1/2}$, Theorem 7.2 can be specialized to the following lemma.

LEMMA 8.4. *A minimal cubature formula of degree $2k-1$, k odd, exists if and only if there is an orthogonal (-1)-circulant matrix $\Gamma_k \in \mathbb{R}^{k+1 \times k+1}$ satisfying* (36), *and*

(i) $(E_k - \Gamma_k) W_k (E_k - \Gamma_k) = 0$,
(ii) $\operatorname{rank}(E_k - \Gamma_k) = [k/2]$.

If k is even, a minimal cubature formula of degree $2k-1$ exists, if there is an orthogonal (-1)-circulant matrix $\Gamma_k \in \mathbb{R}^{k+1 \times k+1}$ satisfying conditions (i) *and* (ii). *The nodes of the formulas are in both cases the common zeros of the polynomials in*

$$(E_k + \Gamma_k) p_k.$$

The special representation of the polynomial vector follows from the fact that $(E_k - \Gamma_k)(E_k + \Gamma_k) = 0$ and from Theorem 7.1. The polynomial entries of the polynomial vector are not linearly independent. Equation (37) is of the form

$$0 = \begin{pmatrix} \gamma_k & 1-\gamma_0 \\ \gamma_0 & -\gamma_1 \\ \vdots & \vdots \\ \gamma_{k-2} & -\gamma_{k-1} \\ \gamma_{k-1}-1 & -\gamma_k \end{pmatrix} \begin{pmatrix} 1-\gamma_0 & -\gamma_1 & \dots & -\gamma_{k-1} & -\gamma_k \\ -\gamma_k & -\gamma_0 & \dots & -\gamma_{k-2} & 1-\gamma_{k-1} \end{pmatrix},$$

which can be written as

$$\begin{aligned} \gamma_k^2 &= (1-\gamma_0)(1-\gamma_{k-1}), \\ \gamma_k\gamma_j &= -\gamma_{j-1}(1-\gamma_0),\ j = 1,2,\dots,k-1, \\ \gamma_i\gamma_j &= \gamma_{i+1}\gamma_{j-1},\ i = 0,1,\dots,k-2,\ j = 1,2,\dots,k-1, \\ \gamma_i\gamma_k &= -\gamma_{i+1}(1-\gamma_{k-1}),\ i = 0,1,\dots,k-2. \end{aligned}$$

The solutions are discussed in three cases.

Case 1. Setting $\gamma_0 = 1$ we find $\gamma_k = 0$ and $\gamma_i = \gamma_1^i$, $i = 1,2,\dots,k-1$. The only solution that generates an orthogonal matrix will be obtained for

$$\gamma_0 = 1,\ \gamma_i = 0,\ i = 1,2,\dots,k.$$

The solutions $\Gamma_k = J_k\Pi_k$ and $\Pi_k J_k$, respectively, satisfy (36) and condition (ii). The corresponding minimal formulas are due to [27], where it has been overseen that these solutions hold for even k, too.

Case 2. Setting $\gamma_0 = 0$, we find two solutions generating an orthogonal matrix. The first, $\Gamma_k = \Pi_k J_k$ has been discussed in Case 1, the second $\Gamma_k = J_k$ satisfies (36), while (ii) holds for even k only. The corresponding minimal formula has been derived in [15].

Case 3. If $0 \neq \gamma_0 \neq 1$, we obtain a class of solutions. They are of the form

$$\gamma_i = \sigma^i\gamma_0,\ i = 0,1,\dots,k-1,\ \gamma_k = (\gamma_0 - 1)/\sigma,$$

where $\sigma \neq 0$ is a free parameter such that

$$(\gamma_0 - 1)/\sigma = \sigma^k\gamma_0 - \sigma \tag{38}$$

holds. If $\sigma = 1$, we obtain

$$\gamma_i = \gamma_0,\ i = 0,1,\dots k-1,\ \gamma_k = \gamma_0 - 1.$$

Hence (38) is satisfied and γ_0 is a free parameter that will be determined to satisfy condition (ii) and (36). From condition (ii) it follows that $\gamma_0 = 2/(k+1)$, (36) is satisfied for k even and odd. The corresponding minimal formulas are due to [1].

For $\sigma = -1$, (38) is satisfied only for odd k. We obtain

$$\gamma_i = (-1)^i\gamma_0,\ i = 0,1,\dots,k-1,\ \gamma_k = 1 - \gamma_0,$$

where γ_0 is a free parameter. For $\sigma^2 \neq 1$ we obtain

$$\gamma_i = \sigma^i\gamma_0,\ i = 0,1,\dots,k-1,\ \gamma_k = (\gamma_0 - 1)/\sigma,$$

and (38). We have to determine the rank of $E_k - \Gamma_k$,

$$(39) \qquad \begin{pmatrix} \gamma_0 - 1 & \sigma\gamma_0 & \dots & \sigma^{k-1}\gamma_0 & \sigma^k\gamma_0 - \sigma \\ \sigma\gamma_0 & \sigma^2\gamma_0 - 1 & \dots & \sigma^k\gamma_0 - \sigma & \gamma_0 \\ \vdots & \vdots & & \vdots & \vdots \\ \sigma^{k-1}\gamma_0 & \sigma^k\gamma_0 - \sigma & \dots & \sigma^{k-3}\gamma_0 - 1 & \sigma^{k-2}\gamma_0 \\ \sigma^k\gamma_0 - \sigma & \gamma_0 & \dots & \sigma^{k-2}\gamma_0 & \sigma^{k-1}\gamma_0 - 1 \end{pmatrix}.$$

Note that from (38) it follows that

$$\gamma_0 - \sigma^{k+1}\gamma_0 + \sigma^2 = 1.$$

Subtracting the $(k-i)$th row multiplied by σ from the $(k-i+1)$th row, $i = 1, 2, \dots, k$, we get for k odd

$$\begin{pmatrix} \gamma_0 - 1 & \sigma\gamma_0 & \dots & \sigma^s\gamma_0 & \sigma^{s+1}\gamma_0 & \dots & \sigma^{k-1}\gamma_0 & \sigma^k\gamma_0 - \sigma \\ \sigma & -1 & \dots & 0 & 0 & \dots & -\sigma & -1 \\ 0 & \sigma & \dots & 0 & 0 & \dots & -1 & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & -1 & -\sigma & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ 0 & -\sigma & \dots & 0 & 0 & \dots & 1 & 0 \\ -\sigma & 1 & \dots & 0 & 0 & \dots & \sigma & 1 \end{pmatrix},$$

and for k even

$$\begin{pmatrix} \gamma_0 - 1 & \sigma\gamma_0 & \dots & \sigma^s\gamma_0 & \dots & \sigma^{k-1}\gamma_0 & \sigma^k\gamma_0 - \sigma \\ \sigma & -1 & \dots & 0 & \dots & -\sigma & -1 \\ 0 & \sigma & \dots & 0 & \dots & -1 & 0 \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & -1-\sigma & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ 0 & -\sigma & \dots & 0 & \dots & -1 & 0 \\ -\sigma & 1 & \dots & 0 & \dots & \sigma & 1 \end{pmatrix}.$$

In both cases we put s = [k/2]. Obviously, the last $[k/2]$ rows are linearly dependent. Erasing them we get for k odd

(40)

$$\begin{pmatrix} \gamma_0 - 1 & \sigma\gamma_0 & \dots & \sigma^{s-1}\gamma_0 & \sigma^s\gamma_0 & \sigma^{s+1}\gamma_0 & \sigma^{s+2}\gamma_0 & \dots & \sigma^{k-1}\gamma_0 & \sigma^k\gamma_0 - \sigma \\ \sigma & -1 & \dots & 0 & 0 & 0 & 0 & \dots & -\sigma & 1 \\ 0 & \sigma & \dots & 0 & 0 & 0 & 0 & \dots & 1 & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & -1 & 0 & 0 & -\sigma & \dots & 0 & 0 \\ 0 & 0 & \dots & \sigma & -1 & -\sigma & 1 & \dots & 0 & 0 \end{pmatrix}$$

and for k even

$$\begin{pmatrix} \gamma_0 - 1 & \sigma\gamma_0 & \dots & \sigma^{s-1}\gamma_0 & \sigma^s\gamma_0 & \sigma^{s+1}\gamma_0 & \dots & \sigma^{k-1}\gamma_0 & \sigma^k\gamma_0 - \sigma \\ \sigma & -1 & \dots & 0 & 0 & 0 & \dots & -\sigma & 1 \\ 0 & \sigma & \dots & 0 & 0 & 0 & \dots & 1 & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & -1 & 0 & -\sigma & \dots & 0 & 0 \\ 0 & 0 & \dots & \sigma & -1-\sigma & 1 & \dots & 0 & 0 \end{pmatrix}.$$

If k is odd and $\sigma = -1$, the rank of (39) will be $[k/2]$ if and only if $\gamma_0 = ([k/2]+1)^{-1}$. This can be computed from (40). If $\sigma^2 \neq 1$ we let

$$\lambda_i = \gamma_0 \frac{\sigma^{2i} - 1}{\sigma^i(\sigma^2 - 1)} - \frac{1}{\sigma^{-i}}, \qquad i = 1, 2, \dots, s,$$

then

$$\lambda_1 = \sigma^{-1}(\gamma_0 - 1) = \sigma^k\gamma_0 - \sigma,$$

and

$$-\lambda_i + \sigma\lambda_{i+1} = \sigma^i\gamma_0, \ -\sigma\lambda_i + \lambda_{i+1} = \sigma^{k-i}\gamma_0, \ i = 1, 2, \dots, s-1.$$

If k is odd, we obtain further

$$-\lambda_s = \sigma^s\gamma_0, \ -\sigma\lambda_s = \sigma^{s+1}\gamma_0,$$

while for k even we get

$$-(\sigma+1)\lambda_s = \sigma^s\gamma_0.$$

Thus we have proved that the rank of (39) is $[k/2]$ in this general case, too. We summarize our application in the following theorem.

THEOREM 8.1. *All minimal cubature formula of degree $2k-1$, k odd, for $I_{1/2}$ are generated by a real ideal*

$$(p_k^t(E_k + \Gamma_k)),$$

where Γ_k is a (-1)-circulant orthogonal matrix of the form

$$\Gamma_k = \begin{pmatrix} \gamma_0 & \sigma\gamma_0 & \dots & \sigma^{k-1}\gamma_0 & \sigma^k\gamma_0 - \sigma \\ \sigma\gamma_0 & \sigma^2\gamma_0 & \dots & \sigma^k\gamma_0 - \sigma & \gamma_0 \\ \vdots & \vdots & & \vdots & \vdots \\ \sigma^{k-1}\gamma_0 & \sigma^k\gamma_0 - \sigma & \dots & \sigma^{k-3}\gamma_0 & \sigma^{k-2}\gamma_0 \\ \sigma^k\gamma_0 - \sigma & \gamma_0 & \dots & \sigma^{k-2}\gamma_0 & \sigma^{k-1}\gamma_0 \end{pmatrix} \tag{41}$$

or of the form

$$J_k\Gamma_k J_k, \tag{42}$$

where

$$\gamma_0 = 2/(k+1), \ \sigma^2 = 1 \quad \text{or} \quad \gamma_0 = \frac{\sigma^2 - 1}{\sigma^{k+1} - 1}, \ \sigma^2 \neq 1, \ \sigma \in \mathbb{R}.$$

THEOREM 8.2. *There are minimal formulas of degree* $2k-1$, k *even, for* $I_{1/2}$ *generated by the real ideal*

$$(p_k^t(E_k+\Gamma_k)),$$

where Γ_k *is a* (-1)*-circulant orthogonal matrix of the form* (41) *or* (42) *and*

$$\gamma_0 = 2/(k+1),\ \sigma = 1, \quad \text{or} \quad \gamma_0 = \frac{\sigma^2-1}{\sigma^{k+1}-1},\ \sigma \neq 1,\ \sigma \in \mathbb{R}.$$

Not all formulas (k even) can be determined by assuming a (-1)-circulant matrix Γ_k. In [1] a minimal formula of degree $2k-1$, k even, has been derived, where the Hankel matrix

$$\Gamma_k = \begin{pmatrix} \gamma_0 & 0 & \gamma_0 & \dots & \gamma_0 & 0 & \gamma_0-1 \\ 0 & \gamma_0 & 0 & \dots & 0 & \gamma_0-1 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & \gamma_0-1 & 0 & \dots & 0 & \gamma_0 & 0 \\ \gamma_0-1 & 0 & \gamma_0 & \dots & \gamma_0 & 0 & \gamma_0 \end{pmatrix}, \qquad \gamma_0 = \frac{4}{k+2},$$

solves the matrix equation of Theorem 7.2 under the necessary constraints.

REFERENCES

[1] R. COOLS AND H. J. SCHMID, *Minimal cubature formulae of degree* $2k-1$, Computing, 43 (1989), pp. 141–157.

[2] ———, *A new lower bound for the number of nodes in cubature formulae of degree* $4n+1$ *for some circularly symmetric integrals*, in Internat. Series Numer. Math. 112, Birkhäuser, Basel, 1993, pp. 57–66.

[3] G. CHOQUET, *Lectures on Analysis* II. *Representation Theory*, W. A. Benjamin, Inc., New York, Amsterdam, 1969.

[4] P. J. DAVIS, *A construction of nonnegative approximate quadratures*, Math. Comp., 21 (1967), pp. 578–587.

[5] P. J. DAVIS AND P. RABINOWITZ, *Methods of Numerical Integration*, Academic Press, New York, 1975.

[6] P. J. DAVIS, *Circulant Matrices*, John Wiley, New York, Chichester, Brisbane, Toronto, 1979.

[7] D. W. DUBOIS AND G. EFROYMSON, *Algebraic theory of real varieties* I, in Studies and Essays Presented to Yu-Why Chen on his sixtieth Birthday, Tapei, Academica Sinica, 1970, pp. 107–135.

[8] G. EFROYMSON, *Local reality on algebraic varieties*, J. Algebra, 29 (1974), pp. 133–142.

[9] H. ENGELS, *Numerical Quadrature and Cubature*, Academic Press, London, New York, Toronto, Sydney, San Francisco, 1980.

[10] G. GODZINA, *Über Kubaturformeln vom Grad* 9 *für Integrale über dem Kreis*, Diplomarbeit, Erlangen, Germany, 1990.

[11] M. A. KOWALSKI, *The recursion formulas for orthogonal polynomials in n variables*, SIAM J. Math. Anal., 13 (1982), pp. 309–315.

[12] ———, *Orthogonality and recursion formulas for orthogonal polynomials in n variables*, SIAM J. Math. Anal., 13 (1982), pp. 316–323.

[13] H. M. MÖLLER, *Polynomideale und Kubaturformeln*, Thesis, University of Dortmund, Germany, 1973.

[14] ———, *Kubaturformeln mit minimaler Knotenzahl*, Numer. Math., 25 (1976), pp. 185–200.

[15] C. R. MORROW AND T. N. L. PATTERSON, *Construction of algebraic cubature rules using polynomial ideal theory*, SIAM J. Numer. Anal., 15 (1978), pp. 953–976.

[16] I. P. MYSOVSKIKH, *Interpolatory cubature formulas* Nauka, Moscow, 1981. (In Russian.)

[17] ———, *Numerical characteristics of orthogonal polynomials in two variables*, Vestnik Leningrad University Math., 3 (1976), pp. 323–332.

[18] J. RADON, *Zur mechanischen Kubatur*, Monatshefte Math., 52 (1948), pp. 286–300.

[19] G. G. RASPUTIN, *Ob uslowijach suschtschestwowanija kubaturnoj formuly gaussowa tipa*, Wiss. Z. PH Potsdam, 31 (1987), pp. 627–633. (In Russian.)

[20] G. G. RASPUTIN, *Zur Konstruktion der Kubaturformel mit geradem algebraischen Grad und minimaler Knotenzahl*, Wiss. Z. PH Potsdam, 23 (1987), pp. 158–165.

[21] G. RENNER, *Über symmetrische Kubaturformeln*, Diplomarbeit, 1981, University Erlangen-Nürnberg, Germany.

[22] ———, *Darstellung von strikt quadratpositiven linearen Funktionalen auf endlichdimensionalen Polynomräumen*, Thesis, University Erlangen-Nürnberg, Germany, 1986.

[23] B. REZNICK, *Sums of even powers of real linear forms*, Memoirs of the AMS, 463 (1992), pp. 1–155.

[24] J. J. RISLER, *Une charactérisation des idéaux des varietés algébriques reéelles*, Note aux CRAS, Paris 272, 1970, pp. 522, 531.

[25] ———, *Un théorème de zéroes en géometrie algébrique et analytique réelles*, in Lecture Notes in Math., 409 (1974), pp. 522–531.

[26] H. J. SCHMID, *On cubature formulae with a minimal number of knots*, Numer. Math., 31 (1978), pp. 282–297.

[27] ———, *Interpolatorische Kubaturformeln*, Diss. Math., CCXX (1983), pp. 1–122.

[28] ———, *On minimal cubature formulae of even degree*, in Internat. Series Numer. Math. 85, Birkhäuser, Basel, 1988, pp. 216–225.

[29] H. J. SCHMID AND Y. XU, *On bivariate Gaussian cubature formulae*, Proc. AMS, 2 (1994), pp. 833–841.

[30] S. L. SOBOLEV, *Introduction to the Theory of Cubature Formulas*, Nauka, Moscow, 1974. (In Russian).

[31] A. H. STROUD, *Approximate Calculation of Multiple Integrals*, Prentice Hall, Englewood Cliffs, NJ, 1971.

[32] V. M. TSCHAKALOV, *Formules de cubature mécaniques à coefficients non negatifs*, Bull. Sci. Math, 61 (1957), pp. 123–134.

[33] P. VERLINDEN AND R. COOLS, *On cubature formulae of degree $4k+1$ attaining Möller's lower bound for integrals with circular symmetry*, Numer. Math., 61 (1992), pp. 395–407.

[34] Y. XU, *Recurrence formulas for multivariate orthogonal polynomials*, Math. Comp., 62 (1994), pp. 687–702.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 922–942, July 1995

014

DISTURBANCE DECOUPLING WITH POLE PLACEMENT FOR STRUCTURED SYSTEMS: A GRAPH-THEORETIC APPROACH*

JACOB VAN DER WOUDE† AND KAZUO MUROTA‡

Abstract. Structured systems are linear systems of which each of the coefficients in the matrices is either fixed to zero or an independent free parameter. In this paper the well-known disturbance decoupling problem with pole placement for such systems is studied and necessary and sufficient conditions for the generic solvability of the problem are derived. Generic solvability here means solvability in almost all cases. The conditions will be stated in terms of weighted matchings in a bipartite graph that easily can be associated with a structured system. The advantage of this is that the conditions then can be verified by means of well-known and efficient combinatorial algorithms.

Key words. disturbance decoupling, pole placement, structured system, bipartite graphs, Dulmage–Mendelsohn decomposition, maximum matching

AMS subject classifications. 93B55, 93C05, 94C15, 05C50, 15A06

1. Introduction.

In this paper we study so-called structured linear systems. These are linear systems with matrices of which we know which entries are fixed to zero, and which entries have some, possibly unknown, value. Hence, we can think of a structured system as being given by the zero-nonzero structure of its matrices. This zero-nonzero structure can be nicely represented by means of graphs. In this paper we formulate a general version of the well-known disturbance decoupling problem with pole placement, in which we may use linear state and disturbance feedback. We consider this problem for structured systems, and develop graph-theoretic conditions that are necessary and sufficient for the so-called generic solvability of the problem. As we will indicate, conditions for the more common version of the problem, in which only linear state feedback may be used, follow easily by a straightforward modification of our reasoning.

The study for structured systems has a long history and may be considered to have been started with [10]. In this reference and in [7], [22] the controllability for structured systems is investigated. The input-output decoupling problem for structured systems is studied in [11] (see also [3]). In [23]–[25] the finite and infinite zeros of structured systems are discussed. Disturbance decoupling problems for structured systems are discussed in [4] and [31], and almost disturbance decoupling problems for such systems are treated in [32]. The latter problems can be formulated in terms of the rank of the transfer matrix of a structured system, which is discussed in [19]. Since we do not try to give a comprehensive list of references in which the relation between certain system theoretic questions and structured systems is discussed, we conclude by referring to two textbooks, [1] and [20], on structured systems.

The starting point in all previous references are linear systems with system matrices of which we know which entries are fixed to zero, and which entries have an arbitrary, often unknown, value. This latter type of entries therefore is considered as independent free parameters. A natural consequence of this point of view is that the structure of the systems can be represented by means of graphs, either signal-flow-type directed graphs or bipartite graphs. When linear systems are described in the

* Received by the editors June 30, 1993; accepted for publication (in revised form) by P. Van Dooren, June 7, 1994.

† Faculty of Technical Mathematics and Informatics, Delft University of Technology, The Netherlands (witajww@dutinfh.tudelft.nl).

‡ Research Institute for Mathematical Sciences, Kyoto University, Kyoto, Japan.

standard form, it is most natural to adopt the directed-graph representation, whereas the bipartite-graph representation is more appropriate for systems in descriptor form.

In [13], linear descriptor systems are treated where it is known in advance which entries in the matrices are fixed to zero, and which are not (more or less as before). However, the nonzero entries are further divided into entries that can be seen as fixed constants, and entries that can be seen as independent free parameters (as above). For example, the ones in a general matrix in companion form can be seen as fixed constants, whereas the remaining nonzero entries can be seen as free parameters. As explained in [13], this more detailed structure of systems can be described by the so-called mixed matrices and their combinatorial structure can be represented by means of a combination of bipartite graphs and linear matroids.

In this paper we do not adopt the above detailed structure, but we only assume to know which of the entries in the system matrices are fixed to zero and which can be seen as free parameters. However, we do consider structured systems in descriptor form, rather than in standard form. The reason for this is that descriptor forms are more appropriate to represent the structural or generic aspects of linear systems, than standard forms. For instance, the class of structured systems in descriptor form includes the class of structured systems in standard form, and if a system in descriptor form can be transformed into a system in standard form, then this transformation in general will destroy the structure present in the system. Motivated by this, we consider here structured systems in descriptor form that are represented by bipartite graphs. We use these graphs to derive conditions for the generic solvability of our version of the above-mentioned decoupling problem. We stress however that for structured systems, which are described in standard form and represented by directed graphs, similar conditions can be found by translating the obtained conditions, although such translation may be cumbersome. We also remark here that for systems with the detailed structure as in [13], a decoupling problem as above can be formulated and solvability conditions can be obtained by an appropriate modification of our reasoning.

The outline of this paper is as follows. In §2 we present some basic facts on our version of the disturbance decoupling problem with pole placement. We bring solvability of the problem in relation with certain matrix equations being solvable over the ring of polynomial matrices as well as over the ring of the proper rational matrices. In §3 we derive some algebraic necessary and sufficient conditions for these two types of solvability. Given these conditions we state in §4 necessary and sufficient conditions for the solvability of our version of the disturbance decoupling problem with pole placement. Also in §4 we introduce structured systems and we formulate what must be understood by the generic solvability of the problem. As indicated, conditions for the solvability of the more common formulation of the problem can be obtained easily by a straightforward modification. The reason for studying our version of the problem (and not the more common one) is that our version (more) naturally fits into the algebraic framework to relate its solvability to certain matrix equations being solvable over the two rings (in fact, principal ideal domains) mentioned before. In §5 we indicate how a structured system can be represented by means of a bipartite graph with two different arc weights, and we introduce the notion of weighted matching. Using these concepts we derive necessary and sufficient conditions for the generic solvability of our version of the disturbance decoupling problem with pole placement in §6. In §7 we illustrate how the results of this paper can be used to investigate the generic solvability of the problem, as well as of the more common version. Moreover, in the last example we consider a structured system in descriptor form, represented by a bipartite graph, that can not be (easily) described as a structured system in

standard form, represented by a directed graph. We conclude this paper with §8 giving some remarks. The Appendix contains some proofs.

2. Problem formulation. We consider the system

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + Bu(t) + Qd(t), \\ z(t) &= Hx(t), \end{aligned} \tag{1}$$

with $x(t) \in \mathbb{R}^n$ the state, $u(t) \in \mathbb{R}^m$ the control, $d(t) \in \mathbb{R}^q$ the disturbance, and $z(t) \in \mathbb{R}^p$ the output (to be controlled). Throughout this paper we assume that the matrix E is real, square (i.e., $n \times n$) and invertible. The matrices A, B, Q, and H are real with suitable dimensions.

In this paper we study the following version of the well-known disturbance decoupling problem with pole placement (cf. [29], [30]).

> Let $p(s)$ be an arbitrary nth order monic polynomial with real coefficients. Find, if possible, a control law $u(t) = Fx(t) + Rd(t)$ such that $H(sE - (A + BF))^{-1}(Q + BR) = 0$ and $\det(sE - (A + BF)) = p(s)$.

We abbreviate the problem as DDPPP′. It is clear that the controllability of system (1) is necessary for the solvability of DDPPP′. We recall that (1) is controllable if and only if the matrix $[A - sE, B]$ has full row rank for all complex s (cf. [8]). Many more characterizations of controllability exist. In the context of this paper we prefer the next algebraic characterization that easily follows from the one given above (see also [21]).

> The system (1) is controllable if and only if the greatest common divisor of all the nth order minors of the matrix pencil $[A - sE, B]$ is identically equal to 1.

To recall some geometric solvability conditions for DDPPP′ we assume for the moment that $E = I$ (the $n \times n$ identity matrix). Then the next result is well-known (cf.[29],[30]).

THEOREM 2.1. DDPPP′ *is solvable for system* (1) *with* $E = I$ *if and only if the system* (1) *is controllable and* $\operatorname{Im} Q \subseteq \mathcal{R}^*(\operatorname{Ker} H) + \operatorname{Im} B$.

In Theorem 2.1 the subspace $\mathcal{R}^*(\operatorname{Ker} H)$ denotes the largest *controllability subspace* in $\operatorname{Ker} H$ (cf.[30]). It is easy to show that

$$\mathcal{R}^*(\operatorname{Ker} H) + \operatorname{Im} B = \mathcal{R}_b^*(\operatorname{Ker} H) \cap (\mathcal{V}^*(\operatorname{Ker} H) + \operatorname{Im} B), \tag{2}$$

where $\mathcal{R}_b^*(\operatorname{Ker} H)$ denotes the largest *almost controllability subspace* of $\operatorname{Ker} H$ and $\mathcal{V}^*(\operatorname{Ker} H)$ the largest *controlled invariant subspace* in $\operatorname{Ker} H$ (cf. [26], [28], [30]). To emphasize that the above subspaces are computed using A, B, and H, we occasionally denote these subspaces as $\mathcal{R}^*(\operatorname{Ker} H; A, B)$, $\mathcal{R}_b^*(\operatorname{Ker} H; A, B)$, and $\mathcal{V}^*(\operatorname{Ker} H; A, B)$. We note that geometric conditions as above can also be derived in case of a general invertible matrix E. Indeed, then we have the following result.

COROLLARY 2.2. DDPPP′ *is solvable for system* (1) *with* E *invertible if and only if the system* (1) *is controllable and* $\operatorname{Im} Q \subseteq E\mathcal{R}^*(\operatorname{Ker} H; E^{-1}A, E^{-1}B) + \operatorname{Im} B$.

Before going on we introduce some notation. We write $\mathbb{R}[s]$ for the set of polynomials with real coefficients and $\mathbb{R}_p(s)$ for the set of proper rational functions with real coefficients. Then $\mathbb{R}^a[s]$ denotes the set of polynomial vectors with a components and $\mathbb{R}_p^a(s)$ denotes the set of proper rational vectors with a components. Furthermore, $\mathbb{R}^{a \times b}[s]$ will denote the set of polynomial matrices with a rows and b columns and $\mathbb{R}_p^{a \times b}(s)$ the set of proper rational matrices with a rows and b columns.

For a general invertible matrix E the following characterizations can easily be deduced from the results in [9], [26], and [28].

PROPOSITION 2.3.

1. *The subspace* $E\mathcal{R}_b^*(\mathrm{Ker}H; E^{-1}A, E^{-1}B)$ *equals the set of all* $x_0 \in \mathbb{R}^n$ *for which there are polynomial vectors* $\xi(s) \in \mathbb{R}^n[s]$ *and* $\omega(s) \in \mathbb{R}^m[s]$ *such that* $x_0 = (sE - A)\xi(s) - B\omega(s)$ *and* $H\xi(s) = 0$ *identically in* s.

2. *The subspace* $E\mathcal{V}^*(\mathrm{Ker}H; E^{-1}A, E^{-1}B) + \mathrm{Im}\, B$ *equals the set of all* $x_0 \in \mathbb{R}^n$ *for which there are proper rational vectors* $\xi(s) \in \mathbb{R}_p^n(s)$ *and* $\omega(s) \in \mathbb{R}_p^m(s)$ *such that* $x_0 = (sE - A)\xi(s) - B\omega(s)$ *and* $H\xi(s) = 0$ *identically in* s.

Next we denote

$$M(s) = \begin{bmatrix} A - sE & B \\ H & 0 \end{bmatrix}, \qquad N(s) = \begin{bmatrix} Q \\ 0 \end{bmatrix}. \tag{3}$$

The characterizations in Proposition 2.3 now imply the following.

COROLLARY 2.4.

1. $\mathrm{Im}\, Q \subseteq E\mathcal{R}_b^*(\mathrm{Ker}H; E^{-1}A, E^{-1}B)$ *if and only if there is a polynomial matrix* $X(s) \in \mathbb{R}^{(n+m)\times q}[s]$ *such that* $M(s)X(s) = N(s)$.

2. $\mathrm{Im}\, Q \subseteq E\mathcal{V}^*(\mathrm{Ker}H; E^{-1}A, E^{-1}B) + \mathrm{Im}\, B$ *if and only if there is a proper rational matrix* $X(s) \in \mathbb{R}_p^{(n+m)\times q}(s)$ *such that* $M(s)X(s) = N(s)$.

Combining the results of Corollaries 2.2 and 2.4 through the subspace equality (2) we obtain the following result.

COROLLARY 2.5. *Let system* (1) *be controllable. Then* DDPPP′ *is solvable if and only if the equation* $M(s)X(s) = N(s)$ *has a proper rational solution as well as a polynomial solution.*

Using Corollary 2.5 we can reformulate the problem of the solvability of DDPPP′ as a problem concerning a certain matrix equation being solvable over the set of polynomial matrices as well as over the set of proper rational matrices. In the next section we present necessary and sufficient conditions for this to be the case.

3. Matrix equations. In this section we summarize some standard results on the solvability of matrix equations. We consider the equation

$$U(s)X(s) = V(s), \tag{4}$$

in $X(s)$ with given polynomial matrices $U(s) \in \mathbb{R}^{a\times b}[s]$ and $V(s) \in \mathbb{R}^{a\times c}[s]$. We say that (4) is solvable over $\mathbb{R}[s]$ if there exists a polynomial matrix $X(s) \in \mathbb{R}^{b\times c}[s]$ that satisfies (4). In the same spirit we say that (4) is solvable over $\mathbb{R}_p(s)$ if there exists a proper rational matrix $X(s) \in \mathbb{R}_p^{b\times c}(s)$ that satisfies (4).

3.1. Polynomial matrices. We call a square polynomial matrix unimodular if the matrix has an inverse that is also a polynomial matrix. As is well known, a polynomial matrix is unimodular if and only if its determinant is a nonzero constant.

For a polynomial matrix $T(s)$ with rank r (as a polynomial or rational matrix) we denote

$\Lambda_r(T(s))$ = the degree of the greatest common divisor of all rth order minors of $T(s)$.

A polynomial matrix $T(s)$ with rank r can be factorized as follows (the so-called Smith normal form) (cf. [6], [18], [21]):

$$T(s) = P(s) \begin{bmatrix} \mathrm{diag}(\alpha_1(s), \ldots, \alpha_r(s)) & 0 \\ 0 & 0 \end{bmatrix} Q(s),$$

where $P(s)$ and $Q(s)$ are unimodular polynomial matrices of suitable dimensions and $\alpha_1(s), \ldots, \alpha_r(s) \in \mathbb{R}[s]$ are monic polynomials such that $\alpha_i(s)$ divides $\alpha_{i+1}(s)$,

$1 \leq i < r$. Note that $\Lambda_r(T(s)) = \sum_{i=1}^{r} \deg(\alpha_i(s))$, where $\deg(\alpha(s))$ denotes the degree of a polynomial $\alpha(s)$.

3.2. Proper rational matrices. We say that a square proper rational matrix is biproper (or bicausal) if the matrix has an inverse that is a proper rational matrix. It is easy to see that a square proper rational matrix is biproper if and only if the value at infinity of its determinant is finite and nonzero.

If $T(s)$ is a rational matrix with rank r we denote

$$\Delta_r(T(s)) = \text{maximum of the degrees of all } r\text{th order minors of } T(s),$$

where the degree of a rational function $f(s) = p(s)/q(s)$ with $p(s)$ and $q(s)$ polynomials, is defined as $\deg(p(s)) - \deg(q(s))$.

A rational matrix $T(s)$ with rank r can be factorized as follows (cf. [5], [18]):

$$T(s) = P(s) \begin{bmatrix} \operatorname{diag}(s^{n_1}, \ldots, s^{n_r}) & 0 \\ 0 & 0 \end{bmatrix} Q(s),$$

where $P(s)$ and $Q(s)$ are biproper matrices of suitable dimensions and $n_1, \ldots, n_r$ are integers such that $n_{i+1} \leq n_i$, $1 \leq i < r$. It follows that $\Delta_r(T(s)) = \sum_{i=1}^{r} n_i$.

3.3. Solvability conditions. We can now state the following fact (see also [18],[27]). The proof of the theorem can be found in the Appendix.

THEOREM 3.1. *For the matrix equation* (4) *the following statements hold.*

1. *The equation* (4) *is solvable over* $\mathbb{R}[s]$ *if and only if* $\operatorname{rank} U(s) = \operatorname{rank}[U(s), V(s)] =: r$, *and* $\Lambda_r(U(s)) = \Lambda_r([U(s), V(s)])$.
2. *The equation* (4) *is solvable over* $\mathbb{R}_p(s)$ *if and only if* $\operatorname{rank} U(s) = \operatorname{rank}[U(s), V(s)] =: r$, *and* $\Delta_r(U(s)) = \Delta_r([U(s), V(s)])$.

In the next section we use Theorem 3.1 to develop necessary and sufficient conditions for the solvability of DDPPP′.

4. Solvability conditions for DDPPP′. In this section we return to system (1). This system is controllable if and only if the greatest common divisor of all the nth order minors of $[A - sE, B]$ is identically equal to 1. By the results of §3 the latter is equivalent to $\operatorname{rank}[A - sE, B] = n$ (as a rational matrix) and $\Lambda_n([A - sE, B]) = 0$. The first statement is always true by the regularity of E, whereas the second statement means that the n polynomials in the Smith normal form are all identically equal to 1.

4.1. Numerically specified systems. The next result is an immediate consequence of the previous remarks.

THEOREM 4.1. *Consider the system* (1) *with* E *invertible and* $M(s)$, $N(s)$ *as in* (3). *Assume that the system is controllable (here equivalent to* $\Lambda_n([A - sE, B]) = 0$*). Then* DDPPP′ *is solvable if and only if* $\operatorname{rank} M(s) = \operatorname{rank}[M(s), N(s)] =: r$, $\Lambda_r(M(s)) = \Lambda_r([M(s), N(s)])$, $\Delta_r(M(s)) = \Delta_r([M(s), N(s)])$.

A similar result can also be obtained for the next more common version of the disturbance decoupling problem with pole placement.

> Let $p(s)$ be an arbitrary nth order monic polynomial with real coefficients. Find, if possible, a control law $u(t) = Fx(t)$ such that $H(sE - (A + BF))^{-1}Q = 0$ and $\det(sE - (A + BF)) = p(s)$.

In [30] the problem is studied in the case that $E = I$; see also [29]. Following [29] we abbreviate the above problem as DDPPP. In the spirit of §2 we now can prove the following result (compare with Corollary 2.5).

COROLLARY 4.2. *Let system* (1) *be controllable. Then* DDPPP *is solvable if and only if the equation* $M(s)X(s) = N(s)$ *has a* strictly *proper rational solution as well as a polynomial solution.*

As in [17] we can now make the observation that there is a *strictly proper* rational matrix $X(s)$ such that $M(s)X(s) = N(s)$ if and only if there is a *proper* rational matrix $\hat{X}(s)$ such that $M(s)\hat{X}(s) = sN(s)$. Hence, with Theorem 3.1, the following theorem is immediate.

THEOREM 4.3. *Under conditions of Theorem* 4.1 *the following holds.* DDPPP *is solvable if and only if* $\text{rank}M(s) = \text{rank}[M(s), N(s)] =: r$, $\Lambda_r(M(s)) = \Lambda_r([M(s), N(s)])$, $\Delta_r(M(s)) = \Delta_r([M(s), sN(s)])$.

Note in Theorems 4.1 and 4.3, matrices $M(s)$, $[M(s), N(s)]$, and $[M(s), sN(s)]$ are matrix pencils. For instance,

$$[M(s), N(s)] = \left[\begin{array}{cc|c} A - sE & B & Q \\ H & 0 & 0 \end{array}\right] = \left[\begin{array}{cc|c} A & B & Q \\ H & 0 & 0 \end{array}\right] - s\left[\begin{array}{cc|c} E & 0 & 0 \\ 0 & 0 & 0 \end{array}\right].$$

Therefore, we are able to check the solvability of DDPPP′ through Theorem 4.1 if we are able to compute the rank r of a general matrix pencil $T(s)$ together with the values of the indices $\Lambda_r(T(s))$ and $\Delta_r(T(s))$. Similar remarks can be made with respect to checking the solvability of DDPPP through Theorem 4.3. However, as the latter will be technically more involved and does not yield any further significant contribution, we concentrate below mainly on the solvability of (a structural version of) DDPPP′.

4.2. Structured systems. In the remainder of this paper we assume that the system is structured. This means that we assume that we regard the nonzero entries in the system matrices as independent parameters. If the number of these parameters is W then all the systems having the same fixed zero entries can be parametrized by a vector $\lambda \in \mathbb{R}^W$. The parameter λ also parametrizes the matrices E, A, B, Q, H, and the pencils $M(s)$, $N(s)$, $[M(s), N(s)]$. We therefore occasionally denote E_λ, A_λ, B_λ, Q_λ, H_λ, and $M_\lambda(s)$, $N_\lambda(s)$ where

$$[M_\lambda(s), N_\lambda(s)] = \left[\begin{array}{cc|c} A_\lambda - sE_\lambda & B_\lambda & Q_\lambda \\ H_\lambda & 0 & 0 \end{array}\right].$$

We say that system (1) is generically controllable if the system is controllable for almost all $\lambda \in \mathbb{R}^W$. Here "almost all" is to be understood as "for all except for those in some proper algebraic variety in $\mathbb{R}^W$" (cf. [30]). A proper algebraic variety is a set of zero Lebesgue measure. Hence, system (1) is generically controllable if the greatest common divisor of all the nth order minors of $[A_\lambda - sE_\lambda, B_\lambda]$ is identically equal to 1 for "almost all" $\lambda \in \mathbb{R}^W$.

Suppose that we have a structured system that is generically controllable. Then inspired by Theorem 4.1 we say that the present version of the disturbance decoupling problem with pole placement, i.e., DDPPP′, is generically solvable precisely when $\text{rank}M_\lambda(s) = \text{rank}[M_\lambda(s), N_\lambda(s)] =: r$, $\Lambda_r(M_\lambda(s)) = \Lambda_r([M_\lambda(s), N_\lambda(s)])$, $\Delta_r(M_\lambda(s)) = \Delta_r([M_\lambda(s), N_\lambda(s)])$ for "almost all " $\lambda \in \mathbb{R}^W$. Hence, DDPPP′ for a structured system is generically solvable if the problem can be solved for almost all values of its nonzero coefficients. Also now we are able to check the generic solvability of DDPPP′ for the structured system if we have methods for the determination of the generic rank r of a general structured matrix pencil $T(s)$ together with the generic values of $\Lambda_r(T(s)$ and $\Delta_r(T(s))$. In the next section we discuss methods for doing these computations.

5. Combinatorial aspects. In this section we consider a matrix pencil $T(s)$ which is structured in the sense that the entries in the two coefficient matrices are either fixed to zero or independent parameters. We are concerned with combinatorial characterizations of the generic values of the rank r and $\Lambda_r(T(s))$ and $\Delta_r(T(s))$ introduced above. We associate with $T(s)$ a bipartite graph and express the above generic values in terms of (weighted) matchings in the bipartite graph. Though the results to be described below are special cases of the more general results of [14] (where also *fixed* nonzero entries in the coefficient matrices are considered), we afford a self-contained simplified argument for the readers' convenience.

5.1. Graphs and matchings. We consider a matrix pencil $T(s) \in \mathbb{R}^{k\times l}[s]$ of which we only know the structure, namely the powers of s appearing in each nonzero entry of $T(s)$. More specifically, a nonzero entry of $T(s)$ is of the form $\alpha s + \beta$, δs, or γ with α, β, δ, and γ independent parameters.

With $T(s) \in \mathbb{R}^{k\times l}[s]$ above we associate a bipartite graph, denoted as $G = (\mathcal{V}, \mathcal{W}, \mathcal{A})$, that consists of the sets $\mathcal{V}$ and $\mathcal{W}$ of vertices and the set $\mathcal{A}$ of arcs directed from $\mathcal{W}$ to $\mathcal{V}$. Hence we have $\mathcal{V} = \{1, 2, \ldots, k\}$, $\mathcal{W} = \{1, 2, \ldots, l\}$ and $\mathcal{A} = \{(j,i) | T_{ij}(s) \not\equiv 0\}$. Here (j,i) denotes the arc from vertex $j \in \mathcal{W}$ to vertex $i \in \mathcal{V}$ and $T_{ij}(s) \not\equiv 0$ indicates that the (i,j) entry of $T(s)$ is not identically equal to zero. An example of a bipartite graph together with the concepts introduced below is given in §7.

In the following we need weights on the arcs of G. These weights are given by two weight functions $\zeta^+, \zeta^- : \mathcal{A} \longrightarrow \{0, 1\}$ defined as follows: $\zeta^+(j,i)$ denotes the exponent of the highest power of s and $\zeta^-(j,i)$ the exponent of the smallest power of s in the nonzero entry $T_{ij}(s)$. To be more specific, for $(j,i) \in \mathcal{A}$ we define $\zeta^+(j,i) = 1$, $\zeta^-(j,i) = 0$ for $T_{ij}(s)$ of the form $\alpha s + \beta$; $\zeta^+(j,i) = \zeta^-(j,i) = 1$ for $T_{ij}(s)$ of the form δs; $\zeta^+(j,i) = \zeta^-(j,i) = 0$ for $T_{ij}(s)$ of the form γ.

If an arc $a \in \mathcal{A}$ is directed from vertex $j \in \mathcal{W}$ to vertex $i \in \mathcal{V}$, the vertex $j \in \mathcal{W}$ is called the initial vertex of a and the vertex $i \in \mathcal{V}$ the terminal vertex. A matching in the graph $G = (\mathcal{V}, \mathcal{W}, \mathcal{A})$ is a subset $\mathcal{M}$ of $\mathcal{A}$ consisting of arcs that pairwise have no vertices in common. Hence, the number of arcs in $\mathcal{M}$, called the order of the matching $\mathcal{M}$, equals the number of initial vertices of the arcs in $\mathcal{M}$ and also the number of terminal vertices of the arcs in $\mathcal{M}$. A matching of maximum order will be simply referred to as a maximum matching. With respect to the arc weight functions ζ^+ and ζ^-, we define the ζ^+-weight of a matching to be the sum of the ζ^+-weights of its arcs and similarly for the ζ^--weight of a matching.

5.2. Generic rank and generic value of $\Delta_r(T(s))$. We say that the generic rank of $T(s)$ equals r (or generic-rank $T(s) = r$) if the rank of $T(s)$ as a polynomial matrix in s equals r for almost all values of the coefficients present in $T(s)$, where "almost all" is to be understood as before. Let generic-rank $T(s) = r$. In a similar way we can define the generic values of $\Lambda_r(T(s))$ and $\Delta_r(T(s))$ as the values that these indices have for almost all values of the coefficients present in $T(s)$.

We are now in the position to state the following results connecting the above-introduced generic values with certain characteristics of the bipartite graph representing the structure of $T(s)$. Proofs of the straightforward results below can be found, for example, in [13].

PROPOSITION 5.1. *The generic rank of $T(s)$ equals the order of a maximum matching in G.*

PROPOSITION 5.2. *Assume that $T(s)$ is square and is generically invertible. The largest power of s in* det $T(s)$ *has an exponent generically equal to the maximum ζ^+-*

weight of a maximum matching in G and the smallest power of s in $\det T(s)$ *has an exponent generically equal to the minimum ζ^--weight of a maximum matching in G.*

It follows from Proposition 5.1 that r equals the order of a maximum matching in G. Recalling the definition of $\Delta_r(T(s))$ from §3 and applying Proposition 5.2 to square submatrices we immediately obtain the following.

THEOREM 5.3. *The value of $\Delta_r(T(s))$ is generically equal to the maximum ζ^+-weight of a maximum matching in G.*

This theorem enables us to compute the generic value of $\Delta_r(T(s))$ using G.

The generic value of $\Lambda_r(T(s))$ will be considered in the next subsection by means of the Dulmage–Mendelsohn-decomposition (DM-decomposition) of the bipartite graph G.

5.3. Generic value of $\Lambda_r(T(s))$. To derive a combinatorial expression for the generic value of $\Lambda_r(T(s))$ we first need to introduce a canonical decomposition of the bipartite graph $G = (\mathcal{V}, \mathcal{W}, \mathcal{A})$ into three interconnected bipartite subgraphs. This is (an aggregation of) the DM-decomposition (cf. [2], [12], [13]).

In the DM-decomposition the vertex sets $\mathcal{V}$ and $\mathcal{W}$ are decomposed as $\mathcal{V} = \mathcal{V}_0 \cup \mathcal{V}_* \cup \mathcal{V}_\infty$ and $\mathcal{W} = \mathcal{W}_0 \cup \mathcal{W}_* \cup \mathcal{W}_\infty$ with $\mathcal{V}_i \cap \mathcal{V}_j = \emptyset$ and $\mathcal{W}_i \cap \mathcal{W}_j = \emptyset$ for all $i, j = 0, *, \infty$ with $i \neq j$. The arc set $\mathcal{A}$ is decomposed as $\mathcal{A} = \mathcal{A}_{00} \cup \mathcal{A}_{0*} \cup \mathcal{A}_{0\infty} \cup \mathcal{A}_{**} \cup \mathcal{A}_{*\infty} \cup \mathcal{A}_{\infty\infty}$. Here the set $\mathcal{A}_{ij}$ contains all arcs in $\mathcal{A}$ from vertices in $\mathcal{W}_j$ to vertices in $\mathcal{V}_i$, $i, j = 0, *, \infty$ with $i \leq j$, where we define an ordering $0 < * < \infty$ among the indices of the components. Furthermore, we have in general $|\mathcal{V}_0| < |\mathcal{W}_0|$, $|\mathcal{V}_*| = |\mathcal{W}_*|$, and $|\mathcal{V}_\infty| > |\mathcal{W}_\infty|$. (We ignore here some special cases such as the case of $|\mathcal{W}_0| = 0$ in which both $\mathcal{V}_0$ and $\mathcal{W}_0$ disappear together with the arc sets $\mathcal{A}_{00}$, $\mathcal{A}_{0*}$, and $\mathcal{A}_{0\infty}$.) The DM-decomposition has the following properties where it should be noted that an empty subset of $\mathcal{A}$ is eligible as a matching.

(i) If $|\mathcal{W}_0| > 0$ the subgraph $G_0 = (\mathcal{V}_0, \mathcal{W}_0, \mathcal{A}_{00})$ contains a maximum matching of order $|\mathcal{V}_0|$ and for every vertex $w \in \mathcal{W}_0$ there is a maximum matching not containing w. The subgraph G_0 is sometimes referred to as the horizontal tail or the minimal inconsistent part.

(ii) If $|\mathcal{V}_\infty| > 0$ the subgraph $G_\infty = (\mathcal{V}_\infty, \mathcal{W}_\infty, \mathcal{A}_{\infty\infty})$ contains a maximum matching of order $|\mathcal{W}_\infty|$ and for every vertex $v \in \mathcal{V}_\infty$ there is a maximum matching not containing v. The subgraph G_∞ is sometimes referred to as the vertical tail or the maximal inconsistent part.

(iii) If $|\mathcal{V}_*|$ ($= |\mathcal{W}_*|$) > 0 the subgraph $G_* = (\mathcal{V}_*, \mathcal{W}_*, \mathcal{A}_{**})$ contains at least one maximum matching of order $|\mathcal{V}_*|$. The subgraph G_* is often referred to as the consistent part.

This decomposition is a rough version of the DM-decomposition. The full DM-decomposition is more refined, but this refinement is not relevant for the present theoretical development. The refinement may be profitable in the actual computations. In [13] an algorithm for computing the (refined) DM-decomposition is described together with the proofs of the above properties.

Recall that the graph G is constructed starting from the matrix pencil $T(s)$. The above-introduced DM-decomposition for G implies that $T(s)$ after some suitable row and/or column permutations can be depicted in a block triangular form as in Fig. 1.

The submatrix $T_{\mathcal{V}_i \mathcal{W}_j}(s)$ in Fig. 1 has dimension $|\mathcal{V}_i| \times |\mathcal{W}_j|$ and consists of the elements of $T(s)$ with row index in $\mathcal{V}_i$ and column index in $\mathcal{W}_j$, $i, j = 0, *, \infty$, $i \leq j$. The zeros denote zero matrices of suitable dimensions. To treat the most general case we assume in the sequel that $|\mathcal{V}_0| > 0$, $|\mathcal{V}_*| = |\mathcal{W}_*| > 0$, $|\mathcal{W}_\infty| > 0$. Other cases, possibly with empty components, can be dealt with in an analogous manner.

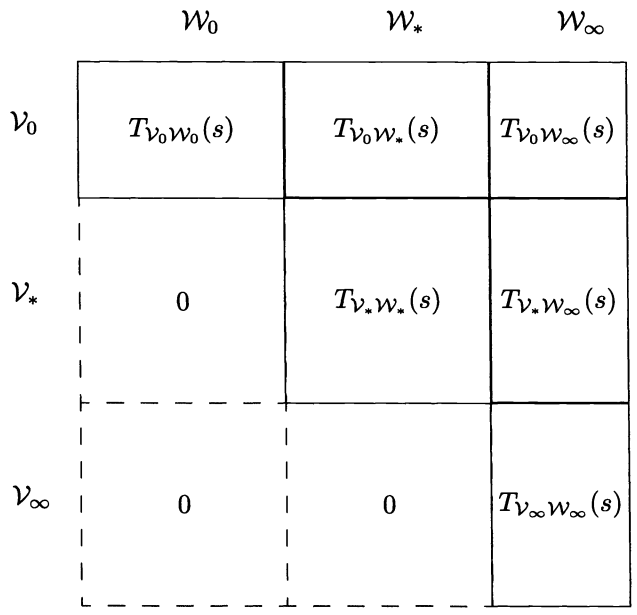


FIG. 1. DM-*decomposition of* $T(s)$.

Using Proposition 5.1 and the properties of the DM-decomposition described above it follows that generic-rank $T_{\mathcal{V}_0\mathcal{W}_0}(s) = |\mathcal{V}_0|$ (full row rank), generic-rank $T_{\mathcal{V}_*\mathcal{W}_*}(s) = |\mathcal{V}_*| = |\mathcal{W}_*|$ (invertible), and generic-rank $T_{\mathcal{V}_\infty\mathcal{W}_\infty}(s) = |\mathcal{W}_\infty|$ (full column rank). Hence, the generic rank r of $T(s)$ is expressed as $r = r_0 + r_* + r_\infty$ with $r_0 = |\mathcal{V}_0|$, $r_* = |\mathcal{V}_*| = |\mathcal{W}_*|$, $r_\infty = |\mathcal{W}_\infty|$.

Now, referring to the above decomposition of $T(s)$, let the index sets $\mathcal{I} \subseteq \mathcal{V}$ and $\mathcal{J} \subseteq \mathcal{W}$ be such that $|\mathcal{I}| = |\mathcal{J}| = r$ and $T_{\mathcal{I}\mathcal{J}}(s)$ is a generically invertible submatrix of $T(s)$. Then $\mathcal{I} = \mathcal{V}_0 \cup \mathcal{V}_* \cup \mathcal{I}_\infty$ and $\mathcal{J} = \mathcal{J}_0 \cup \mathcal{W}_* \cup \mathcal{W}_\infty$ for some index sets $\mathcal{J}_0 \subseteq \mathcal{W}_0$ and $\mathcal{I}_\infty \subseteq \mathcal{V}_\infty$ with $|\mathcal{J}_0| = r_0$ and $|\mathcal{I}_\infty| = r_\infty$. Hence,

$$T_{\mathcal{I}\mathcal{J}}(s) = \begin{pmatrix} T_{\mathcal{V}_0\mathcal{J}_0}(s) & T_{\mathcal{V}_0\mathcal{W}_*}(s) & T_{\mathcal{V}_0\mathcal{W}_\infty}(s) \\ 0 & T_{\mathcal{V}_*\mathcal{W}_*}(s) & T_{\mathcal{V}_*\mathcal{W}_\infty}(s) \\ 0 & 0 & T_{\mathcal{I}_\infty\mathcal{W}_\infty}(s) \end{pmatrix}$$

with square matrices on the diagonal. This shows that the determinant of a generically invertible rth order submatrix of $T(s)$ equals the product of $\det T_{\mathcal{V}_0\mathcal{J}_0}(s)$, $\det T_{\mathcal{V}_*\mathcal{W}_*}(s)$, and $\det T_{\mathcal{I}_\infty\mathcal{W}_\infty}(s)$ for appropriate index sets $\mathcal{J}_0 \subseteq \mathcal{W}_0$ and $\mathcal{I}_\infty \subseteq \mathcal{V}_\infty$ with $|\mathcal{J}_0| = r_0$ and $|\mathcal{I}_\infty| = r_\infty$. Therefore, the greatest common divisor of all rth order minors of $T(s)$ generically equals

$$\left(\gcd_{\mathcal{J}_0 \subseteq \mathcal{W}_0^{r_0}} \det T_{\mathcal{V}_0\mathcal{J}_0}(s) \right) \det T_{\mathcal{V}_*\mathcal{W}_*}(s) \left(\gcd_{\mathcal{I}_\infty \subseteq \mathcal{V}_\infty^{r_\infty}} \det T_{\mathcal{I}_\infty\mathcal{W}_\infty}(s) \right),$$

where gcd stands for the greatest common divisor in $\mathbb{R}[s]$, $\mathcal{W}_0^{r_0} = \{\mathcal{J}_0 \subseteq \mathcal{W}_0 \mid |\mathcal{J}_0| = r_0\}$ and $\mathcal{V}_\infty^{r_\infty} = \{\mathcal{I}_\infty \subseteq \mathcal{V}_\infty \mid |\mathcal{I}_\infty| = r_\infty\}$.

To compute the above-mentioned greatest common divisors we need the following result, stated as Proposition 14.5 in [13]. Since the result is of crucial importance for the development in this paper, we have provided an elementary proof in the Appendix. In this proof the property of the horizontal tail is used that for any vertex w in $\mathcal{W}_0$ there is a maximum matching of order $|\mathcal{V}_0|$ not containing w. By Proposition 5.1 this means that the removal of a single column from $T_{\mathcal{V}_0\mathcal{W}_0}(s)$ does not give a drop in the generic-rank.

PROPOSITION 5.4. *Generically the greatest common divisor of all the r_0th order minors of $T_{\mathcal{V}_0\mathcal{W}_0}(s)$ is a monomial in s.*

The degree of the greatest common divisor of all the r_0th order minors, which is a monomial by Proposition 5.4 above, equals the exponent of the smallest power of s contained in the minors. By Proposition 5.2 the generic value of this degree can be determined by computing in the associated bipartite graph the minimum ζ^--weight that a maximum matching can have. Hence, the greatest common divisors of all the r_0th order minors of $T_{\mathcal{V}_0\mathcal{W}_0}(s)$ can be computed by way of matchings in G_0 as follows (see Theorem 3.1 of [14]).

PROPOSITION 5.5. *The greatest common divisor of all the r_0th order minors of $T_{\mathcal{V}_0\mathcal{W}_0}(s)$ is generically a monomial of degree equal to the minimum ζ^--weight of a maximum matching in G_0.*

It is obvious that similar results can be obtained by dual arguments for $T_{\mathcal{V}_\infty\mathcal{W}_\infty}(s)$ using the bipartite subgraph G_∞. These dual results are omitted here.

We can now indicate how the generic value of $\Lambda_r(T(s))$ can be computed using the graph G. Therefore, observe that $\Lambda_r(T(s))$ equals the *degree* of the greatest common divisor of all the rth order minors of $T(s)$. Hence, $\Lambda_r(T(s))$ is the sum of the degree of the greatest common divisor of all the r_0th order minors of $T_{\mathcal{V}_0\mathcal{W}_0}(s)$, the degree of $\det T_{\mathcal{V}_*\mathcal{W}_*}(s)$ and the degree of the greatest common divisor of all the r_∞th order minors of $T_{\mathcal{V}_\infty\mathcal{W}_\infty}(s)$. In stating the following result we assume that $T(s)$ has the above decomposition and that G is already in the DM-decomposition form, i.e., G can be seen as three interconnected bipartite subgraphs $G_0 = (\mathcal{V}_0, \mathcal{W}_0, \mathcal{A}_{00})$, $G_* = (\mathcal{V}_*, \mathcal{W}_*, \mathcal{A}_{**})$, and $G_\infty = (\mathcal{V}_\infty, \mathcal{W}_\infty, \mathcal{A}_{\infty\infty})$ with the properties mentioned before. Then the following holds.

THEOREM 5.6. *The generic value of $\Lambda_r(T(s))$ is equal to the sum of the minimum ζ^--weight of a maximum matching in G_0, the maximum ζ^+-weight of a maximum matching in G_*, and the minimum ζ^--weight of a maximum matching in G_∞.*

The above theorem enables us to compute the generic value of $\Lambda_r(T(s))$ using the DM-decomposition of the graph G.

REMARK 5.7. If the DM-decomposition is available, the generic value of $\Delta_r(T(s))$ can be obtained in a manner that generally will be more efficient than the direct application of Theorem 5.3. Namely, we utilize the following fact: The value of $\Delta_r(T(s))$ is generically equal to the sum of the maximum ζ^+-weight of a maximum matching in G_0, the maximum ζ^+-weight of a maximum matching in G_*, and the maximum ζ^+-weight of a maximum matching in G_∞.

6. Main results. In this section we use the results in the previous sections to derive necessary and sufficient conditions for the generic solvability of our version of the disturbance decoupling problem with pole placement and describe a graph theoretic procedure for checking the conditions.

We assume that we are given a structured system as in (1) and consider the

bipartite graph G' constructed from the matrix pencil

$$[M(s), N(s)] = \left[\begin{array}{cc|c} A - sE & B & Q \\ H & 0 & 0 \end{array}\right].$$

In G' we can distinguish a bipartite subgraph G corresponding to the pencil $M(s)$. Moreover, we can distinguish in G the bipartite subgraphs G_a and G_b. The subgraph G_a corresponds to the "pencil" E (or sE) and the subgraph G_b to the pencil $[A - sE, B]$.

With G_a we can check whether or not the matrix E is generically invertible, which is the case if and only if the graph G_a contains a matching of order n. If E is generically invertible then we can use the graph G_b to check whether or not system (1) is generically controllable. For this the DM-decomposition of the graph G_b may be used (cf. [13]). If the system (1) is generically controllable we proceed as follows.

GENERIC INDICES OF $M(s)$. First, we find a maximum matching in G and we denote its order by r. Then by Proposition 5.1 the pencil $M(s)$ has generic rank r. Using the obtained maximum matching we now can compute the DM-decomposition for the bipartite graph G. Let it consist of the components (bipartite subgraphs) G_0, G_*, and G_∞ as explained in §5. Next we compute the sum of the minimum ζ^--weight of a maximum matching in G_0, the maximum ζ^+-weight of a maximum matching in G_*, and the minimum ζ^--weight of a maximum matching in G_∞ (see Theorem 5.6). We denote the obtained number by Λ. Finally, in the full graph G we compute the maximum ζ^+-weight of a maximum matching and denote the obtained maximum by Δ (see Theorem 5.3 or Remark 5.7).

GENERIC INDICES OF $[M(s), N(s)]$. Subsequently, we consider the full bipartite graph G' and consider the above matching in G as a matching in G', possibly not being maximum. Using this matching as a starting point we can find a maximum matching in G' and we denote its order by r'. By Proposition 5.1 the pencil $[M(s), N(s)]$ has generic rank r'. Now if $r < r'$ we can stop because our version of the disturbance decoupling problem with pole placement will not be generically solvable (see §4.2). If $r = r'$ we continue with computing the DM-decomposition for G'. Let it consist of the components G'_0, G'_*, and G'_∞. Now in G', as before, we compute the sum of the minimum ζ^--weight of a maximum matching in G'_0, the maximum ζ^+-weight of a maximum matching in G'_*, and the minimum ζ^--weight of a maximum matching in G'_∞. We denote the obtained number by Λ'. Finally, in G' we compute the maximum ζ^+-weight of a maximum matching and denote the obtained maximum by Δ'. We again refer to Theorems 5.3 and 5.6 and Remark 5.7.

Then we have the following graph-theoretic characterization for the generic solvability of DDPPP'.

THEOREM 6.1. *Assume that system* (1) *is generically controllable.* DDPPP' *for system* (1) *is generically solvable if and only if* $r = r'$, $\Lambda = \Lambda'$, *and* $\Delta = \Delta'$.

We stress that we concentrated only on the solvability of the problem and that we were not concerned in doing this in an efficient way. Hence, it may be possible that the computations described above can be done much more efficiently.

7. Illustrative examples. In this section we illustrate the results of this paper.

EXAMPLE 7.1. We consider the structured system of type (1) given by the next matrices.

$$E = \begin{bmatrix} 0 & \lambda_1 & 0 \\ \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 0 & \lambda_4 \\ 0 & \lambda_5 & 0 \\ 0 & 0 & \lambda_6 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ \lambda_7 & 0 \\ 0 & \lambda_8 \end{bmatrix}, \quad Q = \begin{bmatrix} \lambda_9 \\ \lambda_{10} \\ \lambda_{11} \end{bmatrix},$$

$$H = \begin{bmatrix} \lambda_{12} & 0 & 0 \end{bmatrix}.$$

First, observe that E is generically invertible. Next, note that the system is generically controllable and that DDPPP′ is generically solvable. Indeed, premultiplying A, B, and Q by the inverse of E we obtain

$$\hat{A} = E^{-1}A = \begin{bmatrix} 0 & \mu_1 & 0 \\ 0 & 0 & \mu_2 \\ 0 & 0 & \mu_3 \end{bmatrix}, \quad \hat{B} = E^{-1}B = \begin{bmatrix} \mu_4 & 0 \\ 0 & 0 \\ 0 & \mu_5 \end{bmatrix}, \quad \hat{Q} = E^{-1}Q = \begin{bmatrix} \mu_6 \\ \mu_7 \\ \mu_8 \end{bmatrix},$$

$$\hat{H} = \begin{bmatrix} \mu_9 & 0 & 0 \end{bmatrix},$$

with $\mu_1 = \frac{\lambda_5}{\lambda_2}$, $\mu_2 = \frac{\lambda_4}{\lambda_1}$, $\mu_3 = \frac{\lambda_6}{\lambda_3}$, $\mu_4 = \frac{\lambda_7}{\lambda_2}$, $\mu_5 = \frac{\lambda_8}{\lambda_3}$, $\mu_6 = \frac{\lambda_{10}}{\lambda_2}$, $\mu_7 = \frac{\lambda_9}{\lambda_1}$, $\mu_8 = \frac{\lambda_{11}}{\lambda_3}$, $\mu_9 = \lambda_{12}$. The pair $(\hat{A}, \hat{B})$is clearly controllable for almost all λ. Furthermore, consider the feedback matrices

$$F = \begin{bmatrix} -\frac{\alpha}{\mu_4} & -\frac{\mu_1}{\mu_4} & 0 \\ 0 & -\frac{\beta}{\mu_2\mu_5} & -\frac{\gamma+\mu_3}{\mu_5} \end{bmatrix}, \quad R = \begin{bmatrix} -\frac{\mu_6}{\mu_4} \\ 0 \end{bmatrix}.$$

Then

$$\hat{A} + \hat{B}F = \begin{bmatrix} -\alpha & 0 & 0 \\ 0 & 0 & \mu_2 \\ 0 & -\frac{\beta}{\mu_2} & -\gamma \end{bmatrix}, \qquad \hat{Q} + \hat{B}R = \begin{bmatrix} 0 \\ \mu_7 \\ \mu_8 \end{bmatrix},$$

from which it easily follows that $H(sE-(A+BF))^{-1}\,(Q+BR) = H(sI-(\hat{A}+\hat{B}F))^{-1}\,(\hat{Q}+\hat{B}R) = 0$ and $\det\,(sE-(A+BF)) = -\lambda_1\lambda_2\lambda_3\,(s+\alpha)\,(s^2+\gamma s+\beta)$, for almost all λ. Clearly, the system is generically disturbance decoupled and by suitably chosen values for α, β, and γ the poles of the closed loop system can be placed anywhere in the complex plane. Hence, DDPPP′ is generically solvable.

We now show that the generic solvability of DDPPP′ can also be established by means of the graph-theoretic methods in this paper. Therefore, we recall that

$$M(s) = \begin{bmatrix} 0 & \lambda_1 s & \lambda_4 & 0 & 0 \\ \lambda_2 s & \lambda_5 & 0 & \lambda_7 & 0 \\ 0 & 0 & \lambda_3 s + \lambda_6 & 0 & \lambda_8 \\ \lambda_{12} & 0 & 0 & 0 & 0 \end{bmatrix}, \qquad N(s) = \begin{bmatrix} \lambda_9 \\ \lambda_{10} \\ \lambda_{11} \\ 0 \end{bmatrix},$$

and, consequently,

$$[M(s), N(s)] = \begin{bmatrix} 0 & \lambda_1 s & \lambda_4 & 0 & 0 & \lambda_9 \\ \lambda_2 s & \lambda_5 & 0 & \lambda_7 & 0 & \lambda_{10} \\ 0 & 0 & \lambda_3 s + \lambda_6 & 0 & \lambda_8 & \lambda_{11} \\ \lambda_{12} & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Using this matrix pencil the following bipartite graph G' with $\mathcal{W}' = \{1,2,3,4,5,6\}$ and $\mathcal{V}' = \{1,2,3,4\}$ can easily be constructed as in Fig. 2 (see also §5).

In the bipartite graph in Fig. 2 as well as in the bipartite graphs in the remainder of this paper the direction of the arcs is not depicted for reasons of clarity. It should however be kept in mind that in all bipartite (sub)graphs the arcs are directed from (a subset of) $\mathcal{W}'$ (or $\mathcal{W}$) to (a subset of) $\mathcal{V}'$ (or $\mathcal{V}$).

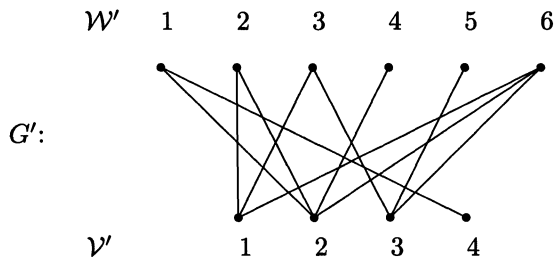


FIG. 2. *Bipartite graph of Example* 7.1.

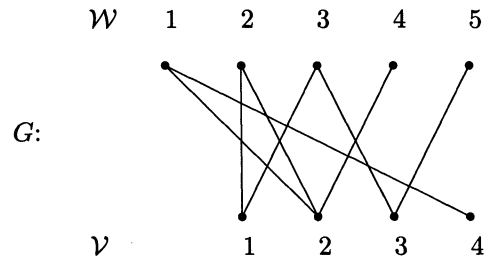


FIG. 3. *Bipartite subgraph of Example* 7.1.

It follows from $[M(s), N(s)]$ that the arc weights in G' are given as in the matrices

$$\Psi^+ = \begin{bmatrix} . & 1 & 0 & . & . & 0 \\ 1 & 0 & . & 0 & . & 0 \\ . & . & 1 & . & 0 & 0 \\ 0 & . & . & . & . & . \end{bmatrix}, \qquad \Psi^- = \begin{bmatrix} . & 1 & 0 & . & . & 0 \\ 1 & 0 & . & 0 & . & 0 \\ . & . & 0 & . & 0 & 0 \\ 0 & . & . & . & . & . \end{bmatrix}.$$

Here the dot (.) indicates that the entry in $[M(s), N(s)]$ is a fixed zero and does not give rise to an arc in the graph. The remaining values in Ψ^+ correspond to the exponents of the highest power in s of the associated entries in $[M(s), N(s)]$. Likewise for Ψ^- and the lowest powers in s.

The graph G' represents the matrix pencil $[M(s), N(s)]$. The graph G representing only the matrix $M(s)$ can be obtained simply from the above graph by deleting the initial vertex 6 in $\mathcal{W}'$ and all arcs starting from this vertex. The graph G with $\mathcal{W} = \{1, 2, 3, 4, 5\}$ and $\mathcal{V} = \mathcal{V}'$ can be depicted as in Fig. 3.

To illustrate the methods of this paper we follow the steps explained in §6. Moreover, to focus on our main results we assume that we already know that the matrix E is generically invertible and that the system is generically controllable. If required, these facts can be easily verified by the graph-theoretic methods in [13].

We note that the set $\{(1,4), (2,1), (3,3), (4,2)\}$ is a matching of order 4 in G as well as in G'. Since the set of terminal edges of G and G', i.e., $\mathcal{V}$ and $\mathcal{V}'$, both consist of four vertices, the above matching is also a maximum matching in both graphs. By Proposition 5.1 we therefore know that generic-rank $M(s) = r = 4$ and

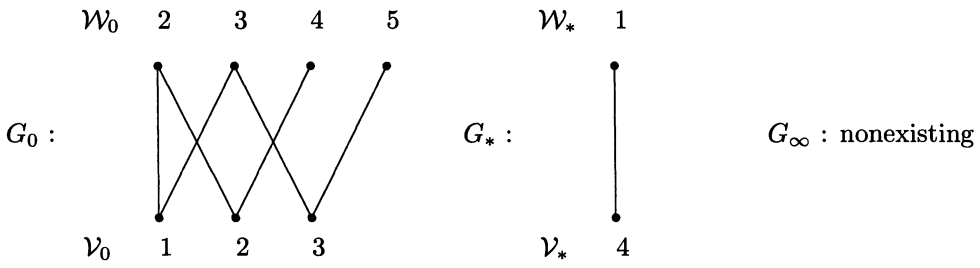


FIG. 4. DM-*decomposition of bipartite subgraph of Example* 7.1.

the generic-rank $[M(s), N(s)] = r' = 4$. Furthermore, we note that the ζ^+-weight of the above matching equals $0 + 1 + 1 + 0 = 2$ and the ζ^--weight $0 + 1 + 0 + 0 = 1$. This simply follows from the weight matrices Ψ^+ and Ψ^- by adding the values on the places $(4,1)$, $(1,2)$, $(3,3)$, and $(2,4)$.

Using the above matching we can also compute the DM-decomposition of both G and G'. For details on this computation we refer to [13]. It turns out that, in terms of §5, the DM-decomposition of G results in a partitioning of $\mathcal{W}$ and $\mathcal{V}$ with $\mathcal{W}_0 = \{2,3,4,5\}$, $\mathcal{W}_* = \{1\}$, $\mathcal{W}_\infty = \emptyset$ and $\mathcal{V}_0 = \{1,2,3\}$, $\mathcal{V}_* = \{4\}$, $\mathcal{V}_\infty = \emptyset$. This implies that the DM-decomposition of G is made up of a horizontal tail G_0 and a consistent part G_*, but not of a vertical tail G_∞. The DM-decomposition of G can be associated with row and column permutations of $M(s)$ that result in the following block triangular matrix:

$$\left[\begin{array}{cccc|c} \lambda_1 s & \lambda_4 & 0 & 0 & 0 \\ \lambda_5 & 0 & \lambda_7 & 0 & \lambda_2 s \\ 0 & \lambda_3 s + \lambda_6 & 0 & \lambda_8 & 0 \\ -- & -- & -- & -- & -- \\ 0 & 0 & 0 & 0 & \lambda_{12} \end{array}\right].$$

(Note that in fact only the first column of $M(s)$ is put as last.) In terms of the bipartite subgraphs G_0, G_*, and G_∞ we have a situation as in Fig. 4.

We recall the property of G_0 that for every initial vertex w in $\mathcal{W}_0$ there is a maximum order (= 3rd order) matching in the graph G_0 that does not contain w. This property is fundamental in Propositions 5.4 and 5.5. (See also the Appendix.)

The weights associated to G_0 and G_* are given by the matrices.

$$\Psi_0^+ = \begin{bmatrix} 1 & 0 & . & . \\ 0 & . & 0 & . \\ . & 1 & . & 0 \end{bmatrix}, \quad \Psi_0^- = \begin{bmatrix} 1 & 0 & . & . \\ 0 & . & 0 & . \\ . & 0 & . & 0 \end{bmatrix}, \quad \Psi_*^+ = 0, \Psi_*^- = 0.$$

The graph G_∞ does not exist and therefore no arc weights need to be specified. We note that there is only one (maximum) matching in G_* and that it has ζ^+-weight 0 and ζ^--weight 0. In G_0 there are several maximum matchings. A simple inspection shows that the maximum ζ^+-weight of a maximum matching in G_0 equals 2 (consider the matching $\{(2,1), (3,3), (4,2)\}$) and the minimum ζ^--weight of a maximum matching is 0 (take the matching $\{(3,1), (4,2), (5,3)\}$). By the results of §5 it now follows that the generic value of $\Lambda_r(M(s))$ is 0 (= Λ) and the generic value of $\Delta_r(M(s))$ is 2 (= Δ) where $r = 4$.

Starting from the previous matching the DM-decomposition of the full bipartite graph G' can be found. It turns out that this DM-decomposition implies that the matrix $[M(s), N(s)]$ can be rearranged as follows

$$\left[\begin{array}{ccccc|cc} \lambda_1 s & \lambda_4 & 0 & 0 & \lambda_9 & | & 0 \\ \lambda_5 & 0 & \lambda_7 & 0 & \lambda_{10} & | & \lambda_2 s \\ 0 & \lambda_3 s + \lambda_6 & 0 & \lambda_8 & \lambda_{11} & | & 0 \\ -- & -- & -- & -- & -- & -- & -- \\ 0 & 0 & 0 & 0 & 0 & | & \lambda_{12} \end{array}\right].$$

(Also now the first column is put as last.) In a similar fashion as above we can now show that the generic value of $\Lambda_{r'}([M(s), N(s)])$ is 0 $(= \Lambda')$ and the generic value of $\Delta_{r'}([M(s), N(s)])$ is 2 $(= \Delta')$ with $r' = 4$.

By Theorem 6.1 it now follows that DDPPP$'$ is generically solvable.

EXAMPLE 7.2. In the present example we study the system of Example 7.1 again. However, now we are interested in the generic solvability of DDPPP as formulated in §4. In the spirit of that section we define the DDPPP to be generically solvable precisely when $\text{rank} M_\lambda(s) = \text{rank}[M_\lambda(s), N_\lambda(s)] =: r$, $\Lambda_r(M_\lambda(s)) = \Lambda_r([M_\lambda(s), N_\lambda(s)])$, $\Delta_r(M_\lambda(s)) = \Delta_r([M_\lambda(s), sN_\lambda(s)])$ for almost all $\lambda \in \mathbb{R}^{12}$. Hence, DDPPP for a structured system is generically solvable if the problem can be solved for almost all values of its nonzero coefficients.

Using the geometric techniques of [30] with $\hat{A}$, $\hat{B}$, $\hat{Q}$, and $\hat{H}$ as in Example 7.1 it follows from $\hat{H}\hat{Q} = \mu_6\mu_9 = \lambda_{10}\lambda_{12}/\lambda_2$ that DDPPP for the system under consideration is generically unsolvable (=not generically solvable).

The latter conclusion can also be obtained with the graph-theoretic methods of this paper. In applying these methods we may use the decompositions of Example 7.1 since the zero-nonzero structure of the pencils $[M(s), N(s)]$ and $[M(s), sN(s)]$ are identical. The only difference in the graphs associated to $[M(s), N(s)]$ and $[M(s), sN(s)]$ is the arc weights. The weights for the arcs of the graph associated to $[M(s), sN(s)]$ follow from

$$[M(s), sN(s)] = \left[\begin{array}{cccccc} 0 & \lambda_1 s & \lambda_4 & 0 & 0 & \lambda_9 s \\ \lambda_2 s & \lambda_5 & 0 & \lambda_7 & 0 & \lambda_{10} s \\ 0 & 0 & \lambda_3 s + \lambda_6 & 0 & \lambda_8 & \lambda_{11} s \\ \lambda_{12} & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

and are given by

$$\tilde{\Psi}^+ = \left[\begin{array}{cccccc} . & 1 & 0 & . & . & 1 \\ 1 & 0 & . & 0 & . & 1 \\ . & . & 1 & . & 0 & 1 \\ 0 & . & . & . & . & . \end{array}\right], \qquad \tilde{\Psi}^- = \left[\begin{array}{cccccc} . & 1 & 0 & . & . & 1 \\ 1 & 0 & . & 0 & . & 1 \\ . & . & 0 & . & 0 & 1 \\ 0 & . & . & . & . & . \end{array}\right].$$

We note that when using the decompositions of the graphs of Example 7.1 the above weights for the graph G coincide with the weights given by Ψ^+ and Ψ^-. The results of example 7.1 therefore yield that the generic-rank $M(s) = r = 4$, the generic value of $\Lambda_r(M(s))$ is 0, and the generic value of $\Delta_r(M(s))$ is 2. Moreover, since it does not depend on the arc weights we also have that the generic-rank $[M(s), sN(s)] = r' = 4$. However, using the weights given by $\tilde{\Psi}^+$ and $\tilde{\Psi}^-$, it follows that the generic value of $\Lambda_{r'}([M(s), sN(s)])$ is 0 and the generic value of $\Delta_{r'}([M(s), sN(s)])$ is 3 with $r' = 4$.

Comparing the obtained generic values we can conclude by the graph-theoretical methods of this paper that DDPPP is generically unsolvable for this system.

EXAMPLE 7.3. As a final example we consider the structured system given by the matrices

$$E = \begin{bmatrix} \lambda_1 & 0 & \lambda_2 \\ 0 & \lambda_3 & 0 \\ \lambda_4 & 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 0 & \lambda_5 \\ 0 & 0 & 0 \\ 0 & \lambda_6 & \lambda_7 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & \lambda_8 \\ \lambda_9 & \lambda_{10} \\ 0 & \lambda_{11} \end{bmatrix}, \quad Q = \begin{bmatrix} \lambda_{12} \\ 0 \\ \lambda_{13} \end{bmatrix},$$

$$H = \begin{bmatrix} \lambda_{14} & \lambda_{15} & 0 \\ \lambda_{16} & 0 & 0 \\ \lambda_{17} & 0 & 0 \end{bmatrix}.$$

We note that E is generically invertible and that the system is generically controllable. Indeed, premultiplying A, B, and Q by the inverse of E we obtain

$$\hat{A} = E^{-1}A = \begin{bmatrix} 0 & \mu_1 & \mu_2 \\ 0 & 0 & 0 \\ 0 & \mu_3 & \mu_4 \end{bmatrix}, \quad \hat{B} = E^{-1}B = \begin{bmatrix} 0 & \mu_5 \\ \mu_6 & \mu_7 \\ 0 & \mu_8 \end{bmatrix}, \quad \hat{Q} = E^{-1}Q = \begin{bmatrix} \mu_9 \\ 0 \\ \mu_{10} \end{bmatrix},$$

$$\hat{H} = \begin{bmatrix} \mu_{11} & \mu_{12} & 0 \\ \mu_{13} & 0 & 0 \\ \mu_{14} & 0 & 0 \end{bmatrix},$$

with $\mu_1 = \frac{\lambda_6}{\lambda_4}$, $\mu_2 = \frac{\lambda_7}{\lambda_4}$, $\mu_3 = -\frac{\lambda_1\lambda_6}{\lambda_2\lambda_4}$, $\mu_4 = \frac{\lambda_5}{\lambda_2} - \frac{\lambda_1\lambda_7}{\lambda_2\lambda_4}$, $\mu_5 = \frac{\lambda_{11}}{\lambda_4}$, $\mu_6 = \frac{\lambda_9}{\lambda_3}$, $\mu_7 = \frac{\lambda_{10}}{\lambda_3}$, $\mu_8 = \frac{\lambda_8}{\lambda_2} - \frac{\lambda_1\lambda_{11}}{\lambda_2\lambda_4}$, $\mu_9 = \frac{\lambda_{13}}{\lambda_4}$, $\mu_{10} = \frac{\lambda_{12}}{\lambda_2} - \frac{\lambda_1\lambda_{13}}{\lambda_2\lambda_4}$, $\mu_{11} = \lambda_{14}$, $\mu_{12} = \lambda_{15}$, $\mu_{13} = \lambda_{16}$, and $\mu_{14} = \lambda_{17}$. By a straightforward calculation we can prove that the pair $(\hat{A}, \hat{B})$ is controllable generically in the original parameters λ_i, $i = 1, 2, \ldots, 17$. Furthermore, from $\hat{A}$, $\hat{B}$, $\hat{Q}$, and $\hat{H}$ it follows that generically $\operatorname{Ker} \hat{H} \cap \operatorname{Im} \hat{B} = \{0\}$, $\operatorname{Ker} \hat{H} + \operatorname{Im} \hat{B} = \mathbb{R}^3$, and $\operatorname{Im} Q \not\subseteq \operatorname{Im} B$ (or $\operatorname{Im} \hat{Q} \not\subseteq \operatorname{Im} \hat{B}$). Using the methods of [30] we can easily prove that generically $\mathcal{V}^*(\operatorname{Ker}\hat{H}; \hat{A}, \hat{B}) = \operatorname{Ker} \hat{H}$ and $\mathcal{R}^*(\operatorname{Ker}\hat{H}; \hat{A}, \hat{B}) = \{0\}$. By Corollary 2.2 it therefore follows that DDPPP′ is generically unsolvable (= not generically solvable). However, DDPPP′ without the pole placement is generically solvable. Following [29] the latter problem is abbreviated as DDP′. Using the well-known results of [30] (see also [29]) it follows easily that DDP′ is indeed generically solvable since $\operatorname{Im} \hat{Q} \subseteq \mathcal{V}^*(\operatorname{Ker}\hat{H}; \hat{A}, \hat{B}) + \operatorname{Im} \hat{B}$ for almost all λ. By the results of §§2 and 3 the latter subspace inclusion holds for an arbitrary λ if and only if $\operatorname{rank} M_\lambda(s) = \operatorname{rank} [M_\lambda(s), N_\lambda(s)] =: r$, and $\Delta_r(M_\lambda(s)) = \Delta_r([M_\lambda(s), N_\lambda(s)])$.

Therefore, it is clear that the previous conclusion on the generic solvability of DDP′ and the generic unsolvability of DDPPP′ can be drawn again by studying the bipartite graphs associated to $M(s)$ and $[M(s), N(s)]$ (see also [17], [31]). Recall that

$$M(s) = \begin{bmatrix} \lambda_1 s & 0 & \lambda_2 s + \lambda_5 & 0 & \lambda_8 \\ 0 & \lambda_3 s & 0 & \lambda_9 & \lambda_{10} \\ \lambda_4 s & \lambda_6 & \lambda_7 & 0 & \lambda_{11} \\ \lambda_{14} & \lambda_{15} & 0 & 0 & 0 \\ \lambda_{16} & 0 & 0 & 0 & 0 \\ \lambda_{17} & 0 & 0 & 0 & 0 \end{bmatrix}, \quad N(s) = \begin{bmatrix} \lambda_{12} \\ 0 \\ \lambda_{13} \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

As before we denote the bipartite graph associated to $M(s)$ by G with $\mathcal{W} = \{1,2,3,4,5\}$ and $\mathcal{V} = \{1, 2, 3, 4, 5, 6\}$. This graph can be depicted as in Fig. 5.

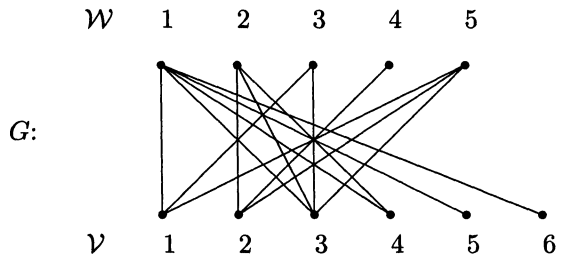


FIG. 5. *Bipartite subgraph of Example* 7.3.

It is easy to see that a maximum matching in G has order 5. The DM-decomposition of G results in a partioning of $\mathcal{W}$ and $\mathcal{V}$ with $\mathcal{W}_0 = \emptyset$, $\mathcal{W}_* = \{2,3,4,5\}$, $\mathcal{W}_\infty = \{1\}$, and $\mathcal{V}_0 = \emptyset$, $\mathcal{V}_* = \{1,2,3,4\}$, $\mathcal{V}_\infty = \{5,6\}$. This means that the DM-decomposition is made of a consistent part G_* and a vertical tail G_∞, but not of a horizontal tail G_0 (see §5). The DM-decomposition of G can be associated with row and column permutations for $M(s)$ that result in the following block triangular matrix

$$\left[\begin{array}{cccc|c} 0 & \lambda_2 s+\lambda_5 & 0 & \lambda_8 & \lambda_1 s \\ \lambda_3 s & 0 & \lambda_9 & \lambda_{10} & 0 \\ \lambda_6 & \lambda_7 & 0 & \lambda_{11} & \lambda_4 s \\ \lambda_{15} & 0 & 0 & 0 & \lambda_{14} \\ \hline 0 & 0 & 0 & 0 & \lambda_{16} \\ 0 & 0 & 0 & 0 & \lambda_{17} \end{array}\right]$$

Using this matrix the weights of the arcs in the consistent part G_* and the vertical tail G_∞ are immediate. With these weights it easily follows that the maximum ζ^+-weight of a maximum (=4th order) matching in G_* equals 1 and the minimum ζ^--weight of such a maximum matching equals 0. For the vertical tail both the maximum ζ^+-weight and the minimum ζ^--weight of a maximum (= 1st order) matching are equal to 0.

Hence, by the results of §5 we can conclude that the generic-rank $M(s) = r = 5$, the generic value of $\Lambda_r(M(s))$ is 1 (= Λ) and the generic value of $\Delta_r(M(s))$ is 1 (= Δ).

Next we consider the pencil

$$[M(s), N(s)] = \left[\begin{array}{cccccc} \lambda_1 s & 0 & \lambda_2 s+\lambda_5 & 0 & \lambda_8 & \lambda_{12} \\ 0 & \lambda_3 s & 0 & \lambda_9 & \lambda_{10} & 0 \\ \lambda_4 s & \lambda_6 & \lambda_7 & 0 & \lambda_{11} & \lambda_{13} \\ \lambda_{14} & \lambda_{15} & 0 & 0 & 0 & 0 \\ \lambda_{16} & 0 & 0 & 0 & 0 & 0 \\ \lambda_{17} & 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

We denote the bipartite graph associated to $[M(s), N(s)]$ by G' with $\mathcal{W}' = \{1,2,3,4,5,6\}$ and $\mathcal{V}' = \{1,2,3,4,5,6\}$. This graph can be depicted as in Fig. 6.

Also here it follows easily that a maximum matching in G' has order 5. The DM-decomposition of G' results in a partioning of $\mathcal{W}'$ and $\mathcal{V}'$ with $\mathcal{W}'_0 = \{3,4,5,6\}$, $\mathcal{W}'_*$

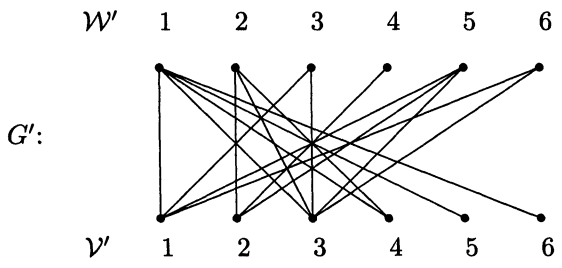


FIG. 6. *Bipartite graph of Example* 7.3.

$= \{2\}$, $\mathcal{W}'_\infty = \{1\}$, and $\mathcal{V}'_0 = \{1,2,3\}$, $\mathcal{V}'_* = \{4\}$, $\mathcal{V}'_\infty = \{5,6\}$. This means that the DM-decomposition of G' is made up of a horizontal tail G'_0, a consistent part G'_* and a vertical tail G'_∞. Note the difference in the DM-decompositions of G and G'. The DM-decomposition of G', can be associated with row and column permutations of $[M(s), N(s)]$ that yield the following block triangular matrix

$$\left[\begin{array}{cccc|c|c|c} \lambda_2 s + \lambda_5 & 0 & \lambda_8 & \lambda_{12} & 0 & \lambda_1 s \\ 0 & \lambda_9 & \lambda_{10} & 0 & \lambda_3 s & 0 \\ \lambda_7 & 0 & \lambda_{11} & \lambda_{13} & \lambda_6 & \lambda_4 s \\ \hline 0 & 0 & 0 & 0 & \lambda_{15} & \lambda_{14} \\ \hline 0 & 0 & 0 & 0 & 0 & \lambda_{16} \\ 0 & 0 & 0 & 0 & 0 & \lambda_{17} \end{array}\right].$$

Using the above matrix the weights of the arcs in the horizontal tail G'_0, the consistent part G'_*, and the vertical tail G'_∞ can be obtained directly. With these weights it easily follows that the maximum ζ^+-weight of a maximum (=3rd order) matching in G'_0 equals 1 and the minimum ζ^--weight of such a maximum matching equals 0. For the consistent part G'_* both the maximum ζ^+-weight and the minimum ζ^--weight of a maximum (=1st order) matching are equal to 0. The same holds true for the vertical tail G'_∞.

By the results of §5 we can conclude that the generic-rank $[M(s), N(s)] = r'$ $= 5$, the generic value of $\Lambda_{r'}([M(s), N(s)])$ is 0 $(= \Lambda')$ and the generic value of $\Delta_{r'}([M(s), N(s)])$ is 1 $(= \Delta')$.

Since $r = r'$, $\Delta = \Delta'$, but $\Lambda \neq \Lambda'$, we can conclude by Theorem 6.1 that DDPPP′ is generically unsolvable. However, because of the first two equalities DDP′ is generically solvable (cf. [17], [31]).

8. Remarks and conclusions. In this paper we have studied a general version of the well-known disturbance decoupling problem for regular descriptor systems of which only the zero-nonzero structure of the system matrices is known. We represent this zero-nonzero structure by means of bipartite graphs. For the development of results, the so-called DM-decomposition for bipartite graphs is used.

The main results of this paper are necessary and sufficient conditions for the generic solvability of our version of the disturbance decoupling problem. Conditions

for the generic solvability of the more common version of the disturbance decoupling problem can be easily obtained in a similar fashion.

We want to stress that the results in this paper only deal with solvability issues. Hence, using the obtained conditions we can only say something on the existence of a feedback solving DDPPP′. No statements are given on how the feedback matrices F and R look or how they can be computed. A similar situation occurs when studying the disturbance decoupling without any pole placement requirement. For an attempt to compute the feedback matrices for the disturbance decoupling problem using the structure of the system as much as possible, we refer to [33].

The developments in this paper are largely based on fundamental results in [13], especially on Proposition 14.5. Basically, this proposition is a special case of the results in [14] in which the Smith normal form for structured polynomial matrices is developed. See also [15] and [16].

By using the results of [14] the arguments of this paper can also be extended (as in [13] and [17]) for systems of which the nonzero elements are divided into fixed constants and free parameters. Then structured matrices (in the present case) are replaced by mixed matrices, bipartite graphs by independent matchings (involving bipartite graphs and linear matroids), and the DM-decomposition by the Combinatorial Canonical form (CCF) of layered mixed matrices.

Appendix. *Proof of Theorem* 3.1. Statements 1 and 2 can each be proved much in a similar way. Therefore, we only prove here statement 1. A proof of statement 2 can also be found in [27].

(Only if part) Denote

$$Q(s) = \begin{bmatrix} I & -X(s) \\ 0 & I \end{bmatrix}$$

with I the identity matrix of suitable dimensions. Since $X(s)$ is a polynomial matrix and the determinant of $Q(s)$ is one, $Q(s)$ is a unimodular polynomial matrix. Note that rank $U(s)$ = rank $[U(s), 0]$ = rank $[U(s), V(s)]Q(s)$ = rank $[U(s), V(s)]$. Let r = rank $U(s)$. Then $\Lambda_r(U(s)) = \Lambda_r([U(s), 0]) = \Lambda_r([U(s), V(s)]Q(s)) = \Lambda_r([U(s), V(s)])$.

(If part) To prove that (4) is solvable over $\mathbb{R}[s]$ we may assume without loss of generality that

$$U(s) = \begin{bmatrix} \operatorname{diag}(\alpha_1(s), \ldots, \alpha_r(s)) & 0 \\ 0 & 0 \end{bmatrix}, \qquad V(s) = \begin{bmatrix} V_1(s) \\ V_2(s) \end{bmatrix}, \tag{5}$$

with polynomial matrices $V_1(s) \in \mathbb{R}^{r \times c}[s]$, $V_2(s) \in \mathbb{R}^{(b-r) \times c}[s]$, and monic polynomials $\alpha_1(s), \ldots, \alpha_r(s)$ such that $\alpha_i(s)$ divides $\alpha_{i+1}(s)$, $1 \leq i < r$. Because rank $U(s)$ = rank $[U(s), V(s)] = r$ it follows that $V_2(s) = 0$. We write $V_1(s) = (v_{ij}(s))$, $1 \leq i \leq r, 1 \leq j \leq c$.

Let us fix i and j ($1 \leq i \leq r$, $1 \leq j \leq c$) and consider the $r \times r$ submatrix of $[U(s), V(s)]$ made up of the rows 1 to r, and the jth column of $V(s)$ and the first r columns of $U(s)$ excepting the ith. This matrix has a determinant equal to $\pm v_{ij}(s) \prod_{k \neq i} \alpha_k(s)$. Another nonzero rth order minor of $[U(s), V(s)]$ is $\prod_{k=1}^{r} \alpha_k(s)$, which corresponds to the $r \times r$ submatrix with the first r rows and columns of $U(s)$. The greatest common divisor of the above two rth order minors is gcd $(\alpha_i(s), v_{ij}(s)) \prod_{k \neq i} \alpha_k(s)$.

Since $\Lambda_r([U(s), V(s)])$ is the degree of the greatest common divisor of *all* the rth order minors of $[U(s), V(s)]$, it follows that $\Lambda_r([U(s), V(s)]) \leq \sum_{k \neq i} \deg \alpha_k(s) + \deg$

$(\gcd(\alpha_i(s), v_{ij}(s))) \leq \sum_{k=1}^{r} \deg \alpha_k(s) = \Lambda_r(U(s))$. However, we have $\Lambda_r([U(s), V(s)])$ $= \Lambda_r(U(s))$ by the assumption. Therefore, $\deg(\gcd(\alpha_i(s), v_{ij}(s))) = \deg \alpha_i(s)$, or in other words, $\alpha_i(s)$ divides $v_{ij}(s)$.

Hence, we can write $v_{ij}(s) = \alpha_i(s)\, X_{ij}(s)$ with $X_{ij}(s) \in \mathbb{R}[s]$, $1 \leq i \leq r, 1 \leq j \leq c$. For i, j with $r < i \leq b, 1 \leq j \leq c$, we let $X_{ij}(s)$ be an arbitrary polynomial. $X(s)$ is a polynomial matrix in $\mathbb{R}^{b \times c}[s]$ satisfying (4) with $U(s)$ and $V(s)$ as in (5). □

Proof of Proposition 5.4. An alternative proof of Proposition 5.4 is given here. In view of the characterization of the horizontal tail mentioned before Proposition 5.4, it suffices to prove the following statement.

PROPOSITION A.1. *Let $T(s) = \Phi + s\Theta$ be a $k \times l$ matrix pencil (with $k < l$) such that the entries in the coefficient matrices Φ and Θ are either fixed to zero or independent parameters. If any $k \times (l-1)$ submatrix of $T(s)$ has generic* rank *k, then generically the greatest common divisor of all the kth order minors of $T(s)$ is a monomial in s.*

Proof. Let ϕ and θ denote the vectors of the nonzero entries of Φ and Θ, respectively. Then a kth order minor, say $f(s)$, of $T(s)$ may be regarded also as a polynomial in s, ϕ, and θ, i.e., $f(s) = f(s, \phi, \theta) \in \mathbb{Q}[s, \phi, \theta]$, where $\mathbb{Q}$ denotes the (field of) rational numbers.

Let us denote by $g_1(s, \phi, \theta)$ the gcd of all the kth order minors when considered in $\mathbb{Q}(\phi, \theta)[s]$ (= set of polynomials in s with rational functions in (ϕ, θ) as the coefficients), and by $g_2(s, \phi, \theta)$ the gcd of those minors when considered in $\mathbb{Q}[s, \phi, \theta]$.

First we observe that $g_1(s, \phi, \theta) = h(\phi, \theta) g_2(s, \phi, \theta)$ for some h, which is a rational function in (ϕ, θ) with rational numbers as the coefficients. Hence it remains to show that $g_2(s, \phi, \theta)$ is a monomial in s.

Second, we claim that $g_2(s, \phi, \theta)$ is free from the parameters ϕ and θ. To prove this, suppose to the contrary that $g_2(s, \phi, \theta)$ contains a parameter, say ϕ_i for concreteness, among ϕ and θ. By the assumption, the submatrix of $T(s)$ that is obtained by deleting that column which contains ϕ_i has generic rank k, and hence there exists a nonzero kth order minor $f(s, \phi, \theta)$ of $T(s)$ that does not contain ϕ_i. This contradicts the fact that f has a factor g_2 which does contain ϕ_i. Thus we have shown that $g_2 \in \mathbb{Q}[s]$.

Finally, we claim that $g_2(s)$ is a monomial with a rational number as the coefficient. To prove this, take a nonzero kth order minor $f(s, \phi, \theta)$ of $T(s)$. Then $f(s, \phi, \theta) = g_2(s) f_1(s, \phi, \theta)$ for some $f_1(s, \phi, \theta) \in \mathbb{Q}[s, \phi, \theta]$. We may regard this relation as an identity in $\mathbb{Q}(s)[\phi, \theta]$. In particular we regard f as a polynomial in (ϕ, θ) with coefficients being rational functions in s. Using the defining expansion of a minor and from the structure of $T(s)$, we see that the coefficient of each term (= product of the elements of ϕ and θ) appearing in f must be a monomial in s and cannot be a general polynomial. Hence $g_2(s)$ cannot be a general polynomial but must be a monomial in s. □

REFERENCES

[1] N. ANDREI, *Sparse Systems, Digraph Approach of Large-Scale Linear Systems Theory*, Verlag TÜV Rheinland, Köln, 1985.

[2] R. A. BRUALDI AND H. J. RYSER, *Combinatorial Matrix Theory*, Cambridge University Press, London, 1991.

[3] C. COMMAULT, J. M. DION, AND M. BENAHCÈNE, *Output feedback disturbance decoupling–graph interpretation for structured systems*, Automatica, 29 (1993), pp. 1463–1472.

[4] C. COMMAULT, J. M. DION, AND A. PEREZ, *Disturbance rejection for structured systems*, IEEE Trans. Automat. Control, AC-36 (1991), pp. 884–887.

[5] J. DESCUSSE AND J. M. DION, *On the structure at infinity of linear square decoupled systems*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 971–974.
[6] F. GANTMACHER, *The Theory of Matrices*, Chelsea, New York, 1959.
[7] K. GLOVER AND L. M. SILVERMAN, *Characterization of structural controllability*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 354–537.
[8] M. L. J. HAUTUS, *Controllability and observability conditions of linear autonomous systems*, Proc. Kon. Ned. Akad. Wetensch., Ser. A, 72 (1969), pp. 443–448.
[9] ———, *(A,B)-invariant and stabilizability subspaces, a frequency domain description*, Automatica, 16 (1980), pp. 703–707.
[10] C. T. LIN, *Structural controllability*, IEEE Trans. Automat. Control, AC-19 (1974), pp. 201–208.
[11] A. LINNEMAN, *Decoupling of structured systems*, Systems Control Lett., 1 (1981), pp. 79–86.
[12] L. LOVÁSZ AND M. PLUMMER, *Matching Theory*, North-Holland, Amsterdam, 1986.
[13] K. MUROTA, *Systems Analysis by Graphs and Matroids–Structural Solvability and Controllability*, Algorithms and Combinatorics, Vol. 3, Springer-Verlag, New York, 1987.
[14] ———, *On the Smith normal form of structured polynomial matrices*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 747–765.
[15] ———, *On the Smith normal form of structured polynomial matrices*, II, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1103–1111.
[16] ———, *Mixed matrices–Irreducibility and decomposition*, in Combinatorial and Graph-Theoretical Problems in Linear Algebra, R. A. Brualdi, S. Friedland, V. Klee, eds., The IMA Volumes in Mathematics and its Applications, Vol. 50, Springer-Verlag, Berlin, 1993, pp. 39–71.
[17] K. MUROTA AND J. W. VAN DER WOUDE, *Structure at infinity of structured descriptor systems and its applications*, SIAM J. Control Optim., 29 (1991), pp. 878–894.
[18] M. NEWMAN, *Integral Matrices*, Academic Press, New York, 1972.
[19] Y. OHTA AND S. KODAMA, *Structural invertibility of transfer functions*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 818–819.
[20] K. J. REINSCHKE, *Multivariable Control, A Graph-theoretic Approach*, Springer-Verlag, New York, 1988.
[21] H. H. ROSENBROCK, *State-Space and Multivariable Theory*, Wiley, New York, 1970.
[22] R. W. SHIELDS AND J. B. PEARSON, *Structural controllability of multi-input linear systems*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 203–212.
[23] W. SÖTE, *Eine graphische Methode zur Ermittlung der Nullstellen in Mehrgrössensystemen*, Reglungstechnik, 28 (1980), pp. 346–348.
[24] N. SUDA, B. WAN, AND I. UENO, *The orders of infinite zeros of structured systems*, Trans. Soc. Instr. Control Engineers, 25 (1989), pp. 1062–1068.
[25] F. SVARICEK, *Graphentheoretische Ermittlung der Anzahl von strukturellen und streng strukturellen invarianten Nullstellen*, Automatisierungstechnik, 34 (1986), pp. 488–497.
[26] H. L. TRENTELMAN, *Almost Invariant Subspaces and High Gain Feedback*, CWI Tracts 29, Amsterdam, 1986.
[27] G. C. VERGHESE, *Infinite frequency behavior in generalized dynamical systems*, Ph.D. thesis, Dept. of Electrical Engineering, Stanford University, Stanford, CA, 1978.
[28] J. C. WILLEMS, *Almost invariant subspaces: an approach to high gain feedback design–Part* I: *almost controlled invariant subspaces*, IEEE Trans. Automat. Control, 26 (1981), pp. 235–252.
[29] J. C. WILLEMS AND C. COMMAULT, *Disturbance decoupling by measurement feedback with stability or pole placement*, SIAM J. Control Optim., 19 (1981), pp. 490–504.
[30] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, 3rd ed., Springer-Verlag, New York, 1985.
[31] J. W. VAN DER WOUDE, *On the structure at infinity of a structured system*, Linear Algebra Appl., 148 (1991), pp. 145–169.
[32] ———, *A graph theoretic characterization for the rank of the transfer matrix of a structured system*, Math. Control Signals Systems, 4 (1991), pp. 33–40.
[33] ———, *Graph theoretic methods for the computation of disturbance decoupling feedback matrices for structured systems*, Linear Algebra Appl., 196 (1994), pp. 139–162.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 943–953, July 1995

015

EXISTENCE OF EMS SOLUTIONS AND A PRIORI ESTIMATES*

GEOFF A. LATHAM†

Abstract. Solvability of the nonlinear EMS (estimate, maximize, smooth) equations in the nonnegative quadrant is established by the use of the Brouwer fixed point theorem and a priori estimates from Perron–Frobenius theory. Existence of solutions and of an a priori estimate are also proven for a generalization of the EMS equations. The a priori estimates illustrate the quantification shortcomings of the EMS algorithm and should be carefully considered both before applying the algorithm and in the choice of smoothing.

Key words. EMS solutions, Perron–Frobenius theory, a priori estimates, fixed points, EMS algorithm, EM algorithm, nonnegative linear systems, nonnegative matrices, maximum likelihood

AMS subject classifications. 15A48, 65F10, 65H10; 15A42, 54H25, 65H20, 65U05, 92C55

1. Introduction.

The need to solve nonnegative linear systems (which are sometimes large, sparse and inconsistent) of the form $P^t\boldsymbol{\theta} = \boldsymbol{n}^*$, where $P \in \mathbb{R}^{B\times D}$ and $\boldsymbol{n}^* \in \mathbb{R}^D$ are both nonnegative, is a recurring one in physical applications [25]. Moreover, any algorithm that naturally guarantees a nonnegative (physical) solution or approximate solution of such systems, without having to take special measures to ensure nonnegativity, has an automatic appeal to the practitioner. The EM (estimate, maximize) algorithm in the particular form

$$\theta_b^{(n+1)} = \frac{\theta_b^{(n)}}{r_b(P)} \sum_{d=1}^{D} \frac{n_d^*\, p_{bd}}{\sum_{\beta=1}^{B} p_{\beta d}\theta_\beta^{(n)}}, \quad b = 1,\dots,B;\ \ n = 0,1,2,\dots, \tag{1}$$

where $r_b(P)$ is the bth row sum of P, is one such algorithm which has been transported to many applications [3], [7], [24], [25] from within its encompassing framework of maximum likelihood estimation. Within this framework, EM gives a much wider methodology [4], however, the form (1) arises from the application of this methodology to the additive Poisson regression problem encountered in emission tomography [11], [22], [26], [23]. In fact, the EM algorithm's use in other disciplines [15], [19] predates this application by some ten years and EM is an instance of the interior method for nonlinear programming described even earlier in [5].

Although EM is easy to use and enjoys global convergence to a nonnegative maximum likelihood solution of the underlying linear system [26], [25], [3], this convergence is slow, noisy and, in the underdetermined case ($B > D$), lacks stability [14]. Even starting from a smooth initial iterate, EM typically produces a very "speckled" approximation when run too near to convergence [26]. As one means of overcoming these drawbacks, the EMS (estimate, maximize, and smooth) algorithm was proposed [23], [18] and entails the ad hoc introduction of a nonnegative smoothing step into (1);

$$\theta_b^{(n+1)} = \sum_{\beta=1}^{B} S_{b\beta}\, \frac{\theta_\beta^{(n)}}{r_\beta(P)} \sum_{d=1}^{D} \frac{n_d^*\, p_{\beta d}}{(P^t\boldsymbol{\theta}^{(n)})_d}, \quad b = 1,\dots,B;\ \ n = 0,1,2,\dots, \tag{2}$$

* Received by the editors August 16, 1993; accepted for publication (in revised form) by L. Kaufman June 27, 1994.

† Centre for Mathematics and Its Applications, Australian National University, Canberra ACT 0200 Australia (gal851@cisr.anu.edu.au).

where $S \in \mathbb{R}^{B \times B}$ is the smoothing matrix, with the broad aim of producing smoothed versions of maximum likelihood approximate solutions.

There is a real danger that, like its parent EM, EMS may be transported to other applications, or even worse, be incorporated in a piece of diagnostic medical equipment, simply because the resulting reconstructions appeal to the practitioner without regard for a proper mathematical assessment. It is the job of the applied mathematician to provide this assessment and so determine an algorithm's usefulness for applications. The work in [13], [14] is aimed precisely at providing a mathematical analysis by which to evaluate the merits of using EMS to obtain approximate nonnegative linear system solutions (EMS solutions) and has highlighted major difficulties in choosing an appropriate class of S which provides stability and accuracy (good quantification). In particular, this work has demonstrated important differences between the classes of reducible and irreducible nonnegative smoothing with respect to uniqueness, stability, and quantification (see §5).

Two major questions surrounding EMS are those of, first, convergence, and second, the nature of the convergence point. The convergence has been proven, for certain types of S, in [10], hence this paper concentrates on the second question of characterizing the approximating properties of EMS solutions. This is explored by using a crude roughness measure to prove existence results for the EMS equations which are the fixed point equations associated with (2). The method of proof is topological and is based on Brouwer's fixed point theorem [2]. This topological method was first used to give an alternative proof of the main results in the Perron–Frobenius theory [17] in the first 1935 edition of [1]. Restricting, as is done here, the class of smoothing matrices in EMS to be nonnegative, it is perhaps not surprising that the Aleksandrov–Hopf method can also be used with success in the nonlinear context of EMS. However, the special structure of the nonlinearity in EMS allows the full exploitation of the Aleksandrov–Hopf method. Furthermore, the same estimation techniques, which are used in the Perron–Frobenius theory, can be applied a priori to EMS solutions, thus producing uniform a priori estimates of the roughness measure which are essential to some of the proofs of existence by the Aleksandrov–Hopf method. These a priori estimates illustrate precise mathematically proven results that demonstrate the EMS algorithm's limited ability to reconstruct nonuniform quantities of physical interest and hence, give limitations on its usefulness in applications. Therefore, from an applied point of view, it is the a priori estimates, used only as a tool for some of the existence proofs, that are important in characterizing the approximating properties of EMS solutions. In addition, the precise form of the estimates permit an identification of those properties of the smoothing matrix that are detrimental to quantification. Moreover, these estimates assume even more relevance given that the EMS solutions, whose existence is proven below, are usually *unique* (see §5). A subtheme of the paper is that EMS provides a good example of an application where a *linear* theory (Perron–Frobenius), when combined with an appropriate topological tool (Brouwer), can provide a useful explicit analysis of a *nonlinear* problem.

Note. In this paper, a priori refers to the use of EMS solutions before they are proven to exist and is not to be confused with the Bayesian notion of a prior distribution which also arises in connection with the EM methodology [9].

Because of the fundamental relationship of the EMS algorithm to Perron–Frobenius theory, some basic notation and facts about nonnegative matrices are summarized in §2.1. The EMS map and equations are defined in §2.2 where the fixed point formulation is also given and the roughness measure is defined. Two existence results

(Theorems 1 and 2) are given in §3. Both of these are proven by using the Aleksandrov–Hopf method, with Theorem 2 making use of an a priori estimate. A generalization of the EMS equations and map, to which the existence results of §3 are easily extended, is considered in §4. The first existence result for the generalized EMS equations (Theorem 3) is proved by the same methods of §3, but the second (Theorem 4) provides an interesting interplay of linear and nonlinear theory and is analogous to the fixed point argument first applied by Schauder [20], [21], [8] for quasilinear elliptic partial differential equations. Like the linear elliptic theory and associated Schauder estimates, it is the linear Perron–Frobenius theory that both identifies a suitable nonlinear continuous mapping whose fixed points are generalized EMS solutions, and provides the necessary uniform a priori estimate. A special case of Theorem 4 produces an a priori estimate for EMS solutions. A brief summary of the implications of the a priori estimates for the limitations of EMS, together with comments on uniqueness and stability, constitutes §5 although [13], [14] should be consulted for a more complete story.

2. Perron–Frobenius theory and EMS.

2.1. Perron–Frobenius theory. A matrix $A \in \mathbb{R}^{m\times n}$ is called *nonnegative* (resp. *positive*) if all its elements satisfy $a_{ij} \geq 0$ (resp. $a_{ij} > 0$). This is written as $A \geq 0$ (resp. $A > 0$). A nonnegative matrix A is *row stochastic* if $r_i(A) = 1$, for all i, where $r_i(A)$ denotes the ith row sum of A. A square matrix $A \in \mathbb{R}^{n\times n}$, $n \geq 2$, is called *irreducible* if there is no permutation matrix Π for which

$$\Pi A \Pi^t = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \quad \text{where } A_{11} \text{ and } A_{22} \text{ are square.}$$

In applications, the distinction between irreducible and reducible S in (2) often corresponds to the difference between using spatial smoothing (i.e., $S_{b\beta}$ is nonzero for all β in some spatial neighbourhood of b) and nonspatial (i.e., diagonal) smoothing. The basic results of Perron–Frobenius theory [17] assert that every (nonzero) nonnegative square matrix A possesses a nonnegative eigenvector $\boldsymbol{x}$ corresponding to the eigenvalue which equals its (positive) spectral radius $\rho(A)$; i.e., $A\boldsymbol{x} = \rho(A)\boldsymbol{x}$. Moreover, if $A \geq 0$ is irreducible, then $\boldsymbol{x} > 0$, $\rho(A)$ is a simple eigenvalue of A, and up to scalar multiples, $\boldsymbol{x}$ is the only nonnegative eigenvector of A. In this case, $\rho(A)$ is called the *maximal eigenvalue* and $\boldsymbol{x}$ the *maximal eigenvector* of A. It is a standard result [17, p. 49] that for any nonnegative square irreducible matrix A with positive trace, there exists a positive integer ν such that $A^\nu > 0$. For such a matrix, let $\mu(A)$ denote the smallest such integer ν. As in the Perron–Frobenius theory, the following elementary inequality [17, p. 26], a proof of which can be found in [16, p. 79] or [13], proves to be useful.

LEMMA 1. *If $q_1, q_2, \ldots, q_n$ are positive real numbers, then*

$$\min_i \frac{p_i}{q_i} \leq \frac{p_1 + p_2 + \cdots + p_n}{q_1 + q_2 + \cdots + q_n} \leq \max_i \frac{p_i}{q_i},$$

for any real numbers $p_1, p_2, \ldots, p_n$. Moreover, equality holds on either side of this inequality if and only if all the ratios p_i/q_i are equal.

2.2. The EMS map, algorithm and equations. Let $K = \{\boldsymbol{\theta} \in \mathbb{R}^B \mid \boldsymbol{\theta} \geq 0\}$ denote the nonnegative cone in $\mathbb{R}^B$, and let $\Omega = \{\boldsymbol{\theta} \in K \mid \sum_{b=1}^B \theta_b = 1\}$ be a hyperplane cross section through K. It is obvious that Ω is convex and compact. For $\boldsymbol{\theta} \in K\backslash\{0\}$, the EMS *map* $\mathcal{F}_S$ is defined by

$$\mathcal{F}_S(\boldsymbol{\theta}) = SF(\boldsymbol{\theta})\boldsymbol{\theta}, \tag{3}$$

where $S \in \mathbb{R}^{B\times B}$ is a nonnegative smoothing matrix and $F(\boldsymbol{\theta}) = \operatorname{diag}(F_1(\boldsymbol{\theta}), \dots, F_B(\boldsymbol{\theta}))$ with

$$F_b(\boldsymbol{\theta}) = \sum_{d=1}^{D} \frac{n_d^* \tilde{p}_{bd}}{(P^t\boldsymbol{\theta})_d}, \quad b = 1, \dots, B, \quad \text{and} \quad \tilde{p}_{bd} = p_{bd}/r_b(P). \tag{4}$$

In (4), $P \in \mathbb{R}^{B\times D}$ and $\boldsymbol{n}^* \in \mathbb{R}^D$ are both nonnegative and it is assumed that $\boldsymbol{n}^*$ is nonzero. In the cases of interest below, either $P \geq 0$ has no row or column of zeros and $\boldsymbol{\theta} \in \operatorname{int}(K)$, the interior of K, or $P > 0$, and hence $P^t\boldsymbol{\theta}$ in (4) has no zero components. The EMS *algorithm* is defined by the nonlinear iteration

$$\boldsymbol{\theta}^{(n+1)} = \mathcal{F}_S(\boldsymbol{\theta}^{(n)}) = SF(\boldsymbol{\theta}^{(n)})\boldsymbol{\theta}^{(n)}, \quad n = 0, 1, 2, \dots, \tag{5}$$

which, for suitable $\boldsymbol{\theta}^{(0)} > 0$, aims to discover a nonnegative fixed point of $\mathcal{F}_S$, or equivalently, a solution of the EMS *equations*

$$\boldsymbol{\theta} = SF(\boldsymbol{\theta})\boldsymbol{\theta}. \tag{6}$$

Nonnegative solutions of (6) are called EMS *solutions* and are by definition the nonnegative fixed points of $\mathcal{F}_S$. When applying the EMS algorithm through the iteration (5) to the "solution" of a nonnegative linear system

$$P^t\boldsymbol{\theta} = \boldsymbol{n}^*, \tag{7}$$

the P and $\boldsymbol{n}^*$ appearing in (4) are precisely those given in (7), while S must be appropriately chosen. It is only for very special S that it is possible to recover a solution of (7) from EMS [12], [14], §5.1, and so usually, the EMS solutions are *not* solutions of (7). If the iteration (5) converges to an EMS solution $\boldsymbol{\theta}^S$, then this convergence point represents a smoothed approximate solution of (7), which may exist irrespective of whether or not (7) has any nonnegative solutions.

For the purposes of obtaining a priori estimates, it will be useful to define the roughness measure

$$\gamma(\boldsymbol{\theta}) = \max_{b,b'} \frac{\theta_b}{\theta_{b'}} \quad \text{for} \quad \boldsymbol{\theta} \in \operatorname{int}(K), \tag{8}$$

and the associated convex and compact set $\Omega_M = \{\boldsymbol{\theta} \in \operatorname{int}(\Omega) \mid \gamma(\boldsymbol{\theta}) \leq M\}$, for a fixed M, $1 \leq M < \infty$. It is convenient to extend γ to $+\infty$ for nonzero $\boldsymbol{\theta} \in K$ with zero components. Clearly, γ is a measure of nonuniformity and gives information on the range of ratios of values present in such a way that near-zero components are highly weighted. It is also convenient to define the *normalized* EMS *map* $\mathcal{F}$ by

$$\mathcal{F}(\boldsymbol{\theta}) = \mathcal{F}_S(\boldsymbol{\theta})/f(\boldsymbol{\theta}), \tag{9}$$

where $f(\boldsymbol{\theta}) = \sum_{b=1}^{B} \big(SF(\boldsymbol{\theta})\boldsymbol{\theta}\big)_b$. If $S = \alpha I$, $\alpha > 0$, then (5) reduces, for $\alpha = 1$, to the EM iteration of Shepp–Vardi from emission tomography [11], [22], [26] (see (1)). For other α, this iteration, also known as the Lucy–Richardson iteration [15], [19], has found applications in diverse image recovery disciplines [24].

Remark 1. Observe that $F(\boldsymbol{\theta})$ in (4) is homogeneous of degree -1; i.e., $F(\lambda\boldsymbol{\theta}) = \lambda^{-1}F(\boldsymbol{\theta})$ for all $\lambda \neq 0$, and, therefore, that $\mathcal{F}_S$ is homogeneous of degree zero.

Remark 2. If for some d_0, $n_{d_0}^* = 0$ in (4), then $F(\boldsymbol{\theta})$ and hence $\mathcal{F}_S(\boldsymbol{\theta})$ are independent of the d_0th column of P. Hence, only the submatrix of P consisting of those columns d for which $n_d^* > 0$ features in $\mathcal{F}_S$. Consequently, only positive data problems need be considered.

Remark 3. If a nonnegative solution $\boldsymbol{\theta}^S$ of (6) exists, and the matrix $SF(\boldsymbol{\theta}^S)$ is nonnegative and irreducible, then the Perron–Frobenius theory (§2.1) automatically implies that $\boldsymbol{\theta}^S > 0$ and $\rho(SF(\boldsymbol{\theta}^S)) = 1$, since $\boldsymbol{\theta}^S$ is the maximal eigenvector of the irreducible matrix $SF(\boldsymbol{\theta}^S)$ corresponding to the maximal eigenvalue 1 [12]. If S is nonnegative and irreducible, the product $SF(\boldsymbol{\theta}^S)$ will inherit these same properties whenever $F_b(\boldsymbol{\theta}^S) > 0$ for all b.

3. Two existence results. The use of topological methods (based on Brouwer's fixed point theorem [2]) for the proof of fundamental results in Perron–Frobenius theory goes back to the first 1935 edition of [1, §12.3, p. 480]. For a survey and generalizations of this method, see [6]. Although the theme here is that this fixed point method easily adapts to the current application to produce nonnegative EMS solutions, other approaches are possible [10], [12]. A direct application of the Aleksandrov–Hopf method produces a quite general existence result for the EMS equations defined in (6).

THEOREM 1. *Assume that $S \geq 0$ has no column of zeros, $P > 0$ and $\boldsymbol{n}^* > 0$. Then $\mathcal{F}_S$ has a fixed point $\boldsymbol{\theta}^S \in K$.*

Proof. Consider the normalized EMS map $\mathcal{F}$ as a map from Ω to itself. Since $P > 0$ and $\boldsymbol{n}^* > 0$, $F_b(\boldsymbol{\theta}) > 0$, $b = 1, \ldots, B$, for every $\boldsymbol{\theta} \in \Omega$ and so S and $SF(\boldsymbol{\theta})$ have the same pattern of nonzero entries. Because S, and therefore $SF(\boldsymbol{\theta})$, has no column of zeros, then for every $\boldsymbol{\theta} \in \Omega$, $\big(SF(\boldsymbol{\theta})\boldsymbol{\theta}\big)_b > 0$ for some b, and hence, $f(\boldsymbol{\theta}) \geq a > 0$ for some constant a, thus making $\mathcal{F}(\boldsymbol{\theta})$ continuous on Ω. By the normalization in (9), $\mathcal{F} : \Omega \to \Omega$, and since Ω is compact and convex, $\mathcal{F}$ has a fixed point $\boldsymbol{\theta}_0 \in \Omega$ by Brouwer's fixed point theorem. Consequently, from (9), $\boldsymbol{\theta}_0 = \mathcal{F}_S(\boldsymbol{\theta}_0)/f(\boldsymbol{\theta}_0)$, or equivalently, $SF(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0 = f_0\boldsymbol{\theta}_0$ where $f_0 = f(\boldsymbol{\theta}_0) = \sum_b (SF(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0)_b$. Now employing Remark 1 gives that $SF(f_0\boldsymbol{\theta}_0)\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0$ and multiplying both sides of this equation by f_0 shows that $\boldsymbol{\theta}^S = f_0\boldsymbol{\theta}_0$ is a nonnegative fixed point of $\mathcal{F}_S$. □

Remark 4. Notice that if $S = \operatorname{diag}(\sigma_1, \ldots, \sigma_B)$, with $\sigma_b > 0$ for all b, in Theorem 1, then there always exists at least B EMS solutions of form $(0, \ldots, 0, \theta_b, 0, \ldots, 0)^t$, where $\theta_b = \sigma_b \sum_d n_d^*$, for each b. The special case $S = \alpha I$, $\alpha > 0$, which includes the EM algorithm, falls into this category.

In the case that P and $\boldsymbol{n}^*$ are derived from a linear system of the form (7), Theorem 1 establishes the solvability of the EMS equations when the EMS algorithm is applied to positive linear systems. Thus, for suitable S, the EMS equations can have nonnegative solutions even if the linear system does not. Although the conditions on S are quite weak, the requirement that $P > 0$ is rather strong. With the help of a uniform a priori estimate, the same method of proof of Theorem 1 produces an existence result where this situation is reversed; namely, P need only be nonnegative provided S is positive.

THEOREM 2. *Assume that $S > 0$, $P \geq 0$ has no row or column of zeros and $\boldsymbol{n}^* > 0$. Then $\mathcal{F}_S$ has a fixed point $\boldsymbol{\theta}^S > 0$, which satisfies*

$$\gamma(\boldsymbol{\theta}^S) \leq \max_{b,b',\beta} \frac{S_{b\beta}}{S_{b'\beta}}. \tag{10}$$

All nonnegative fixed points of $\mathcal{F}_S$ are positive and satisfy (10).

Proof. Consider again the normalized EMS map $\mathcal{F}$ on Ω_M, where M, $1 \leq M < \infty$, shall for the moment remain unspecified. For $\boldsymbol{\theta} \in \Omega_M$, the assumptions on P and $\boldsymbol{n}^*$ imply that $F_b(\boldsymbol{\theta}) > 0$, $b = 1, \ldots, B$ and therefore $SF(\boldsymbol{\theta}) > 0$. Hence, $f(\boldsymbol{\theta}) > 0$ for

all $\boldsymbol{\theta} \in \Omega_M$. For $\boldsymbol{\theta} \in \Omega_M$, an estimate of $\gamma(\mathcal{F}(\boldsymbol{\theta}))$ can be obtained from

$$\frac{\mathcal{F}_b(\boldsymbol{\theta})}{\mathcal{F}_{b'}(\boldsymbol{\theta})} = \frac{\sum_{\beta=1}^{B} S_{b\beta} F_\beta(\boldsymbol{\theta}) \theta_\beta}{\sum_{\beta=1}^{B} S_{b'\beta} F_\beta(\boldsymbol{\theta}) \theta_\beta} \leq \max_\beta \frac{S_{b\beta}}{S_{b'\beta}},$$

where the last inequality comes from Lemma 1 and the fact that all terms of the sums in the quotient are positive. Hence, $\gamma(\mathcal{F}(\boldsymbol{\theta})) \leq \max_{b,b',\beta}(S_{b\beta}/S_{b'\beta})$, and therefore, taking M to be the right-hand side of (10) shows that $\mathcal{F} : \Omega_M \to \Omega_M$. Since $\mathcal{F}$ is clearly continuous and Ω_M is compact and convex, Brouwer's fixed point theorem again gives the existence of a fixed point $\boldsymbol{\theta}_0 \in \Omega_M$, which, in the same manner as in the proof of Theorem 1, gives a fixed point $\boldsymbol{\theta}^S = f_0 \boldsymbol{\theta}_0$ of $\mathcal{F}_S$, where $f_0 = \sum_b (SF(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0)_b$. Because $\gamma(\boldsymbol{\theta}^S) = \gamma(\boldsymbol{\theta}_0)$, $\boldsymbol{\theta}^S$ satisfies the bound in (10). □

In terms of the performance of the EMS algorithm, the estimate (10) shows that the use of strong (spatial) positive smoothing restricts the obtainable nonuniformity and so $\boldsymbol{\theta}^S$ cannot well approximate possibly interesting solutions of (7) whose γ exceeds the right-hand side of (10). Moreover, the bound in (10) is independent of P and $\boldsymbol{n}^*$ illustrating that the severity of this phenomenon is completely determined by S.

4. A generalized EMS map. By abstracting the essential properties of $F(\boldsymbol{\theta})$, which were used in the proof of the above results, it is possible to extend the existence results to more general maps of the type (3). For instance, let $G : K \to \mathbb{R}_+^{B\times B}$ be a given nonnegative matrix-valued function and consider the generalized EMS map

$$\mathcal{G}_S(\boldsymbol{\theta}) = SG(\boldsymbol{\theta})\boldsymbol{\theta} \quad \text{for} \quad \boldsymbol{\theta} \in K\backslash\{0\}. \tag{11}$$

Nonnegative fixed points of $\mathcal{G}_S$ will be called *generalized* EMS *solutions.* The Aleksandrov–Hopf method gives immediate generalizations of Theorems 1 and 2.

THEOREM 3. (a) *Assume that $S \geq 0$ has no column of zeros and $G : K\backslash\{0\} \to \mathbb{R}_+^{B\times B}$ satisfies*

(i) *$G(\boldsymbol{\theta})$ is continuous on Ω,*

(ii) *$G_{bb}(\boldsymbol{\theta}) > 0$ on Ω,*

(iii) *$G(\lambda\boldsymbol{\theta}) = \lambda^{-1}G(\boldsymbol{\theta})$ for all $\lambda \neq 0$ and $\boldsymbol{\theta} \in K\backslash\{0\}$.*

Then $\mathcal{G}_S$ has a fixed point $\boldsymbol{\theta}^S \in K$.

(b) *Assume that $S > 0$, G is diagonal and satisfies*

(i) *$G(\boldsymbol{\theta})$ is continuous on $\operatorname{int}(\Omega)$,*

(ii) *$G_{bb}(\boldsymbol{\theta}) > 0$ on $\operatorname{int}(\Omega)$,*

(iii) *$G(\lambda\boldsymbol{\theta}) = \lambda^{-1}G(\boldsymbol{\theta})$ for all $\lambda \neq 0$ and $\boldsymbol{\theta} \in \operatorname{int}(K)$.*

Then $\mathcal{G}_S$ has a fixed point $\boldsymbol{\theta}^S > 0$ which satisfies the estimate (10). *All nonnegative fixed points of $\mathcal{G}_S$ are positive and satisfy* (10).

Proof. The proof of these results proceeds in exactly the same manner as those of Theorems 1 (for part (a)) and 2 (for part (b)) and rely on the fact that, under the given conditions, $SG(\boldsymbol{\theta})$ has at least as many positive entries as S in at least the same index positions. □

Once again, the fact that the solutions $\boldsymbol{\theta}^S$ of Theorem 3(b) satisfy (10) indicates that, for positive smoothing S, all generalized EMS solutions are restricted in their approximating abilities by this estimate. This shows that generalized EMS algorithms based on (11), with positive S, can be inappropriate if S is not suitably designed.

A more concrete generalization of the EMS map (3) is obtained by supposing that

G is diagonal; i.e., $G = \operatorname{diag}(G_1(\boldsymbol{\theta}), \dots, G_B(\boldsymbol{\theta}))$, where the G_b take the special form

$$G_b(\boldsymbol{\theta}) = \sum_{d=1}^{D} \frac{n_d^* q_{bd}}{(P^t \boldsymbol{\theta})_d}, \quad b = 1, \dots, B, \tag{12}$$

with $\boldsymbol{n}^* > 0$, $Q \geq 0$ having no row of zeros, and $P \geq 0$ having no column of zeros. This G satisfies (i)–(iii) of Theorem 3(b), and if $P > 0$, it satisfies (i)–(iii) of Theorem 3(a), but for irreducible S in (11), much more is true than the conclusions of Theorem 3.

The relevance of irreducibility of S to the stability and quantification properties of the EMS algorithm (5) has been examined in [13], [14]. The primary outcome of this work is that although irreducibility has the disadvantage of reducing quantification (as illustrated below), but the advantage of introducing some stability, the reduction in quantification can be controlled by a careful choice of S [13]. Therefore, irreducibility is a desirable property that ensures stability of the iteration (5) and contributes to the well-posedness of the EMS equations (6) [14]. It is therefore important to consider this class of smoothing in the context of existence results and estimates of γ. The next theorem, has, like Theorem 2, the advantage of producing during the course of its proof, a uniform estimate of $\gamma(\boldsymbol{\theta}^S)$ for all generalized EMS solutions $\boldsymbol{\theta}^S$. The proof constructs what is essentially a uniform a priori bound for $\boldsymbol{\theta}^S$ and is motivated by the use of the infinite dimensional version of Brouwer's theorem; namely, the Schauder fixed point theorem [20], as it was originally applied, in conjunction with Schauder's a priori estimates, to obtain existence of solutions for quasilinear elliptic partial differential equations [21], [8, Chap. 11].

THEOREM 4. *Assume that $S \geq 0$ is irreducible with positive trace, $Q > 0$, $P \geq 0$ has no column of zeros, and $\boldsymbol{n}^* > 0$. Then the generalized* EMS *map $\mathcal{G}_S$ of* (11), *with G given by* (12), *has a fixed point $\boldsymbol{\theta}^S > 0$, which satisfies*

$$\gamma(\boldsymbol{\theta}^S) \leq \left(\max_{b,b',\beta} \frac{S_{b\beta}^{\mu}}{S_{b'\beta}^{\mu}} \right) \left(\max_{b,b',d} \frac{q_{bd}}{q_{b'd}} \right)^{\mu - 1}, \tag{13}$$

where $\mu = \mu(S)$. All nonnegative fixed points of $\mathcal{G}_S$ are positive and satisfy (13).

Proof. Fix $\boldsymbol{\xi} \in \Omega_M$, for some as yet unspecified M, $1 \leq M < \infty$, and consider the nonnegative matrix $SG(\boldsymbol{\xi})$. Under the assumptions on P, Q and $\boldsymbol{n}^*$, $G_b(\boldsymbol{\xi}) > 0$ for $b = 1, \dots, B$, and hence, $SG(\boldsymbol{\xi})$ which has the same zero pattern as S, is also an irreducible matrix with positive trace. Moreover, $(SG(\boldsymbol{\xi}))^{\mu} > 0$ where $\mu = \mu(S)$. From the Perron–Frobenius theory, $SG(\boldsymbol{\xi})$ has a unique positive eigenvector $\boldsymbol{\theta}_{\xi} \in \operatorname{int}(\Omega)$ corresponding to the eigenvalue equal to its finite spectral radius $\rho(\boldsymbol{\xi})$; i.e., $SG(\boldsymbol{\xi})\boldsymbol{\theta}_{\xi} = \rho(\boldsymbol{\xi})\boldsymbol{\theta}_{\xi}$. Denote by $\mathcal{P}$ the nonlinear map which defines this correspondence; i.e., $\mathcal{P}(\boldsymbol{\xi}) = \boldsymbol{\theta}_{\xi}$ is by definition the map $\boldsymbol{\xi} \mapsto \boldsymbol{\theta}_{\xi}$. To estimate $\gamma(\boldsymbol{\theta}_{\xi})$, the consequential relation $(SG(\boldsymbol{\xi}))^{\mu}\boldsymbol{\theta}_{\xi} = \rho^{\mu}(\boldsymbol{\xi})\boldsymbol{\theta}_{\xi}$ leads to the quotient

$$\frac{(\theta_{\xi})_b}{(\theta_{\xi})_{b'}} = \frac{\sum_{\beta=1}^{B} (SG(\boldsymbol{\xi}))_{b\beta}^{\mu} (\theta_{\xi})_{\beta}}{\sum_{\beta=1}^{B} (SG(\boldsymbol{\xi}))_{b'\beta}^{\mu} (\theta_{\xi})_{\beta}}$$

in which all terms in the numerator and denominator sums are positive. The elementary estimates

$$(\min_b G_b(\boldsymbol{\xi}))^{\mu-1} S_{b\beta}^{\mu} G_{\beta}(\boldsymbol{\xi}) \leq (SG(\boldsymbol{\xi}))_{b\beta}^{\mu} \leq (\max_b G_b(\boldsymbol{\xi}))^{\mu-1} S_{b\beta}^{\mu} G_{\beta}(\boldsymbol{\xi})$$

now follow easily [13] and, hence,

$$\frac{(\theta_\xi)_b}{(\theta_\xi)_{b'}} \le \left(\max_{b,b'} \frac{G_b(\boldsymbol{\xi})}{G_{b'}(\boldsymbol{\xi})}\right)^{\mu-1} \frac{\sum_{\beta=1}^B S_{b\beta}^\mu G_\beta(\boldsymbol{\xi})(\theta_\xi)_\beta}{\sum_{\beta=1}^B S_{b'\beta}^\mu G_\beta(\boldsymbol{\xi})(\theta_\xi)_\beta}. \tag{14}$$

Applying Lemma 1 to the first factor in this estimate gives

$$\frac{G_b(\boldsymbol{\xi})}{G_{b'}(\boldsymbol{\xi})} = \frac{\sum_{d=1}^D n_d^* q_{bd}/(P^t\boldsymbol{\xi})_d}{\sum_{d=1}^D n_d^* q_{b'd}/(P^t\boldsymbol{\xi})_d} \le \max_d \frac{q_{bd}}{q_{b'd}},$$

and hence $\max_{b,b'}(G_b/G_{b'}) \le \max_{b,b',d}(q_{bd}/q_{b'd})$. This estimate, together with Lemma 1 applied to the second factor in (14), gives the final uniform estimate

$$\gamma(\mathcal{P}(\boldsymbol{\xi})) = \gamma(\boldsymbol{\theta}_\xi) \le \left(\max_{b,b',\beta} \frac{S_{b\beta}^\mu}{S_{b'\beta}^\mu}\right)\left(\max_{b,b',d} \frac{q_{bd}}{q_{b'd}}\right)^{\mu-1}. \tag{15}$$

Setting M to be the right-hand side of (15) then shows that $\mathcal{P} : \Omega_M \to \Omega_M$. Since the entries of $\boldsymbol{\theta}_\xi = \mathcal{P}(\boldsymbol{\xi})$ depend continuously on those of $SG(\boldsymbol{\xi})$, which in turn depend continuously on $\boldsymbol{\xi}$, the mapping $\mathcal{P}$ is continuous. By Brouwer's fixed point theorem, $\mathcal{P}$ has a fixed point $\boldsymbol{\theta}_0 \in \Omega_M$; that is, $\boldsymbol{\theta}_0$ satisfies $SG(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0 = \rho(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0$. Using again Remark 1, applied to G, to write this as $SG(\rho_0\boldsymbol{\theta}_0)\rho_0\boldsymbol{\theta}_0 = \rho_0\boldsymbol{\theta}_0$, where $\rho_0 = \rho(\boldsymbol{\theta}_0)$, then shows that $\boldsymbol{\theta}^S = \rho_0\boldsymbol{\theta}_0$ is a positive fixed point of $\mathcal{G}_S$ which satisfies the estimate (13) since $\gamma(\boldsymbol{\theta}^S) = \gamma(\boldsymbol{\theta}_0)$. □

Remark 5. Note that the main use of the assumption $Q > 0$ is to obtain a uniform estimate of the ratio $G_b/G_{b'}$ from Lemma 1. Uniform a priori estimates of the type represented by (13) were first obtained in [13] during the course of a quantification study for the EMS algorithm. Note that the estimate (13) is again independent of P and $\boldsymbol{n}^*$.

Remark 6. The generalization of the EMS map (3) to that of (11) and (12) allows the necessary conditions on P to be considerably relaxed, only requiring that P be nonnegative (cf. Theorem 1). A generalized EMS algorithm based on (11) and (12) may therefore provide an alternative for the solution of nonnegative linear systems of the type (7).

The special case $Q = \tilde{P}$ (the matrix with entries $\tilde{p}_{bd}$) in Theorem 4 gives an existence result, together with an a priori estimate for all fixed points, for the original EMS map $\mathcal{F}_S$. However, the requirement for the positivity of Q is now transferred to P.

COROLLARY 1. *Assume that $S \ge 0$ is irreducible with positive trace, $P > 0$, and $\boldsymbol{n}^* > 0$. Then $\mathcal{F}_S$ has a fixed point $\boldsymbol{\theta}^S > 0$ which satisfies*

$$\gamma(\boldsymbol{\theta}^S) \le \left(\max_{b,b',\beta} \frac{S_{b\beta}^\mu}{S_{b'\beta}^\mu}\right)\left(\max_{b,b',d} \frac{\tilde{p}_{bd}}{\tilde{p}_{b'd}}\right)^{\mu-1}, \tag{16}$$

where $\mu = \mu(S)$. All nonnegative fixed points of $\mathcal{F}_S$ are positive and satisfy (16).

Remark 7. Corollary 1 provides a companion result to Theorem 1 with the a priori estimate (16) now applying to the EMS solution. For $Q = \tilde{P}$ in Theorem 4, two applications of Lemma 1 to the first factor on the right-hand side in (14) produces the alternative (but weaker) estimate

$$\gamma(\boldsymbol{\theta}^S) \le \left(\max_{b,b',\beta} \frac{S_{b\beta}^\mu}{S_{b'\beta}^\mu}\right)\left(\max_{b,b'}\left(\max_d \frac{p_{bd}}{p_{b'd}}\right)\left(\max_d \frac{p_{b'd}}{p_{bd}}\right)\right)^{\mu-1}$$

for $\boldsymbol{\theta}^S$ in Corollary 1.

Note that, although the restrictive estimates (13) and (16) now hold for nonnegative irreducible (spatial) smoothing, the bounds in them now consist of two factors (cf. (10)). In the case of (13), the second factor is determined by Q, and μ, which is a measure of the sparsity of S. (Typically, for spatial local averaging smoothing, μ is of the order of the dimension of S, B in this paper.) By using a suitable Q and S, the size of this estimate can be controlled independently of the linear system matrix P. On the other hand, for the usual EMS algorithm (5), $Q = \tilde{P}$ and so the ability to control the right-hand side of the estimate in (16) is limited to only the choice of S. In either case, S should be chosen so as to maximize the size of the upper bounds. The added flexibility afforded in controlling the size of the right-hand side of (13) through Q, and the fact that Theorem 4 applies to the practical case in which $S \geq 0$ is irreducible and $P \geq 0$ may both have many zeros, gives some support to using a generalized EMS algorithm based on (11) and (12) for the solution of (7) instead of (5). Other work [13] indicates that estimates similar to (16) apply for the EMS algorithm (5) for certain specially structured problems in which the assumption $P > 0$ can be relaxed to $P \geq 0$.

5. Summary. Here we make a few remarks to place the above results in context and summarize their relevance to an assessment of the EMS algorithm. As already mentioned, the analysis of Theorem 4 provides a rare example of where a linear theory (Perron–Frobenius) gives a useful result for a fully nonlinear problem (EMS), and not the linearization of that problem, but the focus here is on the consequences of the a priori estimates. In particular, the following remarks highlight a dichotomy between irreducible (spatial) and reducible (nonspatial) smoothing where we use the loose association already mentioned in §2.1.

5.1. Exact solutions. If $\boldsymbol{\xi} \geq 0$ is a solution of (7) with $\boldsymbol{n}^* > 0$ and $S \geq 0$ is chosen so that the eigencondition $S\boldsymbol{\xi} = \boldsymbol{\xi}$ holds, then, as is easily seen from (6), $\boldsymbol{\theta}^S = \boldsymbol{\xi}$ is an EMS solution [12]. Since this would require that the solution $\boldsymbol{\xi}$ be known beforehand, it is not practical to choose S on this basis, however, it illustrates that EMS can in theory obtain exact solutions of (7). In general though, EMS solutions are *not* solutions of (7). The same eigencondition implies that exact solutions can be recovered from the generalized EMS algorithm based on (11) and (12) provided Q is row stochastic.

5.2. Quantification. From Remark 3, $\boldsymbol{\theta}^S > 0$ whenever S is irreducible (spatial). This implies that for irreducible smoothing, no EMS solution can well approximate a desired vector $\boldsymbol{\theta}$ with zero components—a type of vector that often occurs in practice. In fact, the estimates (10), (13), and (16) demonstrate the inability of EMS to well approximate any desired vector $\boldsymbol{\theta}$ for which $\gamma(\boldsymbol{\theta})$ exceeds the right-hand sides of these estimates. The assumptions for these estimates deserve comment. Namely,

(i) for (10): $S > 0$ and $P > 0$;
(ii) for (13): $S \geq 0$ is irreducible and has positive trace, $P \geq 0$ and $Q > 0$;
(iii) for (16): $S \geq 0$ is irreducible and has positive trace and $P > 0$.

For (i), if the (spatial) smoothing is positive, it alone determines the upper bound for $\gamma(\boldsymbol{\theta}^S)$ irrespective of the underlying system matrix P. In the case of the generalized EMS algorithm in (ii), the more practical conditions on S and P appear, but then Q features in the estimate in (13). Finally, for (iii), which applies to the EMS algorithm (5), the estimate (16) holds for the restricted class of problems for which $P > 0$. Note however, that the existence of the upper bound for $\gamma(\boldsymbol{\theta}^S)$ derived for this restricted

class of problems cannot be eliminated by considering the wider class of problems for which $P \geq 0$ only. Indeed, estimates of the type (16) do hold for certain problems within the latter class [14]. The usefulness of the explicit forms of the bounds in (10), (13), and (16) derives from now being able, for a given P, to choose S (and Q) so that they are maximized and thereby restrict $\gamma(\boldsymbol{\theta}^S)$ as little as possible.

5.3. Uniqueness. When combined with the results of [10], the fact that $\boldsymbol{\theta}^S > 0$ whenever S is irreducible, implies that an EMS solution is *unique*. Hence, in Theorems 2 and 4 and Corollary 1, the estimates (10), (13), and (16) respectively, characterize the *unique* EMS solution. Furthermore, in the more general case where $S \geq 0$ is irreducible (with positive trace) and $P \geq 0$, the estimates in [14] would also characterize the unique EMS solution.

5.4. Stability. It was shown in [14] that irreducibility of S is sufficient for minimal stability (see [14] for a definition) and the elimination of extraneous structural solutions which are a cause of nonuniqueness when using a reducible S. Hence, from the point of view of (5) being a numerically stable iteration, irreducibility is a desirable property.

5.5. Synopsis. For irreducible S, one gains stability, uniqueness, and improved convergence but loses some quantifying ability ($\boldsymbol{\theta}^S > 0$). EMS solutions almost never coincide with those of the underlying linear system. The considerations in §§5.1 and 5.2 above may enable the design of a suitable irreducible S. The reader is urged to consult [10] and [12]–[14] for a full evaluation of EMS, but the short message is to exercise diligent *caution* in its use.

Acknowledgments. The author would like to express sincere thanks to Bob Anderssen for reading a draft of the paper, to Shane Latham for a reminder about the Schauder theory, and to Mario Bertero for a pointer to the papers [15] and [19]. The author is also grateful to the referees for their suggestions and to the associate editor for further comments, and the link of EM with [5], all of which greatly improved the presentation.

REFERENCES

[1] P. Aleksandrov and H. Hopf, *Topologie*, corrected reprinting, Springer-Verlag, Berlin, 1974.
[2] L.E.J. Brouwer, *Über Abbildungen von Mannigfaltigkeiten*, Math. Ann., 71 (1912), pp. 97–115.
[3] C.L. Byrne, *Iterative image reconstruction algorithms based on cross-entropy minimization*, IEEE Trans. Image Processing, 2 (1993), pp. 96–103.
[4] A.P. Demptser, N.M. Laird, and D.B. Rubin, *Maximum likelihood for incomplete data via the EM algorithm (with discussion)*, J. R. Statist. Soc. B, 39 (1977), pp. 1–38.
[5] I.I. Dikin, *Iterative solution of problems of linear and quadratic programming*, Sov. Math. Dokl., 8 (1967), pp. 674–675.
[6] Ky Fan, *Topological proofs for certain theorems on matrices with nonnegative elements*, Monats. Math., 62 (1958), pp. 219–237.
[7] A.M. Fraser and A. Dimitriadis, *Forecasting probability densities by using hidden markov models with mixed states*, in Time series prediction: forecasting the future and understanding the past, A.S. Weigend and N.A. Gershenfeld, eds., Addison-Wesley, Reading, MA, 1994, pp. 265–282.
[8] D. Gilbarg and N.S. Trudinger, *Elliptic partial differential equations of second order*, 2nd edition, Springer-Verlag, Berlin, Heidelberg, 1983.
[9] P.J. Green, *On use of the* EM *algorithm for penalized likelihood estimation*, J. R. Statist. Soc. B, 52 (1990), pp. 443–452.
[10] J.W. Kay, *On the convergence of the* EMS *algorithm*, to appear.

[11] K. LANGE AND R. CARSON, EM *reconstruction algorithms for emission and transmission tomography*, J. Comp. Assist. Tomog., 8 (1984), pp. 306–316.

[12] G.A. LATHAM AND R.S. ANDERSSEN, *A hyperplane approach to the* EMS *algorithm*, Appl. Math. Lett., 5:5 (1992), pp. 71–74.

[13] ———, *Assessing quantification for the* EMS *algorithm*, Linear Algebra Appl., 210 (1994), pp. 89–122.

[14] ———, *On the stabilization inherent in the* EMS *algorithm*, Inverse Problems, 10 (1994), pp. 161–183.

[15] L. LUCY, *An iterative technique for the restoration of observed distributions*, Astron. J., 79 (1974), pp. 745–754.

[16] M. MARCUS AND H. MINC, *Modern University Algebra*, Macmillan, New York, 1966.

[17] H. MINC, *Nonnegative matrices*, John Wiley and Sons, New York, 1988.

[18] D. NYCHKA, *Some properties of adding a smoothing step to the EM algorithm*, Statist. Probab. Lett., 9 (1990), pp. 187–193.

[19] W.H. RICHARDSON, *Bayesian-based iterative method of image restoration*, J. Optical Soc. Am., 62 (1972), pp. 55–59.

[20] J. SCHAUDER, *Der Fixpunktsatz in Funktionalräumen*, Studia Math., 2 (1930), pp. 171–180.

[21] ———, *Über das Dirichletsche Problem im Großen für nicht-linear elliptische Differentialgleichungen*, Math. Z., 37 (1933), pp. 623–634, 768.

[22] L.A. SHEPP AND Y. VARDI, *Maximum likelihood reconstruction for emission tomography*, IEEE Trans. Med. Imaging, 1 (1982), pp. 113–122.

[23] B.W. SILVERMAN, M.C. JONES, D.W. NYCHKA, AND J.D. WILSON, *A smoothed* EM *approach to indirect estimation problems, with particular reference to stereology and emission tomography (with discussion)*, J. R. Statist. Soc. B, 52 (1990), pp. 271–324.

[24] D.L. SNYDER, A.M. HAMMOUD, AND R.L. WHITE, *Image recovery from data acquired with a charge-coupled-device camera*, J. Optical Soc. Am. A, 10 (1993), pp. 1014–1023.

[25] Y. VARDI AND D. LEE, *From image deblurring to optimal investments: maximum likelihood solution of positive linear inverse problems*, J. R. Statist. Soc. B, 55 (1993), pp. 560–612.

[26] Y. VARDI, L.A. SHEPP, AND L. KAUFMAN, *A statistical model for positron emission tomography*, J. Amer. Statist. Assoc., 80 (1985), pp. 8–37.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 954–963, July 1995

ACCURATE COMPUTATION OF THE FUNDAMENTAL MATRIX OF A MARKOV CHAIN *

DANIEL P. HEYMAN†

Abstract. Associated with every stochastic matrix is another matrix called the fundamental matrix. The fundamental matrix can be used to obtain mean first-passage-times and other interesting operating characteristics. The fundamental matrix is defined as a matrix inverse, and computing it from the definition can be fraught with numerical errors. We establish a new representation of the fundamental matrix where matrix inversion is replaced by multiplying and then adding a pair of matrices. The representation requires the solution of a system of linear equations, and we show that that can be done via back and forward substitution from numbers that have already been calculated when the GTH algorithm is used to compute the steady-state probabilities. An algorithm based on this representation is given. The time complexity of the faster implementation is 75% of the time complexity of using Gaussian elimination.

Key words. Gaussian elimination, direct methods, first-passage-times

AMS subject classifications. 60J10, 65F15, 65G05

1. Introduction.

The purpose of this paper is to present a numerically stable method of computing the fundamental matrix of a Markov chain. The method is based on a new representation of the fundamental matrix that does not involve an explicit inverse, and on exploiting the UL factorization implicit in the GTH algorithm for obtaining the stationary distribution.

Let P be a regular (i.e., irreducible and aperiodic with no transient states) finite stochastic matrix, π' be its unique stationary (row) vector, $\mathbf{e}$ be a column vector of ones with the same length as π', and $W = \mathbf{e}\pi'$. The matrix

$$Z = (I - P + W)^{-1} \tag{1}$$

exists and is called the *fundamental matrix* of a Markov chain with transition matrix P (see Kemeny and Snell (1960) [11]). An alternative to the fundamental matrix is the *group inverse* described in Meyer (1975) [13]. Let $A^{\#}$ denote the group inverse; then $Z = A^{\#} + W$. Our method also applies to computing $A^{\#}$.

The fundamental matrix has several important applications that are described by Kemeny and Snell. One is to compute the matrix of mean first-passage-times, M say, from the formula

$$M = (I - Z + EZ_{dg})D, \tag{2}$$

where Z_{dg} is the diagonal matrix with elements Z_{ii} and D is the diagonal matrix with elements $1/\pi_i$, and E is the matrix of all ones. Another use depends on the following interpretation of the elements of Z. Let $v_{ij}(n)$ be the expected number of visits to state j, starting in state i, after n transitions; by convention, $v_{ii}(0) = 1$ for all i. Renewal theory establishes that $v_{ij}(n)/n \to \pi_j$: Theorem 4.3.4 in Kemeny and Snell [11] states that

$$\lim_{n\to\infty} v_{ij}(n) - n\pi_j \to z_{ij} - \pi_j = a_{ij}^{\#}. \tag{3}$$

* Received by the editors November 23, 1993; accepted for publication (in revised form) by C. Meyer, July 5, 1994.

† Bellcore, 331 Newman Springs Road, Red Bank, New Jersey 07701 (dph@bellcore.com).

This "bias term" is important in comparing Markov chains with rewards and in Markov decision processes; see, e.g., Heyman and Sobel (1984) [6, §§4–6].

Accurate computation of the inverse in (1) can be difficult. Heyman and Reeves (1989) [8] give two examples where straightforward matrix inversion leads to very inaccurate computation of the mean first-passage-times. They provide an algorithm for obtaining the matrix of mean first-passage-times that avoids an explicit calculation of Z, but their method has time complexity $O(n^4)$, where n is the order of P. When Z is at hand, the time complexity to compute M from Z is $O(n^2)$, so accurate computation of Z in time $O(n^3)$ is much faster than the method of Heyman and Reeves. The method proposed in this paper takes $\Theta(n^3)$ multiplications and the same number of additions including the work to compute π (which is $\Theta(n^3/3)$. This is the same work that would be done if (1) were inverted by Gaussian elimination after π was computed. Meyer shows that $A^{\#}$ can be computed with $\Theta(n^3)$ additions and multiplications without first computing π. However, applications (e.g., (2) and (3)) typically require π.

The representation theorem is given in §2, followed by an explanation in §3 of how the GTH algorithm provides information that permits the linear system that defines a matrix used in the representation to be solved by back and forward substitution. Computational issues are addressed in the penultimate section. Section 5 contains three numerical examples. Throughout this paper, matrices are denoted by upper case Roman letters and column vectors by lower case Roman letters. The elements of matrices are denoted by the same letter in lower case with two subscripts. The elements of vectors are denoted by the same letter with one subscript.

2. A representation of the fundamental matrix. The existence of the inverse in (1) is established in Theorem 4.3.1 in Kemeny and Snell (1960) [11]. More general inverses are shown to exist in Kemeny (1981) [12] and Hunter (1982) [9] and (1983) [10]. Our characterization of Z is given by the following theorem.

THEOREM 1. *Let P be a regular finite stochastic matrix, π' be its steady-state distribution, and $W = e\pi'$. Then*

$$Z \triangleq (I - P + W)^{-1} = W + (I - W)X, \tag{4}$$

where X is any solution of

$$(I - P)X = I - W. \tag{5}$$

Proof. A constructive proof that (5) has solutions is given in §3. Observe that $W^2 = e(\pi' e)\pi' = W$ and $PW = (Pe)\pi' = W$. For any matrix X, we have

$$\begin{aligned} Y &\triangleq (I - P + W)[W + (I - W)X] \\ &= W + (I - W - P + PW + W - W^2)X \\ &= W + (I - P)X. \end{aligned}$$

Thus (5) makes $Y = I$ and we are done. □

The second term on the right side of (4), $(I - W)X$, is the group inverse $A^{\#}$, so it is slightly easier to compute than the fundamental matrix. It will be argued in §4 and demonstrated by an example in §5 that this theorem leads to more accurate computations than Gaussian elimination. Although the proof of the theorem is easy, its assertion and computational implications may not be obvious.

Remark 1. When P is regular, it is well known that the rank $I - P$ is one less than the dimension of P, so if (5) has solutions, one element in each column of X

can be chosen arbitrarily. The vector $x = Xe$ consists of the row sums of X. Post multiplying both sides of (5) by e yields $(I-P)x = 0$. The only solutions of this system of equations is for x to have equal components (see, e.g., Lemma 7-3 in Heyman and Sobel (1982) [5]). We can obtain a particular solution of (5) by setting some row of X (the first is convenient) equal to zero. Postmultiplying the outsides of (4) by e makes manifest the known result that $Ze = e$.

3. Factorization of $I-P$. Here we show how the GTH algorithm for computing π' produces a UL factorization of $I - P$. This factorization can be used to solve (5). The GTH algorithm was introduced in Grassmann, Taksar, and Heyman (1985) [2]. It is a variant of Gaussian elimination that accurately computes the stationary vector of a regular stochastic matrix. Empirical evidence of its accuracy is given in Heyman (1987) [7], and analytic evidence in O'Cinneide (1993) [14].

To make this paper self-contained, and to introduce some notation, we display the GTH algorithm here. The states of the Markov chain are numbered from 1 to n, and p_{ij} is the (i,j)th element of P.

ALGORITHM GTH

1. (State reduction) For $k = n, n-1, \ldots, 2$, do the following:

   (a) Let $S_k = \sum_{j=0}^{k-1} p_{kj}$.

   (b) Let $p_{ik} \leftarrow p_{ik}/S_k, i < k$.

   (c) Let $p_{ij} \leftarrow p_{ij} + p_{ik}p_{kj}, i,j < k$.

2. (Back substitution) Initialize $TOT = 1$ and $\pi_1 = 1$.
   For $j = 2, 3, \ldots, n$ do the following:

   (a) Let $\pi_j = p_{1j} + \sum_{k=2}^{j-1} \pi_k p_{kj}$.

   (b) Let $TOT \leftarrow TOT + \pi_j$.

3. (Normalization) Let $\pi_j \leftarrow \pi_j/TOT$, $j = 1, 2, \ldots, n$.

The calculations in step 1 overwrite the elements of P, so let $\bar{P} = (\bar{p}_{ij})$ be the contents of the array when the algorithm terminates. Define f_{ij} and g_{ij} by

$$f_{ij} = \bar{p}_{ij} \quad \text{for } i < j \quad \text{and} \quad g_{ij} = \bar{p}_{ij} \quad \text{for } i > j,$$

so

$$\bar{P} = F + G + (I - S),$$

where

$$F = \begin{Bmatrix} f_{ij} & \text{if } i < j \\ 0 & \text{if } i \geq j \end{Bmatrix} \quad G = \begin{Bmatrix} g_{ij} & \text{if } i \geq j \\ 0 & \text{if } i < j \end{Bmatrix} \quad \text{and} \quad S = \operatorname{diag}(S_j),$$

where $S_1 = 0$. It is shown in Grassmann (1993) [3] that (our f_{ij} is $a_{ij}/S_j, j > 0$ in Grassmann's notation)

$$I - P = (F - I)(G - S). \tag{6}$$

The first term on the right side of (6) is upper triangular and the second term is lower triangular, so we have a UL factorization of $I - P$. The reason we obtain a UL factorization instead of the usual LU factorization is because the GTH algorithm eliminates the last equation first, while Gaussian elimination eliminates the first equation first. It is straightforward to show that both $U = F - I$ and $L = G - S$ have all row sums equal to zero, and that the top row of L is identically zero.

The UL factorization gives an interpretation of the back substitution step 2(a) that will be needed subsequently. To solve $\pi'(I-P) = \pi' UL = 0$, one first solves $z'L = 0$. Since the top row of L is zero, $e_1' = (1, 0, 0, \ldots, 0)$ is a solution; it corresponds to setting $\pi_1 = 1$ before normalization. Next one solves $\pi' U = z' = e_1'$ which yields $\pi' = e_1' U^{-1}$. (Notice that $\det(U) = 1$ so U^{-1} exists.) Thus $\pi_j = u_{1j}^{-1}$ is the unnormalized solution of $\pi'(I - P) = 0$ and

$$\pi_j = \frac{u_{1j}^{-1}}{\sum_{k=1}^{n} u_{ik}^{-1}}, \qquad j = 1, 2, \ldots, n \tag{7}$$

is the normalized solution.

Remark 2. The UL factorization can be used to solve (5). The first step to solve $UY = I - W$ and the second step is to solve $LX = Y$. Since the top row of L is identically zero, $LX = Y$ has no solution unless the top row of Y is identically zero. Since

$$y_{1j} = u_{1j}^{-1} - \pi_j \sum_{k=1}^{n} u_{1j}^{-1},$$

(7) shows that $y_{1j} = 0$ for all j.

4. Computing the fundamental matrix. Here we present an algorithm for computing the fundamental matrix based on Theorem 1. A faster implementation is given in §4.2; this version makes it easy to present a constructive proof that (5) has solutions. Partition L as follows:

$$L = \begin{pmatrix} 0 & \omega' \\ l & L_1 \end{pmatrix}, \tag{8}$$

where ω' is a $1 \times n - 1$ row vector of zeros and l is the first column of L with the top element discarded.

Remark 3. Grassmann, Taksar, and Heyman show that $S_k > 0$ for $k > 1$, so L_1 is nonsingular.

Partition matrices X and Y conformally,

$$X = \begin{pmatrix} 0 & \omega' \\ x & X_1 \end{pmatrix}, \quad \text{and} \quad Y = \begin{pmatrix} 0 & \omega' \\ y & Y_1 \end{pmatrix},$$

and set $\hat{X} = (x, X)$ and $\hat{Y} = (y, Y)$. The following algorithm, FUND, computes the fundamental matrix from (4) and (5).

ALGORITHM FUND

0. Use the GTH algorithm to compute π', L, and U.
1. (Solve (5).)
   - (a) Solve $UY = I - W$ by back substitution.
   - (b) Solve $L_1 \hat{X} = \hat{Y}$ by forward substitution.
   - (c) Let $x_{1j} = 0, j = 1, 2, \ldots, n$.

2. (Compute Z from (4)). For $j = 1, 2, \ldots, n$ do the following:

$$\text{(a)} \quad p \leftarrow \sum_{k=1}^{n} \pi_k x_{kj}.$$

$$\text{(b)} \quad z_{ij} = \pi_i + x_{ij} - p, \qquad i = 1, 2, \ldots, n.$$

Remarks 2 and 3 show that the equation in step 1(b) has a unique solution, and Remark 1 shows that step 1(c) produces a solution of (5). The fact that the top row of Y must equal zero provides an accuracy check for step 1(a). The fact that the row sums of Z must equal one can provide an accuracy check for step 2(b).

4.1. Operation counts. Now we compare operation counts for computing Z via FUND and by using Gaussian elimination to compute a matrix inverse. Recall that a flop is the work required to implement $s \leftarrow s + a_{ik}b_{kj}$, and n is the dimension of P, and the Θ function gives the exact order of growth in an asymptotic formula. The work of Stewart (1973) [16] contains multiplication counts for the classical algorithms of matrix algebra. Since a flop has one multiplication, our count for flops is identical to Stewart's count for multiplications. Matrix addition takes $\Theta(n^2)$ flops, and that will be dominated by other factors.

The GTH algorithm requires $\Theta(n^3/3)$ flops to produce π' and L and U are byproducts that come for free. Solving a triangular system requires $\Theta(n^2/2)$ flops. Solving (5) involves solving $2n-1$ triangular systems, so that solving (5) by back and forward substitution would contribute $\Theta(n^3)$ to computing Z. The matrix multiplication in (4) requires $O(n^2)$ flops because W has rank one, so computing Z via FUND can be done with $\Theta(4n^3/3)$ flops, excluding the work to compute π. Matrix inversion via Gaussian elimination requires $\Theta(n^3)$ flops, so FUND takes the same work as does the computation of π followed by matrix inversion of (1). Some speedup in FUND is possible. Adjacent right sides of the upper triangular systems in (5) are adjacent columns of $I - W$, so all but two adjacent elements are the same. The acceleration given by Heyman and Reeves to exploit this can be used to reduce the computational burden of step 1(a) by a factor of one-fourth at the expense of a significant increase in memory requirements.

We can save one-third of the work to do step 1 by exploiting the special structure of $I-W$. (I am indebted to D. P. O'Leary for her assistance with this.) We can compute U^{-1} in $n^3/6$ flops by using back substitution repeatedly. U^{-1} inherits the property of being upper triangular with minus ones on the diagonal, and has nonpositive entries above the diagonal. Matrix multiplication yields (note that u_{ij}^{-1} is the (i,j)th element of U^{-1})

$$[U^{-1}(I - W)]_{ij} = u_{ij}^{-1} - \pi_i' \sum_{k=i}^{n} u_{ik}^{-1},$$

which requires just $n - i$ additions and one multiplication. Thus, the computing cost to obtain $U^{-1}(I-W)$ is inconsequential compared to the cost of obtaining U^{-1}. Step 1 can be replaced by

1(a′). Compute U^{-1} and L_1^{-1} by back and forward substitution, respectively.

1(b′). Set $x_{1j} = 0, j = 1, 2, \ldots, n$.

1(c′). Solve $LX = U^{-1}(I - W)$ by forward substitution.

With this enhancement, FUND requires $\Theta(n^3)$ flops, which is three-fourths of the work required by computing π and then inverting (1) by Gaussian elimination.

4.2. Accuracy assessment. The accuracy of the fundamental matrix computed via FUND is primarily determined by the accuracy of the solutions of (5). In the next section we present some empirical evidence that accurate solutions are obtained; here we give some analytic support for the contention that FUND solves (5) accurately. O'Cinneide (1993) [14] gives an analytic demonstration that the GTH algorithm produces an accurate π'. Corollary 4 in O'Cinneide (1994) [15] shows that the GTH algorithm produces an accurate UL decomposition of $I - P$. Since "*the solutions of triangular systems are usually computed to high accuracy*" (Stewart (1973), [16, p. 150]) accurate solutions of (5) should be achieved. An error analysis that demonstrates that (5) is solved accurately has yet to be constructed. When some elements of X are large, there is the possibility of subtractive cancellation in step 2(b).

There is another way to solve (5) for which the error analysis can be completed; this method is inspired by O'Cinneide (1994) [15]. Partition $I - P$ and recall the partition of X:

$$I - P = \begin{pmatrix} 1 - p_{11} & r_1' \\ -c_1 & I - Q \end{pmatrix} \quad X = \begin{pmatrix} 0 & 0 \\ x & X_1 \end{pmatrix},$$

where (p_{11}, c_1) and (p_{11}, r_1') are the first column and first row of P, respectively. Then $(I - Q)^{-1}$ is the fundamental matrix of the absorbing Markov chain whose transition matrix agrees with P except for the first row, where the absorbing chain has $p_{11} = 1$. This inverse is known to exist (Kemeny and Snell (1960) [11, Chap. III]), and is of interest in itself. Simple matrix algebra leads to

$$X_1 = (I - Q)^{-1}(I - W_1) \quad \text{and} \quad x = -\pi_1'(I - Q)^{-1}e,$$

where W_1 is the $(n-1) \times (n-1)$ southeast corner of W and $\pi_1' = (\pi_2, \ldots, \pi_n)$.

In terms of the unit roundoff error, $\mathbf{u}$ say, of the computer, Theorem 2 in O'Cinneide (1993) [14] shows that the relative error in the elements of π' is at most $2n^3/3\mathbf{u}$ when the GTH algorithm is used to compute them. Corollary 3 in O'Cinneide (1994) [15] shows that the relative error in the elements of $(I - Q)^{-1}$ is at most $2n^3/3\mathbf{u}$ when the GTH algorithm is used to compute them. For $\mathbf{u} = 10^{-14}$ (a conservative choice for current computers), when $n = 1{,}000$ these matrices have at least five accurate digits. The matrices on the rightmost side of (4) are computed accurately, and the errors from matrix addition and multiplication can be assessed in the usual way; see, e.g., Vandergraft (1983), [17, §2.4]. Since there are some subtractions, there is no bound on the relative errors. The algorithm of computing $(I - Q)^{-1}$ considered by O'Conneide requires $O(n^4)$ flops, so proceeding in this way will not yield an algorithm that is time-competitive with FUND. When $(I - Q)^{-1}$ is computed from the UL factorization, X_1 is given by (1c′) above.

4.3. Individual elements. There are situations where only a few elements of Z are needed. For example, if a particular mean first-passage-time is wanted, $m_{ij} (1 \neq j)$ say, (2) shows that only z_{jj} and z_{ij} are needed. (Since $m_{ii} = 1/\pi_i$ no further calculations are needed to compute it.) It is wasteful to compute the entire matrix Z when only a few elements are needed. Algorithm FUND is easily modified to compute a given z_{ij}; we proceed as follows.

Let e_j be the jth unit vector and x_j be the jth column of X. Writing (4) in scalar form yields

$$z_{ij} = \pi_j + x_{ij} - \pi' x_j. \tag{9}$$

The following algorithm computes any subset of the elements of the jth column of Z.

ALGORITHM FUNDIJ

0. Use the GTH algorithm to compute π', L, and U.
1. Solve $Uy = e_j - \mathbf{e}\pi_j$ by back substitution.
2. Solve $Lx_j = y$ by forward substitution.
3. Compute z_{ij} from (9).

$$\text{(a)} \quad p \leftarrow \sum_{k=1}^{n} \pi_k x_{kj}.$$

(b) For any i of interest, $z_{ij} = \pi_j + x_{ij} - p$.

The work in steps 1 and 2 each take $\Theta(n^2/2)$ flops, and the work in step 3 takes $\Theta(n^2)$ flops, so FUNDIJ requires negligible work (beyond computing π) to obtain any number of $\{z_{ij}\}$ (with j fixed).

5. Numerical examples. We present three numerical examples. The first can be done by hand and illustrates the algorithm FUND. The second requires a computer and demonstrates that FUND can have more accuracy than matrix inversion. The third shows that using (2) to compute M can produce very inaccurate results even when Z is computed accurately.

Example 1. This is the Land of Oz example from Kemeny and Snell. The transition matrix is

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}.$$

The GTH algorithm yields $S = \operatorname{diag}(0, 3/4, 1/2)$, $\pi' = (2/5, 1/5, 2/5)$, and

$$\bar{P} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{3}{4} & \frac{1}{4} & 1 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}$$

The UL factorization is

$$UL = \begin{pmatrix} -1 & \frac{1}{2} & \frac{1}{2} \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ \frac{3}{4} & -\frac{3}{4} & 0 \\ \frac{1}{4} & \frac{1}{4} & -\frac{1}{2} \end{pmatrix}.$$

Solving $UY = I - W$ by back substitution and then solving $LX = Y$ by forward substitution, setting the top row of X equal to zero, yields

$$Y = \frac{1}{5} \begin{pmatrix} 0 & 0 & 0 \\ 4 & -3 & -1 \\ 2 & 1 & -3 \end{pmatrix} \quad \text{and} \quad X = \frac{1}{15} \begin{pmatrix} 0 & 0 & 0 \\ -16 & 12 & 4 \\ -20 & 0 & 20 \end{pmatrix}.$$

Evaluating Z via (4) yields

$$Z = \frac{1}{75} \begin{pmatrix} 86 & 3 & -14 \\ 6 & 63 & 6 \\ -14 & 3 & 86 \end{pmatrix},$$

which is the known solution.

Example 2. This is test problem 3 from Harrod and Plemmons (1984) [4]. In Heyman and Reeves it is shown that this example causes numerical difficulties when $(I-P+W)^{-1}$ is computed via ordinary Gaussian elimination. The transition matrix is

$$P = \begin{pmatrix} 0.9999990 & 0.0000001 & 0.0000002 & 0.0000003 & 0.0000004 \\ 0.4 & 0.3 & 0 & 0 & 0.3 \\ 0.0000005 & 0 & 0.9999990 & 0 & 0.0000005 \\ 0.0000005 & 0 & 0 & 0.9999990 & 0.0000005 \\ 0.0000002 & 0.0000003 & 0.0000001 & 0.0000002 & 0.9999990 \end{pmatrix}.$$

The fundamental matrix was computed in double precision arithmetic via FUND and by matrix inversion using LINPACK. See, e.g., Coleman and Van Loan (1988) [1] for a description of LINPACK. Both methods agreed to ten decimal digits (which is all that were checked), so the following can be taken as the seven-decimal digit exact answer

$$Z = \begin{pmatrix} 499291.4 & -0.04741659 & -26036.96 & -196184.9 & -277068.5 \\ 173204.0 & 1.474332 & -58645.80 & -163576 & 49017.89 \\ -196360.7 & -0.1343731 & 837731.2 & -393286.3 & -248083.0 \\ -196360.7 & -0.1343731 & -162268.8 & 606713.7 & 248083.0 \\ -261578.2 & 0.1699747 & -102123.9 & -120098.0 & 483801.0 \end{pmatrix}.$$

The LINPACK subroutine DGECO gives 1.5×10^6 as the estimate of the condition number of the matrix being inverted. When π' is computed in double precision and the inverse is computed with single precision LINPACK subroutines, we obtain

$$Z_{\text{inv}} = \begin{pmatrix} 493850.9 & -0.04736168 & -23740.84 & -194982.2 & -275126.9 \\ 171866.7 & 1.473395 & -57133.48 & -162034.5 & 47300.76 \\ -195574.3 & -0.1320674 & 825822.5 & -387281.9 & -242965.2 \\ -195574.3 & -0.1320674 & -161072.5 & 599613.1 & 242965.2 \\ -257444.5 & 0.1677386 & -101656.7 & -118103.4 & 477205.4 \end{pmatrix}.$$

We see that Z_{inv} agrees with Z usually for two digits; sometimes three and sometimes one. This is consistent with a condition number of order 10^6 and eight-decimal digit accuracy in single precision.

When π' and the fundamental matrix calculated via FUND are computed in single precision, we obtain

$$Z_{\text{new}} = \begin{pmatrix} 499291.4 & -0.04741660 & -26036.96 & -196184.9 & -277068.5 \\ 173204.0 & 1.474332 & -58645.80 & -163576 & 49017.91 \\ -196360.7 & -0.1343731 & 837731.1 & -393286.3 & -248083.0 \\ -196360.7 & -0.1343731 & -162268.8 & 606713.7 & 248083.0 \\ -261578.2 & 0.1699747 & -102123.9 & -120098.0 & 483801.0 \end{pmatrix}.$$

All but three elements of Z_{new} are exact; one element is off by 2 and the other two are off by 1 in the seventh digit (counting from left to right).

Consider now computing the matrix of mean first-passage-times, M. Once Z is obtained, calculating M requires only matrix addition and subtraction, and multiplication involving a diagonal matrix. These operations can be done in $\Theta(n^2)$ flops each, so they are negligible compared to computing Z. Heyman and Reeves show that when single precision computation is used, for this example (2) achieves only three-digit accuracy when Z is computed by Gaussian elimination. Slightly more than seven-digit

accuracy is achieved when Z is computed via FUND, which is what is achieved by the algorithm given by Heyman and Reeves.

Example 3. The accuracy of M achieved in Example 2 does not always occur. The following transition matrix is test problem 4 from Harrod and Plemmons.

$$P = \begin{pmatrix} .1-\varepsilon & .3 & .1 & .2 & .3 & \varepsilon & 0 & 0 & 0 & 0 \\ .2 & .1 & .1 & .2 & .4 & 0 & 0 & 0 & 0 & 0 \\ .1 & .2 & .2 & .4 & .1 & 0 & 0 & 0 & 0 & 0 \\ .4 & .2 & .1 & .2 & .1 & 0 & 0 & 0 & 0 & 0 \\ .6 & .3 & 0 & 0 & .1 & 0 & 0 & 0 & 0 & 0 \\ \varepsilon & 0 & 0 & 0 & 0 & .1-\varepsilon & .2 & .2 & .4 & .1 \\ 0 & 0 & 0 & 0 & 0 & .2 & .2 & .1 & .3 & .2 \\ 0 & 0 & 0 & 0 & 0 & .1 & .5 & 0 & .2 & .2 \\ 0 & 0 & 0 & 0 & 0 & .5 & .2 & .1 & 0 & .2 \\ 0 & 0 & 0 & 0 & 0 & .1 & .2 & .2 & .3 & .2 \end{pmatrix}.$$

When $\varepsilon = 10^{-7}$, FUND produces at least seven accurate digits for each element of Z. However, substituting Z into (2) yields very poor accuracy for those elements of M in the second and fourth quadrants, and very good accuracy for the others. For example, there are no accurate digits in $m_{10,9}$ and eight accurate digits in $m_{10,1}$. The inaccuracy is caused by subtractive cancellation. This can be predicted from the structure of the transition matrix.

This P is *nearly completely decomposable* (NCD), which means that if P were written in block partition form, the blocks on the diagonal would have row sums that are close to one, and the other blocks would have elements that are close to zero. Here, states $\{1,2,3,4,5\}$ form a cluster, and states $\{6,7,8,9,10\}$ form another cluster. Transitions within a cluster occur frequently, and transitions between clusters occur with probability less than ε. For i and j in the same cluster, transitions from i to j occur at a rate that is commensurate with one; when i and j are in different clusters, transitions from i to j occur at a rate that is commensurate with ε. Consequently, when i and j are in the same cluster, the mean number of transitions from i to j that occur before the cluster is left is commensurate with $1/\varepsilon$.

Recall that $v_{ij}(n)$ is the expected number of visits to state j after n transition, starting in state i. When i and j are in the same cluster, $v_{ij}(n)$ will be much larger than if the initial state were chosen by chance according to π; the argument above shows that the increase is commensurate with $1/\varepsilon$. Thus (3) shows that z_{ij} is in the vicinity of $1/\varepsilon$. The scalar form of (2) is

$$m_{ij} = \frac{\delta_{ij} - z_{ij} + z_{jj}}{\pi_i}, \tag{10}$$

where δ_{ij} is one when $i = j$ and is zero otherwise. The second and fourth quadrants of M correspond to states in the same cluster, so (10) shows that these $\{m_{ij}\}$ are subject to catastrophic subtractive cancellation. When i and j are in different clusters, the bias in $v_{ij}(n)$ is large in magnitude and negative in sign, so (10) shows that the leading digits will be computed accurately.

The transition matrix in Example 2 is also NCD, but the clusters have only one state in them, so the subtractive cancellation in (10) is precisely right. This analysis suggests that (10) will produce accurate results when the matrix is not NCD because in those Markov chains, the bias terms, and hence z_{ij} and z_{jj}, will not be

extraordinarily large. Finally, we note that if only M is desired, some computational benefit is achieved by substituting (9) into (10), producing

$$m_{ij} = \frac{\delta_{ij} - x_{ij} + x_{jj}}{\pi_j},$$

which is (2) with X replacing Z.

Acknowledgments. I thank C. A. O'Cinneide for sending me advance copies of his papers, and the referees for their helpful comments. D. P. O'Leary made particularly useful comments on early drafts.

REFERENCES

[1] T. F. COLEMAN AND C. VAN LOAN, *Handbook for Matrix Computations*, Society for Industrial and Applied Mathematics, Philadelphia, 1988.

[2] W. K. GRASSMANN, M. I. TAKSAR, AND D. P. HEYMAN, *Regenerative analysis and steady-state distributions for Markov chains*, Ops. Res., 33 (1985), pp. 1107–1116.

[3] W. K. GRASSMANN, *Means and variances in Markov reward systems*, in Linear Algebra, Markov Chains and Queuing Models, C. D. Meyer and R. J. Plemmons, eds., Springer-Verlag, New York, 1993, pp. 193–204.

[4] W. J. HARROD AND R. J. PLEMMONS, *Comparisons of some direct methods for computing stationary distributions of Markov chains*, SIAM J. Sci. Statist. Comput., 33 (1984), pp. 453–469.

[5] D. P. HEYMAN AND M. J. SOBEL, *Stochastic Models in Operations Research*, Vol. 1, McGraw-Hill, New York, 1982.

[6] D. P. HEYMAN AND M. J. SOBEL, *Stochastic Models in Operations Research*, Vol. II, McGraw-Hill, New York, 1984.

[7] D. P. HEYMAN, *Further comparisons of some direct methods for computing stationary distributions of Markov chains*, SIAM J. Alg. and Disc. Meth., 8 (1987), pp. 226–232.

[8] D. P. HEYMAN AND A. REEVES, *Numerical solution of linear equations arising in Markov chain models*, ORSA J. Comput., 1 (1989), pp. 52–60.

[9] J. HUNTER, *Generalized inverses and their application to applied probability problems*, Linear Algebra Appl., 45 (1982), pp. 193–206.

[10] J. HUNTER, *Mathematical Techniques of Applied Probability*, Vol. II, Academic Press, New York, 1983.

[11] J. G. KEMENY AND J. L. SNELL, *Finite Markov Chains*, Van Nostrand, New York, 1960.

[12] J. G. KEMENY, *Generalizations of a fundamental matrix*, Linear Algebra Appl., 38 (1981), pp. 193–206.

[13] C. D. MEYER, *The role of the group generalized inverse in the theory of finite Markov chains*, SIAM Rev., 17 (1975), pp. 443–464.

[14] C. A. O'CINNEIDE, *Entrywise perturbation theory and error analysis for Markov chains*, Numer. Math., 65 (1993), pp. 109–120.

[15] C. A. O'CINNEIDE, *Relative-error bounds for direct algorithms for Markov chains*, Working paper, School of Industrial Engineering, Purdue University, West Lafayette, IN, 1994.

[16] G. W. STEWART, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[17] J. S. VANDERGRAFT, *Introduction to Numerical Computations*, Academic Press, New York, 1983.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 964–976, July 1995

017

PATTERN PROPERTIES AND SPECTRAL INEQUALITIES IN MAX ALGEBRA *

R. B. BAPAT†, DAVID P. STANFORD‡, AND P. VAN DEN DRIESSCHE§

Abstract. The max algebra consists of the set of real numbers, along with negative infinity, equipped with two binary operations, maximization and addition. This algebra is useful in describing certain conventionally nonlinear systems in a linear fashion. Properties of eigenvalues and eigenvectors over the max algebra that depend solely on the pattern of finite and infinite entries in the matrix are studied. Inequalities for the maximal eigenvalue of a matrix over the max algebra, motivated by those for the Perron root of a nonnegative matrix, are proved.

Key words. max algebra, eigenvalue, eigenvector, circuit mean, Frobenius normal form

AMS subject classifications. 15A18, 15A42, 05C38

1. Introduction.

The algebraic system called "max algebra" has been used to describe, in a linear fashion, phenomena that are nonlinear in the conventional algebra. Examples include transportation networks, machine scheduling, and parallel computation. A system in which one component must wait for results from other components (a "discrete event dynamic system") can be modeled in max algebra. See [13, Chap. 1] for a detailed description of such systems. As described there, the question of regularizing a system, that is, of initiating a system in such a way that all components begin cycles at the same time, is answered by solving the eigenproblem in max algebra.

An early exposition of max algebra is the monograph of Cuninghame-Green [13]. Related works are Carré [7, Chap. 3] and Gondran and Minoux [19], that discuss more general "path algebras" and describe Gaussian and related solutions of linear systems over path algebras. Currently, work on max algebra systems is progressing in many directions; see [1], [6], [11], [18], [24]. Over the max algebra, eigenproblems for irreducible matrices were studied in [13] and for reducible matrices in [8] and [18].

The max algebra consists of the set $\boldsymbol{M} = \mathbf{R} \cup \{-\infty\}$, where $\mathbf{R}$ is the set of real numbers, equipped with two binary operations, addition and multiplication, denoted by $\oplus$ and $\otimes$, respectively. The operations are defined as follows:

$$a \oplus b = \max(a, b), \text{ the maximum of } a \text{ and } b$$

and

$$a \otimes b = a + b.$$

* Received by the editors July 9, 1993; accepted (in revised form) by R. Cottle July 5, 1994. The research of these authors was undertaken at the University of Victoria.

†Indian Statistical Institute, New Delhi, 110016, India.

‡Department of Mathematics, College of William and Mary, Williamsburg, Virginia, 23185. This research was partially supported by a College of William and Mary Faculty Research Grant. This author presented some of the results of this article at the Second Symposium on Matrix Analysis and Applications held on 22–23 October 1993 at Western Michigan University.

§Department of Mathematics and Statistics, University of Victoria, Victoria, B.C., V8W 3P4, Canada. This research was partially supported by Natural Sciences and Engineering Research Council of Canada grant A-8965 and the University of Victoria Committee on Faculty Research and Travel (pvdd@smart.math.uvic.ca).

Clearly, $-\infty$ and 0 serve as identity elements for the operations $\oplus$ and $\otimes$, respectively. We denote $x_1 \oplus \cdots \oplus x_n$ by

$$\sum_{\oplus i=1}^{n} x_i,$$

or by $\sum_{\oplus} x_i$ when the range of summation of the index i is clear from the context.

We deal with vectors and matrices over the max algebra. Basic operations on matrices are defined in the natural way. Thus, if $A = [a_{ij}], B = [b_{ij}]$ are $m \times n$ matrices over $\boldsymbol{M}$, then $A \oplus B$ is the $m \times n$ matrix with (i,j)-entry $a_{ij} \oplus b_{ij}$. If $k \in \boldsymbol{M}$, then $k \otimes A$ is the matrix $[k \otimes a_{ij}] = [k + a_{ij}]$. If A is $m \times n$ and B is $n \times p$, then $A \otimes B$ is the $m \times p$ matrix with (i,j)-entry.

$$\sum_{k=1}^{n}{}_{\oplus} a_{ik} \otimes b_{kj} = \max_k (a_{ik} + b_{kj}).$$

It is easily verified that matrix multiplication is associative and that it distributes over matrix addition.

The transpose of the matrix A is denoted by A^T. The $n \times n$ matrix with each diagonal entry zero and each off-diagonal entry $-\infty$ is the identity matrix over the max algebra. If we permute the rows (and/or columns) of the identity matrix, then we obtain a permutation matrix over the max algebra. If A, B are $m \times n$ matrices over $\boldsymbol{M}$, then $A \geq B$ means that $a_{ij} \geq b_{ij}$ for all i, j. Similarly, $A > B$ means that $a_{ij} > b_{ij}$ for all i, j. A column or row vector x over $\boldsymbol{M}$ is said to be finite if each component x_i of the vector is finite. A vector is called partly infinite if it has a finite component as well as an infinite component. A matrix or vector with each component $-\infty$ is called infinite and we denote it by $-\infty$ as well; this should not cause any confusion.

The exponential function provides a natural one-to-one map from $\boldsymbol{M}$ onto the nonnegative reals. Under this correspondence, matrices over max algebra correspond to nonnegative matrices over the reals, and much of our work is motivated by the theory of nonnegative matrices. Techniques of proof for max algebra sometimes reflect those for conventional algebra. In particular, the directed graph of a matrix, which provides much information in the study of nonnegative matrices, plays an even more central role in matrices over max algebra; see the definition of $\mu(A)$ below.

Let A be an $n \times n$ matrix over $\boldsymbol{M}$. We associate a directed graph (digraph) $G(A)$ with A as follows. The vertices of $G(A)$ are $1, 2, \ldots, n$. There is an edge from vertex i to vertex j, denoted by (i,j), if a_{ij} is finite and in that case we say that a_{ij} is the weight of the edge (i,j). We use standard terminology from the theory of digraphs. Thus a path of length l in a digraph is a sequence of edges $(i_1, i_2), (i_2, i_3), \ldots, (i_l, i_{l+1})$, also denoted by $i_1 \to i_2 \to \cdots \to i_l \to i_{l+1}$; here the vertices are not necessarily distinct. The weight of a path is the sum of the weights of the edges in the path. The average weight of the path $i_1 \to i_2 \to \cdots \to i_l \to i_{l+1}$ is defined as

$$\frac{a_{i_1 i_2} + a_{i_2 i_3} + \cdots + a_{i_l i_{l+1}}}{l}.$$

A circuit τ of length l is a closed path $i_1 \to i_2 \to \cdots \to i_l \to i_1$, where $i_1, \ldots, i_l$ are distinct. A circuit of length one is a loop. We denote the set of circuits in $G(A)$, or in A, by $\boldsymbol{C}(A)$. If $\tau \in \boldsymbol{C}(A)$ then the average weight of τ is called the mean of the

circuit τ, denoted by $M_A(\tau)$. We define the maximal circuit mean of A, denoted by $\mu(A)$, as

$$\mu(A) = \max_{\tau \in \boldsymbol{C}(A)} M_A(\tau)$$

if $\boldsymbol{C}(A) \neq \phi$, and we set $\mu(A) = -\infty$ otherwise. A circuit $\tau \in \boldsymbol{C}(A)$ is called a critical circuit if $M_A(\tau) = \mu(A)$. The set of all critical circuits in A is denoted by $\tilde{\boldsymbol{C}}(A)$. The critical graph of A is a digraph with vertices $1, 2, \ldots, n$, defined as follows. For $i, j \in \{1, 2, \ldots, n\}$, edge (i, j) is in the critical graph of A if and only if it belongs to a critical circuit in $\boldsymbol{C}(A)$.

A digraph is strongly connected if there exists a path from any vertex to any other vertex. We say that the matrix A is irreducible if $G(A)$ is strongly connected. If A is not irreducible then we say that it is reducible. If A is an $n \times n$ matrix over $\boldsymbol{M}$ then clearly A is irreducible if and only if $[e^{a_{ij}}]$ is a nonnegative, irreducible matrix in the usual sense (see, e.g., [4]). We also remark that A is reducible if and only if either A is 1×1 containing $-\infty$ or there exists a permutation matrix Q_1 over the max algebra such that

$$Q_1 \otimes A \otimes Q_1^T = \begin{bmatrix} A_{11} & -\infty \\ A_{21} & A_{22} \end{bmatrix},$$

where A_{11} and A_{22} are square matrices of order at least one. For A reducible and not 1×1 containing $-\infty$, there exist $q \geq 2$ and a permutation matrix Q over the max algebra such that

$$Q \otimes A \otimes Q^T = \begin{bmatrix} A_{11} & -\infty & \cdots & -\infty \\ A_{21} & A_{22} & \cdots & -\infty \\ \vdots & \vdots & \ddots & \vdots \\ A_{q1} & A_{q2} & \cdots & A_{qq} \end{bmatrix}, \tag{1.1}$$

where each A_{ii} is either square and irreducible or is 1×1 containing $-\infty$. This is the Frobenius normal form of A.

In §2 we give the basic definitions and state results for eigenvalues and eigenvectors of general square matrices over the max algebra. Proofs of these results can be found in the literature. In applications to discrete event dynamic systems such as machine scheduling or parallel computing, it may be useful to obtain information about eigenvalues and eigenvectors given only partial information concerning the entries of the matrix. In particular, it may be known which components of the system must wait for input from which other components, while the waiting times are unknown. It will then be known where the finite entries of the matrix of interest occur, but their magnitudes will be unknown; that is, only the "pattern" of the matrix will be specified. In §3 we obtain results concerning eigenvalues and eigenvectors that depend only on the pattern of the given matrix. In §4 we present new inequalities concerning the maximal circuit mean of a matrix over the max algebra. Most of these are motivated by known corresponding inequalities for the spectral radius of a nonnegative matrix.

2. Eigenvalues and eigenvectors. Let A be an $n \times n$ matrix over $\boldsymbol{M}$, then $\lambda \in \boldsymbol{M}$ is an eigenvalue of A if there exists a vector $x \neq -\infty$ such that

$$A \otimes x = \lambda \otimes x.$$

In this case, x is an eigenvector of A corresponding to the eigenvalue λ. Furthermore, we call (λ, x) an eigenpair of A. Note that (λ, x) is an eigenpair of A if and only if

$x \neq -\infty$ and $\max_j(a_{ij} + x_j) = \lambda + x_i, i = 1, 2, \ldots, n$. For example, if

$$A = \begin{bmatrix} 3 & -\infty \\ 2 & 4 \end{bmatrix},$$

then

$$A \otimes \begin{bmatrix} -\infty \\ 0 \end{bmatrix} = 4 \otimes \begin{bmatrix} -\infty \\ 0 \end{bmatrix},$$

thus 4 is an eigenvalue of A. It can be checked that 4 is the only eigenvalue of A. Note that A^T has both 3 and 4 as eigenvalues.

If Q is a permutation matrix over the max algebra and $\lambda \in \boldsymbol{M}$ then (λ, x) is an eigenpair of A if and only if $(\lambda, Q \otimes x)$ is an eigenpair of $Q \otimes A \otimes Q^T$. In particular, A and $Q \otimes A \otimes Q^T$ have the same eigenvalues. In view of these observations we often find it convenient to deal with the Frobenius normal form of A in (1.1) instead of the matrix A itself. Note that $G(A)$ and $G(Q \otimes A \otimes Q^T)$ are identical except for labeling of the vertices.

We need the following basic spectral results, which can be found in [1], [8], [10], [12]–[14], [18]. Detailed proofs are also given in [3]. The first result deals with the occurrence of $-\infty$ as an eigenvalue, the other results deal with $\mu(A)$ as an eigenvalue, with A irreducible in the third result.

THEOREM 2.1. *Let A be an $n \times n$ matrix over $\boldsymbol{M}$. Then,*

(i) *$-\infty$ is an eigenvalue of A if and only if A has an infinite column, and*

(ii) *$-\infty$ is the only eigenvalue of A if and only if $\boldsymbol{C}(A) = \phi$.*

THEOREM 2.2. *Let A be an $n \times n$ matrix over $\boldsymbol{M}$. Then $\mu(A)$ is an eigenvalue of A. Moreover, if (λ, x) is an eigenpair with x finite, then $\lambda = \mu(A)$.*

THEOREM 2.3. *Let A be an $n \times n$ irreducible matrix over $\boldsymbol{M}$. Then,*

(i) *$\mu(A)$ is the only eigenvalue of A, and every eigenvector of A is finite,*

(ii) *A has a unique eigenvector (up to scalar multiple over the max algebra) if and only if the critical graph of A is strongly connected.*

Now suppose that A is reducible and is in Frobenius normal form (1.1). For $k = 1, 2, \ldots, q$, let V_k denote the set of indices of rows in A that intersect the diagonal block A_{kk}. The sets V_k are called the classes of A. If V_i and V_j are classes, we say V_j has access to V_i provided either $i = j$ or there is a $u \in V_j$ and a $v \in V_i$ such that there is a path from u to v in $G(A)$. Since each A_{jj} is either irreducible or $[-\infty]$, the relation "has access to" is reflexive and transitive. If $\mu(A_{jj}) > \mu(A_{ii})$ then we say that class V_j dominates class V_i. These definitions are used in the following result to specify the eigenvalues of A, for proofs see [3], [8, Thm. 1], [18, Chap. 4, Coro. 2.2.5].

THEOREM 2.4. *Let A be an $n \times n$ matrix over $\boldsymbol{M}$, which is in Frobenius normal form (1.1), and let $\lambda \in \boldsymbol{M}$. Then λ is an eigenvalue of A if and only if there is an i such that $\mu(A_{ii}) = \lambda$ and no class which dominates V_i has access to V_i.*

3. Pattern properties in max algebra. In this section we investigate spectral properties that depend only on the placement of finite and infinite entries in the matrix, and not on the magnitudes of the finite entries. Such properties are called "pattern properties" of the matrix.

A (square) pattern is an $n \times n$ array $P = [p_{ij}]$ of symbols chosen from $\{*, -\infty\}$. If A is an $n \times n$ matrix over $\boldsymbol{M}$, we write $A \in P$ provided

$$a_{ij} \in \mathbf{R} \text{ if } p_{ij} = *, \qquad a_{ij} = -\infty \text{ if } p_{ij} = -\infty.$$

Following [21], a pattern P is said to allow a particular property if there is a matrix $A \in P$ which has the property. P is said to require the property if every matrix $A \in P$

has the property. We determine which patterns allow, and which patterns require, various spectral properties in the max algebra.

The digraph $G(P)$ of an $n \times n$ pattern P has vertices $\{1, 2, \ldots, n\}$, and an edge from i to j if and only if $p_{ij} = *$. We denote the set of circuits in $G(P)$ by $\boldsymbol{C}(P)$. The concept of reducibility of a square matrix, introduced in §2, extends in an obvious way to patterns. Pattern P is irreducible if and only if $G(P)$ is strongly connected. It follows that P is reducible if and only if P is 1×1 containing $-\infty$, or if by an identical permutation of rows and columns P can be brought to the form

$$\begin{bmatrix} P_{11} & -\infty \\ P_{21} & P_{22} \end{bmatrix},$$

where P_{11} and P_{22} are square with order at least one. We will also deal with the Frobenius normal form of the pattern, defined analogously to that of a matrix; see (1.1).

We first discuss properties of the eigenvalues of a matrix determined by its pattern.

LEMMA 3.1. *Let P be a pattern. The following are equivalent.*

(i) *P requires a finite eigenvalue.*

(ii) *P allows a finite eigenvalue.*

(iii) *$\boldsymbol{C}(P)$ is not empty.*

Proof. The proof follows easily from Theorems 2.1 and 2.2. □

LEMMA 3.2. *Let P be a pattern. The following are equivalent.*

(i) *P requires $-\infty$ as an eigenvalue.*

(ii) *P allows $-\infty$ as an eigenvalue.*

(iii) *P has an infinite column.*

Proof. The proof follows immediately from Theorem 2.1. □

The following corollary is an immediate consequence of Lemmas 3.1 and 3.2.

COROLLARY 3.3. *Let P be a pattern.*

(i) *P requires that $-\infty$ be the only eigenvalue if and only if P allows the same property, and this occurs if and only if $\boldsymbol{C}(P)$ is empty.*

(ii) *P requires that all eigenvalues be finite if and only if P allows the same property, and this occurs if and only if P has no infinite column.*

THEOREM 3.4. *Let P be a pattern.*

(i) *P requires a unique and finite eigenvalue if and only if P has no infinite column and the Frobenius normal form of P has exactly one irreducible diagonal block.*

(ii) *P allows a unique and finite eigenvalue if and only if P has no infinite column.*

Proof. (i). We may assume without loss of generality that P is in Frobenius normal form. Suppose P requires a unique and finite eigenvalue. By Lemma 3.2, P has no infinite column. If P had a 1×1 diagonal block $[-\infty]$ in the lower right corner, P would have an infinite column. Hence the lower right diagonal block is irreducible. If P had another irreducible diagonal block, a matrix $A \in P$ could be constructed with the lower right diagonal block having eigenvalue 0 and another irreducible diagonal block having a positive eigenvalue. It follows from Theorem 2.4 that A would have two eigenvalues, one 0 and one positive, violating the fact that P requires a unique eigenvalue.

Now suppose that P has no infinite column and exactly one irreducible block, which then must be P_{qq}, the lower right block. Let $A \in P$. By Theorem 2.1, $-\infty$ is not an eigenvalue of A. By Theorem 2.2, A has an eigenvalue which, by Theorem 2.4, is $\mu(A_{ii})$ for some diagonal block A_{ii} in A. Since A_{qq} is the only irreducible diagonal block in A, $\mu(A_{qq}) > -\infty$ is the only eigenvalue of A.

(ii) If P allows a unique and finite eigenvalue, then P does not require $-\infty$ as an eigenvalue, so by Lemma 3.2 P has no infinite column. Conversely, if P has no infinite column, then the matrix $A \in P$ which has 0 in all the $*$ positions has the unique and finite eigenvalue 0. □

We now turn to pattern properties concerning the eigenvectors of a matrix. We obtain necessary and sufficient conditions on a pattern that it allow (or require) all (or some) eigenvectors to be finite (or partly infinite). Some of the results parallel those concerning partly zero eigenvectors in the conventional algebra presented in [23]. We remark that in the context of a discrete event dynamical system, the existence of a finite eigenvector implies that the system can be regularized. Note that the eigenpairs of the matrix (pattern) with each entry $-\infty$ are of the form $(-\infty, x)$ with $x \neq -\infty$. We exclude that pattern from consideration in the following.

THEOREM 3.5. *Let P be a pattern with at least one $*$. Then P requires that all eigenvectors be partly infinite if and only if P has an infinite row.*

Proof. First suppose P has no infinite row. Let $A \in P$ be obtained by replacing all $*$'s with 0's. Then the vector of all 0's is a finite eigenvector of A corresponding to the eigenvalue 0. Hence P does not require that all eigenvectors be partly infinite.

Now suppose that row i of P is infinite, but that $A \in P$ has a finite eigenvector x corresponding to eigenvalue λ. Then entry i of $A \otimes x$ is $-\infty$, so $\lambda \otimes x_i$ is $-\infty$. Since x_i is finite, $\lambda = -\infty$. Now if a_{jk} is finite, then entry j of $A \otimes x$ is finite, whereas entry j of $\lambda \otimes x = -\infty$. Hence $A = -\infty$, so $P = -\infty$, a contradiction. Therefore if P has an infinite row, then P requires that all eigenvectors be partly infinite. □

COROLLARY 3.6. *Let P be a pattern with at least one $*$. Then P allows a finite eigenvector if and only if P has no infinite row.*

THEOREM 3.7. *Let P be a pattern with at least one $*$. The following are equivalent.*

(i) *P is irreducible.*

(ii) *P requires that all eigenvectors be finite.*

(iii) *P allows all eigenvectors to be finite.*

Proof. (i) $\Rightarrow$ (ii). If $A \in P$ then A is irreducible, so by Theorem 2.3, all eigenvectors of A are finite. Therefore (i) $\Rightarrow$ (ii).

(ii) $\Rightarrow$ (iii) is trivial.

(iii) $\Rightarrow$ (i). Suppose that P is reducible, so that without loss of generality we may assume $P = \begin{bmatrix} P_{11} & -\infty \\ P_{21} & P_{22} \end{bmatrix}$. Let $A = \begin{bmatrix} A_{11} & -\infty \\ A_{21} & A_{22} \end{bmatrix} \in P$ be partitioned as P is. Let $x_{(2)}$ be an eigenvector of A_{22} corresponding to $\mu(A_{22})$, and let $x = \begin{bmatrix} -\infty \\ x_{(2)} \end{bmatrix}$. Then $A \otimes x = \mu(A_{22}) \otimes x$, so x is an eigenvector of A which is partly infinite. Hence P does not allow all eigenvectors to be finite. Therefore (iii) $\Rightarrow$ (i). □

COROLLARY 3.8. *Let P be a pattern with at least one $*$. The following are equivalent.*

(i) *P is reducible.*

(ii) *P allows a partly infinite eigenvector.*

(iii) *P requires a partly infinite eigenvector.*

Proof. The equivalence of (i) through (iii) in Theorem 3.7 implies the corollary. □

THEOREM 3.9. *Let P be a pattern with at least one $*$. Then P requires a finite eigenvector if and only if P has no infinite row and the Frobenius normal form of P has exactly one irreducible diagonal block.*

Proof. We may assume without loss of generality that P is in Frobenius normal form. Suppose P requires a finite eigenvector. By Theorem 3.5, P has no infinite

row. Therefore the upper left diagonal block P_{11} in P is irreducible. Suppose there is a $k > 1$ such that P_{kk} is irreducible. We will construct a matrix $A \in P$ with all eigenvectors partly infinite, contradicting the hypothesis on P. To do this, let

$$U_1 = [P_{21}^T \quad P_{31}^T \quad \cdots \quad P_{q1}^T]^T \quad \text{and} \quad U_2 = \begin{bmatrix} P_{22} & -\infty & -\infty & \cdots & -\infty \\ P_{32} & P_{33} & -\infty & \cdots & -\infty \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ P_{q2} & P_{q3} & \cdots & \cdots & P_{qq} \end{bmatrix}.$$

Then $P = \begin{bmatrix} P_{11} & -\infty \\ U_1 & U_2 \end{bmatrix}$, and P_{kk} is one of the diagonal blocks in U_2. Since P_{kk} is irreducible, P_{kk} has a circuit. Select a circuit in P_{kk} and set all its $*$ entries equal to 1. Set the other $*$ entries in U_2 to 0 to create a matrix $A_2 \in U_2$. Set all $*$ entries in P_{11} and U_1 to 0 to complete $A = \begin{bmatrix} A_{11} & -\infty \\ A_1 & A_2 \end{bmatrix} \in P$ with $\mu(A_{11}) = 0$ and $\mu(A) = \mu(A_2) = 1$. Suppose A has a finite eigenvector $x = \begin{bmatrix} x_{(1)} \\ x_{(2)} \end{bmatrix}$ partitioned to conform to the partition of A above. Since x is finite, the corresponding eigenvalue must be $\mu(A)$ by Theorem 2.2. But then $A_{11} \otimes x_{(1)} = 1 \otimes x_{(1)}$, which is impossible because the only eigenvalue of A_{11} is 0. Hence A cannot have a finite eigenvector and the desired contradiction is reached.

Now suppose P has no infinite row and has exactly one irreducible diagonal block, which must then be P_{11}. If $P = P_{11}$, that is if P is irreducible, then P requires all eigenvectors finite and we are through. Otherwise, let $q \geq 2$ be the number of diagonal blocks in P. Let $A \in P$ be partitioned as P is. We construct a finite eigenvector of A inductively as follows. Since A_{11} is irreducible, A_{11} has a finite eigenvector $x_{(1)}$ corresponding to its eigenvalue $\mu(A_{11})$. Let $x_2 = A_{21} \otimes x_{(1)} - \mu(A_{11})$, and for $2 \leq i < q$, let $x_{i+1} = [A_{i+1,1} A_{i+1,2} \ldots A_{i+1,i}] \otimes [x_{(1)}^T x_2 \ldots x_i]^T - \mu(A_{11})$, a finite member of $\boldsymbol{M}$. It then follows that $x = [x_{(1)}^T x_2 \ldots x_q]^T$ is a finite eigenvector of A corresponding to $\mu(A_{11})$, so P requires a finite eigenvector. □

COROLLARY 3.10. *Let P be a pattern with at least one $*$. Then P allows all eigenvectors to be partly infinite if and only if P has an infinite row or the Frobenius normal form of P has two irreducible diagonal blocks.*

Proof. Upon observing that a pattern with no infinite row must have a Frobenius normal form with the upper left diagonal block irreducible, the corollary follows immediately from Theorem 3.9. □

THEOREM 3.11. *Let P be a pattern. Then P allows a unique and finite eigenvector if and only if P is irreducible.*

Proof. Assume P is irreducible. Let $a_{ij} = 0$ whenever $p_{ij} = *$. Then A has a unique and finite eigenvector by Theorem 2.3. Assume P is reducible, then P requires a partly infinite eigenvector by Corollary 3.8. Thus P does not allow a unique and finite eigenvector. □

THEOREM 3.12. *Let P be a pattern. Then P requires a unique and finite eigenvector if and only if P is irreducible and the directed graph $G = G(P)$ does not contain two vertex-disjoint circuits.*

Proof. Assume P is irreducible and G does not have two vertex-disjoint circuits. Let $A \in P$. Then by Theorem 2.3, A has a unique eigenvalue which is $\mu(A)$, each eigenvector of A is finite, and A has a unique eigenvector if and only if the critical graph $\boldsymbol{C}$ of A is strongly connected. Now $\boldsymbol{C}$ is a subgraph of G and is a union of circuits. Since G does not have two vertex-disjoint circuits, $\boldsymbol{C}$ does not have two vertex-disjoint circuits. If i and j are vertices in $\boldsymbol{C}$, then i lies on a circuit $\boldsymbol{C}_i$ and j lies on a circuit $\boldsymbol{C}_j$. If $\boldsymbol{C}_i = \boldsymbol{C}_j$ there are paths from i to j and from j to i in

$\boldsymbol{C}$. Otherwise $\boldsymbol{C}_i$ and $\boldsymbol{C}_j$ share a vertex, and again there are paths from i to j and from j to i in $\boldsymbol{C}$. Hence $\boldsymbol{C}$ is strongly connected and the eigenvector of A is unique up to scalar multiples in the max algebra. Therefore P requires a unique and finite eigenvector.

Now assume P requires a unique and finite eigenvector. If P were reducible, then by Corollary 3.8 P would allow a partly infinite eigenvector. Hence P is irreducible. Suppose G has two vertex-disjoint circuits. Then we may select two vertex disjoint circuits in G and construct a matrix $A \in P$ which has 1 in the positions belonging to either of the two circuits and 0 and $-\infty$ elsewhere. Then the circuit means are 1 on each of the two circuits and less than 1 on each other circuit, so the critical graph of A is the union of the two disjoint circuits and is not strongly connected. Hence the eigenvector of A is not unique, contradicting the assumption on P. Hence G does not have two vertex-disjoint circuits. □

4. Inequalities. Many of the results in this section are motivated by known inequalities for the spectral radius (or the Perron root) $\rho(B)$ of a nonnegative matrix B. Thus, Lemma 4.1 and Corollary 4.2 are analogs of well-known bounds for the Perron root; see, for example, [4, p. 28] and [25, p. 31]. Theorem 4.3 is the max algebra version of a result due to Birkhoff and Varga [5]. The parallels between inequalities for $\mu(A)$, where A is a matrix over $\boldsymbol{M}$ and $\rho(B)$, where B is a nonnegative matrix, are quite striking and remain to be fully explored. Theorem 4.9 is yet another result in this direction. Let A be an $n \times n$ matrix over $\boldsymbol{M}$ and let B be the Hadamard exponential of A, i.e., $b_{ij} = e^{a_{ij}}$ for all i, j. Then $e^{\mu(A)}$ is the maximal circuit geometric mean of the nonnegative matrix B. We remark that the maximal circuit geometric mean of a nonnegative matrix has been considered in the literature; see, e.g., [15], [17], [22].

The following lemma is stated and proved in [18, Chap. 4, Lemmas 1.3.8, 1.3.9].

LEMMA 4.1. *Let A be an $n \times n$ matrix over $\boldsymbol{M}$ and $\eta \in \boldsymbol{M}$. Then $\mu(A) \geq \eta$, if and only if there exists a vector $z \neq -\infty$ such that $A \otimes z \geq \eta \otimes z$. Furthermore, if A is irreducible, then $\mu(A) \leq \eta$, if and only if there exists a vector $z \neq -\infty$ such that $A \otimes z \leq \eta \otimes z$.*

COROLLARY 4.2. *Let A be an $n \times n$ matrix over $\boldsymbol{M}$. Then*

$$\min_i \max_j a_{ij} \leq \mu(A) \leq \max_{i,j} a_{ij}.$$

Proof. Let $\alpha = \min_i \max_j a_{ij}$ and let 0 denote the vector with each component zero. Then $A \otimes 0 \geq \alpha \otimes 0$. It follows from Lemma 4.1 that $\mu(A) \geq \alpha$. It is easy to see that for any $\sigma \in C(A)$, $M_A(\sigma) \leq \max_{i,j} a_{ij}$, and hence $\mu(A) \leq \max_{i,j} a_{ij}$, giving the second inequality. □

Let A be an $n \times n$ matrix over $\boldsymbol{M}$. By Theorem 2.2, $\mu(A)$ is an eigenvalue of A and there is a vector $x \neq -\infty$ such that $A \otimes x = \mu(A) \otimes x$. We refer to x as a right eigenvector of A corresponding to $\mu(A)$. Since $\mu(A) = \mu(A^T)$, there is a vector $y \neq -\infty$ as a left eigenvector of A corresponding to $\mu(A)$. We note that (by Theorem 2.3) if A is irreducible, then x and y are finite and $\mu(A)$ is the only eigenvalue of A.

THEOREM 4.3. *Let A be an $n \times n$ irreducible matrix over $\boldsymbol{M}$. Then the following assertions hold.*

(i) $\mu(A) = \max_{x > -\infty} \min_{y > -\infty} (y^T \otimes A \otimes x - y^T \otimes x)$.

(ii) $\mu(A) = \min_{y > -\infty} \max_{x > -\infty} (y^T \otimes A \otimes x - y^T \otimes x)$.

Proof. For any finite vectors x, y, we have

$$\begin{aligned} y^T \otimes A \otimes x &= \max_{i,j}(a_{ij} + y_i + x_j) \\ &= \max_{i,j}(a_{ij} + x_j - x_i + y_i + x_i) \\ &\geq \min_i \max_j (a_{ij} + x_j - x_i) + y^T \otimes x. \end{aligned}$$

Therefore,

$$y^T \otimes A \otimes x - y^T \otimes x \geq \min_i \max_j (a_{ij} + x_j - x_i). \tag{4.1}$$

Suppose

$$\min_i \max_j (a_{ij} + x_j - x_i) = \max_j (a_{kj} + x_j - x_k).$$

Let z be the vector with $z_k = -x_k$, with the remaining components chosen finite and so that $z^T \otimes x = 0$ and satisfying

$$\max_j (a_{ij} + z_i + x_j) \leq \max_j (a_{kj} + z_k + x_j), \qquad i = 1, 2, \ldots, n.$$

When we set $y = z$, equality holds in (4.1) and hence we have shown that for any finite x,

$$\min_{y > -\infty} (y^T \otimes A \otimes x - y^T \otimes x)$$

exists. Thus by (4.1)

$$\min_{y > -\infty} (y^T \otimes A \otimes x - y^T \otimes x) = \min_i \max_j (a_{ij} + x_j - x_i).$$

Let $S = [a_{ij} + x_j - x_i]$. Then $\mu(A) = \mu(S)$ and by Corollary 4.2

$$\min_i \max_j (a_{ij} + x_j - x_i) \leq \mu(S).$$

Therefore, we conclude that

$$\mu(A) \geq \sup_{x > -\infty} \min_{y > -\infty} (y^T \otimes A \otimes x - y^T \otimes x). \tag{4.2}$$

When we set x to be a right eigenvector of A, we see that for any finite $y, y^T \otimes A \otimes x - y^T \otimes x = \mu(A)$. Thus, (i) follows from (4.2). The proof of (ii) is similar. □

We next give an easy inequality, and then characterize the case of equality.

LEMMA 4.4. *Let X, Y be $n \times n$ matrices over $\boldsymbol{M}$ such that $X \geq Y$. Then $\mu(X) \geq \mu(Y)$.*

Proof. The result is obvious if $\boldsymbol{C}(Y) = \phi$, since in that case, $\mu(Y) = -\infty$. So suppose $\boldsymbol{C}(Y) \neq \phi$. For any $\sigma \in \tilde{\boldsymbol{C}}(Y)$,

$$\mu(Y) = M_Y(\sigma) \leq M_X(\sigma) \leq \mu(X)$$

and the proof is complete. □

Observe that Lemma 4.4 shows that if Z is a principal submatrix of X, then $\mu(X) \geq \mu(Z)$.

To discuss the case of equality in Lemma 4.4, we now introduce some notation. Suppose σ is the circuit $(i_1 \quad i_2 \quad \cdots \quad i_k)$; in this notation we assume i_1 to be the least integer among $i_1, i_2, \ldots, i_k$ and this convention makes the representation of the circuit uniquely determined. If X is an $n \times n$ matrix and if $\sigma = (i_1 \quad i_2 \quad \cdots \quad i_k) \in \boldsymbol{C}(X)$, then we define $X(\sigma)$ as the vector

$$[x_{i_1 i_2} \quad x_{i_2 i_3} \quad \cdots \quad x_{i_k i_1}]^T.$$

LEMMA 4.5. *Let X, Y be $n \times n$ matrices over $\boldsymbol{M}$ such that $X \geq Y$ and suppose $\mu(Y)$ is finite. Then the following conditions are equivalent.*

(i) $\mu(X) = \mu(Y)$.

(ii) *There exists* $\sigma \in \tilde{\boldsymbol{C}}(X) \cap \tilde{\boldsymbol{C}}(Y)$ *such that* $M_X(\sigma) = M_Y(\sigma)$.

(iii) *There exists* $\sigma \in \tilde{\boldsymbol{C}}(X)$ *such that* $M_X(\sigma) = M_Y(\sigma)$.

(iv) $\tilde{\boldsymbol{C}}(Y) \subset \tilde{\boldsymbol{C}}(X)$ *and for all* $\sigma \in \tilde{\boldsymbol{C}}(Y), X(\sigma) = Y(\sigma)$.

(v) $\tilde{\boldsymbol{C}}(Y) \subset \tilde{\boldsymbol{C}}(X)$ *and there exists* $\sigma \in \tilde{\boldsymbol{C}}(Y)$ *such that* $X(\sigma) = Y(\sigma)$.

Proof. First observe that since $\mu(Y)$ is finite, and $X \geq Y$, $\mu(X)$ is finite and $C(Y), C(X)$ are nonempty.

(i) $\Rightarrow$ (ii). Let $\sigma \in \tilde{\boldsymbol{C}}(Y)$. Then

$$\mu(Y) = M_Y(\sigma) \leq M_X(\sigma) \leq \mu(X) \tag{4.3}$$

and since $\mu(X) = \mu(Y)$, equality holds throughout in (4.3). It follows that $\sigma \in \tilde{\boldsymbol{C}}(X) \cap \tilde{\boldsymbol{C}}(Y)$ and $M_X(\sigma) = M_Y(\sigma)$.

(iii) $\Rightarrow$ (i). Let $\sigma \in \tilde{\boldsymbol{C}}(X)$ such that $M_X(\sigma) = M_Y(\sigma)$. Then $\mu(X) = M_X(\sigma) = M_Y(\sigma) \leq \mu(Y) \leq \mu(X)$, and hence $\mu(X) = \mu(Y)$.

(i) $\Rightarrow$ (iv). Let $\sigma \in \tilde{\boldsymbol{C}}(Y)$. As in the proof of (i) $\Rightarrow$ (ii), equality holds throughout in (4.3). It follows that $\sigma \in \tilde{\boldsymbol{C}}(X)$ and $M_X(\sigma) = M_Y(\sigma)$. Since $X \geq Y$, we have $X(\sigma) \geq Y(\sigma)$. If $X(\sigma) \neq Y(\sigma)$, then it will follow, after taking the sum of the entries in $X(\sigma), Y(\sigma)$, that $M_X(\sigma) > M_Y(\sigma)$, which is a contradiction. Thus $X(\sigma) = Y(\sigma)$.

It is easy to see that (ii) $\Rightarrow$ (iii), (iv) $\Rightarrow$ (v), and (v) $\Rightarrow$ (i). That completes the proof. □

THEOREM 4.6. *Let $X_1, \ldots, X_m$ be $n \times n$ matrices and let $X = \sum_{\oplus} X_i$. Then*

$$\mu(X) \geq \sum\nolimits_{\oplus} \mu(X_i). \tag{4.4}$$

Furthermore, equality holds in (4.4) *if and only if one of the following conditions is satisfied.*

(i) $\mu(X) = -\infty$.

(ii) $\mu(X)$ *is finite and there exists* $\sigma \in \tilde{\boldsymbol{C}}(X)$ *and* $k \in \{1, 2, \ldots, m\}$ *such that* $X_k(\sigma) \geq X_i(\sigma), i = 1, 2, \ldots, m$.

Proof. If $\mu(X) = -\infty$, then $\mu(X_i) = -\infty, i = 1, 2, \ldots, m$ and both sides in (4.4) are $-\infty$. So we assume that $\mu(X)$ is finite. Since $X \geq X_i, i = 1, 2, \ldots, m$, by Lemma 4.4, we have $\mu(X) \geq \mu(X_i), i = 1, 2, \ldots, m$ and hence (4.4) holds.

If equality holds in (4.4) then there exists $k \in \{1, 2, \ldots, m\}$ such that $\mu(X) = \mu(X_k)$. Thus $\mu(X_k)$ is finite. By Lemma 4.5 (see (i) $\Rightarrow$ (v)), there exists $\sigma \in \tilde{\boldsymbol{C}}(X)$ such that $X(\sigma) = X_k(\sigma)$. It follows that $X_k(\sigma) \geq X_i(\sigma), i = 1, 2, \ldots, m$.

Conversely, if (ii) holds, then $X(\sigma) = X_k(\sigma)$. Thus $\mu(X_k)$ is finite. Set $Y = X_k$ and use implication (iii) $\Rightarrow$ (i) of Lemma 4.5 to conclude that equality holds in (4.4). □

A square matrix D is a diagonal matrix over the max algebra if $d_{ij} = -\infty$ for all $i \neq j$. A well-known result due to Cohen [9] (see also [20, p. 364]) asserts that the Perron root of a nonnegative matrix B is a convex function of the diagonal entries of B. In this context the next result is somewhat surprising since it says that $\mu(A)$, considered as a function of the diagonal entries of A, is linear over the max algebra.

THEOREM 4.7. *Let X be an $n \times n$ matrix over $\boldsymbol{M}$ and let $D_1, \ldots, D_m$ be $n \times n$ diagonal matrices over the max algebra. Then*

$$\mu\left(X \oplus \sum\nolimits_{\oplus} D_j\right) = \sum\nolimits_{\oplus} \mu(X \oplus D_j). \tag{4.5}$$

Proof. Let $X_j = X \oplus D_j, j = 1, 2, \ldots, m$. Then $X \oplus \sum_{\oplus} D_j = \sum_{\oplus} X_j$. If $\mu(X \oplus \sum_{\oplus} D_j) = -\infty$, then (4.5) is true by Theorem 4.6. So let $\mu(X \oplus \sum_{\oplus} D_j)$ be finite. If there exists $\sigma \in \tilde{\boldsymbol{C}}(X \oplus \sum_{\oplus} D_j)$ of length more than one, then $\sigma \in \tilde{\boldsymbol{C}}(X \oplus D_j)$ $j = 1, 2, \ldots, m$ and (4.5) is proved, in view of (ii) $\Rightarrow$ (i) of Lemma 4.5. So suppose that every circuit in $\tilde{\boldsymbol{C}}(X \oplus \sum_{\oplus} D_j)$ is of length one, and let σ be one such. Clearly, there exists $k \in \{1, 2, \ldots, m\}$ such that $D_k(\sigma) \geq D_j(\sigma)$, and hence $X_k(\sigma) \geq X_i(\sigma), i = 1, 2, \ldots, m$. Thus (ii), Theorem 4.6 is satisfied, and (4.5) holds. □

Let $C \neq -\infty$ be an $n \times n$ matrix over $\boldsymbol{M}$ and

$$\Omega(C) = \left\{(i, j) : c_{ij} = \max_{k,l} c_{kl}\right\}.$$

Construct a $(0, 1)$ matrix $\hat{C} = [\hat{c}_{ij}]$ by setting $\hat{c}_{ij} = 1$ if $(i, j) \in \Omega(C)$ and $\hat{c}_{ij} = 0$ otherwise. Let $\gamma = \sum_{s=1}^{n} \sum_{t=1}^{n} \hat{c}_{st}$, and for $i, j = 1, 2, \ldots, n$, let

$$\alpha_i(C) = \frac{1}{\gamma} \sum_{t=1}^{n} \hat{c}_{it} \quad \text{and} \quad \beta_j(C) = \frac{1}{\gamma} \sum_{s=1}^{n} \hat{c}_{sj}.$$

With this notation, we have the following result, which is the max algebra analog of [2, Thm. 3].

LEMMA 4.8. *Let A be an $n \times n$ matrix over $\boldsymbol{M}$, with $A \neq -\infty$, let u, v, w, z be vectors over $\boldsymbol{M}$ with w and z finite, and let $C = [a_{ij} \otimes z_i \otimes w_j]$. Then*

$$v^T \otimes A \otimes u - z^T \otimes A \otimes w \geq \sum_{i=1}^{n} \alpha_i(C)(v_i - z_i) + \sum_{j=1}^{n} \beta_j(C)(u_j - w_j).$$

Proof. For any i, j, we have

$$a_{ij} \otimes v_i \otimes u_j - a_{ij} \otimes z_i \otimes w_j = v_i - z_i + u_j - w_j. \tag{4.6}$$

If $(i, j) \in \Omega(C)$, then $a_{ij} \otimes z_i \otimes w_j = z^T \otimes A \otimes w$. Apply (4.6) to each $(i, j) \in \Omega(C)$ and add the resulting equations to get

$$\sum_{(i,j) \in \Omega(C)} a_{ij} \otimes v_i \otimes u_j - \gamma(z^T \otimes A \otimes w) = \sum_{(i,j) \in \Omega(C)} (v_i - z_i) + \sum_{(i,j) \in \Omega(C)} (u_j - w_j). \tag{4.7}$$

Now

$$\sum_{(i,j) \in \Omega(C)} (v_i - z_i) = \sum_{i=1}^{n} \left\{ (v_i - z_i) \sum_{j=1}^{n} \hat{c}_{ij} \right\} = \gamma \sum_{i=1}^{n} \alpha_i(C)(v_i - z_i),$$

and similarly

$$\sum_{(i,j)\in\Omega(C)} (u_j - w_j) = \gamma \sum_{j=1}^{n} \beta_j(C)(u_j - w_j).$$

Since $\sum_{(i,j)\in\Omega(C)} a_{ij} \otimes v_i \otimes u_j \leq \gamma(v^T \otimes A \otimes u)$ the result follows from (4.7) after a trivial simplification. □

Let B be an $n \times n$ nonnegative, irreducible matrix. Then it is known, see [16], that

$$f^T B g \geq \rho(B) f^T g,$$

where f and g are right and left Perron eigenvectors of B, respectively. We now obtain a max algebra analog of this result. In the special case of an irreducible matrix A with $G(A)$ having a unique critical circuit, a proof based on Lemma 4.8 is contained in [3]. For the more general result, we give an alternative proof that was suggested by an anonymous referee.

THEOREM 4.9. *Let A be an $n \times n$ matrix over $\boldsymbol{M}$ and let x and y be right and left eigenvectors, respectively, of A corresponding to the eigenvalue $\mu(A)$. Let u, v be n-vectors over $\boldsymbol{M}$ such that $u_i \otimes v_i = x_i \otimes y_i$ for $i = 1, 2, \ldots, n$. Then $v^T \otimes A \otimes u \geq \mu(A) \otimes y^T \otimes x$. In particular, $x^T \otimes A \otimes y \geq \mu(A) \otimes y^T \otimes x$.*

Proof. The result is trivial if $\mu(A) = -\infty$. Assume then that $\mu(A)$ is finite, and so there is a critical circuit in $G(A)$. Let $F = \{i\colon x_i \text{ is finite}\}$ and let H be the digraph with vertex set F and edge set $E = \{(i,j)\colon i, j \in F \text{ and } a_{ij} + x_j - x_i = \mu(A)\}$. Thus, from the right eigenequation,

$$\max_j (a_{ij} + x_j) = \mu(A) + x_i,$$

every vertex in H has outdegree at least one in H. Furthermore, for each $i \in F$, there is a path from i to a circuit τ_i in H, which must be a critical circuit in $G(A)$. The left eigenequation gives

$$a_{ij} + y_i \leq \mu(A) + y_j,$$

for each $(i,j) \in E$. Hence $x_i + y_i \leq x_j + y_j$ for each $(i,j) \in E$. Thus if τ is a circuit of length $|\tau|$ in H, there is a number k_τ such that $x_i + y_i = k_\tau$ for all vertices i lying along the circuit τ. Also if $i \in F$, then $x_i + y_i \leq k_{\tau_i} \leq \max_{\tau\in\Gamma} k_\tau$, where Γ denotes the set of all circuits in H, thus $\Gamma \subseteq \tilde{\boldsymbol{C}}(A)$. We have

$$\begin{aligned}
v^T \otimes A \otimes u &= \max_{i,j}(v_i + a_{ij} + u_j) = \max_{i,j}(x_i + y_i - u_i + a_{ij} + u_j) \\
&\geq \max_{\tau\in\Gamma}\{\max_{(i,j)\in\tau}(x_i + y_i + a_{ij} - u_i + u_j)\} \\
&\geq \max_{\tau\in\Gamma}\left\{\frac{1}{|\tau|}\sum_{(i,j)\in\tau}(x_i + y_i + a_{ij} - u_i + u_j)\right\} \\
&= \max_{\tau\in\Gamma}\left\{\frac{1}{|\tau|}\sum_{(i,j)\in\tau}(x_i + y_i + a_{ij})\right\} = \mu(A) + \max_{\tau\in\Gamma} k_\tau \\
&= \mu(A) + \max_{i\in F}(x_i + y_i) = \mu(A) \otimes y^T \otimes x.
\end{aligned}$$

The second inequality in the theorem follows by setting $v = x, u = y$. □

Acknowledgments. We thank Dr. Stéphane Gaubert of INRIA, France, for sending us a copy of his thesis [18]. We also thank anonymous referees for pointing out reference [8], and for their constructive comments that led to improvements, especially in Theorem 4.9.

REFERENCES

[1] F. L. BACCELLI, G. COHEN, G. J. OLSDER, AND J.-P. QUADRAT, *Synchronization and Linearity: An Algebra for Discrete Event Systems*, Wiley, Chichester, 1992.

[2] R. B. BAPAT, *Applications of an inequality in information theory to matrices*, Linear Algebra Appl., 78 (1986), pp. 107–117.

[3] R. B. BAPAT, D. P. STANFORD, AND P. VAN DEN DRIESSCHE, *The eigenproblem in max algebra*, DMS-631-IR, University of Victoria, British Columbia, 1993.

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

[5] G. BIRKHOFF AND R. S. VARGA, *Reactor criticality and nonnegative matrices*, SIAM J., 6 (1958), pp. 354–377.

[6] J. G. BRAKER AND G. J. OLSDER, *The power algorithm in max algebra*, Linear Algebra Appl., 182 (1993), pp. 67–89.

[7] B. CARRÉ, *Graphs and Networks*, Clarendon Press, Oxford, 1979.

[8] W. CHEN, X. QI, AND S. DENG, *The eigen-problem and period analysis of the discrete-event system*, Systems Sci. Math. Sci., 3 (1990), pp. 243–260.

[9] J. E. COHEN, *Convexity of the dominant eigenvalue of an essentially nonnegative matrix*, Proc. Amer. Math. Soc., 81 (1981), pp. 657–658.

[10] G. COHEN, D. DUBOIS, J.-P. QUADRAT, AND M. VIOT, *Analyse du comportment périodique des systèmes de production par la théorie des dioïdes*, INRIA Rep. 191, Le Chesnay, France, 1983.

[11] ———, *A linear-system-theoretic view of discrete-event processes and its use for performance evaluation in manufacturing*, IEEE Trans. Automat. Control, 30 (1985), pp. 210–220.

[12] G. COHEN, P. MOLLER, J.-P. QUADRAT, AND M. VIOT, *Algebraic tools for the performance evaluation of discrete event systems*, Proc. IEEE, 77 (1989), pp. 39–58.

[13] R. A. CUNNINGHAME-GREEN, *Minimax Algebra*, Lecture Notes in Economics and Math. Systems, 166, Springer-Verlag, 1979.

[14] P. I. DUDNIKOV AND S. N. SAMBORSKII, *Endomorphisms of finitely generated free semimodules*, in Idempotent Analysis, V. P. Maslov and S. N. Samborskii, eds., Advances Soviet Math., 13 (1992), pp. 65–85.

[15] G. M. ENGEL AND H. SCHNEIDER, *Diagonal similarity and equivalence for matrices over groups with 0*, Czechoslovak Math. J., 25 (1975), pp. 389–403.

[16] M. FIEDLER, C. R. JOHNSON, T. MARKHAM, AND M. NEUMANN, *A trace inequality for M-matrices and the symmetrizability of a real matrix by a positive diagonal matrix*, Linear Algebra Appl., 71 (1985), pp. 81–94.

[17] S. FRIEDLAND, *Limit eigenvalues of nonnegative matrices*, Linear Algebra Appl., 74 (1986), pp. 173–178.

[18] S. GAUBERT, *Théorie des Systèmes Linéaires dans des Dioïdes*, Thèse, L'École des Mines de Paris, Paris, 1992.

[19] M. GONDRAN AND M. MINOUX, *Graphs and Algorithms*, John Wiley & Sons Ltd., 1984.

[20] R. A. HORN AND C. R. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, 1991.

[21] C. R. JOHNSON, *Combinatorial matrix analysis: an overview*, Linear Algebra Appl., 107 (1988), pp. 3–15.

[22] S. KARLIN AND F. OST, *Some monotonicity properties of Schur powers of matrices and related inequalities*, Linear Algebra Appl., 68 (1985), pp. 47–65.

[23] J. S. MAYBEE, D. D. OLESKY, AND P. VAN DEN DRIESSCHE, *Partly zero eigenvectors*, Linear Multilinear Algebra, 28 (1990), pp. 83–92.

[24] G. J. OLSDER AND C. ROOS, *Cramer and Cayley–Hamilton in max algebra*, Linear Algebra Appl., 101 (1988), pp. 87–108.

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

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 977–1003, July 1995

018

ACCURATE EIGENSYSTEM COMPUTATIONS BY JACOBI METHODS*

ROY MATHIAS†

Abstract. Demmel and Veselić showed that, subject to a minor proviso, Jacobi's method computes the eigenvalues and eigenvectors of a positive definite matrix more accurately than methods that first tridiagonalize the matrix. We extend their analysis and thereby:

1. We remove the minor proviso in their results and thus guarantee the accuracy of Jacobi's method.
2. We show how to cheaply check, a posteriori, whether tridiagonalizing a particular matrix has caused a large relative perturbation in the eigenvalues on the matrix. This can be useful when dealing with graded matrices.
3. We derive hybrid Jacobi algorithms that have the same accuracy of Jacobi's method but are faster, at least on a serial computer.
4. We show that if G is an $m \times n$ matrix and $m >> n$ then Jacobi's method computes the singular values almost as quickly as standard methods, but potentially much more accurately.

Key words. Jacobi, symmetric eigenvalue problem, singular value decomposition, graded matrix, error analysis, Hilbert matrix

AMS subject classifications. 65F15, 65G05, 15A18, 15A42, 15A48

1. Introduction.

Jacobi's method computes the eigenvalues and eigenvectors of a positive definite matrix more accurately than methods based on first tridiagonalizing the matrix. Indeed, there is a sense in which Jacobi's method computes the eigenvalues and eigenvectors to optimal accuracy. These results were proved again in [4] and proved again in [14]; some precursors were presented in [1]. We refer to [4], [17], [15] for a complete survey of the literature. The purpose of this paper is to strengthen and generalize the results in [4] and to simplify some of the proofs there. This paper deals with both the one-sided and two-sided Jacobi algorithms. However, we make only passing reference to threshold criteria and orderings as the results here are independent of such considerations.

Given a positive definite matrix H, let S_H denote the positive diagonal scaling matrix such that the main diagonal entries of $S_H H S_H$ are all 1. That is,

$$S_H = \operatorname{diag}(H_{11}^{-1/2}, H_{22}^{-1/2}, \ldots, H_{nn}^{-1/2}). \tag{1.1}$$

A key idea in [4] was that if H is a positive definite matrix then the *relative* perturbation of the eigenvalues of H caused by the perturbation δH is bounded by

$$\frac{\|S_H \delta H S_H\|}{\lambda_{\min}(S_H H S_H)}. \tag{1.2}$$

The standard bound is

$$\frac{\|\delta H\|}{\lambda_{\min}(H)}, \tag{1.3}$$

* Received by the editors April 26, 1993; accepted for publication (in revised form) by F. T. Luk July 15, 1994.

† Department of Mathematics, College of William and Mary, Williamsburg, Virginia 23187 (na.mathias@na-net.ornl.gov). This research was supported in part by National Science Foundation grant DMS-9201586 and was done while the author was visiting the Institute for Mathematics and Its Applications at the University of Minnesota.

which is potentially much larger. It was also shown in [4] that the bound (1.2) is optimal, up to factors of n. The other key idea in [4] is that if a two-sided Jacobi transformation is applied to H using arithmetic of precision ϵ, then the result is the same as if a Jacobi transformation were applied in exact arithmetic to $H + \delta H$ where $\|S_H \delta H S_H\| = O(\epsilon)$. Combining this backward error bound with the new perturbation bound (1.2) shows that one step of Jacobi's method causes a relative perturbation in each of the eigenvalues of the order of $\epsilon \lambda_n^{-1}(S_H H S_H)$ at the most. Merely introducing a relative perturbation of size ϵ in each of the entries of H, as we may do in entering H into the computer, could cause a similar relative perturbation in its eigenvalues. Thus Jacobi's method is "as accurate as we can hope for" subject to the ratio in (1.5) not being too large. The standard bound on the relative error in the eigenvalues computed by other methods (tridiagonalization-based methods or Jacobi with the old stopping criterion) is

$$\frac{|\lambda_i(H) - \hat{\lambda}_i|}{\lambda_i(H)} \leq c_n \epsilon \frac{\|H\|}{\lambda_i(H)},$$

which can be as large as $c_n \epsilon \kappa(H)$, which is potentially much larger. Here c_n is a modestly growing function of n.

We generalize their result to show that a similar bound holds for a wider class of orthogonal transformations that are applied to two rows and the corresponding columns. This result shows that if one computes an eigendecomposition of a positive definite matrix by QR (or some other method based on a preliminary tridiagonalization) and refines it by Jacobi, then the resulting decomposition is as accurate as the Jacobi algorithms in [4] but considerably faster on a serial computer.

The error bounds for Jacobi's method presented in [4] contain a factor

$$\max_{0 \leq i \leq M} \lambda_n^{-1}(S_{H_i} H_i S_{H_i}), \tag{1.4}$$

where $H_0 = H$ is the $n \times n$ positive definite matrix whose eigenvalues we wish to compute and the H_i are the iterates in Jacobi's method.[1] However, as already mentioned, the perturbation bounds only contain the factor $\lambda_n^{-1}(S_H H S_H)$. It is conceivable that the ratio

$$\frac{\max_{0 \leq i \leq M} \lambda_n^{-1}(S_{H_i} H_i S_{H_i})}{\lambda_n^{-1}(S_H H S_H)} \tag{1.5}$$

is very large, although in practice this has not been observed. We show that if we use the one-sided Jacobi method applied to the Cholesky factor of H (as proposed in [4]), then we can replace the factor $\max_{0 \leq i \leq M} \lambda_n^{-1}(S_{H_i} H_i S_{H_i})$ by $\lambda_n^{-1}(S_{H_0} H_0 S_{H_0})$ (Theorem 3.3). That is, the one-sided Jacobi method computes the eigenvalues as accurately as we can hope, up to factors of n. This is perhaps the most important result in this paper. The fact that one can replace the factor (1.4) by $\lambda_n(S_H H S_H)$ has been shown independently by Drmač [6]. There are situations where one would use an algorithm that has the factor (1.4) in the bound—such a situation is described at the beginning of §4—so it is still of interest to study the ratio (1.5).

[1] This factor arises because the error at each step is bounded $\lambda_n^{-1}(S_{H_i} H_i S_{H_i})\epsilon$. Since the sequence $H_i, i = 0, 1, 2, \ldots$ converges to a diagonal matrix, it follows that the sequence $\lambda_n^{-1}(S_{H_i} H_i S_{H_i}), i = 0, 1, 2, \ldots$ converges to 1, but it is possible that it increases before decreasing to 1.

One of the themes of this paper is that the one-sided algorithm is easier to analyze and has better error bounds. The one-sided algorithm was preferred in [4] for parallelism and efficiency of data movement, and was used as the basis for a Jacobi method to accurately compute the eigenvalues of an indefinite symmetric matrix in [16], and a relative error analysis of the method was presented in [15]. Another theme is the analysis of algorithms at the matrix level rather than the scalar level. This makes the proofs much shorter and easier to understand. It also suggests how these results can be generalized. For example, by essentially the same method we can analyze the relative errors in the eigenvalues introduced by Householder tridiagonalization of a positive definite matrix.

An idea that has not been observed until now is that taking a Cholesky factorization, then computing the singular values of the Cholesky factor, by bidiagonalization for example, and finally squaring them yields the eigenvalues of the original matrix to higher relative accuracy than computing the eigenvalues of the positive definite matrix directly by tridiagonalization. This approach has been suggested in conjunction with Jacobi's method in [4]—there the reason was that if one computes the factorization with complete pivoting then Jacobi's method tends to converge more rapidly; accuracy was not the primary consideration.

Let us briefly compare Jacobi's method with tridiagonalization methods. The relative error bound for Jacobi's method will be much better than for tridiagonalization methods if and only if

$$\kappa(S_H H S_H) << \kappa(H). \tag{1.6}$$

It is easy to check that $\kappa(H) \leq \kappa(S_H)^2 \kappa(S_H H S_H)$. So (1.6) will be true only if $\kappa(S_H)$ is large—loosely speaking, if H is graded. It is natural to try to accelerate Jacobi's method in the case that H is graded because generally Jacobi's method is much slower than tridiagonalization followed by the QR iteration. We discuss this idea in §7. Demmel and Veselić have presented experimental evidence to show that as $\kappa(S_H)$ grows (for fixed n and $\kappa(S_H H S_H)$) the number of Jacobi sweeps required decreases [4, p. 1243], and they present a heuristic explanation for this observation that is valid for one-sided Jacobi. In [13] is shown that one can compute the eigenvalues of a *strongly* graded positive definite matrix to high relative accuracy at a cost that is only slightly greater than the cost of computing its Cholesky factorization. Thus, in this situation, one can obtain the accuracy of Jacobi's method at a cost that is less than that of tridiagonalizing the matrix.

We now give an outline of the paper and then conclude this section with some notation. In §2 we present the necessary preliminaries—some error bounds for floating point computations, some perturbation bounds for scaled matrices, and a review of the improved stopping criterion for Jacobi's method proposed in [4].

In §3 we analyze one variant of one-sided Jacobi. This is the simpler case where we apply the transformations on the left and the columns of the matrix are close to orthogonal. This allows us to avoid the factor (1.4). We also consider the reduction of rectangular matrices to square in this section.

In §4 we consider the other case, where we apply the orthogonal transformations and the scaling matrix on the same side, and the closely related two-sided Jacobi algorithm. We also use the analysis of §4 in §5 to show how one can check at each stage, while tridiagonalizing a positive definite matrix, whether large relative perturbations are being introduced in the eigenvalues. This is useful when working with graded positive definite matrices.

In §6 we give an application to the computation of the the eigenvalues of the Hilbert matrix. The main point here is that if one uses one-sided Jacobi applied to the Cholesky factor of a positive definite matrix, then the relative errors in the computed eigenvalues are due almost entirely to the errors in computing the Cholesky factor. This was observed in the numerical results presented [4] and hinted at in [17]—we give an explanation. The Hilbert matrix provides a dramatic illustration of this fact.

In §7 we present some hybrid algorithms that may be viewed as QR with Jacobi refinement. These algorithms deliver the accuracy of Jacobi's method but require considerably less time on a serial computer (but more time than plain QR). These new algorithms, though not completely parallelizable, are of interest since Jacobi's method is the fastest known way to compute all the eigenvalues of a positive definite matrix to maximum relative accuracy, even on a serial computer.

Let $M_{m,n}$ denote the space of $m \times n$ matrices and let $M_n \equiv M_{n,n}$. We only consider real matrices, but our results generalize to complex matrices in the obvious way since we have bounds of the form (1.7) for complex arithmetic. For a symmetric matrix H we let $\lambda_1(H) \geq \lambda_2(H) \geq \cdots \geq \lambda_n(H)$ denote its eigenvalues ordered in decreasing order. For $G \in M_{m,n}$ we let $\sigma_1(G) \geq \sigma_2(G) \geq \cdots \geq \sigma_{\min\{m,n\}}(G)$ denote its ordered singular values, and let $G_{\cdot i}$ denote its ith column $i = 1, \ldots, n$. For matrices let $\|\cdot\|$ denote the spectral (or 2-) norm and let $\|\cdot\|_F$ denote the Frobenius norm, i.e.,

$$\|X\| = \sigma_1(X), \qquad \|X\|_F^2 = \operatorname{trace}(X^T X).$$

We use κ to denote the condition number with respect to the spectral norm, i.e.,

$$\kappa(X) = \|X\| \, \|X^{-1}\|.$$

For vectors, $\|\cdot\|$ denotes the Euclidean norm. We always use $|X|$ to denote the entry-wise absolute value of a matrix or vector X. For $G \in M_{m,n}$ we define R_G (C_G) to be the positive diagonal matrix such that the rows (columns) of $R_G G$ (GC_G) have unit Euclidean length. Given a perturbation δH of a positive definite matrix H we refer to $S_H \delta H S_H$ as the *scaled perturbation*, and similarly for perturbations of matrices when we are considering row or column scalings. Recall that $S_H = \operatorname{diag}(H_{11}^{-1/2}, H_{22}^{-1/2}, \ldots, H_{nn}^{-1/2})$. We use the term *elementary orthogonal matrix* to mean a orthogonal matrix that differs from the identity in only two rows and columns. We use $R_{ij}(t)$ to denote the elementary orthogonal matrix that is the identity except that

$$[R_{ij}(t)]_{ii} = [R_{ij}(t)]_{jj} = c = (1 + t^2)^{-1/2}$$

and

$$[R_{ij}(t)]_{ij} = -[R_{ij}(t)]_{ji} = s = t(1 + t^2)^{-1/2}.$$

Here c and s are the cosine and sine of the rotation angle. Any elementary orthogonal matrix is either a rotation (like $R_{ij}(t)$) or the product of a rotation and a signed permutation. Since multiplication by a signed permutation does not cause any rounding error, for purposes of error analysis, it is sufficient to consider only matrices of the form $R_{ij}(t)$. We refer to an algorithm that applies a sequence of elementary orthogonal matrices as a *generalized Jacobi algorithm.* In a generalized Jacobi algorithm it is not necessary that each transformation orthogonalize a pair of columns (in the one-sided case) or zero a particular element (in the two-sided case).

We use the model of finite precision arithmetic with precision ϵ that does not assume the use of a guard digit:

$$
\begin{aligned}
fl(a * b) &= a * b(1+\epsilon_1) && * = \cdot \text{ or } / \\
fl(a * b) &= a(1+\epsilon_2) \pm b(1+\epsilon_3) && * = + \text{ or } - \\
fl(\sqrt{a}) &= \sqrt{a}(1+\epsilon_4),
\end{aligned} \tag{1.7}
$$

where $|\epsilon_i| \leq \epsilon$. Here $fl(\cdot)$ denotes the computed value. As in [4] the use of a guard digit does not significantly improve our error bounds. All our results are to first order in ϵ. In the statements of our results we include the term $O(\epsilon^2)$ to remind the reader of this, however, in the proofs we drop second-order terms for convenience.

As observed earlier in this section the relative perturbation in the eigenvalues of a positive definite matrix H caused by a perturbation δH is bounded by $\|S_H \delta H S_H\| \lambda_n^{-1}(S_H H S_H)$. However, even though this was proved by Demmel and Veselić in [4] they use the weaker bound

$$\|S_H \delta H S_H\| \kappa(S_H H S_H) = \|S_H \delta H S_H\| \lambda_1(S_H H S_H) \lambda_n^{-1}(S_H H S_H).$$

Because the main diagonal entries of $S_H H S_H$ are one and it is positive definite, its norm is at most n, and so the two bounds are the same, up to a factor of n. We use the stronger bound and so many of our results will appear to be rather different from those in [4], [15] though in fact they are essentially the same.

2. Preliminaries. In this section we give two error bounds for finite precision computations, two perturbation bounds from [4], the stopping criteria for Jacobi's method proposed in [4], and a bound on the scaled backward error in computing the Cholesky factorization. This section may be skipped by a reader who is only interested in the results in the rest of the paper. In the interest of brevity we omit the proofs of Lemmas 2.1 and 2.3—they are standard error analysis.

LEMMA 2.1. *Let $w, x, y \in R^m$ where $m > 1$. Let ϵ be the precision and let ϵ_i denote a real number of absolute value at most ϵ. Then assuming that the inner product $x^T y$ is computed as*

$$x^T y = x_1 y_1 + (x_2 y_2 + (\cdots)), \tag{2.1}$$

then

$$fl(x^T y) = x^T y + 2(m-1)\epsilon_1 |x|^T |y| + O(\epsilon^2). \tag{2.2}$$

If $\|w\| = \sqrt{2}$ then

$$\|fl(y - (w^T y)w) - (y - (w^T y)w)\| \leq 4m\epsilon_2 \|y\| + O(\epsilon^2). \tag{2.3}$$

If $Q \in M_m$ is orthogonal then

$$\|fl(Qy) - Qy\| \leq 2(m-1)m^{1/2}\epsilon_3 \|y\| + O(\epsilon^2). \tag{2.4}$$

For any $A, \in M_{l,m}$ and $B \in M_{m,n}$, with $m > 1$

$$|fl(AB) - AB| \leq 2(m-1)\epsilon |A|\,|B| + O(\epsilon^2). \tag{2.5}$$

The importance of the next result will be apparent from the discussion following Theorems 4.2 and 4.4. The quantity $t = t(a,b,c)$ is such that the 2×2 orthogonal matrix $R_{12}(t)$ diagonalizes the matrix in (2.6).

LEMMA 2.2. *Let*

$$\begin{pmatrix} a & c \\ c & b \end{pmatrix} \tag{2.6}$$

be positive definite and $c \neq 0$. *Let*

$$\zeta = \frac{b-a}{2c}, \qquad t(a,b,c) = \frac{\operatorname{sign}(\zeta)}{|\zeta| + \sqrt{1+\zeta^2}}.$$

Then

$$|t(a,b,c)| \max\left\{\sqrt{\frac{a}{b}}, \sqrt{\frac{b}{a}}\right\} \leq 1. \tag{2.7}$$

Proof. Without loss of generality we may assume that $a \leq b$ and that $c \geq 0$. Then $\zeta \geq 0$ and hence $t = (\zeta + \sqrt{1+\zeta^2})^{-1}$ is a decreasing function of ζ. Because the matrix is positive definite it follows that $\zeta = (b-a)/2c \geq (b-a)/2\sqrt{ab}$. Substituting this lower bound on ζ we obtain $t(a,b,c) \leq \sqrt{a/b}$. The inequality (2.7) follows from this bound. □

For error bounds it may be important that (2.7) hold for the *computed* value of t and it is possible that the computed value of t contains a large relative error in the absence of a guard digit and so, in order that (2.7) still hold for the computed value of $t(a,b,c)$, we may take the tangent of the rotation angle to be

$$\hat{t} = \min\left\{fl(t(a,b,c)), \sqrt{\frac{a}{b}}, \sqrt{\frac{b}{a}}\right\}.$$

Since we often need to compute $fl(1+t^2)$ and $fl(\sqrt{1+t^2})$, it is worth stating the error bounds in a lemma.

LEMMA 2.3. *Let* $t \in R$. *Then in finite precision arithmetic with precision* ϵ,

$$fl(1+t^2) = (1+t^2)(1+2\epsilon_1) + O(\epsilon^2)$$

and

$$fl(\sqrt{1+t^2}) = \sqrt{1+t^2}(1+2\epsilon_2) + O(\epsilon^2),$$

where $|\epsilon_i| \leq \epsilon$.

The next two scaled perturbation bounds are very useful results from [4].

THEOREM 2.4. [4, Theorem 2.3] *Let* H *be a positive definite matrix. Let* $\delta H =$ *be such that* $\|S_H \delta H S_H\| < \lambda_{\min}(A)$. *Then*

$$\frac{|\lambda_i(H) - \lambda_i(H+\delta H)|}{\lambda_i(H)} \leq \frac{\|S_H \delta H S_H\|}{\lambda_{\min}(S_H H S_H)}. \tag{2.8}$$

THEOREM 2.5. [4, Theorem 2.14] *Let* $G \in M_{m,n}$ *be of full column rank. Let* δG *be such that* $\|\delta G C_G\| < \sigma_n(GC_G)$. *Then*

$$\frac{|\sigma_i(G) - \sigma_i(G+\delta G)|}{\sigma_i(G)} \leq \frac{\|\delta G C_B\|}{\sigma_n(GC_G)}. \tag{2.9}$$

In the Jacobi algorithm we generate a sequence of matrices H_k that converge to diagonal. We stop when the off-diagonal elements are sufficiently small and then take the diagonal elements as the eigenvalues. How small is small enough? Suppose that $H_M = \Lambda + E$ where H_M is positive definite, Λ is diagonal, and E has zero diagonal. Suppose that

$$|S_{H_M} E S_{H_M}|_{ij} \leq \text{tol} \quad i,j = 1,\ldots,n \tag{2.10}$$

and that tol $<<$ 1. This implies that $\|S_{H_M} H_M S_{H_M}\| \leq (n-1)$ tol and hence that

$$\lambda_n(S_{H_M} H_M S_{H_M}) = \lambda_n(I + S_{H_M} E S_{H_M}) \geq 1 - (n-1)\text{tol}$$

and so by Theorem 2.4

$$\frac{|\lambda_i(H_M) - \lambda_i(\Lambda)|}{\lambda_i(H_M)} \leq \frac{(n-1)\text{tol}}{1-(n-1)\text{tol}} + O(\text{tol}^2) = (n-1)\text{tol} + O(\text{tol}^2).$$

The relative error introduced in at least one of the eigenvalues computed using Jacobi's method will generally be of the order of $\epsilon\lambda_n^{-1}(S_H H S_H)$. If one is content to compute all the eigenvalues of H to this relative accuracy, it would be appropriate to use the stopping criterion (2.10) with tol $\approx \epsilon\lambda_n^{-1}(S_H H S_H)$, which may be much larger than ϵ. In [4] it was proposed that one take tol $\approx c\epsilon$ for some constant that was independent of the matrix H. However, as just argued, tol $\approx \epsilon\lambda_n^{-1}(S_H H S_H)$, which will generally be larger and so result in earlier termination, will give the same accuracy for the computed eigenvalues (the eigenvectors may not be orthogonal to full working precision however). On the other hand, it was shown in [4, Proposition 2.4] that there is a relative condition number associated with *each* eigenvalue and that some eigenvalues may be more sensitive to small componentwise perturbations than others. It is possible, though it was not shown in [4], that Jacobi's method with tol $= c\epsilon$ will compute each eigenvalue to a relative accuracy that reflects its relative condition number. To summarize, taking tol $= \epsilon\lambda_n^{-1}(S_H H S_H)$ is sufficient to compute each eigenvalue to the relative accuracy one can expect for the most sensitive eigenvalue, but tol $= c\epsilon$ may compute each eigenvalue to the relative accuracy that it deserves.

We are *not* proposing that one set h_{ij} to zero based on the stopping criterion (2.10). If $|h_{ij}| \leq \text{tol}\sqrt{h_{ii}h_{jj}}$ for some pair of indices i,j, but not for all pairs, then we cannot be sure that $\lambda_n(S_H H S_H) \approx 1$ as we can when (2.10) holds for all pairs of indices. Thus in the interest of obtaining maximal accuracy one should not set off diagonal elements to zero—one should either perform the Jacobi transformation or, if $|h_{ij}|$ is small, one should not perform the transformation and move to the next pair of indices leaving h_{ij} unchanged. (One could, if one wished, set h_{ij} to zero if $|h_{ij}| \leq \epsilon\sqrt{h_{ii}h_{jj}}$ (note ϵ not tol), however, there is no evidence that faster convergence would result.)

Similarly, if we are applying right-handed Jacobi to compute the singular values of a matrix $G \in M_{m,n}, n \leq m$ to high relative accuracy, then it would be appropriate to stop when

$$\frac{|G_{\cdot i}^T G_{\cdot j}|}{\sqrt{(G_{\cdot i}^T G_{\cdot i})(G_{\cdot j}^T G_{\cdot j})}} \leq \text{tol}, \qquad i,j = 1,\ldots,n \tag{2.11}$$

for a tolerance tol $\approx \epsilon\sigma_n^{-1}(GC_G)$.

The criterion (2.10) itself is not new, but this justification for it was first presented in [4]. Until then the generally accepted termination criterion was

$$|E_{ij}| \leq \text{tol}\|H\| \quad i,j = 1,\ldots,n.$$

It is easy to check that this criterion is more easily satisfied than (2.10).

It will be useful to have the following backward error bound for the Cholesky factorization from [4], [5].

LEMMA 2.6. [5, Lemma 4.14] *Let L be the Cholesky factor of the positive definite matrix H computed by Algorithm* 4.3 *in* [5] (*or* [4]) *in finite precision arithmetic with precision ϵ. Then $LL^T = H + E$ where $|E_{ij}| \leq (n+5)\epsilon\sqrt{H_{ii}H_{jj}}$.*

The proof of this lemma is in [5] and is straightforward. Algorithm 4.3 in [5] is just the gaxpy version of the Cholesky factorization; one can expect a similar result for the outer product algorithm.

3. The easy case. In this section we consider the case where the scaling matrix and the orthogonal transformation are applied on different sides.

Note that if $Q \in M_m$ is orthogonal and $G \in M_{m,n}$ then the Euclidean lengths of the columns of QG are the same as those of G, and hence that

$$\sigma_n(QGC_{QG}) = \sigma_n(GC_{QG}) = \sigma_n(GC_G). \tag{3.1}$$

This simple fact makes the analysis of the case where we apply the orthogonal transformations and the scaling matrix on different sides much easier than when they are applied on the same side. The identity (3.1) is what allows us to avoid the factor (1.4), since at each step the singular values of the scaled matrix are the same.

THEOREM 3.1. *Let $Q \in M_m$ be orthogonal and let $G \in M_{m,n}$. Let ϵ be the arithmetic precision. Then in finite precision arithmetic with precision ϵ*

$$\|(QG - fl(QG))C_G\| \leq c\sqrt{n}\epsilon, \tag{3.2}$$

where $c = 2(m-1)m^{1/2}$. If Q is an elementary orthogonal matrix then (3.2) *holds with $c = 2\cdot 2^{1/2} \leq 3$ and if Q is a Householder reflection (applied as $G_1 = G - v(Gv)^T$) then* (3.2) *holds with $c = 4m$.*

Proof. The results are a straightforward application of Lemma 2.1. □

When computing the singular values of a rectangular matrix, in the interests of computational efficiency, one would like to reduce the matrix to a square matrix with the same singular values before applying some other method. We now show that this can be done without losing relative accuracy of the computed singular values.

THEOREM 3.2. *Let $G \in M_{m,n}$ $(m \geq n)$ have* rank n. *Let*

$$\begin{pmatrix} \hat{R} \\ 0 \end{pmatrix}$$

be the computed upper triangular factor in the QR *factorization of G computed by Householder* QR *in finite precision arithmetic with precision ϵ. Assume that for each Householder transformation $I - ww^T$, the computed value of w differs from the true value in norm by at most $c\epsilon$. Then to first order in ϵ*

$$\sigma_n(\hat{R}C_{\hat{R}}) = \sigma_n(GC_G) \tag{3.3}$$

and

$$\left|\frac{\sigma_i(G) - \sigma_i(\hat{R})}{\sigma_i(G)}\right| \leq (3mn^{3/2} + 2\sqrt{2}cn)\sigma_n^{-1}(GC_G)\epsilon + O(\epsilon^2). \tag{3.4}$$

We have been intentionally vague in our definition of c in order that the result is not dependent on the way in which w is computed. If one uses the method outlined in [18, pp. 153–155] then from [18, p. 155, (39.23)] one can take $c = 1.501$.

Proof. The statement (3.3) follows from the discussion just before Theorem 3.1.

Now let us consider (3.4)—first assuming that the Householder vectors w are computed exactly, but applied in finite precision arithmetic. Let $\hat{G}_i$ denote the computed matrix after the ith column has been reduced to upper triangular form, and let G_i denote the corresponding exact quantity. Let $E_i = \hat{G}_i - G_i$. Let Q_i be product of the first i Householder. By Lemma 2.1 the columns of $E_1 C_G$ have norms at most $4m\epsilon$. Let $\tilde{E}_2$ be the matrix or errors incurred at the second step, that is, $\tilde{E}_2 = fl(Q_2\hat{G}_1) - Q_2\hat{G}_1$. Then by substituting for $\hat{G}_1$ we have

$$E_2 = Q_2 E_1 + \tilde{E}_2.$$

Note that the first column of $\tilde{E}_2 C_G$ is zero and that the norms of the remaining columns are bounded by $4m\epsilon$. Thus the norm of the first column of $E_2 C_G$ is at most $4m\epsilon$, while the norms of the remaining columns are at most twice this quantity. The final error matrix is E_n. Continuing in this manner we see that the ith column of $E_n C_G$ has norm at most $4mi\epsilon$. Thus

$$\|E_n C_G\| \le \|E_n C_G\|_F = \|E_n C_G\|_F \le 4m\epsilon \sqrt{\sum_{i=1}^{n} i^2} \le 4m\epsilon\sqrt{n^3/3} \le 3mn^{3/2}\epsilon.$$

The bound

$$\left| \frac{\sigma_i(G) - \sigma_i(\hat{R})}{\sigma_i(G)} \right| \le 3mn^{3/2}\sigma_n^{-1}(GC_G)\epsilon + O(\epsilon^2)$$

on singular values follows from this and Theorem 2.5.

The additional term $2\sqrt{2}cn$ arises due to the inaccuracy in computing w. □

THEOREM 3.3. *Let $G \in M_n$. Suppose that a sequence of elementary orthogonal matrices when applied on the left to G produces G_M, after M transformations, and that G_M^T satisfies the termination criterion* (2.11). *Let s_i be the ordered computed (not exact), Euclidean row lengths of G_M. Then*

$$\left| \frac{s_i - \sigma_i(G)}{\sigma_i(G)} \right| \le 3M\sqrt{n}\sigma_n^{-1}(GC_G)\epsilon + n \cdot \text{tol} + 2n^2\epsilon + O(\epsilon^2). \tag{3.5}$$

Proof. Now combine Theorem 3.1 and the bound in Theorem 2.5, noting that $\sigma_n(GC_G) = \sigma_n(G_i C_{G_i})$ where G_i is the result after i transformations. The $n \cdot$ tol term arises because the columns of G_M are not exactly orthogonal and the $2n^2\epsilon$ arises because of possible errors in evaluating the $G_{\cdot i}^T G_{\cdot j}$ terms and s_i terms. □

Note that if we apply general orthogonal transformations rather than elementary orthogonal matrices, then we have a similar bound except that the factor 3 in (3.5) is replaced by $2(m-1)m^{1/2}$.

From this result, the previous result, and the discussion between (2.10) and (2.11) we can see that one can compute the singular values of $G \in M_{m,n}$ $(m \ge n)$ to a relative accuracy of the order of $\sigma_m^{-1}(GC_G)\epsilon$ by the following algorithm.

ALGORITHM 3.4. Given $G \in M_{m,n}$, $m \ge n$;

1. reduce G to upper triangular form $QG = \binom{R}{0}$;
2. obtain a lower bound σ on $\sigma_n(RC_R)$;
3. set tol $= \sigma^{-1}\epsilon$;
4. apply left-handed Jacobi to R.

The idea of setting tol $= \sigma^{-1}\epsilon$ (rather that tol $= c\epsilon$) will save at most one sweep of Jacobi if $\sigma_n(RC_R) \geq \sqrt{\epsilon}$ (as will typically be the case). This is not an enormous saving, but more than compensates for the small cost ($O(n^2)$) of estimating $\sigma_n(RC_R)$. Note also that since we have an estimate of $\sigma_n(GC_G) = \sigma_n(RC_R)$, we have an upper bound on the relative error in the computed singular values. However, because we are using tol $= \sigma^{-1}\epsilon$ rather than tol $= c_n\epsilon$ (where c_n is a modest function of n) in the stopping criterion, the computed singular vectors will not be orthogonal to full working precision. If this is a concern then one should take tol $= c_n\epsilon$ and do the (small amount of) extra work that this entails.

Since it is possible to reduce a rectangular matrix to a smaller square matrix without causing large relative perturbations in its singular values, we only consider square matrices hereafter.

Suppose that $G \in M_{m,n}$ with $m >> n$. The standard approach to finding the singular values of G is to first reduce G to an $n \times n$ upper triangular matrix with the same singular values, then bidiagonalize the $n \times n$ matrix and finally compute the singular values of the bidiagonal matrix. Most of the computational effort in this procedure is in the initial QR factorization (since $m >> n$). Consequently, Algorithm 3.4 is not significantly more expensive than the standard approach in this situation, but yields the singular values to maximal relative accuracy.

Demmel and Veselić [4] obtained a similar result for left-handed Jacobi applied to $n \times n$ matrices, except that because they applied the orthogonal transformations and the scaling matrices on the same side, they had the factor

$$\max_{0 \leq i \leq M} \sigma_n^{-1}(R_{G_i} G_i) \tag{3.6}$$

in their bounds rather than just $\sigma_n^{-1}(R_G G)$. Here G_i are the iterates produced by the left-handed Jacobi. The quantity in (3.6) is harder to compute exactly than $\sigma_n^{-1}(GC_G)$. However, in practice it was found to be equal to $\sigma_n^{-1}(R_{G_0} G_0)$, or not much larger.

As a corollary to Theorem 3.3, we can obtain an easily computed bound on the relative error in the eigenvalues of a positive definite matrix computed by one-sided generalized Jacobi applied to the Cholesky factor of the matrix. In [4] it was observed that doing Cholesky with complete pivoting tends to accelerate convergence, but there is no need for pivoting for this result to be valid, nor for the validity of the corresponding result in [4].

THEOREM 3.5. *Let $H \in M_n$ be positive definite and let $\hat{L}$ be its computed Cholesky factor. Assume that after M steps of right-handed generalized Jacobi applied to $\hat{L}$ the resulting matrix L_M satisfies* (2.11). *Let s_i be the ordered column lengths of L_M, which we take to be the computed singular values. Then*

$$\left| \frac{s_i^2 - \lambda_i(H)}{\lambda_i(H)} \right| \leq \left[\frac{n^2 + 5n}{\lambda_n(S_H H S_H)} + \frac{6M\sqrt{n}}{\sqrt{\lambda_n(S_H H S_H)}} \right] \epsilon + 4n^2 \cdot \text{tol} + O(\epsilon^2). \tag{3.7}$$

Proof. It is easy to check that $\sigma_n^{-1}(\hat{L} C_{\hat{L}}) = \sqrt{\lambda_n^{-1}(S_H H S_H)}$, at least to first order in ϵ. The result follows from combining the bound on the backward error in

the computation of $\hat{L}$ (Lem. 2.6) with that resulting from errors in Jacobi's method (Thm. 3.3). Note that when one squares a quantity the relative error is doubled. □

If $\lambda_n^{-1}(S_H H S_H)$ is large then the error in the eigenvalues is almost entirely due to the errors during the Cholesky decomposition. This fact was observed empirically in [4, §7.4], but no explanation was given. If the matrix H has special structure so that one can compute its Cholesky factor to higher accuracy than indicated by Lemma 2.6, then there will be a corresponding increase in the accuracy in the computed singular values. See §6 where we use the fact that the Hilbert matrix has a closed form Cholesky factor that can be evaluated very accurately to show that the eigenvalues can also be computed very accurately.

Since most of the error is caused by the Cholesky factorization, while most of the work is done in the generalized Jacobi updates, it may make sense to compute the Cholesky factorization in higher precision. This was suggested in [17, p. 632], but no theoretical justification was given.

Since $\sigma_n(R_{\hat{L}}\hat{L}) = \sqrt{\lambda_n(S_H H S_H)}$, one can easily estimate $\lambda_n(S_H H S_H)$, using a condition estimator based on a few steps of the power method or the Lanczos algorithm since $\hat{L}$ is triangular. Thus the bound (3.7) can be evaluated cheaply in practice—unlike the bounds where one applies the transformations and the scaling on the same side.

The algorithm outlined in Theorem 3.5 is essentially Algorithm 4.4 in [4]. Our error bound does not involve a factor of the form

$$\max_{0 \le i \le M} \lambda_n^{-1}(S_{H_i} H_i S_{H_i}) \tag{3.8}$$

as all the bounds in [4] do. Thus we have shown that this algorithm computes the eigenvalues of a positive definite matrix to optimal relative accuracy without the proviso in [4] that the quantity in (3.8) not be much larger than $\lambda_n^{-1}(S_H H S_H)$.

Combining all these ideas we have the following algorithm that computes the eigenvalues of a positive definite matrix H to a relative accuracy of the order of

$$\epsilon \lambda_n^{-1}(S_H H S_H). \tag{3.9}$$

This algorithm has several nice properties. First, it computes the eigenvalues to as high a relative accuracy as we can hope. Second, it estimates the relative error (3.9) in the computed eigenvalues at little additional cost. Finally, it may be a little cheaper than Algorithm 4.4 in [4] because of the choice of stopping criterion. (Algorithm 4.4 in [4] is very similar to Algorithm 3.6 here.)

ALGORITHM 3.6. Given a positive definite matrix H:

1. compute $LL^T = H$ (Cholesky with complete pivoting);
2. find a lower bound $\sigma > 0$ on $\sigma_n(S_H L)$;
3. set tol $= \sigma^{-2}\epsilon$ for the stopping criterion;
4. compute the singular values of L by right-hand Jacobi with stopping criterion (2.11);
5. square singular values to give eigenvalues of H.

The pivoting in step 1 is not necessary for the accuracy of the algorithm. It is included to produce faster convergence of the Jacobi process in step 4.

The idea of first taking a Cholesky factorization $H = LL^T$ can be useful when trying to compute the eigenvalues of a graded positive definite matrix H to a high

relative accuracy using bidiagonalization methods. Suppose that we use bidiagonalization followed by the QR iteration to compute the singular values of $\hat{L}$, the computed Cholesky factor of H. The resulting singular values $\hat{\sigma}_i$ satisfy

$$\frac{|\hat{\sigma}_i - \sigma_i(\hat{L})|}{\sigma_i(\hat{L})} \leq c_n \kappa(\hat{L})\epsilon \approx c_n \sqrt{\kappa(H)}\epsilon,$$

where c_n grows modestly with n. From the backward error bound for the Cholesky decomposition in Lemma 2.6 and the forward error bound in Theorem 2.4 we have

$$\frac{|\lambda_i(H) - \sigma_i^2(\hat{L})|}{\lambda_i(H)} \leq (n^2 + 5n)\lambda_n^{-1}(S_H H S_H)\epsilon.$$

The sum[2] of these two bounds can be less than the standard bound in (1.3), i.e., $\kappa(H)\epsilon$, by a factor of as much as $\sqrt{\kappa(H)}$. Note also, that if it should happen that $\kappa(\hat{L}) \approx \sqrt{\kappa(H)} \leq \lambda_n^{-1}(S_H H S_H)$, then this Cholesky/bidiagonalization algorithm will compute the eigenvalues as accurately as Jacobi's method, but at considerably less cost. The statement $\sqrt{\kappa(H)} \leq \lambda_n^{-1}(S_H H S_H)$ may be interpreted as saying that at most half of the ill conditioning of H is due to the grading of H.

4. The harder case. In this section we consider the case where the orthogonal transformations are applied on the same side as the scaling matrix. This makes the analysis harder and the resulting bound is harder to compute because it involves a factor $\max_{0\leq i\leq M} \sigma_n^{-1}(R_{G_i}G_i)$—and to compute or bound this, one must estimate the condition number of M matrices not just one. However, as we explain in the next paragraph, this is the algorithm of choice if one is prepared to make the assumption that

$$\max_{0\leq i\leq M} \sigma_n^{-1}(R_{G_i}G_i) \approx \sigma_n^{-1}(R_{G_0}G_0). \tag{4.1}$$

There is considerable experimental evidence for this assumption; see [4, §7.4] and [15, Chap. 5, Table 4]. Mascarenhas has given a family of examples that shows that the left-hand side of (4.1) can be $n/2$ times larger than the right-hand side [12]. This is the worst known growth.

Let $G \in M_n$. Then we may compute the singular values of G by left-handed Jacobi or right-handed Jacobi. Which should we choose? Suppose that $\sigma_n^{-1}(R_G G) << \sigma_n^{-1}(GC_G)$. Then the rows of G are much closer to orthogonality than the columns. Multiplying G by an orthogonal on the left leaves the angles between the columns of G unchanged but changes the angles between the rows. Since our goal is to apply a sequence of orthogonal transformations until either the rows or columns of the transformed matrix are orthogonal, one would expect that it would be more efficient to transform the matrix so that the rows are orthogonal rather than the columns. That is, it is more efficient to use left-handed Jacobi.[3] In practice if $\sigma_n^{-1}(R_G G) << \sigma_n^{-1}(GC_G)$, then one observes that left-handed Jacobi does indeed converge more quickly than right-handed Jacobi, just as one would expect from this argument (we give examples in the next two paragraphs). However, the error bounds in the previous section are in terms of $\sigma_n^{-1}(GC_G)$. Naturally we would like bounds in terms of $\sigma_n^{-1}(R_G G)$, which is considerably smaller. Such bounds, and the related problem of bounds for two-sided

[2] Actually we must take twice the bound on the relative error in $\hat{\sigma}_i$ as we are squaring it.

[3] This is a heuristic argument and one can construct counterexamples.

Jacobi, where we necessarily apply the orthogonal transformation and the scaling matrix on the same side, are the subjects of this section. The next two paragraphs present instances where left-handed Jacobi converges more rapidly than right-handed Jacobi. The reader who is not interested in these details may omit them and go directly to Lemma 4.1.

We now give a couple of specific instances where $G = DB$ with B well conditioned and D diagonal where right-handed Jacobi is much slower than left-handed Jacobi. The first example is where B is a random 10×10 orthogonal matrix and $D = \mathrm{diag}(1, 10^{-1/8}, 10^{-2/8}, \ldots, 10^{-9/8})$. Using MATLAB ($\epsilon_M \approx 2 \times 10^{-16}$) and $\mathrm{tol} = 10^{-12}$ (which gives a relatively lax stopping criterion) typically between five and seven sweeps of right-handed Jacobi are required for convergence. Of course, since the rows of G are orthogonal, the termination criterion for left-handed Jacobi is satisfied even before the first sweep.

In the previous example D was not particularly ill conditioned ($\kappa(D) \approx 13$). In our next example, which is due to an anonymous referee, we take D very ill conditioned and *rounding errors* will cause right-handed Jacobi to be much slower than left-handed Jacobi. In exact arithmetic both left- and right-handed Jacobi would converge in one iteration. Let B be a 2×2 reasonably conditioned matrix with all elements about the same size (say $O(1)$) and $|b_{11}| > |b_{12}|$. Let $D = \mathrm{diag}(d_1, d_2)$ where the d_i are positive and $d_1 >> d_2$ (it is necessary that $d_2/d_1 > \epsilon^j$, and the larger j is, the slower right-handed Jacobi is; thus this example requires rather extreme ill conditioning). Since G is 2×2 each sweep consists of just one Jacobi rotation. Let $G_1, G_2, \ldots$ be the sequence of matrices generated by right-handed Jacobi applied in finite precision arithmetic with precision ϵ. One can check that $(G_1)_{11}$ will be approximately $d_1\sqrt{b_{11}^2 + b_{12}^2}$ (since $d_1 >> d_2$ and $|b_{11}| > |b_{12}|$) and that *due to rounding errors* $(G_1)_{12}$ could be as large as $O(\epsilon d_1 |G_{12}|) = O(\epsilon d_1)$. So now, using the fact that $d_1 >> d_2$ we have

$$|(G_1)_{\cdot 1}^T (G_1)_{\cdot 2}| \approx |(G_1)_{11}(G_1)_{12}| \approx \|(G_1)_{\cdot 1}\| \, \|(G_1)_{\cdot 2}\|.$$

Thus, even though in exact arithmetic the columns of G_1 would be orthogonal, the columns of the computed G_1 are far from orthogonal, in fact they are almost parallel! One can check that in successive iterations we will have $(G_k)_{11} \approx d_1\sqrt{b_{11}^2 + b_{22}^2}$ (unchanged) and $(G_k)_{12}$ could be as large as $O(\epsilon^k d_1)$, $k = 2, 3, \ldots$. Because the elements in the second row of G_k are at most $O(d_2)$ one can show that the columns of G_k will not satisfy the stopping criterion (2.11) unless

$$|(G_k)_{12}| \leq \mathrm{tol} \cdot d_2. \tag{4.2}$$

We can only be sure of this for

$$k = (\log(d_2/d_1) + \log(\mathrm{tol}))/\log(\epsilon) \equiv k_{\max}.$$

By making d_1/d_2 large one can make $k_{\max}$ arbitrarily large. To summarize, if G is a 2×2 matrix as above then it is *possible* that, as a result of rounding errors, $k_{\max}$ sweeps of right-handed Jacobi will be required before the termination criterion (2.11) is satisfied. If one chooses B randomly then typically after a few tries one obtains a matrix G for which $k_{\max}$ sweeps are indeed required for convergence. However, for such a matrix only one sweep of left-handed Jacobi is necessary because the effect of rounding errors will not be so serious.

We now give a rather general lemma which we then specialize to the case of multiplication by an elementary orthogonal matrix.

LEMMA 4.1. *Let $G \in M_{m,n}$ and let $J \in M_n$ be nonsingular. Then in finite precision arithmetic with precision ϵ*

$$fl(GJ) = (G + \Delta G)J, \tag{4.3}$$

where

$$\|\Delta G C_G\| \leq 2\sqrt{n}(n-1)\epsilon \|C_G\, |J|\, |J^{-1}|\, C_G^{-1}\|. \tag{4.4}$$

Proof. From (4.3) and (2.5) it follows that

$$\begin{aligned} |\Delta G| &= |[fl(GJ) - GJ]J^{-1}| \\ &\leq |fl(GJ) - GJ|\, |J^{-1}| \\ &\leq 2(n-1)\epsilon |G|\, |J|\, |J^{-1}|. \end{aligned}$$

So

$$\begin{aligned} \|\Delta G C_G\| &\leq \| \, |\Delta G C_G| \, \| \\ &\leq 2(n-1)\epsilon \| \, |G|\, |J|\, |J^{-1}|\, |C_G| \, \| \\ &= 2(n-1)\epsilon \|(|G|C_G)(C_G^{-1}|J|\, |J^{-1}|\, C_G)\| \\ &\leq 2(n-1)\epsilon \| \, |G|C_G \, \|_F \, \| \, C_G^{-1}|J|\, |J^{-1}|\, C_G \, \| \\ &\leq 2(n-1)\epsilon\sqrt{n} \, \| \, C_G^{-1}|J|\, |J^{-1}|\, C_G \, \|. \end{aligned}$$

We have used that fact that the columns of GC_G have unit length for the final inequality. □

The following bound on the scaled backward error due to multiplication by an elementary orthogonal matrix is a generalization of [4, Theorem 3.3.3]. There it was assumed that the elementary orthogonal matrix was chosen to orthogonalize the two columns that it affects. The added generality does not make the proof of this result any more complicated. It is useful in that it shows that the high relative accuracy of Jacobi's method does not depend on the rotation angle being computed very exactly; it is sufficient that α not be too large.

THEOREM 4.2. *Let $G \in M_{m,n}$, and let $B = GC_G$. Let G_1 denote the matrix obtained by applying the rotation $R_{ij}(t)$ in finite precision arithmetic with precision ϵ by multiplying columns i and j of G by the matrix*

$$J = \begin{pmatrix} 1 & t \\ -t & 1 \end{pmatrix}$$

and then dividing them by $\sqrt{1+t^2}$. Let

$$\alpha = |sc| \max\left\{ \frac{\|G_{\cdot i}\|}{\|G_{\cdot j}\|}, \frac{\|G_{\cdot j}\|}{\|G_{\cdot i}\|} \right\}. \tag{4.5}$$

Then

$$G_1 = (G + \Delta G)R_{ij}(t) + \Delta G_1 \tag{4.6}$$

where the arithmetic on the right-hand side is performed exactly and

$$\|\Delta G C_G\| \leq 2\sqrt{2}(1+2\alpha)\epsilon + O(\epsilon^2), \qquad \|\Delta G_1 C_{G_1}\| \leq 3\sqrt{2}\epsilon + O(\epsilon^2). \tag{4.7}$$

Proof. It is sufficient to consider the case where G has only two columns. It is easily seen that

$$|J|\,|J^{-1}| = \begin{pmatrix} 1 & 2|sc| \\ 2|sc| & 1 \end{pmatrix},$$

where s and c are the sine and cosine of the rotation angle. So

$$C_G\,|J|\,|J^{-1}|\,C_G^{-1} = \begin{pmatrix} 1 & 2|sc|\|G_{\cdot i}\|/\|G_{\cdot j}\| \\ 2|sc|\|G_{\cdot j}\|/\|G_{\cdot i}\| & 1 \end{pmatrix},$$

which has norm at most $(1+2\alpha)$. Thus, by Lemma 4.1, $fl(GJ) = (G+E)J$, where $\|EC_G\| \le 2\sqrt{2}(1+2\alpha)$. Now, dividing by $\sqrt{1+t^2}$, which is itself computed in finite precision arithmetic, causes a relative error of at most 3ϵ (Lemma 2.3 and (1.7)) in each component of G_1, from which the second bound in (4.7) follows. □

Note that if, as in regular one-sided Jacobi, one choses t to be such that $R_{ij}(t)$ orthogonalizes the columns i and j, i.e., $t = t(\|G_{\cdot i}\|^2, \|G_{\cdot j}\|^2, G_{\cdot i}^* G_{\cdot j})$, then by Lemma 2.2 (and the ensuing discussion), we have

$$\alpha = |sc| \max\left\{\frac{\|G_{\cdot i}\|}{\|G_{\cdot j}\|}, \frac{\|G_{\cdot j}\|}{\|G_{\cdot i}\|}\right\} \le |t| \max\left\{\frac{\|G_{\cdot i}\|}{\|G_{\cdot j}\|}, \frac{\|G_{\cdot j}\|}{\|G_{\cdot i}\|}\right\} \le 1.$$

In view of Theorem 2.5 this says that

$$\frac{|\sigma_i(G) - \sigma_i(G_1)|}{\sigma_i(G)} \le (6\sqrt{2}\sigma_n^{-1}(GC_G) + 3\sqrt{2}\sigma_n^{-1}(G_1 C_{G_1}))\epsilon + O(\epsilon^2). \tag{4.8}$$

This is slightly stronger than the corresponding results in [4], [15]. Actually one can further strengthen (4.8) by a more careful argument, but our purpose here is just to show that our results imply results similar to those in [4], [15].

Because we have done computations at the matrix level rather than the scalar level and have proved some preliminary lemmas, this proof is considerably simpler than those in [15, Theorem 3.3.3] and [4, Theorem 4.1], even though the result is more general. Note also that we have avoided the necessity of dividing the proof into two cases as was done in these other proofs.

There is another minor advantage of this result over those in [4], [15]. When implementing one-sided Jacobi one needs $\|G_{\cdot i}\|$, $\|G_{\cdot j}\|$, and $G_{\cdot i}^T G_{\cdot j}$ to compute the rotation angle. In the interests of computational efficiency one usually does not compute $\|G_{\cdot i}\|$ and $\|G_{\cdot j}\|$ explicitly, at a cost of $O(n)$ flops, but rather updates them at each step at a cost of merely $O(1)$. This causes a gradual loss of accuracy in the computed rotation angle which translates into a larger error bound. (The scaled backward error bound is still $O(\epsilon)$ but the constant is larger.) The error in the rotation angle does not affect our analysis, as we do not require that the rotation orthogonalize two columns, merely that it have a small corresponding value of α.

The idea of combining forward and backward error bounds has been used to good effect in [7, Corollary 1]. There it was apparently essential in obtaining a simple proof of a strong error bound. Here it is not essential; we could have easily converted the forward error (resulting from the division by $\sqrt{1+t^2}$) into a backward error at the cost of a slightly larger error bound.

Now let us consider the scaled backward error for the two-sided transformation. First, we need an analog of Lemma 4.1 in the two-sided case.

LEMMA 4.3. *Let $H \in M_n$ be positive definite. Let $J \in M_n$ be nonsingular. Suppose that we compute J^THJ in finite precision arithmetic with precision ϵ by*

$$(J^THJ)_{ij} = (J_{\cdot i})^T(HJ_{\cdot j}) \tag{4.9}$$

for $i \le j$ and by symmetry for $i > j$. Then $fl(J^THJ)$ is symmetric and

$$fl(J^THJ) = J^T(H + \Delta H)J, \tag{4.10}$$

where

$$\|S_H \Delta H S_H\| \le 4n(n-1)\epsilon \|S_H^{-1}\,|J|\,|J^{-1}|\,S_H\|^2, \tag{4.11}$$

assuming that $n > 1$.

Proof. The proof is similar to Lemma 4.1. □

The next result is a generalization of Theorem 3.3 in [4].

THEOREM 4.4. *Let $H \in M_n$ be positive definite. Given a pair of indices i, j and a scalar t. Let $\hat{H}$ be the value of $R_{ij}(t)^T H R_{ij}(t)$ computed by first forming J^THJ, where $J = \sqrt{1+t^2}R_{ij}(t)$, and then dividing the $(i,i),(i,j),(j,i)$, and (j,j) entries by $1+t^2$ and the remaining entries in the ith and jth rows and columns by $\sqrt{1+t^2}$. Assume the computations are done in finite precision arithmetic with precision ϵ. Let*

$$\alpha = |sc| \max\left\{ \sqrt{\frac{H_{ii}}{H_{jj}}}, \sqrt{\frac{H_{jj}}{H_{ii}}} \right\}, \tag{4.12}$$

where s and c are the sine and cosine of the rotation angle. Then

$$\hat{H} = R_{ij}(t)^T(H + \delta H)R_{ij}(t) + \Delta\hat{H}, \tag{4.13}$$

where

$$\|S_H(\Delta H)S_H\|_2 \le f(n,\alpha)\epsilon, \quad \text{and} \quad \|S_{\hat{H}}\Delta\hat{H}S_{\hat{H}}\| \le 3(\sqrt{2n-4}+2)\epsilon \tag{4.14}$$

and $f(n,\alpha) = 8(1+2\alpha)^2 + 2\sqrt{2n-4}(1+2\alpha)$.

Proof. Without loss of generality we may assume that $(i,j) = (1,2)$. Let $A = S_H H S_H$. Partition H, A, and S_H as

$$H = \begin{pmatrix} H_1 & G^T \\ G & H_2 \end{pmatrix}, \quad H_1 = \begin{pmatrix} a & c \\ c & b \end{pmatrix}, \quad A = \begin{pmatrix} A_{11} & A_{21}^* \\ A_{21} & A_{22} \end{pmatrix}, \quad S_H = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}.$$

Partition $\hat{H}$ in the same way as H. Because the results depend only the first and second rows and columns of H it follows that

$$\Delta H = \begin{pmatrix} \delta H_1 & \delta G^* \\ \delta G & 0 \end{pmatrix},$$

and so

$$S_H \delta H S_H = \begin{pmatrix} 0 & D_1\delta G^* D_2 \\ D_2\delta G D_1 & 0 \end{pmatrix} + \begin{pmatrix} D_1\delta H_{11} D_1 & 0 \\ 0 & 0 \end{pmatrix}.$$

We will first bound $\|D_2\delta G D_1\|_2$. As in the proof of Theorem 4.2

$$|D_2(\delta G)D_1| \le 2\epsilon(D_2|G|D_1)(D_1^{-1}|J||J^{-1}|D_1) \le 2\epsilon|A_{21}|(D_1^{-1}|J||J^{-1}|D_1).$$

Because A_{21} is the off-diagonal block of A, a positive definite matrix with ones on the diagonal, it follows that the $2n-4$ entries of $|A_{21}|$ are at most one, and hence that

$$\|A_{21}\| \leq \| \, |A_{21}| \, \|_F \leq \sqrt{2n-4}.$$

Thus, taking norms and using the definition of α we have

$$\|D_2 \delta G D_1\| \leq 2\epsilon\sqrt{2n-4}(1+2\alpha).$$

So

$$\left\| \begin{pmatrix} 0 & D_1 \delta G^* D_2 \\ D_2 \delta G D_1 & 0 \end{pmatrix} \right\| = \|D_2 \delta G D_1\| \leq 2\sqrt{2n-4}(1+2\alpha)\epsilon. \tag{4.15}$$

Now let us consider δH_1. From Lemma 4.3, using the same ideas as in the proof of Theorem 4.4, we have

$$\|D_1 \delta H_1 D_1\| \leq 4(2-1)2\epsilon \|D_1|J||J^{-1}|D_1^{-1}\|^2 \leq 8\epsilon(1+2\alpha)^2.$$

Adding these two bounds gives the first in equality in (4.14).

By Lemma 2.3 and (1.7), computing $\sqrt{1+t^2}$ or $(1+t^2)$ and dividing by it causes a relative error of at most 3ϵ. One can check that a matrix with entries bounded by 1 and nonzero entries in the first two rows and columns only has norm at most $2+\sqrt{2n-4}$. The second inequality in (4.14) follows from this. □

The proof could have been considerably simplified if we had not taken advantage of the fact that only two rows and two columns of H are changed by an elementary orthogonal transformation, since in this case we would not have to partition H or any of the other matrices in the proof. The cost of this simplification would have been to increase $f(n,\alpha)$ to

$$f(n,\alpha) = 8n(1+2\alpha)^2,$$

which would increase the error bound by a factor of about $\sqrt{8n}(1+2\alpha)$ for large n and moderate α.

If $\hat{H}$ is the computed result of applying one Jacobi rotation to H, then, by Lemma 2.2, $\alpha \leq 1$. So from Theorem 4.4 and the perturbation bound in Theorem 2.4 we have

$$\frac{|\lambda_i(\hat{H}) - \lambda_i(H)|}{\lambda_i(H)} \leq \frac{75+9\sqrt{2n-4}}{\lambda_{\min}(S_H H S_H)}\,\epsilon.$$

This bound is slightly stronger than those in [4, Theorem 3.1] and [15, Theorem 3.2.1, trigonometric case].

Note that the Jacobi step we use in this theorem is *not* the standard Jacobi step. Here, even if we are choosing J to annihilate the i,j element we explicitly apply the matrix J to the 2×2 submatrix in rows and columns i and j and, as a result, the i,j element may not be exactly zero. The standard algorithm would compute the ii and jj elements by formulae (and not by matrix multiplication) and set the i,j and j,i entries to 0.[4] So, strictly speaking, our analysis does not apply to the standard two-sided Jacobi algorithm. The analysis in [4] gives a backward error analysis of the formulae used to compute the ii and jj elements. This is more complicated than the

[4] Both algorithms are identical except in the way that they compute the ii, ij, ji, and jj elements. Both algorithms require essentially the same amount of computation and are about as accurate.

backward error analysis of matrix multiplication that we used. This is part of the reason why the proofs in [4] are more complicated than those here.

The idea of not setting the ij and ji elements to zero but explicitly computing them was suggested in [11]. There the reason was to improve accuracy; here we do it to simplify the analysis.

Note that in the one-sided case (Theorem 4.2) the highest power of α in the bound is α, while here it is α^2. This will be relevant when general orthogonal transformations are used and we do not have the bound $\alpha < 1$.

We will not state the generalizations of all the results in [4], [15] but merely give one specific example of how they extend to the generalized Jacobi methods.

THEOREM 4.5. *Let $H = H_0$ be a positive definite matrix and let $H_m, m = 1, \ldots, M$ be the sequence of matrices generated by applying elementary orthogonal similarities. Let α_m be the value of α, as defined in* (4.12), *for the mth transformation. Then*

$$\left|\frac{\lambda_i(H) - \lambda_i(H_M)}{\lambda_i(H)}\right| \leq \epsilon \cdot M \cdot \max_{0 \leq m \leq M} \frac{f(n, \alpha_m) + 3(\sqrt{2n-4}+2)}{\lambda_n(S_{H_m} H_m S_{H_m})} + O(\epsilon^2),$$

where $f(n, \alpha)$ is defined in Theorem 4.4.

The proof is identical to those in [4], [15]. Note that it is not necessary that H_M, the final matrix, be diagonal or almost diagonal. In §5 we apply this result with H_M tridiagonal. If H_M is almost diagonal in the sense that it satisfies the termination criterion (2.10), and if we let $\hat{\lambda}_i, i = 1, \ldots, n$, be the ordered diagonal entries of H_M then

$$\left|\frac{\lambda_i(H) - \hat{\lambda}_i}{\lambda_i(H)}\right| \leq \epsilon \cdot M \cdot \max_{0 \leq m \leq M} \frac{f(n, \alpha_m) + 3(\sqrt{2n-4}+2)}{\lambda_n(S_{H_m} H_M S_{H_m})} + n \cdot \text{tol} + O(\epsilon^2).$$

We have said very little about the accuracy of the computed eigenvectors. That is because the proofs in [4] when applied to this situation show that generalized two-sided Jacobi computes the eigenvectors to high normwise accuracy [4, Theorem 3.3], and even the components of the eigenvectors to high relative accuracy[5] [4, Theorem 3.4].

Note that the proofs in [4] of results on eigenvectors depend only on a bound on the scaled backward error and not on the fact that the Jacobi transformation annihilates the ij element. The results in §4 of [4] on one-sided Jacobi also generalize in the same way to the case where we apply the transformations on one side only and have a bound on the corresponding α's.

5. Stability of transformations of positive definite matrices. Lemma 4.3 and Theorem 4.4 shed light on the errors caused by tridiagonalizing a positive definite matrix. This section is devoted to the accuracy of various tridiagonalization methods. We could have equally well discussed the bidiagonalization of a general matrix.

The following example was presented in [4, p. 1238]:

$$H = \begin{pmatrix} 10^{40} & 10^{19} & 10^{19} \\ 10^{19} & 10^{20} & 10^{9} \\ 10^{19} & 10^{9} & 1 \end{pmatrix}. \tag{5.1}$$

[5] The componentwise bound on the error in the eigenvectors in [4] contains a factor that is possibly exponential in M, the number of Jacobi rotations required for convergence. See [14] for a proof that the exponential growth factor in [4] can be replaced by a linear growth factor.

This is a graded matrix, and so one might hope that tridiagonalization followed by QR will compute all its eigenvalues to high relative accuracy. However, one can check that no matter how one permutes the rows and columns of H, the eigenvalues of H as computed by tridiagonalization followed by QR will contain at least one negative eigenvalue (at least when one uses a particular version of MATLAB; see [4, p. 1238]). It was not made clear in [4] whether this inaccuracy is due to the error incurred in tridiagonalizing the matrix or the inaccuracy of the QR algorithm when applied to the resulting tridiagonal matrices. In [3] Demmel showed that implicit QR is inherently inaccurate for some symmetric tridiagonal matrices. However, in this instance the reason that tridiagonalization followed by QR is inaccurate is that merely tridiagonalizing H (or any permutation of H) can cause a large relative perturbation in the eigenvalues. This can be seen from Theorem 4.4. For example, suppose that we do a rotation in rows and columns 2 and 3 to zero the 3,1 and 1,3 elements. Then for this rotation $|t| = 1$ and so

$$\alpha = \frac{1}{2} \cdot \max\left\{\sqrt{\frac{1}{10^{20}}}, \sqrt{\frac{10^{20}}{1}}\right\} = 5 \times 10^9.$$

The backward error bound in Theorem 4.4 contains the term $8\epsilon\alpha^2 \approx 4 \times 10^4$. (One might say that the reason for is that h_{21} is too small in relation to h_{31} *and* the grading of the matrix, i.e., $(S_H H S_H)_{21}$ is too small in relation to $(S_H H S_H)_{31}$.) Thus this bound cannot guarantee us any relative accuracy. One can repeat this for every permutation of the rows and columns of H and check that the right-hand side of the bound (4.14) is greater than one in every case.[6]

As noted in [4] this example shows that tridiagonalization followed by the QR iteration does not necessarily compute the eigenvalues of a graded matrix to high relative accuracy. There are algorithms that will compute the eigenvalues of a positive definite tridiagonal matrix T to a relative accuracy of $\epsilon\kappa(S_T T S_T)$—for example, bisection or the qd algorithm in [7]. If one has a graded positive definite matrix H and reduces it to a tridiagonal T by Givens rotations (possibly fast rotations), one can monitor the possible loss of accuracy by computing the value of α for each of the transformations applied. If the maximum of these α's is not large, then the eigenvalues of T are close to those of H (in the relative sense) and so applying bisection (or some other method that guarantees high relative accuracy of the computed eigenvalues) to T will give eigenvalues of H with close to optimal relative accuracy. Here, as always, we are assuming that $\min \lambda_n(S_{H_i} H_i S_{H_i}) \approx \lambda_n(H)$, where the H_i are the intermediate matrices in the computation.

A more accurate, though more expensive, way to compute the eigenvalues of a graded matrix without using Jacobi's method is to compute the Cholesky factorization $H = LL^T$ and then bidiagonalize L. The reason this is better is that the relative error is now contains the factor $\kappa(L) = \sqrt{\kappa(H)}$ rather than $\kappa(H)$. See the discussion after Algorithm 3.6 for further details. Furthermore, one will only get the factor α in the error bounds rather than α^2.

Now let us consider the accuracy of tridiagonalization by Householder transformations. We will apply Lemma 4.3 with $J = I - ww^T$ where w is a vector such that

[6] Of course, one could do a rotation in rows and columns 1 and 3 to eliminate the 1, 3 and 3, 1 elements, and since this would just be a Jacobi transformation it causes only a small relative change in the eigenvalues. However, this idea is only available in the 3×3 case, and is not the standard method of tridiagonalization. One can construct a 4×4 positive definite matrix such that no permutation of it can be stably tridiagonalized by three Givens rotations.

$w^T w = 2$. For simplicity of analysis we assume that we form J explicitly and then apply it as outlined in Lemma 4.3. Of course, in practice we would exploit the special structure of J to evaluate $J^T H J$ in $O(n^2)$ flops rather that $O(n^3)$ flops; see, e.g., [9, §8.2.1] for details. An analysis of this more efficient tridiagonalization is more complicated but yields essentially the same bound (actually the bound is slightly stronger in that it has one less factor of n).

It is easy to see that

$$|J| = |J^{-1}| \leq I + |w||w|^T$$

and so

$$\|S_H^{-1}|J||J^{-1}|S_H\| \leq \|S_H^{-1}(I + 4|w||w|^T)S_H\| \leq 1 + 4\|S_H^{-1}w\|\|S_H w\|.$$

Thus if we apply the Householder transformation J to H in finite precision arithmetic with precision ϵ then

$$fl(J^T HT) - J^T H J = J^T(H + \Delta H)J,$$

where

$$\|S_H \Delta H S_H\| \leq 4n(n-1)(1 + 4\|S_H w\|\|S_H^{-1}w\|)^2\epsilon. \tag{5.2}$$

This bound can be evaluated easily and cheaply since all we need do is form $S_H w$ and $S_H^{-1}w$ and compute their norms.

In view of the 3×3 example presented earlier in the section one might expect that if the entries $h_{i,j}, i = j+2, \ldots, n$ are sufficiently small with respect to $h_{i+1,i}$ and the scaling on the matrix, then eliminating $h_{i,j}, i = j+2, \ldots, n$ by Givens rotations in rows and columns $i, j, i = j+2, \ldots, n$ will preserve the high relative accuracy of the eigenvalues. This is indeed the case, and it can be seen by checking that in this case α will be small. It is natural to conjecture that in this situation tridiagonalizing by Householder transformations, in finite precision arithmetic, will not cause a large relative error in the eigenvalues. This is indeed the case and we now make this precise.

Suppose that H is a positive definite matrix with first column that is "nicely scaled." That is

$$H = \begin{pmatrix} d_1^2 & \alpha d_1 d_2 & d_1(D_3 r)^T \\ \alpha d_1 d_2 & d_2^2 & * \\ d_1 D_3 r & * & H_3 \end{pmatrix}, \qquad S_H = \begin{pmatrix} d_1^{-1} & & \\ & d_2^{-1} & \\ & & D_3^{-1} \end{pmatrix}, \tag{5.3}$$

where $H_3 \in M_{n-2}, r \in R^{n-2}$ and $\|r\| < |\alpha| < 1$, and furthermore, the main diagonal of H is decreasingly ordered. One can show that for Householder transformation that "tridiagonalizes" the first column of H

$$\|S_H w\| \, \|S_H^{-1} w\| \leq 8n^2,$$

and consequently that the scaled backward error caused by applying it in finite precision arithmetic is $O(\epsilon)$. The proof is a rather tedious computation based on the fact that $w = \sqrt{2}z/\|z\|$, where

$$z = \begin{pmatrix} 0 \\ \alpha d_1 d_2 (1 + \sqrt{1 + \|D_3 r\|^2/(\alpha d_2)^2}) \\ d_1 D_3 r \end{pmatrix}.$$

Of course, in order that the entire tridiagonalization procedure not cause a large relative error in the eigenvalues of H it is necessary that at the ith stage the i column of the current H is "nicely scaled." It is not clear how one can guarantee this a priori. For example, if we take

$$H = \begin{pmatrix} 10 & 1 & 10^{-10} & 10^{-10} \\ 1 & 1 & 0 & 0 \\ 10^{-10} & 0 & 10^{-10} & 0 \\ 10^{-10} & 0 & 0 & 10^{-20} \end{pmatrix},$$

then tridiagonalizing the first column of H we have the matrix

$$H^{(1)} = \begin{pmatrix} 10 & 1 & 0 & 0 \\ 1 & 1 & 10^{-10} & 10^{-10} \\ 0 & 10^{-10} & 10^{-10} & 10^{-20} \\ 0 & 10^{-10} & 10^{-20} & 2 \times 10^{-20} \end{pmatrix},$$

where the figures are correct to nine decimal places. The second column of this matrix is no longer "nicely scaled" in the sense described above. So one would expect that after we apply one more Householder transformation to zero the $4,2$ element then the resulting matrix $H^{(2)}$ will have eigenvalues that will not be close to those of H in the relative sense if we do the computations in finite precision arithmetic. This is indeed the case using MATLAB ($\epsilon \approx 2 \times 10^{-16}$). We computed $\hat{H}^{(1)}$ and $\hat{H}^{(2)}$, the computed values of $H^{(1)}$ and $H^{(2)}$. The relative difference between the eigenvalues of H and $\hat{H}^{(1)}$ was bounded 2×10^{-15} while relative difference between the smallest eigenvalue of H and $\hat{H}^{(2)}$ was about 5×10^{-8}. One can check that $\lambda_4(S_H H S_H) > .5$ and so the eigenvalues of H can be computed to high relative accuracy by Jacobi's method.

Now consider $J = U$ the matrix of eigenvectors. This situation may arise when we want to compute the eigendecomposition of $H(t)$ for $t = k\delta$, $\delta = 0, 1, 2, \ldots$ where $H(t)$ and δ are such that $U(t)$, the matrix of eigenvectors of $H(t)$, is approximately $U(t+\delta)$. In this case we can compute the eigendecomposition of $H(t+\delta)$ from that of $U(t)^T H(t+\delta) U(t)$, which is nearly diagonal. Again we need a bound on

$$\| S_H^{-1} |J| |J^{-1}| S_H \|.$$

From [14] we have

$$|U_{ij}| \leq \kappa^{1/2}(S_H H S_H) \min \left\{ \sqrt{\frac{H_{ii}}{H_{jj}}}, \sqrt{\frac{H_{jj}}{H_{ii}}} \right\}.$$

The same inequality with $\kappa^{1/2}$ replaced by $\kappa^{3/2}$ is given in [1, Prop. 6] or [4, Prop. 2.8]. Since the diagonal entries of S_H are $H_{ii}^{-1/2}$ it follows that

$$|U| \leq \kappa^{1/2}(S_H H S_H) |S_H E S_H^{-1}| \quad \text{and} \quad |U|^T \leq \kappa^{1/2}(S_H H S_H) |S_H E S_H^{-1}|,$$

where E is the $n \times n$ matrix of ones. Substituting these bounds on U and U^T we have

$$\| S_H^{-1} |J| |J^{-1}| S_H \| \leq n^2 \kappa(S_H H S_H),$$

which would imply that the scaled backward error in computing $U^T H U$ is about $\kappa^2(S_H H S_H)\epsilon$ and therefore that the relative perturbation in the eigenvalues of the

computed value of U^THU is bounded by approximately $\kappa^2(S_HHS_H)\epsilon$—possibly too large to be a useful bound.

One can do better by considering forward error directly. Let $\Lambda = U^THU$, the diagonal matrix of eigenvalues of H. From [14] we have

$$|U| \leq \lambda_n^{-1/2}(S_HHS_H)S_HE\Lambda^{1/2}.$$

Using this bound for the second inequality and Lemma 2.1 for the first inequality we have

$$\begin{aligned}|fl(U^THU) - U^THU| &\leq 4n^2|U|^T|H||U|\epsilon \\ &\leq 4n^2\lambda_n^{-1}(S_HHS_H)\Lambda^{1/2}E^3\Lambda^{1/2}\epsilon.\end{aligned}$$

So now, since $S_\Lambda\Lambda S_\Lambda = I$ has all eigenvalues equal to one, Theorem 2.4 gives

$$\frac{|\lambda_i(fl(U^THU)) - \lambda_i(H)|}{\lambda_i(H)} = \frac{|\lambda_i(fl(U^THU)) - \lambda_i(U^THU)|}{\lambda_i(U^THU)} \leq 4n^4\lambda_n^{-1}(S_HHS_H)\epsilon.$$

The term n^4 may be a worry but the conventional wisdom is that such factors of n are rarely a problem in practice. Thus, using the matrix of eigenvectors as a preconditioner does not cause a larger relative error in the eigenvalues than does merely introducing a relative perturbation of size $O(\epsilon)$ in each of the entries of H. Notice that we have obtained a good bound using a forward error approach in this instance. However, if we try to use a forward error approach to bound the relative perturbation in the eigenvalues caused by Jacobi's method, we would get a factor $\lambda_n^{-2}(S_{H_1}H_1S_{H_1})$ (H_1 is the matrix obtained after one Jacobi rotation) rather that $\lambda_n^{-1}(S_HHS_H)$. The reason that this is not a problem here is that $H_1 = U^THU$ is approximately diagonal and so when scaled has smallest eigenvalue one, which when squared is again approximately one. (Whereas, after one Jacobi rotation the resulting matrix is not nearly diagonal.)

In the case that U is merely an approximation to the matrix of eigenvectors of H one can expect a similar result and, once one has computed $H_1 = fl(U^THU)$, one can determine, a posteriori, a bound on the relative error in the eigenvalues of H_1. The bound is

$$\frac{|\lambda_i(H_1) - \lambda_i(H)|}{\lambda_i(H)} \leq 4n^2\|S_{H_1}^{-1}US_{H_1}\|^2\lambda_n^{-1}(S_{H_1}H_1S_{H_1})\epsilon. \tag{5.4}$$

Because U is an approximation to the matrix of eigenvalues of H we may expect that $\|S_{H_1}^{-1}US_{H_1}\|^2$ is not much larger than n^2 and that $\lambda_n(S_{H_1}H_1S_{H_1})$ is not much smaller than one. The idea of using approximate eigenvectors was suggested in [16, footnote 8], but no error analysis was given there.

6. The eigenvalues of the Hilbert matrix. The Hilbert matrix of order n, denoted H_n, is the $n\times n$ matrix with i,j entry $(i+j-1)^{-1}$. It is a well-known example of an ill-conditioned positive definite matrix. Since H_n is ill conditioned, Householder tridiagonalization followed by the QR iteration cannot compute the smallest eigenvalues of H_n to much relative accuracy. The main diagonal entries of H_n differ by a factor of at most n, so it follows that $\lambda_n^{-1}(S_{H_n}H_nS_{H_n}) \geq n^{-1}\lambda_n^{-1}(H_n)$ and thus the bound on the relative error in the eigenvalues computed by the two-sided Jacobi method or Cholesky followed by one-sided Jacobi will not be much better than that for tridiagonal QR. However, there is a closed form for the Cholesky factor of H_n, and

the entries of this closed form can be evaluated to high relative accuracy. In particular if we take L_n to be the lower triangular matrix with i, j entry

$$\frac{\sqrt{2j-1}((i-1)!)^2}{(i+j-1)!(i-j)!} \tag{6.1}$$

for $i \geq j$, then one can check that $L_n L_n^T = H_n$. This fact, along with a wealth of other results on the Hilbert matrix, is in [2]. This Cholesky factor is not what one would get by doing Cholesky with complete pivoting, but nonetheless it is relatively well conditioned after column scaling, so left-handed Jacobi will compute the singular values of L_n to high relative accuracy. Right-handed Jacobi will also get them to high relative accuracy provided that $\max_{0 \leq i \leq M} \sigma_n^{-1}(G_i R_{G_i})$ is not much larger than $\sigma_n^{-1}(GR_G)$, where $G = G_0 = L_n$.

Let us illustrate this with a numerical example in the case $n = 12$. Numerical calculation shows that $\kappa(H_{12}) \approx 2 \times 10^{16}$ and $\sigma_{12}(L_{12}C_{L_{12}}) \approx 2 \times 10^3$. We used MATHEMATICA to compute the eigenvalues of H_{12} to a relative accuracy of 10^{-16}. We then used MATLAB ($\epsilon \approx 2 \times 10^{-16}$) to compute the eigenvalues of H_{12} by the tridiagonalization followed by QR. The rounding errors in merely forming H_{12} will cause a relative perturbation of order ϵ in the entries of H_{12} and this can cause a relative perturbation of order $\epsilon\kappa(H_{12}) \approx 4$ in the smallest eigenvalue computed by tridiagonalization followed by QR. In fact the largest relative error in a computed eigenvalue of H_{12} using tridiagonalization and QR was 2×10^{-1}. Applying left-handed Jacobi to the closed form for the Cholesky factor and then squaring the computed singular values (again on MATLAB) can be expected to give each of the eigenvalues to a relative accuracy of about $\epsilon\lambda_{12}^{-1}(L_{12}C_{L_{12}}) \approx 10^{-13}$. In fact, when we did the calculation we got each of the eigenvalues to within 7×10^{-15}, and four sweeps were required for convergence. This extra accuracy is due both to the accuracy of one-sided Jacobi in this situation *and* the fact that we can evaluate the entries of L_{12} to high relative accuracy.

A positive definite Cauchy matrix is a matrix of the form $[(\alpha_i + \alpha_j)^{-1}]_{i,j=1}^n$, where the α_i are distinct and positive. There is a similar closed form for the Cholesky factor of such matrices [8] and so one can compute their eigenvalues to a high relative accuracy by one-sided Jacobi applied to the Cholesky factor.

7. Hybrid Jacobi methods. The major drawback of Jacobi's method is that it is several times more expensive than tridiagonalization followed by QR.[7] According to [9, §8.5.8] two Jacobi sweeps without accumulating the transformations cost about the same as computing all the eigenvectors and eigenvalues by tridiagonal QR. See Table 1 for a more detailed comparison.[8] Table 2 in [4] gives the number of sweeps of two-sided Jacobi and right-handed Jacobi (applied to the Cholesky factor, computed with complete pivoting, of the positive definite matrix) required for convergence. Right-handed Jacobi always converged more quickly, and for 50×50 matrices it required

[7] In this paper we use QR as the standard of comparison. However, divide-and-conquer has recently been shown to be stable and faster than QR for computing the eigenvalues and eigenvectors of a symmetric tridiagonal matrix [10]. So if we compare Jacobi against divide-and-conquer it will appear to be even slower.

[8] The flop counts for tridiagonalization and QR are taken from [9, p. 424]. The remaining figures are easily verified. The reason that one sided Jacobi requires more flops per sweep is that one must evaluate the inner product of two columns to find the rotation angle to orthogonalize the columns. One also needs to know the column lengths for this, but these can be computed once at the beginning and then updated cheaply.

TABLE 1
Approximate flop counts.

	Eigenvalues only	Eigenvalues and Eigenvectors
Tridiagonalization and QR (entire process)	$4n^3/3$	$9n^3$
One-sided Jacobi (1 sweep)	$4n^3$	$7n^3$
Two-sided Jacobi (1 sweep)	$3n^3$	$6n^3$
One-sided Jacobi fast rotations (1 sweep)	$3n^3$	$5n^3$
Two-sided Jacobi fast rotations (1 sweep)	$2n^3$	$4n^3$

between three and six sweeps depending on $\kappa(S_H H S_H)$ and $\kappa(S_H)$. Two-sided Jacobi required between about six and seventeen sweeps.[9] The disadvantage of right-handed Jacobi applied to the Cholesky factor is that it does not compute the eigenvectors to as high accuracy as two-sided Jacobi with accumulation of the transformations. One can compute the eigenvectors to high relative accuracy by applying left-handed Jacobi to the Cholesky factor and accumulating the transformations. However, this is essentially the same as two-sided Jacobi and requires more or less the same number of sweeps for convergence. Thus, on a serial machine there is a considerable difference between the time required by Jacobi's method and methods based on tridiagonalization. If one wants the eigenvectors also to high accuracy and so does two sided Jacobi accumulating the transformations then the difference is greater. This difference grows as the size of the problems grows. Thus it would be useful to have hybrid algorithms that are faster than Jacobi, but still have the same high relative accuracy. This is the subject of this section.

It is not unreasonable to consider Jacobi's method implemented on a serial machine since Jacobi's method is the fastest way we know to compute the eigenvalues of a dense positive definite matrix to full maximum relative accuracy. The only other known ways are bisection and inverse iteration applied to the *full* matrix [4, §5] and so cost $O(n^4)$ to compute all the eigenvalues and eigenvalues to maximal accuracy.

Typically, several Jacobi sweeps are required before Jacobi's method starts to converge quadratically. We describe two preconditioning strategies that may be expected to hasten the onset of quadratic convergence but that do not destroy the high relative accuracy of Jacobi's method. There are many other preconditioning ideas that can be produced by combining the backward error bounds in this paper with the perturbation bounds in [4], [14]. After any of these preconditionings Jacobi's method should

[9] The number of sweeps of two sided Jacobi required to diagonalize a matrix can be greatly reduced by sorting the main diagonal entries *once* before starting Jacobi's method. This has been observed often. The cost of this sorting is essentially the same as the cost of the sorting associated with Cholesky with complete pivoting. Demmel and Veselić did not sort the main diagonal prior to running two sided Jacobi. Had they done so, two sided Jacobi typically would have required at one or two sweeps more than one sided Jacobi on the Cholesky factor computed with complete pivoting for the matrices generated by the method outlined in [4, §7].

converge in one sweep, or, at most, two sweeps. These preconditioning strategies are based on computing a singular value (or eigenvalue) decomposition of the matrix by bidiagonalization (or tridiagonalization) and then using QR or some other algorithm that is specially suited to bidiagonal (or tridiagonal) matrices.

One would like preconditioners that can be applied efficiently in parallel. Unfortunately that is not the case with the preconditioners that we present. Thus the algorithms that we give in this section are useful only in the situation where one is using a serial computer and wants to compute the eigenvalues to high relative accuracy. Nonetheless they do show that one can compute the eigendecomposition of a positive definite matrix to full accuracy faster than by Jacobi's method.

Our first preconditioner is for the situation where we want to compute the singular values of $G \in M_n$ to maximal relative accuracy and $\sigma_n^{-1}(GC_G) \leq \sigma_n^{-1}(R_G G)$.

ALGORITHM 7.1. Given a matrix G:

1. quickly compute U such that $G = U\Sigma V^T$;
   (by bidiagonalization and QR, for example)
2. set $G_0 = U^T G$;
3. apply left-handed Jacobi to G_0 to compute its singular values $\hat{\sigma}_i, i = 1, \ldots, n$.

The matrix U is the product of many orthogonal matrices U_i. There is no need to explicitly form the product U in step 1 and then apply it to G in the next step. We could apply the U_i to another copy of G directly and thereby save a little computation. In Algorithm 7.1 we separated the two steps for clarity.

The matrix U in step 1 need not be computed very accurately since it is only being used as a preconditioner—it is however essential that it be close to orthogonal. In [13] we discuss how one can compute approximate eigenvectors (that are orthogonal) more quickly than by bidiagonalization.

When applied in serial this algorithm is, in some situations, faster than Jacobi applied in serial to compute the singular values of G because we need only bidiagonalize G and then find its singular values accumulating only the transformations applied on the left and finally do one sweep or, at most, two sweeps of Jacobi. When applied in serial to compute only the singular values this algorithm does not compare favorably in terms of efficiency with QR or other bidiagonalization approaches because one must apply the U_i in addition to doing one or more Jacobi sweeps. So one would only use it if one required the singular values to high relative accuracy on a serial computer.

Algorithm 7.1 computes singular values of G to a relative accuracy of $c\sigma_n^{-1}(GC_G)\epsilon$ (for a modest constant c). This can be seen by combining Theorems 3.1 and 3.3. In Theorem 3.1 we assume that Q is orthogonal, but the U computed in step 1 may have singular values differing from 1 by $c_n\epsilon$, where c_n is a modest function of n. This is not a problem as it causes only a $c_n\epsilon$ relative perturbation in the singular values of UG.

Algorithm 7.1 can be modified in various ways to compute the eigenvalues of a positive definite matrix. For example, we have the following algorithm.

ALGORITHM 7.2.

1. (a) Compute $H = GG^T$ (Cholesky).
   (b) Quickly compute U such that $H = U\Lambda U^T$.
   (by tridiagonalization and QR, for example).
2. Continue with Algorithm 7.1 starting at step 2.

One can check that this algorithm requires about 15–$19n^3$ flops plus the cost of a Cholesky factorization if we exploit the triangularity of G when applying U^T to it.

The higher figure occurs when two sweeps are required and is less likely than $15n^3$. Thus when Cholesky followed by one-sided Jacobi (Algorithm 3.6) requires more than 4–5 iterations our algorithm will be faster. From Table 2 in [4] it can be seen that Algorithm 3.6 is faster for $n \leq 16$ while for $n = 50$ they are about the same, depending on the choice of $\kappa(S_H)$ and $\kappa(S_H H S_H)$. Thus one may expect that Algorithm 7.2 will be better for larger n.

Algorithm 7.1 does not compute the left singular vectors to componentwise high accuracy. We know that we can compute the left singular vectors to high componentwise accuracy by using left-handed Jacobi and accumulating the transformations, but we would like to do it more quickly than by regular left-handed Jacobi. The idea is that if one applies Jacobi accumulating transformations, then the resulting orthogonal matrix gives the eigenvectors/singular vectors to a high componentwise relative accuracy. (See [4, Theorem 3.4] for a proof.) One can show that the algorithm below also has this property by using the idea at the end of §6.

ALGORITHM 7.3.

1. Quickly compute U such that $G = U\Sigma V^T$
   (by bidiagonalization and QR, for example).
2. Set $G_1 = U^T G$.
3. Apply left-handed Jacobi to G_1 accumulating the transformations in U.

One could use right-handed Jacobi (not accumulating transformations) to compute U in step 1. The resulting algorithm would still be faster than left-handed Jacobi with accumulation of transformations applied to G, and would be simple to implement in parallel.

It has been observed that generally the number of sweeps required for the convergence of right-handed Jacobi (for example) decreases with $\sigma_n^{-1}(GC_G)$, since this is a measure of how far the columns of G are from orthogonality. (Of course, one can construct examples where $\sigma_n^{-1}(GC_G)$ is arbitrarily large but Jacobi's method converges in one sweep.) The algorithms we presented in this section were designed to accelerate convergence in a rather crude way—by merely computing a singular value decomposition quickly and possibly inaccurately and then applying it to orthogonalize the columns of the matrix. This is rather inefficient since it is possible to greatly reduce $\sigma_n^{-1}(GC_G)$ by just considering a few columns of G or the singular vectors corresponding to a few of the smallest singular vectors of GC_G. This is the subject of further research [13].

8. Conclusions. We have shown that Jacobi's method is guaranteed to compute the eigenvalues of a positive definite matrix to the maximum possible relative accuracy. This was shown by using the fact that for the *singular value* problem if we apply the orthogonal matrices on one side and the scaling matrix on the other side then the scaled condition number remains constant. This result also allowed us to show that a rectangular matrix can be reduced to a square (upper triangular) matrix without causing large relative errors in the singular values.

We have extended the error analysis of [4], [15] to show that high relative accuracy of the eigenvalues is maintained provided that we use a transformation algorithm that only uses orthogonal transformations for which α (as defined in (4.5) or (4.12) as appropriate) is bounded by a modest constant. For Jacobi transformations α is never larger than 1 so the results in [4], [15] are a special case of the results in §4. Our generalization allows one to derive algorithms that when implemented in serial are faster than Jacobi but more accurate than tridiagonalization-based methods.

Our error analysis is at the matrix level so it is much simpler than that in [4], [15]. Also, it is easily generalized to give an improved understanding of the relative perturbations caused by tridiagonalizing a positive definite matrix by Givens or Householder transformations, especially for graded matrices. This technique may well be useful in studying the scaled backward errors in transforming indefinite matrices to tridiagonal form and nonsymmetric matrices to upper Hessenberg form.

Finally, we explicitly showed that if one computes the eigenvalues of a positive definite matrix by computing the singular values of its Cholesky factor then most of the error is due to the Cholesky factorization, and presented several instances where this idea is useful.

Acknowledgments. M.-D. Choi alerted me to the closed form for the Cholesky factor of the Hilbert matrix and Jim Demmel pointed out a serious error in my analysis of Householder transformations.

REFERENCES

[1] J. BARLOW AND J. DEMMEL, *Computing accurate eigensystems of scaled diagonally dominant matrices*, SIAM J. Numer. Anal., 27 (1990), pp. 762–791.

[2] M.-D. CHOI, *Tricks or treats with the Hilbert matrix*, Amer. Math. Monthly, 90 (1983), pp. 310–312.

[3] J. DEMMEL, *On the inherent inaccuracy of implicit tridiagonal QR*, Tech. Report 983, Institute for Mathemetics and its Applications, University of Minnesota, Minneapolis, April 92.

[4] J. DEMMEL AND K. VESELIĆ, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.

[5] ———, *Jacobi's method is more accurate than QR*, Tech. Report 468, Department of Computer Science, Courant Institute, New York, October 1989. (LAPACK Working Note # 15).

[6] Z. DRMAČ, *The Generalized Singular Value Problem*, Ph.D. thesis, FernUniversitat, Hagen, Germany, 1994.

[7] K. V. FERNANDO AND B. PARLETT, *Accurate singular values and differential qd algorithms*, Tech. Report PAM-544, Center for Pure and Applied Mathematics, University of Calfornia, Berkeley, 1992. Numer. Math., to appear.

[8] I. GOHBERG AND I. KOLTRACHT, *Error analysis for triangular factorization of Cauchy and Vandermonde matrices*, preprint, 1993.

[9] G. GOLUB AND C. VAN LOAN, *Matrix computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, 1989.

[10] MING GU AND STANLEY C. EISENSTAT, *A divide-and-conquer algorithm for the symmetric tridiagonal eigenproblem*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 172–191.

[11] V. HARI, *On Cyclic Jacobi Methods for the Generalized Eigenvalue Problem*, Ph.D. thesis, Fernuniversitat, Hagen, Germany, 1984.

[12] W. F. MASCARENHAS, *A note on Jacobi being more accurate then QR*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 215–218.

[13] R. MATHIAS, *Fast accurate eigenvalue computations using the Cholesky factorization*, 1993, manuscript.

[14] ———, *Spectral perturbation bounds for graded positive definite matrices*, 1993, manuscript.

[15] I. SLAPNIČAR, *Accurate Symmetric Eigenreduction by a Jacobi Method*, Ph.D. thesis, Fernuniversität, Hagen, Germany, 1992.

[16] K. VESELIĆ, *A Jacobi eigenreduction algorithm for definite matrix pairs*, Numer. Math., 64 (1993), pp. 241–269.

[17] K. VESELIĆ AND V. HARI, *A note on a one sided Jacobi algorithm*, Numer. Math., 56 (1989), pp. 627–633.

[18] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 1004–1012, July 1995

019

PRINCIPAL SUBMATRICES, GEOMETRIC MULTIPLICITIES, AND STRUCTURED EIGENVECTORS *

CHARLES R. JOHNSON† AND BRENDA K. KROSCHEL†

Abstract. It is a straightforward matrix calculation that if λ is an eigenvalue of A, x an associated eigenvector and α the set of positions in which x has nonzero entries, then λ is also an eigenvalue of the submatrix of A that lies in the rows and columns indexed by α. A converse is presented that is the most general possible in terms of the data we use. Several corollaries are obtained by applying the main result to normal and Hermitian matrices. These corollaries lead to results concerning the case of equality in the interlacing inequalities for Hermitian matrices, and to the problem of the relationship among eigenvalue multiplicities in various principal submatrices.

Key words. interlacing inequalities, geometric multiplicity, principal submatrix, structured eigenvector

AMS subject classifications. 15A18, 15A57

For $\varnothing \neq \alpha \subseteq N \equiv \{1, 2, \ldots, n\}$ and $A \in M_n(F)$, denote the principal submatrix of A lying in the rows and columns indexed by α as $A[\alpha]$ and the complementary principal submatrix, resulting from the deletion of the rows and columns α, as $A(\alpha)$. It is a straightforward partitioned matrix calculation that if λ is an eigenvalue of A, x an associated eigenvector, and α the set of positions in which x has entries not equal to zero, then λ is also an eigenvalue of $A[\alpha]$. Converses to this statement are known in certain special situations. For example, several people have recently noted that if $A \in M_n(C)$ is Hermitian, $|\alpha| = n - 1$, and $\lambda \in R$ is an eigenvalue of both A and $A[\alpha]$, i.e., a case of equality in the interlacing inequalities, then there is an eigenvector $x = (x_1, x_2, \ldots, x_n)^T$ of A associated with λ such that if $i \notin \alpha$ then $x_i = 0$. For a general matrix $A \in M_n(F)$ and λ an eigenvalue of A with geometric multiplicity k, the rank of $A - \lambda I$ is $n - k$. Then for $|\alpha| > n - k$ the rank of $A[\alpha] - \lambda I$ is at most $n - k$ and λ is also an eigenvalue of $A[\alpha]$. Moreover, it is implicit in the proof of Theorem 1.4.9 in [HJ] that there is an eigenvector of A associated with λ all of whose components indexed by α^c are zero. It is our purpose here to give a converse to the opening statement that is the most general possible in terms of the data we use. A variety of statements, including those just mentioned, may then be easily recognized as special cases.

The general converse, as well as some special cases, will be valid over a general field F. For $x \in F^n$ and $\alpha \subseteq N$, let $x[\alpha]$ be the subvector of x containing the components of x indexed by α, and let $x(\alpha)$ be the complementary subvector. For $A \in M_n(F)$, let $\sigma(A)$ denote the set of all eigenvalues of A, some of which may lie only in an extension field of F, and for $\lambda \in \sigma(A)$, denote the geometric multiplicity of λ in A by $g_\lambda(A)$.

The most optimistic converse to the opening statement would be that if λ is an eigenvalue of both A and $A[\alpha]$, then there is an eigenvector x (of A associated with λ) in which all components of $x(\alpha)$ are zero. However, this is not always the case.

* Received by the editors April 21, 1994; accepted for publication (in revised form) by T. Ando, July 22, 1994.

†Department of Mathematics, College of William and Mary, Williamsburg, Virginia 23187-8795. (kroschel@cs.wm.edu) The work of Dr. Johnson was supported, in part, by National Science Foundation grant DMS-92-00899 and Office of Naval Research contract N00014-90-J-1739.

Consider

$$A = \left[\begin{array}{cc|cc} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{array}\right]$$

and the set $\alpha = \{1, 2\}$. This matrix has zero as an eigenvalue, as does $A[\alpha]$, but any eigenvector of A associated with zero is of the form $[a \quad 0 \quad 0 \quad -a]^T$. The converse cannot, therefore, be as general as one might hope.

Before stating a converse that is as general as it can be, several definitions are needed. The main result will be stated in terms of the dimensions of special subspaces, of the left and right eigenspaces of a general matrix A associated with λ, in which the vectors have support among the components indexed by α. These special subspaces (of the eigenspaces) are defined as follows:

$$LE_\alpha^\lambda(A) = \{y \in F^n | y^T A = \lambda y^T, y(\alpha) = 0\},$$
$$RE_\alpha^\lambda(A) = \{x \in F^n | Ax = \lambda x, x(\alpha) = 0\}.$$

Similarly, let $LN(A)$ and $RN(A)$ denote the left and right nullspaces of A and define the special subspaces (of the nullspaces) $LN_\alpha(A) = LE_\alpha^0(A)$ and $RN_\alpha(A) = RE_\alpha^0(A)$. It is clear that the dimensions of all these spaces are permutation similarity invariant, and this fact will be exploited repeatedly without further mention. If x is an eigenvector of A associated with λ, then x is an eigenvector of $A - \lambda I$ associated with the eigenvalue zero. For this reason, results concerning the special nullspaces underlie observations concerning the special eigenspaces.

For contrast to the main result, we note some preliminary facts that indicate circumstances under which both the left and right special subspaces are nonempty. It is first observed that for general matrices, when the rank deficiency (the rank deficiency of a matrix A is $n -$ rank $(A) = g_0(A)$) of a principal submatrix is sufficiently large, then the dimensions of the left and right nullspaces are positive. Suppose that the submatrix $A[\alpha]$ is such that its rank deficiency is greater than the number of rows or columns deleted from A to obtain $A[\alpha]$. That is, for $|\alpha| = n - k, g_0(A[\alpha]) > k$. In this case, the rank of $A[\alpha]$ is $n - k - g_0(A[\alpha])$ and the rank of A can be at most $2k$ more than the rank of $A[\alpha]$. But then the rank deficiency of A is at least $g_0(A[\alpha]) - k$. Since this number is positive, A is rank deficient and the left and right nullspaces of A are both nonempty. The lemma below states that, in fact, the left and right *special* nullspaces of A are both nonempty.

LEMMA 0. *Let $A \in M_n(F)$ and let $\alpha \subseteq N$ be such that $|\alpha| = n - k$.*

(i) *If $g_0(A[\alpha]) > k$, then $\dim(LN_\alpha(A)), \dim(RN_\alpha(A)) \geq g_0(A[\alpha]) - k$.*

(ii) *Let $0 \leq g_0 \leq \min\{k, |\alpha|\}$ be given. Then there is a matrix B such that $g_0(B[\alpha]) = g_0$ and $\dim(LN_\alpha(B)) = \dim(RN_\alpha(B)) = 0$.*

Proof. We assume, without loss of generality, that $\alpha = \{1, 2, \ldots, n - k\}$. Then A has the partitioned form

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

in which $A_{11} = A[\alpha]$. In this case, if x is in $RN_\alpha(A)$ it is of the form $x = \begin{bmatrix} \hat{x} \\ 0 \end{bmatrix}$ in which $\hat{x} \in F^{n-k}$. Similarly, any vector $y^T \in LN_\alpha(A)$ is of the form $y^T = [\hat{y}^T \quad 0]$ in which $\hat{y} \in F^{n-k}$.

Transformation of A by an appropriate equivalence will not affect

$$g_0(A), g_0(A[\alpha]) = g_0(A_{11}),$$

or the form of the nullvectors of A; so, choose $S, T \in M_{n-k}(F)$ nonsingular matrices such that

$$(1) \qquad \begin{bmatrix} S & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} T & 0 \\ 0 & I \end{bmatrix} = \left[\begin{array}{cc|c} 0 & 0 & Y_1 \\ 0 & I & Y_2 \\ \hline X_1 & X_2 & A_{22} \end{array}\right] = \hat{A},$$

in which the upper left zero block of $\hat{A}$ is $g_0(A_{11})$-by-$g_0(A_{11})$, $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = SA_{12}$, and $[X_1 \quad X_2] = A_{21}T$. Because of the identity block in $\hat{A}$, a vector x in $RN_\alpha(\hat{A})$ must be of the form

$$x = \begin{bmatrix} x_1 \\ 0 \\ 0 \end{bmatrix}, x_1 \in F^{g_0(A_{11})}.$$

In addition, x_1 must be in the right nullspace of the submatrix X_1. Conversely, for every vector in the right nullspace of X_1, there is a vector of the form indicated above in $RN_\alpha(\hat{A})$ and $\dim(RN_\alpha(\hat{A})) = \dim(RN(X_1))$. Moreover, any vector in $RN_\alpha(\hat{A})$ corresponds to a vector in $RN_\alpha(A)$ of the form

$$\begin{bmatrix} T & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x_1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} T \begin{bmatrix} x_1 \\ 0 \end{bmatrix} \\ 0 \end{bmatrix} = \begin{bmatrix} \hat{x} \\ 0 \end{bmatrix},$$

in which $\hat{x} = T \begin{bmatrix} x_1 \\ 0 \end{bmatrix} \in F^{n-k}$. Therefore,

$$\dim(RN_\alpha(A)) = \dim(RN_\alpha(\hat{A})) = \dim(RN(X_1)).$$

By similar arguments for the left nullspace

$$\dim(LN_\alpha(A)) = \dim(LN_\alpha(\hat{A})) = \dim(LN(Y_1)).$$

A second equivalence will zero out X_2 and Y_2:

$$(2) \qquad \left[\begin{array}{cc|c} I & 0 & 0 \\ 0 & I & 0 \\ \hline 0 & -X_2 & I \end{array}\right] \left[\begin{array}{cc|c} 0 & 0 & Y_1 \\ 0 & I & Y_2 \\ \hline X_1 & X_2 & A_{22} \end{array}\right] \left[\begin{array}{cc|c} I & 0 & 0 \\ 0 & I & -Y_2 \\ \hline 0 & 0 & I \end{array}\right]$$

$$= \left[\begin{array}{cc|c} 0 & 0 & Y_1 \\ 0 & I & 0 \\ \hline X_1 & 0 & \tilde{A}_{22} \end{array}\right] = \tilde{A}.$$

Note that this equivalence does not change the form of the nullvectors discussed above and the dimensional equalities still hold.

Now, suppose that $g_0(A_{11}) > k$, as assumed in part (i) of the lemma. Since Y_1 and X_1 are $g_0(A_{11})$-by-k and k-by-$g_0(A_{11})$, respectively,

$$\dim(LN(Y_1)),\ \dim(RN(X_1)) \geq g_0(A_{11}) - k.$$

But, $\dim(LN_\alpha(A)) = \dim(LN(Y_1))$ and $\dim(RN_\alpha(A)) = \dim(RN(X_1))$, so that part (i) of the lemma is verified.

For part (ii) consider the matrix

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \left[\begin{array}{cc|cc} 0 & 0 & I_{g_0} & 0 \\ 0 & I_{n-k-g_0} & 0 & 0 \\ \hline I_{g_0} & 0 & * & 0 \\ 0 & 0 & 0 & \hat{B}_{22} \end{array}\right], \tag{3}$$

in which B_{11} is $(n-k)$-by-$(n-k)$ and $g_0(B_{11}) = g_0$. For this matrix, $0 \leq g_0 \leq k$, but there are no nonzero vectors in either $LN_\alpha(B)$ or $RN_\alpha(B)$, and part (ii) of the lemma is also proved. □

Replacement of A with $A - \lambda I$ in Lemma 0 gives the following.

THEOREM 0. *Let $A \in M_n(F)$ and let $\alpha \subseteq N$ be such that $|\alpha| = n - k$.*

(i) *If $g_\lambda(A[\alpha]) > k$, then $\dim(LE_\alpha^\lambda(A)), \dim(RE_\alpha^\lambda(A)) \geq g_\lambda(A[\alpha]) - k$.*

(ii) *Let $0 \leq g_\alpha \leq \min\{k, |\alpha|\}$ be given. Then there is a matrix B such that $g_\lambda(B[\alpha]) = g_\alpha$ and $\dim(LE_\alpha^\lambda(B)) = \dim(RE_\alpha^\lambda(B)) = 0$.*

Statement (i) in Theorem 0 is best possible when left and right eigenspaces are considered separately. By considering the left and right eigenspaces simultaneously, one arrives at a general converse to the opening statement. This main result will first be stated in terms of the special nullspaces.

LEMMA 1. *Let $A \in M_n(F)$; then for $\alpha \subseteq N$ with $|\alpha| = n - k$,*

(i) $\dim(LN_\alpha(A)) + \dim(RN_\alpha(A)) \geq g_0(A) + g_0(A[\alpha]) - k$.

(ii) *Let g and g_α such that $0 \leq g \leq n, 0 \leq g_\alpha \leq |\alpha|$, and $|g - g_\alpha| \leq k$ be given. Then, if $g + g_\alpha - k > 0$ there is a matrix B such that $g_0(B) = g, g_0(B[\alpha]) = g_\alpha$ and*

$$\dim(LN_\alpha(B)) + \dim(RN_\alpha(B)) = g_0(B) + g_0(B[\alpha]) - k.$$

If $g + g_\alpha - k \leq 0$, then there is a matrix B, with the given parameters, such that

$$\dim(LN_\alpha(B)) = \dim(RN_\alpha(B)) = 0.$$

Proof. Begin the proof of Lemma 1 by performing the equivalences in (1) and (2) as in the proof of Lemma 0. The matrices Y_1 and X_1 are of order $g_0(A_{11})$-by-k and k-by-$g_0(A_{11})$, respectively. By basic linear algebra $\dim(LN(Y_1)) = g_0(A_{11}) - \operatorname{rank}(Y_1)$ and $\dim(RN(X_1)) = g_0(A_{11}) - \operatorname{rank}(X_1)$. Addition of these two equations results in

$$\dim(LN(Y_1)) + \dim(RN(X_1)) = 2g_0(A_{11}) - \operatorname{rank}(Y_1) - \operatorname{rank}(X_1). \tag{4}$$

The equivalence transformations performed on A in the proof of Lemma 0 do not change the rank of A and, since

$$\operatorname{rank} \begin{bmatrix} 0 & Y_1 \\ X_1 & \tilde{A}_{22} \end{bmatrix} \geq \operatorname{rank}(Y_1) + \operatorname{rank}(X_1),$$

we have

$$\begin{aligned} \operatorname{rank}(A) = \operatorname{rank}(\tilde{A}) &= \operatorname{rank}(A_{11}) + \operatorname{rank}\begin{bmatrix} 0 & Y_1 \\ X_1 & \tilde{A}_{22} \end{bmatrix} \\ &\geq \operatorname{rank}(A_{11}) + \operatorname{rank}(Y_1) + \operatorname{rank}(X_1). \end{aligned} \tag{5}$$

Combining (5) and (4) results in

$$\begin{aligned} \dim(RN(X_1)) + \dim(LN(Y_1)) &\geq 2g_0(A_{11}) - \operatorname{rank}(A) + \operatorname{rank}(A_{11}) \\ &= g_0(A) + g_0(A_{11}) - k. \end{aligned} \tag{6}$$

From the discussion in the proof of Lemma 0 $\dim(LN_\alpha(A)) = \dim(LN(Y_1))$ and $\dim(RN_\alpha(A)) = \dim(RN(X_1))$ so that

$$\begin{aligned} \dim(LN_\alpha(A)) + \dim(RN_\alpha(A)) &= \dim(LN(Y_1)) + \dim(RN(X_1)) \\ &\geq g_0(A) + g_0(A_{11}) - k, \end{aligned}$$

and part (i) of Lemma 1 is proved.

There are two cases to consider in proving part (ii) of Lemma 1. To begin, consider the case in which $g + g_\alpha - k \leq 0$. Note that for this to be the case, g_α must be less than or equal to k. For the matrix B in (3), if $g_\alpha = g_0$, then $g_0(B_{11}) = g_\alpha$ and the submatrix $\hat{B}_{22}$ is $(k - g_\alpha)$-by-$(k - g_\alpha)$. This submatrix can be chosen so that B has rank deficiency, g, from 0 to $k - g_\alpha$. Thus, B has the appropriate parameters, and, as mentioned in the proof of Lemma 0, B has $\dim(LN_\alpha(B)) = \dim(RN_\alpha(B)) = 0$.

For the case in which $g + g_\alpha - k > 0$, consider

$$B = \left[\begin{array}{ccc} 0 & 0 & Y_1 \\ 0 & I_{n-k-g_\alpha} & 0 \\ \hline X_1 & 0 & 0 \end{array}\right].$$

The submatrices Y_1 and X_1 can independently be chosen to have rank from zero to $\min(g_\alpha, k)$, inclusive, which gives B a rank deficiency, g, from $|g_\alpha - k|$ to $g_\alpha + k$, inclusive. Now, note that in (5) if $\tilde{A}_{22} = 0$, then

$$\operatorname{rank}(A) = \operatorname{rank}(A_{11}) + \operatorname{rank}(Y_1) + \operatorname{rank}(X_1)$$

and there is equality in (6). Because B is of this form, the equality holds and $\dim(LN_\alpha(B)) + \dim(RN_\alpha(B)) = g_0(B) + g_0(B_{11}) - k$, which proves the lemma. □

Our main result, the proof of which follows from Lemma 1 by translation, is then:

THEOREM 1. *Let $A \in M_n(F)$; then for $a \subseteq N$ with $|\alpha| = n - k$*

(i) $\dim(LE_\alpha^\lambda(A)) + \dim(RE_\alpha^\lambda(A)) \geq g_\lambda(A) + g_\lambda(A[\alpha]) - k$.

(ii) *Let g and g_α such that $0 \leq g \leq n, 0 \leq g_\alpha \leq |\alpha|$, and $|g - g_\alpha| \leq k$ be given. Then, if $g + g_\alpha - k > 0$ there is a matrix B such that $g_\lambda(B) = g, g_\lambda(B[\alpha]) = g_\alpha$ and*

$$\dim(LE_\alpha^\lambda(B)) + \dim(RE_\alpha^\lambda(B)) = g_\lambda(B) + g_\lambda(B[\alpha]) - k.$$

If $g + g_\alpha - k \leq 0$, then there exists a matrix B, with the given parameters, such that

$$\dim(LE_\alpha^\lambda(B)) = \dim(RE_\alpha^\lambda(B)) = 0.$$

In each of Lemmas 0 and 1 and Theorems 0 and 1, statement (ii) indicates that statement (i) is best possible. The restrictions regarding α only avoid logical impossibilities and, otherwise, all situations not covered by statement (i) are covered in statement (ii).

At this point we make two general observations that are direct consequences of Theorem 1.

(i) If $A \in M_n(F)$ and $|\alpha| = n-1$, then $\lambda \in \sigma(A) \cap \sigma(A[\alpha])$ if and only if there is either a left or a right eigenvector of A (associated with λ) whose α^c component is zero.

(ii) If $A \in M_n(F), \lambda \in \sigma(A)$ and $\alpha \subseteq N$ with $|\alpha| = n - k$ are such that $\dim(LE_\alpha^\lambda(A)) = \dim(RE_\alpha^\lambda(A))$, then each of

$$\dim(LE_\alpha^\lambda(A)), \dim(RE_\alpha^\lambda(A)) \geq \frac{g_\lambda(A) + g_\lambda(A[\alpha]) - k}{2}.$$

In this event, if $g_\lambda(A) + g_\lambda(A[\alpha]) > k$, then both $\dim(LE_\alpha^\lambda(A))$ and $\dim(RE_\alpha^\lambda(A))$ are positive.

Note that statement (i) does *not* follow from Theorem 0 and that statement (i) cannot be improved, as it may be that there is not both a left special eigenvector and a right special eigenvector. For example,

$$A = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$$

does not have the property assumed in (ii) for $0 \in \sigma(A)$, and $g_0(A) = 1 = g_0(A[\{1,2\}])$. Thus, as every right null vector of A is a multiple of $(1,1,1)^T$, A has no special right eigenvector associated with 0, while it, of course, has a left such eigenvector, e.g., $(1,1,0)$, because of statement (i). Similarly, for many values of $g_\lambda(A)$ and $g_\lambda(A[\alpha])$, the conclusion of (ii) does not follow from Theorem 0, and, for further values, the estimates that follow from Theorem 0 are weaker. For example, the statement about Hermitian matrices in the opening paragraph does not follow from Theorem 0.

We may now give several specific corollaries to Theorem 1. First, note that if $A \in M_n(C)$ is normal, then, as $UAU^* = D$, with U unitary and D diagonal, any left eigenspace of A is the conjugate transpose of a right eigenspace. Thus, the hypothesis of (ii) above is satisfied for each λ and α. From this observation we can conclude the following.

COROLLARY 1. *Let $A \in M_n(C)$ be a normal matrix. For $\alpha \subseteq N$ with $|\alpha| = n - k$*

$$\dim(LE_\alpha^\lambda(A)), \dim(RE_\alpha^\lambda(A)) \geq \frac{g_\lambda(A) + g_\lambda(A[\alpha]) - k}{2}.$$

Of course Hermitian matrices are normal so the following is a special case of Corollary 1.

COROLLARY 2. *Let $A \in M_n(C)$ be Hermitian. For $\alpha \subseteq N$ with $|\alpha| = n - k$*

$$\dim(LE_\alpha^\lambda(A)), \dim(RE_\alpha^\lambda(A)) \geq \frac{g_\lambda(A) + g_\lambda(A[\alpha]) - k}{2}.$$

In the opening paragraph we mentioned that if A is Hermitian, $\lambda \in \sigma(A) \cap \sigma(A[\alpha])$, and $|\alpha| = n - 1$, then there is an eigenvector x (of A associated with λ) in which

$x(\alpha) = 0$. But then $g_\lambda(A), g_\lambda(A[\alpha]) \geq 1$ which results in a positive right-hand side in Corollary 2. In this case, both the left and the right special eigenspaces are nonempty, which proves the following corollary.

COROLLARY 3. *Let $A \in M_n(C)$ be Hermitian, let $\alpha \subseteq N$ be such that $|\alpha| = n-1$, and let $\lambda \in R$ be an eigenvalue of A. Then, there is an eigenvector x of A associated with λ such that $x(\alpha) = 0$ if and only if $\lambda \in \sigma(A[\alpha])$.*

Thus, the general scheme adopted here provides an algebraic proof to the statement in the opening paragraph.

In the case that A is Hermitian, the interlacing inequalities [HJ, Thm. 4.3.8] hold and, since any principal submatrix of an Hermitian matrix is Hermitian, Corollary 3 may be applied at each "level" of interlacing. Sequential application of Corollary 3 will lead to the corollaries below, but first several definitions are needed. For the following discussion, let $A \in M_n(C)$ be Hermitian. Suppose λ is an eigenvalue of A, then A is said to have *interlacing equality at λ of breadth k* if there are exactly k distinct index sets $\alpha_1, \alpha_2, \ldots, \alpha_k \subseteq N$ in which $|\alpha_i| = n-1$ and $\lambda \in \sigma(A[\alpha_i]), i = 1, 2, \ldots, k$. If A is such that $g_\lambda(A) = 1$, then the breadth of interlacing equality at λ is just the number of zero components in an eigenvector (because of Corollary 3). The matrix A is said to have *interlacing equality at λ of depth k* if $\lambda \in \sigma(A[\beta_j])$ for some index sets $\beta_0, B_1, \ldots, B_k \subseteq N$ such that $\beta_{j+1} \subset \beta_j, j = 0, 1, \ldots, k-1, |\beta_j| = n-j, j = 0, 1, \ldots, k$ and k is a maximum. If, in addition, $g_\lambda(A[\beta_{j+1}]) \geq g_\lambda(A[\beta_j]), j = 0, 1, \ldots, k-1$, then A is said to have *interlacing equality at λ of* ***restricted*** *depth k*. Here, k is the number of principal submatrices in the nested sequence for which the geometric multiplicity of λ is nondecreasing, so that the depth of interlacing equality may be greater than the *restricted* depth. The following corollaries relate these concepts.

COROLLARY 4. *Let $A \in M_n(C)$ be Hermitian and be such that $g_\lambda(A) = 1$. If A has interlacing equality at λ of breadth k, then A has interlacing equality at λ of depth at least k.*

Proof. If A has interlacing equality at λ of breadth k, then there are k distinct principal submatrices $A[\alpha_i]$ such that $\lambda \in \sigma(A[\alpha_i])$ and $|\alpha_i| = n-1$. In this case, $g_\lambda(A[\alpha_i]) \geq 1$ and, by assumption, $g_\lambda(A) = 1$. Thus, by Corollary 3, for each α_i there is an eigenvector y_i of A associated with λ, such that $y_i(\alpha_i) = 0$. However, since $g_\lambda(A) = 1$, the (right) eigenspace of A associated with λ is one dimensional, so that each of the y_i's may be taken to be the same, x. It follows that $x(\alpha_1 \cap \cdots \cap \alpha_k) = 0$. By the partitioned calculation mentioned in the opening paragraph $\beta_0 = N$, and $\beta_i = \alpha_1 \cap \cdots \cap \alpha_i, i = 1, \ldots, k$, exhibit that A has interlacing equality at λ of depth at least k. □

Corollary 4 is stated in the Hermitian case for parallelism to the corollaries that follow. However, it should be noted that the argument is equally valid in the normal case (using Corollary 1 in place of Corollary 3 with an obvious generalization of the definitions), so that Corollary 4 may be generalized by replacing "Hermitian" in the hypothesis with "normal." On the other hand, Corollary 4 is not valid for general matrices, as exhibited by the example

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix},$$

in which 0 is an eigenvalue of breadth 2, while its depth is only 1.

The converse to Corollary 4 does not hold. A counterexample is given by the

matrix

$$A = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix},$$

which has interlacing equality at 0 of depth 3 $(A(\{4\}), A(\{3,4\}), A(\{2,3,4\}))$, but interlacing equality at 0 of breadth only 2 $(A(\{3\}), A(\{4\}))$. However, the geometric multiplicities of the principal submatrices that yield interlacing equality at 0 of depth 3 are

$$g_0(A(\{4\})) = 1,$$
$$g_0(A(\{3,4\})) = 2,$$
$$g_0(A(\{2,3,4\})) = 1.$$

In fact, the restricted depth of interlacing equality at 0 is only 2 and this is exactly the breadth of interlacing equality at 0. As indicated in the following corollary, the breadth of interlacing equality at λ must be at least that of the restricted depth.

COROLLARY 5. *Let $A \in M_n(C)$ be Hermitian and suppose $\lambda \in \sigma(A)$. If A has interlacing equality at λ of restricted depth k, then A has interlacing equality at λ of breadth at least k.*

Proof. If $g_\lambda(A) > 1$, the breadth at λ is n (see discussion later, if necessary) and the conclusion is automatically valid. Thus, we suppose $g_\lambda(A) = 1$. If A has interlacing equality at λ of restricted depth k, then there is some nested sequence of $k+1$ principal submatrices $A[\beta_i]$, such that $|\beta_i| = n - i, \lambda \in \sigma(A[\beta_i]), i = 0, 1, \ldots, k$, and $g_\lambda(A[\beta_{i+1}]) \geq g_\lambda(A[\beta_i]), i = 0, 1, \ldots, k-1$. Assume, without loss of generality, that the rows and columns of $A[\beta_i]$ are numbered 1 to $n - i$. Note that $n - i$ is the index of the row and column deleted from $A[\beta_i]$ to obtain $A[\beta_{i+1}]$. By Corollary 2

$$\begin{aligned} \dim(LE^\lambda_{\beta_{i+1}}(A[\beta_i])), \dim(RE^\lambda_{\beta_{i+1}}(A[\beta_i])) &\geq \frac{g_\lambda(A[\beta_i]) + g_\lambda(A[\beta_{i+1}]) - 1}{2} \\ &\geq g_\lambda(A[\beta_i]) - \frac{1}{2} \end{aligned}$$

since $g_\lambda(A[\beta_i]) \leq g_\lambda(A[\beta_{i+1}])$. Both dimensions must be integral; so, the dimensions of the special eigenspaces must both be at least $g_\lambda(A[\beta_i])$. Then, every (left and right) eigenvector of $A[\beta_i]$ associated with λ is in the special (left and right) eigenspace and, thus, component $n - i$ of each of these vectors is 0.

Let x be an eigenvector (essentially unique) of A associated with λ. Since

$$g_\lambda(A) = g_\lambda(A[\beta_0]) = 1 \quad \text{and} \quad g_\lambda(A[\beta_1]) \geq 1,$$

by Corollary 3, $x(\beta_1) = 0$. By the preceding paragraph, if $i = 1$, then *every* eigenvector of $A[\beta_1]$ associated with λ, *including* $x[\beta_1]$, has a zero in the $n-1$ component. Thus, $x(\beta_1 \cap \beta_2) = x(\beta_2) = 0$.

Continuing in this manner, for each $i = 0, 1, \ldots, k-1, x[\beta_i]$ is an eigenvector of $A[\beta_i]$ associated with λ with a zero in the $n - i$ component so that

$$x(\beta_1 \cap \beta_2 \cap \cdots \cap \beta_{i+1}) = x(\beta_{i+1}) = 0.$$

Then, $x(\beta_k) = 0$ and for each $j \notin \beta_k, x(\{j\})$ is an eigenvector of $A(\{j\})$ associated with λ. Thus, $\alpha_j = N - \{n + 1 - j\}, j = 1, \ldots, k$, exhibits that A has interlacing equality at λ of breadth at least k. □

Note that the breadth of interlacing equality can be strictly greater than the *restricted* depth of interlacing equality. For example, the matrix

$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

has interlacing equality at 0 of restricted depth 1, but the breadth of interlacing equality at 0 is 2.

If the matrix A is such that $g_\lambda(A[\alpha]) \leq 1$ for every index set $\alpha \subseteq N$, and A has interlacing equality at λ of depth k, then A also has interlacing equality at λ of restricted depth k. In this case, by Corollary 5, A has interlacing equality at λ of breadth at least k. Combining Corollaries 4 and 5 then yields the following.

COROLLARY 6. *Let $A \in M_n(C)$ be Hermitian and suppose for every index set $\alpha \subseteq N$ that $g_\lambda(A[\alpha]) \leq 1$ with $g_\lambda(A) = 1$. Then, A has interlacing equality at λ of breadth k if and only if A has interlacing equality at λ of depth k.*

Let $A \in M_n(C)$ be Hermitian. Due to classical interlacing, when $g_\lambda(A) > 1, \lambda \in \sigma(A[\alpha])$ for any $\alpha \subseteq N$ such that $|\alpha| = n - 1$. In addition, when $g_\lambda(A) > 1$ there is for each such α an eigenvector, z, of A associated with λ such that $z(\alpha) = 0$. This may be seen in an elementary way by noting that, given any two linearly independent eigenvectors x, y in the eigenspace, there is a linear combination with a zero in any specified position. Such an A has interlacing equality at λ of breadth n, but may have depth at λ as little as 1. For example, the matrix

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

has interlacing equality at 0 of breadth 3, while the depth at 0 is only 1. Thus, the assumption in Corollaries 4 and 6 that $g_\lambda(A) = 1$ is necessary.

REFERENCES

[HJ] R. HORN AND C. R. JOHNSON *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 1013–1023, July 1995

020

THE SCHUR COMPLEMENT INTERLACING THEOREM*

SHU-AN HU† AND RONALD L. SMITH†

Abstract. The following analogue of the converse of the Cauchy Interlacing Theorem is proved: if $\lambda_1, \lambda_2, \ldots, \lambda_n, \mu_1, \mu_2, \ldots, \mu_{n-r}$ are $2n-r$ real numbers satisfying $\frac{1}{\lambda_1} \geq \frac{1}{\lambda_2} \geq \cdots \geq \frac{1}{\lambda_n}$, $\frac{1}{\mu_1} \geq \frac{1}{\mu_2} \geq \cdots \geq \frac{1}{\mu_{n-r}}$, and $\frac{1}{\lambda_i} \geq \frac{1}{\mu_i} \geq \frac{1}{\lambda_{i+r}}$, $1 \leq i \leq n-r$, and if the sets $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ and $\{\mu_1, \mu_2, \ldots, \mu_{n-r}\}$ have the same number of zeros, then there exists an $n \times n$ hermitian matrix H with an $r \times r$ nonsingular principal submatrix A such that the spectra of H and H/A (the Schur complement of H with respect to A) are $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ and $\{\mu_1, \mu_2, \ldots, \mu_{n-r}\}$, respectively. Here, the reciprocal of zero is defined to be zero. This result is then used to prove an analogue for semidefinite matrices.

Key words. eigenvalues, hermitian, inertia, interlacing, principal submatrix, Schur complement, semidefinite

AMS subject classifications. 15A42, 15A57

1. Introduction.

The *Cauchy interlacing theorem* was first proved in [2] for real symmetric matrices. The theorem states that the eigenvalues of a hermitian matrix are interlaced by the eigenvalues of any principal submatrix, i.e., any submatrix obtained from the original matrix by deleting the same rows and columns; it is usually proved using the Courant–Fischer minimax characterization of the eigenvalues of a hermitian matrix. The precise statement of the Cauchy interlacing theorem follows.

THEOREM 1. *Let H be an $n \times n$ hermitian matrix with partitioned form*

$$H = \begin{bmatrix} A & B \\ B^* & D \end{bmatrix},$$

where A has order r. Order the eigenvalues of H and A so that $\lambda_1(H) \geq \lambda_2(H) \geq \cdots \geq \lambda_n(H)$ and $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_r(A)$. Then, $\lambda_i(H) \geq \lambda_i(A) \geq \lambda_{i+n-r}(H)$, $i = 1, 2, \ldots, r$.

The converse was proved by Fan and Pall [4].

This classical separation theorem was generalized by Kantorovic [6], Wielandt [10], and Bauer [1] in developing results important to the rates of convergence of certain iterative methods of solving systems of equations, and also in obtaining error bounds in the use of direct methods.

In [8], an analogue of the Cauchy interlacing theorem was proven. More specifically, it was shown that if H is an $n \times n$ hermitian matrix and A is an $r \times r$ nonsingular principal submatrix, then the eigenvalues of H^+, the Moore–Penrose inverse of H, are interlaced by the eigenvalues of $(H/A)^+$, i.e., $\lambda_i(H^+) \geq \lambda_i((H/A)^+) \geq \lambda_{i+r}(H^+)$, $1 \leq i \leq n-r$. Here, H/A denotes the Schur complement of H with respect to A [5], and the reciprocal of zero is defined to be zero. In a private communication [9], T. Y.

*Received by the editors February 25, 1994; accepted for publication (in revised form) by R. A. Horn August 2, 1994.

†Department of Mathematics, The University of Tennessee at Chattanooga, Chattanooga, Tennessee 37403-2504 (shu@utcvm.utc.edu) and (rsmith@utcvm.utc.edu). The research of the first author was supported in part by a grant from The University of Chattanooga Foundation. The research of the second author was supported in part by a Summer Fellowship from The University of Chattanooga Foundation.

Tam proposed the converse of this analogue. That is, given that $\frac{1}{\lambda_1} \geq \frac{1}{\lambda_2} \geq \cdots \geq \frac{1}{\lambda_n}$, $\frac{1}{\mu_1} \geq \frac{1}{\mu_2} \geq \cdots \geq \frac{1}{\mu_{n-r}}$, and $\frac{1}{\lambda_i} \geq \frac{1}{\mu_i} \geq \frac{1}{\lambda_{i+r}}$, $1 \leq i \leq n-r$, does there exist an $n \times n$ hermitian matrix H with an $r \times r$ nonsingular principal submatrix A such that the spectra of H and H/A are $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ and $\{\mu_1, \mu_2, \ldots, \mu_{n-r}\}$, respectively? Observe that by the well-known Inertia Theorem [5], the sets $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ and $\{\mu_1, \mu_2, \ldots, \mu_{n-r}\}$ must have the same number of zeros. In this paper, we use a constructive process to answer Tam's question in the affirmative.

In [8], it was also shown that if H is an $n \times n$ hermitian semidefinite matrix and A is an $r \times r$ nonsingular principal submatrix of H, then the eigenvalues of H are interlaced by the eigenvalues of H/A, i.e., $\lambda_i(H) \geq \lambda_i(H/A) \geq \lambda_{i+r}(H)$, $i = 1, 2, \ldots, n-r$. We show that the converse of this result holds also.

Schur complements of positive definite matrices have been used to provide a priori estimates for regular solutions of the n-metaharmonic differential equation in [7].

2. Notation. We use the following notation.

$\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_{n-1}, \lambda_n)$,

$\mu = (\mu_1, \mu_2, \ldots, \mu_{n-1}, \mu_n)$,

$\mu^{in} = (\mu_1, \mu_2, \ldots, \mu_{i-1}, \mu_{i+1}, \ldots, \mu_{n-1})$, $i = 1, 2, \ldots, n-1$, and

$\mu^{ijn} = (\mu_1, \mu_2, \ldots, \mu_{i-1}, \mu_{i+1}, \ldots, \mu_{j-1}, \mu_{j+1}, \ldots, \mu_{n-1})$, $1 \leq i, j \leq n-1$.

Also, let $E_j^k(x_1, \ldots, x_k)$ denote the jth elementary symmetric function of k variables, where $1 \leq j \leq k$. Define $E_0^k(x_1, \ldots, x_k) = 1$, $k \geq 1$ for convenience.

For a complex matrix A, let A^T denote the transpose of A, and let A^* denote the conjugate transpose of A. If A is square, let $|A|$ denote the determinant of A, and let $M[i_1, i_2, \ldots, i_k]$ denote the principal minor of A obtained by selecting the rows and columns of A indexed by the strictly increasing sequence $i_1, i_2, \ldots, i_k$.

3. Main results. Our main result is the following theorem.

THEOREM 2. *Suppose $\lambda_1, \lambda_2, \ldots, \lambda_n$ and $\mu_1, \mu_2, \ldots, \mu_{n-r}$ are $2n-r$ real numbers satisfying $\frac{1}{\lambda_1} \geq \frac{1}{\lambda_2} \geq \cdots \geq \frac{1}{\lambda_n}$, $\frac{1}{\mu_1} \geq \frac{1}{\mu_2} \geq \cdots \geq \frac{1}{\mu_{n-r}}$. Here, the reciprocal of zero is defined to be zero. Then there exists an $n \times n$ hermitian matrix H with an $r \times r$ nonsingular principal submatrix A such that the spectra of H and H/A are $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ and $\{\mu_1, \mu_2, \ldots, \mu_{n-r}\}$, respectively, if and only if the following are true.*

(i) *The sets $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ and $\{\mu_1, \mu_2, \ldots, \mu_{n-r}\}$ have the same number of zeros, and*

(ii) $\frac{1}{\lambda_i} \geq \frac{1}{\mu_i} \geq \frac{1}{\lambda_{i+r}}$, $1 \leq i \leq n-r$.

In fact, if (i) and (ii) hold, there is a real symmetric matrix H with an $r \times r$ nonsingular submatrix A such that the spectra of H and H/A are $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ and $\{\mu_1, \mu_2, \ldots, \mu_{n-r}\}$, respectively.

The theorem allows us to obtain the following analogous result for semidefinite matrices.

COROLLARY. *Suppose $\lambda_1, \lambda_2, \ldots, \lambda_n$ and $\mu_1, \mu_2, \ldots, \mu_{n-r}$ are $2n-r$ real numbers satisfying* (a) $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$, $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_{n-r} \geq 0$ *or* (b) $0 \geq \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, $0 \geq \mu_1 \geq \mu_2 \geq \cdots \geq \mu_{n-r}$. *Then there exists an $n \times n$ hermitian semidefinite matrix H with an $r \times r$ nonsingular principal submatrix A such that the spectra of H and H/A are $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ and $\{\mu_1, \mu_2, \ldots, \mu_{n-r}\}$, respectively, if and only if the following are true.*

(i) *The sets $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ and $\{\mu_1, \mu_2, \ldots, \mu_{n-r}\}$ have the same number of zeros, and*

(ii) $\lambda_i \geq \mu_i \geq \lambda_{i+r}$, $1 \leq i \leq n-r$.

In fact, if (i) *and* (ii) *hold, there is a real symmetric semidefinite matrix* H *with an* $r \times r$ *nonsingular submatrix* A *such that the spectra of* H *and* H/A *are* $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ *and* $\{\mu_1, \mu_2, \ldots, \mu_{n-r}\}$, *respectively.*

To facilitate the proof of both Theorem 2 and the corollary, we first prove several lemmas.

LEMMA 1. *Let* $\Lambda = \operatorname{diag}(\mu_1, \mu_2, \ldots, \mu_k)$ *and* $C = [c_1, c_2, \ldots, c_k]$. *Then,*

$$|\Lambda + C^T C| = \prod_{j=1}^{k} \mu_j + \sum_{i=1}^{k} \left(\prod_{1 \le j \le k}^{j \ne i} \mu_j \right) c_i^2.$$

Proof. Since

$$\begin{bmatrix} 1 & 0 \\ C^T & I \end{bmatrix} \begin{bmatrix} -1 & C \\ C^T & \Lambda \end{bmatrix} \begin{bmatrix} 1 & C \\ 0 & I \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & \Lambda + C^T C \end{bmatrix},$$

$$|\Lambda + C^T C| = -\begin{vmatrix} -1 & C \\ C^T & \Lambda \end{vmatrix} = -\begin{vmatrix} \Lambda & C^T \\ C & -1 \end{vmatrix} = -\begin{vmatrix} \mu_1 & \cdots & 0 & 0 & c_1 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & \mu_{k-1} & 0 & c_{k-1} \\ 0 & \cdots & 0 & \mu_k & c_k \\ c_1 & \cdots & c_{k-1} & c_k & -1 \end{vmatrix} = -|D|,$$

where $D = \begin{bmatrix} \Lambda & C^T \\ C & -1 \end{bmatrix}$.

If some μ_j equals zero, say $\mu_i = 0$, then expand successively by the ith column of D and the ith row of the resulting determinant to obtain

$$|\Lambda + C^T C| = -|D| = \left(\prod_{1 \le j \le k}^{j \ne i} \mu_j \right) c_i^2,$$

and the theorem holds.

So, assume no μ_j equals zero. Then, by Schur's formula [3],

$$\begin{aligned} |\Lambda + C^T C| = -|D| = -|\Lambda| \cdot |D/\Lambda| &= -\left(\prod_{j=1}^{k} \mu_j \right) \left(-1 - \sum_{i=1}^{k} \frac{1}{\mu_i} c_i^2 \right) \\ &= \prod_{j=1}^{k} \mu_j + \sum_{i=1}^{k} \left(\prod_{1 \le j \le k}^{j \ne i} \mu_j \right) c_i^2, \end{aligned}$$

which completes the proof. □

LEMMA 2. *Let*

$$H = \begin{bmatrix} a & b_1 & b_2 & \cdots & b_{n-1} \\ b_1 & \mu_1 + \frac{b_1^2}{a} & \frac{b_1 b_2}{a} & \cdots & \frac{b_1 b_{n-1}}{a} \\ b_2 & \frac{b_2 b_1}{a} & \mu_2 + \frac{b_2^2}{a} & \cdots & \frac{b_2 b_{n-1}}{a} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{n-1} & \frac{b_{n-1} b_1}{a} & \frac{b_{n-1} b_3}{a} & \cdots & \mu_{n-1} + \frac{b_{n-1}^2}{a} \end{bmatrix},$$

where $a = \mu_n$ is nonzero. Let M_j denote the sum of jth order principal minors of H. Then,

$$M_j = E_j^n(\mu) + \frac{1}{a}\sum_{i=1}^{n-1} E_{j-1}^{n-2}(\mu^{in})b_i^2, \quad j = 1, 2, \ldots, n-1, \text{ and}$$

$$M_n = E_n^n(\mu) = \mu_1\mu_2 \ldots \mu_{n-1}\mu_n = a\mu_1\mu_2 \ldots \mu_{n-1}.$$

Proof. We prove the formula for M_n first, because the same argument is used to prove the formula for the other M_k.

Observe that

$$H = \begin{bmatrix} a & B \\ B^T & \Lambda + \frac{1}{a}B^TB \end{bmatrix},$$

where $\Lambda = \operatorname{diag}(\mu_1, \mu_2, \ldots, \mu_{n-1})$ and $B = [b_1, b_2, \ldots, b_{n-1}]$. We can use the following well-known identity:

$$\begin{bmatrix} 1 & 0 \\ \frac{-1}{a}B^T & I \end{bmatrix}\begin{bmatrix} a & B \\ B^T & \Lambda + \frac{1}{a}B^TB \end{bmatrix}\begin{bmatrix} 1 & \frac{-1}{a}B \\ 0 & I \end{bmatrix} = \begin{bmatrix} a & 0 \\ 0 & \Lambda + \frac{1}{a}B^TB - \frac{1}{a}B^TB \end{bmatrix}$$

$$= \begin{bmatrix} a & 0 \\ 0 & \Lambda \end{bmatrix}.$$

Therefore, $M_n = |H| = a\mu_1\mu_2 \ldots \mu_{n-1}$.

For any jth order principal minor containing the first row and column of H, say $M[1, i_1+1, \ldots, i_{j-1}+1]$, where $1 \le i_1 < i_2 < \cdots < i_{j-1} \le n-1$, we can use the same argument as for M_n to show that it equals $a\mu_{i_1}\mu_{i_2} \ldots \mu_{i_{j-1}} = \mu_{i_1}\mu_{i_2} \ldots \mu_{i_{j-1}}\mu_n$.

For any jth order principal minor that does not contain the first row and column of H, say $M[i_1+1, i_2+1, \ldots, i_j+1]$, where $1 \le i_1 < i_2 < \cdots < i_j \le n-1$, we have

$$M[i_1+1, i_2+1, \ldots, i_j+1] = \begin{vmatrix} \mu_{i_1} + \frac{b_{i_1}^2}{a} & \frac{b_{i_1}b_{i_2}}{a} & \cdots & \frac{b_{i_1}b_{i_j}}{a} \\ \frac{b_{i_2}b_{i_1}}{a} & \mu_{i_2} + \frac{b_{i_2}^2}{a} & \cdots & \frac{b_{i_2}b_{i_j}}{a} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{b_{i_j}b_{i_1}}{a} & \frac{b_{i_j}b_{i_2}}{a} & \cdots & \mu_{i_j} + \frac{b_{i_j}^2}{a} \end{vmatrix}.$$

If we let $c_{i_k} = b_{i_k}/\sqrt{a}$, where $\sqrt{a}$ is the principal square root of a, $k = 1, \ldots, j$, then Lemma 1 provides

$$M[i_1+1, i_2+1, \ldots, i_j+1] = \prod_{t=1}^{j}\mu_{i_t} + \frac{1}{a}\sum_{k=1}^{j}\left(\prod_{1\le t\le j}^{t\ne k}\mu_{i_t}\right)b_{i_k}^2.$$

Taking the summation of all jth order principal minors of H, we have

$$M_j = E_j^n(\mu) + \frac{1}{a}\sum_{i=1}^{n-1} E_{j-1}^{n-2}(\mu^{in})b_i^2, \quad j = 1, 2, \ldots, n-1. \qquad \square$$

LEMMA 3. *Let*

$$S = \begin{bmatrix} E_0^{n-2}(\mu^{1n}) & E_0^{n-2}(\mu^{2n}) & E_0^{n-2}(\mu^{3n}) & \cdots & E_0^{n-2}(\mu^{(n-1)n}) \\ E_1^{n-2}(\mu^{1n}) & E_1^{n-2}(\mu^{2n}) & E_1^{n-2}(\mu^{3n}) & \cdots & E_1^{n-2}(\mu^{(n-1)n}) \\ E_2^{n-2}(\mu^{1n}) & E_2^{n-2}(\mu^{2n}) & E_2^{n-2}(\mu^{3n}) & \cdots & E_2^{n-2}(\mu^{(n-1)n}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ E_{n-2}^{n-2}(\mu^{1n}) & E_{n-2}^{n-2}(\mu^{2n}) & E_{n-2}^{n-2}(\mu^{3n}) & \cdots & E_{n-2}^{n-2}(\mu^{(n-1)n}) \end{bmatrix}.$$

Then,

$$|S| = \prod_{1\le i<j\le n-1} (\mu_i - \mu_j) \text{ and}$$

$$|S(t;k)| = \mu_k^{n-1-t} \prod_{1\le i<j\le n-1}^{i\neq k, j\neq k} (\mu_i - \mu_j),\ 1 \le t, k \le n-1,$$

where $S(i;j)$ denotes the submatrix obtained from S by deleting the ith row and the jth column.

Proof. To prove the first identity, observe that the determinant on the right is actually a homogeneous polynomial in the variables $\mu_1, \mu_2, \ldots, \mu_{n-1}$. Its total degree is $0+1+2+\cdots+(n-2) = \frac{(n-2)(n-1)}{2}$. If $\mu_i = \mu_j$, then for any i, j, $1 \le i < j \le n-1$, we see that the ith column and jth column are identical. This implies that $(\mu_i - \mu_j)$ is a factor of the polynomial. There are $\binom{n-1}{2} = \frac{(n-2)(n-1)}{2}$ such factors. Hence, the determinant is a scalar multiple of $\prod_{1\le i<j\le n-1}(\mu_i - \mu_j)$. Since the coefficient of $\mu_1^{n-2}\mu_2^{n-3}\mu_3^{n-4}\ldots\mu_{n-3}^2\mu_{n-2}$ in the determinant is 1, the scalar must be 1. Therefore, the first identity holds.

To prove the second identity, first note that an argument identical to that for the first one yields

$$|S(n-1;k)| = \prod_{1\le i<j\le n-1}^{i\neq k, j\neq k} (\mu_i - \mu_j),\ 1 \le k \le n-1.$$

So assume $1 \le t < n-1$ and observe that for any j, $j \neq k$, $1 \le j \le n-1$, we have

$$E_{n-2}^{n-2}(\mu^{jn}) = \mu_k E_{n-3}^{n-3}(\mu^{kln}),$$

and

$$E_r^{n-2}(\mu^{jn}) - E_r^{n-3}(\mu^{kjn}) = \mu_k E_{r-1}^{n-3}(\mu^{kjn}), \text{ where } 1 \le r < n-2.$$

Start with the last row and proceed upward, using the above observation. Factor out μ_k and then subtract the resulting row from the one above it, ending when μ_k is factored from the tth row. We obtain

$$|S(t;k)| = \begin{vmatrix} E_0^{n-2}(\mu^{1n}) & \cdots & E_0^{n-2}(\mu^{(k-1)n}) & E_0^{n-2}(\mu^{(k+1)n}) & \cdots & E_0^{n-2}(\mu^{(n-1)n}) \\ E_1^{n-2}(\mu^{1n}) & \cdots & E_1^{n-2}(\mu^{(k-1)n}) & E_1^{n-2}(\mu^{(k+1)n}) & \cdots & E_1^{n-2}(\mu^{(n-1)n}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ E_{t-2}^{n-2}(\mu^{1n}) & \cdots & E_{t-2}^{n-2}(\mu^{(k-1)n}) & E_{t-2}^{n-2}(\mu^{(k+1)n}) & \cdots & E_{t-2}^{n-2}(\mu^{(n-1)n}) \\ E_t^{n-2}(\mu^{1n}) & \cdots & E_t^{n-2}(\mu^{(k-1)n}) & E_t^{n-2}(\mu^{(k+1)n}) & \cdots & E_t^{n-2}(\mu^{(n-1)n}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ E_{n-2}^{n-2}(\mu^{1n}) & \cdots & E_{n-2}^{n-2}(\mu^{(k-1)n}) & E_{n-2}^{n-2}(\mu^{(k+1)n}) & \cdots & E_{n-2}^{n-2}(\mu^{(n-1)n}) \end{vmatrix}$$

which is equal to μ_k^{n-1-t} times

$$\begin{vmatrix} E_0^{n-2}(\mu^{1n}) & \cdots & E_0^{n-2}(\mu^{(k-1)n}) & E_0^{n-2}(\mu^{(k+1)n}) & \cdots & E_0^{n-2}(\mu^{(n-1)n}) \\ E_1^{n-2}(\mu^{1n}) & \cdots & E_1^{n-2}(\mu^{(k-1)n}) & E_1^{n-2}(\mu^{(k+1)n}) & \cdots & E_1^{n-2}(\mu^{(n-1)n}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ E_{t-2}^{n-2}(\mu^{1n}) & \cdots & E_{t-2}^{n-2}(\mu^{(k-1)n}) & E_{t-2}^{n-2}(\mu^{(k+1)kn}) & \cdots & E_{t-2}^{n-2}(\mu^{(n-1)n}) \\ E_{t-1}^{n-3}(\mu^{1kn}) & \cdots & E_{t-1}^{n-3}(\mu^{(k-1)kn}) & E_{t-1}^{n-3}(\mu^{(k+1)kn}) & \cdots & E_{t-1}^{n-3}(\mu^{(n-1)kn}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ E_{n-3}^{n-3}(\mu^{1kn}) & \cdots & E_{n-3}^{n-3}(\mu^{(k-1)kn}) & E_{n-3}^{n-3}(\mu^{(k+1)kn}) & \cdots & E_{n-3}^{n-3}(\mu^{(n-1)kn}) \end{vmatrix}.$$

The determinant on the right is again a homogeneous polynomial in the variables μ_1, $\mu_2, \ldots, \mu_{n-1}$. By a similar argument to that given in the first part, it can be shown that this determinant equals to

$$\prod_{1\le i<j\le n-1}^{i\neq k, j\neq k} (\mu_i - \mu_j).$$

Thus,

$$|S(t;k)| = \mu_k^{n-1-t} \prod_{1\le i<j\le n-1}^{i\neq k, j\neq k} (\mu_i - \mu_j),\ 1 \le t, k \le n-1,$$

and the second identity holds. □

In the next lemma, we prove the theorem for $r = 1$.

LEMMA 4. *If* $\lambda_1, \lambda_2, \ldots, \lambda_n$ *and* $\mu_1, \mu_2, \ldots, \mu_{n-1}$ *are* $2n-1$ *real numbers satisfying* $\frac{1}{\lambda_1} \ge \frac{1}{\mu_1} \ge \frac{1}{\lambda_2} \ge \frac{1}{\mu_2} \ge \cdots \ge \frac{1}{\lambda_{n-1}} \ge \frac{1}{\mu_{n-1}} \ge \frac{1}{\lambda_n}$, *and if the sets* $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ *and* $\{\mu_1, \mu_2, \ldots, \mu_{n-1}\}$ *have the same number of zeros, then there exists a real symmetric matrix*

$$H = \begin{bmatrix} a & B \\ B^T & D \end{bmatrix},$$

such that its spectrum is $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ *and the spectrum of its Schur complement* $H/a = D - \frac{1}{a}B^TB$ *is* $\{\mu_1, \mu_2, \ldots, \mu_{n-1}\}$.

Proof. We first prove the lemma for the case where $\lambda_1, \lambda_2, \ldots, \lambda_n, \mu_1, \mu_2, \ldots, \mu_{n-1}$ are nonzero and interlacing inequalities are strict. Let $a = \frac{\lambda_1\lambda_2\cdots\lambda_n}{\mu_1\mu_2\cdots\mu_{n-1}} = \mu_n$ and let b_i, $i = 1, \ldots, n-1$, be $n-1$ unknowns. Furthermore, let

$$H = \begin{bmatrix} a & B \\ B^T & D \end{bmatrix},$$

where $D = \Lambda + \frac{1}{a}B^TB$, $\Lambda = \operatorname{diag}(\mu_1, \mu_2, \ldots, \mu_{n-1})$ and $B = [b_1, b_2, \ldots, b_{n-1}]$. It is apparent that the Schur complement $H/a = D - \frac{1}{a}B^TB = \Lambda$ has spectrum $\{\mu_1, \mu_2, \ldots, \mu_{n-1}\}$. Therefore, it suffices to find a real solution for the b_i, $i = 1, \ldots, n-1$, such that the spectrum of H is $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$.

The characteristic polynomial of H is

$$\lambda^n - M_1\lambda^{n-1} + M_2\lambda^{n-2} + \ldots + (-1)^{n-1}M_{n-1}\lambda + (-1)^n M_n,$$

where M_j is the sum of jth order principal minors of H. From Lemma 2,

$$M_j = E_j^n(\mu) + \frac{1}{a}\sum_{i=1}^{n-1} E_{j-1}^{n-2}(\mu^{in})b_i^2, \ j = 1, 2, \ldots, n-1, \text{ and}$$

$$M_n = |H| = a\mu_1\mu_2 \ldots \mu_{n-1}.$$

Assuming that $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ is the set of roots of the polynomial, we have the system of equations $M_j = E_j^n(\lambda)$, $1 \le j \le n$. The last equation $M_n = E_n^n(\lambda)$ does not involve b_i. Actually, it was used to determine a at the beginning of the proof. The other $n-1$ equations give a linear system of $n-1$ unknowns

$$x_i = \frac{b_i^2}{a}, \ i = 1, \ldots, n-1.$$

Rewriting the system in matrix form, we have

$$\begin{bmatrix} E_0^{n-2}(\mu^{1n}) & E_0^{n-2}(\mu^{2n}) & \cdots & E_0^{n-2}(\mu^{(n-1)n}) \\ E_1^{n-2}(\mu^{1n}) & E_1^{n-2}(\mu^{2n}) & \cdots & E_1^{n-2}(\mu^{(n-1)n}) \\ E_2^{n-2}(\mu^{1n}) & E_2^{n-2}(\mu^{2n}) & \cdots & E_2^{n-2}(\mu^{(n-1)n}) \\ \vdots & \vdots & \ddots & \vdots \\ E_{n-2}^{n-2}(\mu^{1n}) & E_{n-2}^{n-2}(\mu^{2n}) & \cdots & E_{n-2}^{n-2}(\mu^{(n-1)n}) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \end{bmatrix} = \begin{bmatrix} E_1^n(\lambda) - E_1^n(\mu) \\ E_2^n(\lambda) - E_2^n(\mu) \\ E_3^n(\lambda) - E_3^n(\mu) \\ \vdots \\ E_{n-1}^n(\lambda) - E_{n-1}^n(\mu) \end{bmatrix}$$

We use Cramer's rule to solve the system. By the first identity of Lemma 3,

$$\Delta = \prod_{1 \le i < j \le n-1} (\mu_i - \mu_j).$$

Furthermore, applying the second identity of Lemma 3 and expanding along the kth column, we have

$$\begin{aligned} \Delta_k &= \sum_{t=1}^{n-1} (-1)^{t+k}(E_t^n(\lambda) - E_t^n(\mu))|S(t;k)| \\ &= \sum_{t=1}^{n-1} (-1)^{t+k}(E_t^n(\lambda) - E_t^n(\mu)) \left[\mu_k^{n-1-t} \prod_{1 \le i < j \le n-1}^{i \ne k, j \ne k} (\mu_i - \mu_j) \right] \\ &= (-1)^k \left(\frac{1}{\mu_k}\right) \sum_{t=1}^{n-1} (-1)^t (E_t^n(\lambda) - E_t^n(\mu)) \left[\mu_k^{n-t} \prod_{1 \le i < j \le n-1}^{i \ne k, j \ne k} (\mu_i - \mu_j) \right] \\ &= (-1)^k \left(\frac{1}{\mu_k}\right) \sum_{t=0}^{n-1} (-1)^t (E_t^n(\lambda) - E_t^n(\mu)) \left[\mu_k^{n-t} \prod_{1 \le i < j \le n-1}^{i \ne k, j \ne k} (\mu_i - \mu_j) \right] \end{aligned}$$

$$= (-1)^k \left(\frac{1}{\mu_k}\right) \left[\sum_{t=0}^{n-1} (-1)^t (E_t^n(\lambda)\mu_k^{n-t} - E_t^n(\mu))\mu_k^{n-t}\right] \prod_{1\le i<j\le n-1}^{i\neq k, j\neq k} (\mu_i - \mu_j)$$

$$= (-1)^k \left(\frac{1}{\mu_k}\right) \left[\prod_{s=1}^{n}(\mu_k - \lambda_s) - \prod_{s=1}^{n}(\mu_k - \mu_s)\right] \prod_{1\le i<j\le n-1}^{i\neq k, j\neq k} (\mu_i - \mu_j)$$

$$= (-1)^k \left(\frac{1}{\mu_k}\right) \prod_{s=1}^{n}(\mu_k - \lambda_s) \prod_{1\le i<j\le n-1}^{i\neq k, j\neq k} (\mu_i - \mu_j).$$

In step 4 observe that $E_n^n(\lambda) = E_n^n(\mu)$ (since $\mu_n = \frac{\lambda_1\lambda_2\cdots\lambda_n}{\mu_1\mu_2\cdots\mu_{n-1}}$) and $E_0^n(\lambda) = E_0^n(\mu) = 1$.

Therefore, the system has the unique solution:

$$x_k = \frac{b_k^2}{a} = \frac{\Delta_k}{\Delta}.$$

This implies

$$\begin{aligned} b_k^2 &= \frac{a[(-1)^k \prod_{s=1}^n(\mu_k - \lambda_s)]}{\mu_k \prod_{1\le i<k}(\mu_i - \mu_k) \prod_{k<j\le n-1}(\mu_k - \mu_j)} \\ &= \frac{(-1)^k a\mu_k [\prod_{s=1}^n \lambda_s][\prod_{s=1}^n (\frac{1}{\lambda_s} - \frac{1}{\mu_k})]}{[\prod_{1\le i<k}(\frac{1}{\mu_k} - \frac{1}{\mu_i})][\prod_{k<j\le n-1}(\frac{1}{\mu_j} - \frac{1}{\mu_k})][\prod_{1\le i\le n-1}^{i\neq k} \mu_i]} \\ &= \frac{(-1)^k a^2 \mu_k^2 \prod_{s=1}^n (\frac{1}{\lambda_s} - \frac{1}{\mu_k})}{[\prod_{1\le i<k}(\frac{1}{\mu_k} - \frac{1}{\mu_i})][\prod_{k<j\le n-1}(\frac{1}{\mu_j} - \frac{1}{\mu_k})]}. \end{aligned}$$

The last equality follows from the fact that $a = \frac{\lambda_1\lambda_2\cdots\lambda_n}{\mu_1\mu_2\cdots\mu_{n-1}}$.

According to the given interlacing inequalities, there are $n-k$ negative factors in the product in the numerator and $n-2$ negative factors in the two products in the denominator. In all, there are $(n-k)+(n-2)+k = 2n-2$ negative factors. Therefore, the expression is positive and we can find a real solution for b_k, $k = 1, \ldots, n-1$.

For the general case, renumber the sequences $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_{n-r})$ so that

$$\frac{1}{\lambda_1} > \frac{1}{\mu_1} > \frac{1}{\lambda_2} > \frac{1}{\mu_2} > \cdots \frac{1}{\lambda_{m-1}} > \frac{1}{\mu_{m-1}} > \frac{1}{\lambda_m} \text{ and } \frac{1}{\lambda_j} = \frac{1}{\mu_{j-1}},$$

$m+1 \le j \le n$. By assumption, $\lambda_i \neq 0$ and $\mu_{i-1} \neq 0$ for $1 \le i \le m$.

By the first part of the proof, there is a real symmetric matrix

$$H_1 = \begin{bmatrix} a & B_1 \\ B_1^T & D_1 \end{bmatrix}$$

such that the spectrum of H_1 is $\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ and the spectrum of $H_1/a = D_1 - \frac{1}{a}B_1^T B_1$ is $\{\mu_1, \mu_2, \ldots, \mu_{m-1}\}$. Now let

$$H = \begin{bmatrix} H_1 & 0 \\ 0 & D_2 \end{bmatrix} = \begin{bmatrix} a & B_1 & 0 \\ B_1^T & D_1 & 0 \\ 0 & 0 & D_2 \end{bmatrix} = \begin{bmatrix} a & B \\ B^T & D \end{bmatrix},$$

where $D_2 = \operatorname{diag}(\lambda_{m+1}, \lambda_{m+2}, \ldots, \lambda_n) = \operatorname{diag}(\mu_m, \mu_{m+1}, \ldots, \mu_{n-1})$,

$$D = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}$$

and $B = [B_1 \quad 0]$. The spectrum of H is $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ while the spectrum of

$$H/a = \begin{bmatrix} D_1 - \frac{1}{a}B_1^T B_1 & 0 \\ 0 & D_2 \end{bmatrix}$$

is $\{\mu_1, \mu_2, \ldots, \mu_{n-1}\}$ which completes the proof. □

Example. The following example serves to illustrate Lemma 4. Suppose we have the interlacing real numbers $\frac{1}{\lambda_1} = \frac{1}{\mu_1} = \frac{1}{\lambda_2} = 1 > \frac{1}{\mu_2} = \frac{1}{2} > \frac{1}{\lambda_3} = \frac{1}{4} > \frac{1}{\mu_3} = \frac{1}{\lambda_4} = 0 > \frac{1}{\mu_4} = -\frac{1}{4} > \frac{1}{\lambda_5} = -1 > \frac{1}{\mu_5} = \frac{1}{\lambda_6} = -2$.

Renumber the λ's and μ's so that $\frac{1}{\lambda_1} = 1 > \frac{1}{\mu_1} = \frac{1}{2} > \frac{1}{\lambda_2} = \frac{1}{4} > \frac{1}{\mu_2} = \frac{1}{4} > \frac{1}{\lambda_3} = -1$, $\frac{1}{\lambda_4} = \frac{1}{\mu_3} = 1$, $\frac{1}{\lambda_5} = \frac{1}{\mu_4} = 0$, and $\frac{1}{\lambda_6} = \frac{1}{\mu_5} = -2$. Then $\lambda_1 = 1$, $\lambda_2 = 4$, $\lambda_3 = -1$, $\mu_1 = 2$, $\mu_2 = -4$, and $\mu_3 = a = \frac{1 \cdot 4 \cdot (-1)}{2 \cdot (-4)} = \frac{1}{2}$. Also, $\lambda_4 = \mu_3 = 1$, $\lambda_5 = \mu_4 = 0$, and $\lambda_6 = \mu_5 = -\frac{1}{2}$.

Furthermore, $M_1 = E_1^3(\mu_1 = 2, \mu_2 = -4, \mu_3 = \frac{1}{2}) + 2(b_1^2 + b_2^2) = -\frac{3}{2} + 2b_1^2 + 2b_2^2$, $M_2 = E_2^3(\mu_1 = 2, \mu_2 = -4, \mu_3 = \frac{1}{2}) + 2(E_1^1(\mu_2 = -4)b_1^2 + E_1^1(\mu_1 = 2)b_2^2) = -9 - 8b_1^2 + 4b_2^2$, $E_1^3(\lambda) = E_1^3(\lambda_1 = 1, \lambda_2 = 4, \lambda_3 = -1) = 4$, and $E_2^3(\lambda) = -1$.

Solving the system $M_1 = E_1^3(\lambda)$ and $M_2 = E_2^3(\lambda)$, we obtain $b_1^2 = \frac{1}{4}$ and $b_2^2 = \frac{5}{2}$. So we can choose $b_1 = \frac{1}{2}$ and $b_2 = \frac{\sqrt{10}}{2}$. Then, let

$$H_1 = \begin{bmatrix} a & B_1 \\ B_1^T & \Lambda_1 + \frac{1}{a}B_1^T B_1 \end{bmatrix},$$

where $\Lambda_1 = \operatorname{diag}(2, -4)$ and $B_1 = [\frac{1}{2} \quad \frac{\sqrt{10}}{2}]$. Finally, let

$$H = \begin{bmatrix} H_1 & 0 \\ 0 & D_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{\sqrt{10}}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{5}{2} & \frac{\sqrt{10}}{2} & 0 & 0 & 0 \\ \frac{\sqrt{10}}{2} & \frac{\sqrt{10}}{2} & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} \end{bmatrix},$$

where $D_2 = \operatorname{diag}(1, 0, -\frac{1}{2})$. H is the desired real symmetric matrix such that the spectra of H and H/a are $\{4, 1, 1, 0, -\frac{1}{2}, -1\}$ and $\{2, 1, 0, -\frac{1}{2}, -4\}$, respectively.

Remark. The proof of Lemma 4 is the essence of the the proof of the theorem since we will show that we can insert $r - 1$ intermediate sequences from μ to λ such that (1) each intermediate sequence is interlaced by the previous one in the manner of Lemma 4 and (2) the last inserted sequence interlaces λ in the manner of Lemma 4.

Constructively, we could then apply Lemma 4 r times in order to obtain the desired real symmetric matrix H.

Proof of Theorem 2. Suppose H is an $n \times n$ hermitian matrix with an $r \times r$ nonsingular principal submatrix A such that the spectra of H and H/A are $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ and $\{\mu_1, \mu_2, \ldots, \mu_{n-r}\}$, respectively. Then (i) follows from the well-known Inertia Theorem [5] and (ii) was proved in [8].

For the converse, suppose (i) and (ii) hold. If $r = 1$, we are done by Lemma 4. Inductively, assume the theorem holds for all positive integers less than r where $2 \le r \le n-1$.

Insert an intermediate sequence $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_{n-r+1})$ as follows. First select γ_1 such that $\frac{1}{\lambda_1} \ge \frac{1}{\gamma_1} \ge \max\{\frac{1}{\mu_1}, \frac{1}{\lambda_r}\}$. Then $\frac{1}{\lambda_1} \ge \frac{1}{\gamma_1} \ge \frac{1}{\lambda_r}$ and $\frac{1}{\gamma_1} \ge \frac{1}{\mu_1}$. Observe that $\min\{\frac{1}{\lambda_s}, \frac{1}{\mu_{s-1}}\} \ge \max\{\frac{1}{\mu_s}, \frac{1}{\lambda_{r+s-1}}\}$ for $s = 2, 3, \dots, n-r$. So for $s = 2, 3, \dots, n-r$, select γ_s such that $\min\{\frac{1}{\lambda_s}, \frac{1}{\mu_{s-1}}\} \ge \frac{1}{\gamma_s} \ge \max\{\frac{1}{\mu_s}, \frac{1}{\lambda_{r+s-1}}\}$. Then $\frac{1}{\lambda_s} \ge \frac{1}{\gamma_s} \ge \frac{1}{\lambda_{r+s-1}}$ and $\frac{1}{\mu_{s-1}} \ge \frac{1}{\gamma_s} \ge \frac{1}{\mu_s}$, $2 \le s \le n-r$. Finally, select γ_{n-r+1} such that $\min\{\frac{1}{\lambda_{n-r+1}}, \frac{1}{\mu_{n-r}}\} \ge \frac{1}{\gamma_{n-r+1}} \ge \frac{1}{\lambda_n}$. Then $\frac{1}{\lambda_{n-r+1}} \ge \frac{1}{\gamma_{n-r+1}} \ge \frac{1}{\lambda_n}$ and $\frac{1}{\mu_{n-r}} \ge \frac{1}{\gamma_{n-r+1}}$. In summary, the intermediate sequence γ satisfies $\frac{1}{\gamma_1} \ge \frac{1}{\mu_1} \ge \frac{1}{\gamma_2} \ge \frac{1}{\mu_2} \ge \cdots \ge \frac{1}{\gamma_{n-r}} \ge \frac{1}{\mu_{n-r}} \ge \frac{1}{\gamma_{n-r+1}}$ and $\frac{1}{\lambda_i} \ge \frac{1}{\gamma_i} \ge \frac{1}{\lambda_{i+r-1}}$, $1 \le i \le n-r+1$.

Let $\Lambda_0 = \operatorname{diag}(\mu_1, \dots, \mu_{n-r})$, $\Lambda_1 = \operatorname{diag}(\gamma_1, \dots, \gamma_{n-r+1})$, and $\Lambda_2 = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. By the induction assumption, there are real symmetric matrices

$$C_1 = \begin{bmatrix} A_1 & B_1 \\ B_1^T & \Lambda_0 + B_1^T A_1^{-1} B_1 \end{bmatrix} = U_1^T \Lambda_1 U_1 \text{ and}$$

$$C_2 = \begin{bmatrix} A_2 & B_2 \\ B_2^T & \Lambda_1 + B_2^T A_2^{-1} B_2 \end{bmatrix} = U_2^T \Lambda_2 U_2,$$

where A_1 is nonsingular of order 1, where A_2 is nonsingular of order $r-1$, B_1 is $1 \times (n-r)$, B_2 is $(r-1) \times (n-r+1)$, U_1 is orthogonal of order $n-r+1$, and U_2 is orthogonal of order n. Note that

$$\begin{aligned} C_2 &= \begin{bmatrix} A_2 & B_2 \\ B_2^T & U_1 C_1 U_1^T + B_2^T A_2^{-1} B_2 \end{bmatrix} \\ &= \begin{bmatrix} I & O \\ O & U_1 \end{bmatrix} \begin{bmatrix} A_2 & B_2 U_1 \\ U_1^T B_2^T & C_1 + U_1^T B_2^T A_2^{-1} B_2 U_1 \end{bmatrix} \begin{bmatrix} I & O \\ O & U_1^T \end{bmatrix} \\ &= \begin{bmatrix} I & O \\ O & U_1 \end{bmatrix} \begin{bmatrix} A_2 & B_{21} & B_{22} \\ B_{21}^T & A_1 + B_{21}^T A_2^{-1} B_{21} & B_1 + B_{21}^T A_2^{-1} B_{22} \\ B_{22}^T & B_1^T + B_{22}^T A_2^{-1} B_{21} & \Lambda_0 + B_1^T A_1^{-1} B_1 + B_{22}^T A_2^{-1} B_{22} \end{bmatrix} \begin{bmatrix} I & O \\ O & U_1^T \end{bmatrix} \\ &= \begin{bmatrix} I & O \\ O & U_1 \end{bmatrix} H \begin{bmatrix} I & O \\ O & U_1^T \end{bmatrix}, \end{aligned}$$

where $B_2 U_1 = [\, B_{21} \quad B_{22} \,]$ is partitioned conformally with C_1.

Thus,

$$H = \begin{bmatrix} I & O \\ O & U_1^T \end{bmatrix} U_2^T \Lambda_2 U_2 \begin{bmatrix} I & O \\ O & U_1 \end{bmatrix}$$

is a real symmetric matrix with an $r \times r$ nonsingular principal submatrix

$$A = \begin{bmatrix} A_2 & B_{21} \\ B_{21}^T & A_1 + B_{21}^T A_2^{-1} B_{21} \end{bmatrix}.$$

$H/A = \Lambda_0$ since $H/A = (H/A_2)/(A/A_2)$ (see [3]). Therefore, the spectra of H and H/A are $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ and $\{\mu_1, \mu_2, \dots, \mu_{n-r}\}$, respectively, which completes the proof. □

Proof of Corollary. Let $\lambda_1, \lambda_2, \dots, \lambda_n$ and $\mu_1, \mu_2, \dots, \mu_{n-1}$ be $2n-r$ nonzero real numbers satisfying (a) or (b) and let $\alpha_i = \lambda_{n-i+1}$, $1 \le i \le n$, and $\beta_i = \mu_{n-r-i+1}$,

$1 \le i \le n-r$. That is, α and β are the sequences λ and μ, respectively, in reverse order. Then $\frac{1}{\alpha_1} \ge \frac{1}{\alpha_2} \ge \cdots \ge \frac{1}{\alpha_n}$ and $\frac{1}{\beta_1} \ge \frac{1}{\beta_2} \ge \cdots \ge \frac{1}{\beta_{n-r}}$.

Assume H is a hermitian semidefinite matrix with $r \times r$ nonsingular principal submatrix A and the spectra of H and H/A are $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ and $\{\mu_1, \mu_2, \ldots, \mu_{n-r}\}$, respectively. By the theorem, (i) holds and $\frac{1}{\alpha_i} \ge \frac{1}{\beta_i} \ge \frac{1}{\alpha_{i+r}}$, $1 \le i \le n-r$. The latter inequality is equivalent to $\alpha_{i+r} \ge \beta_i \ge \alpha_i$, $1 \le i \le n-r$, which in turn is equivalent to (ii).

Conversely, suppose (i) and (ii) hold. (ii) implies $\frac{1}{\lambda_i} \ge \frac{1}{\mu_i} \ge \frac{1}{\lambda_{n+r}}$, $1 \le i \le n-r$, or equivalently, $\frac{1}{\alpha_{n-r-i+1}} \ge \frac{1}{\beta_{n-r-i+1}} \ge \frac{1}{\alpha_{n-i+1}}$, $1 \le i \le n-r$. The latter inequalities are equivalent to $\frac{1}{\alpha_i} \ge \frac{1}{\beta_i} \ge \frac{1}{\alpha_{i+r}}$, $1 \le i \le n-r$.

Applying the theorem, there is a hermitian (in fact, real symmetric) semidefinite matrix H with an $r \times r$ nonsingular principal submatrix A such that the spectra of H and H/A are $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ and $\{\mu_1, \mu_2, \ldots, \mu_{n-r}\}$, respectively. □

Acknowledgments The authors would like to express their gratitude to Professor T. Y. Tam for introducing them to this problem. Also, the authors would like to thank Professor Roger Horn and the referees for their diligence in reviewing the manuscript; their useful suggestions led to significant improvement in the presentation.

REFERENCES

[1] F. L. BAUER, *A further generalization of Kantorovič inequality*, Numer. Math., 3 (1961), pp. 117–119.

[2] A. L. CAUCHY, *Sur l'équation à l'aide de laquelle on détemine les inégalités séculaires*, in Oeuvres Complètes de A. L. Cauchy, 2nd Ser., Vol IX, Gauthier–Villars, Paris, 1891, pp. 174–195.

[3] D. CRABTREE AND E. V. HAYNSWORTH, *An identity for the Schur complement of a matrix*, Proc. Amer. Math. Soc., 22 (1969), pp. 364–366.

[4] K. FAN AND G. PALL, *Imbedding conditions for hermitian and normal matrices*, Canada J. Math., 9 (1957), pp. 298–304.

[5] E. V. HAYNSWORTH, *Determination of the inertia of a partitioned hermitian matrix*, Linear Algebra Appl., 1 (1968), pp. 73–81.

[6] L. V. KANTOROVIC, *Funkcional'nyi analiz i prikladnaja matematika*, Uspehi Mat. Nauk., 3 (28) (1948), pp. 89–185.

[7] E. NICHOLS AND R. L. SMITH, *On Schur complements and n-metaharmonic functions*, Proc. Roy. Soc. Edinburgh, 119A (1991), pp. 233–240.

[8] R. L. SMITH, *Some interlacing properties of the Schur complement of a hermitian matrix*, Linear Algebra Appl., 177 (1992), pp. 137–144.

[9] T. Y. TAM, *private communication*, 1993.

[10] H. WIELANDT, *Inclusion theorems for eigenvalues*, Nat. Bur. Standards Appl. Math. Ser., 29 (1953), pp. 75–78.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 3, pp. 1024–1041, July 1995

021

ORDERING GIVENS ROTATIONS FOR SPARSE *QR* FACTORIZATION *

M. I. GILLESPIE† AND D. D. OLESKY†

Abstract. The QR factorization of a large, sparse matrix A is frequently computed using Givens rotations. The precise order in which the rotations are applied can affect the amount of storage required. We present an ordering for the Givens rotations that, when A has the Hall property, is optimal with regard to storage for Q (a so-called "tight" ordering) and that preserves sparsity by restricting fill to those locations in R that are necessarily nonzero. This ordering is of particular interest when A does not have the strong Hall property and is not permuted into block upper trapezoidal form.

We describe a bipartite graph model of sparse matrix structures and summarize the characterization of the structures of the factors Q and R. We define the product of structures of matrices, determine the product of the structures of a sequence of Givens rotations, and specify a tight ordering for these transformations.

Key words. QR factorization, sparse matrix, Givens rotation, tight ordering, bipartite graph

AMS subject classifications. 65F25, 65F50, 68R10

1. Introduction.

Let A be a real $m \times n$ matrix, where $m \geq n$. The QR factorization of A is defined either by $A = QR$, where Q is an $m \times n$ matrix with orthonormal columns and R is an $n \times n$ upper triangular matrix, or by $A = \hat{Q}\hat{R}$, where

$$\hat{Q} = [\,Q \quad Q_2\,]$$

is an orthogonal $m \times m$ matrix, Q is $m \times n$, Q_2 is $m \times (m-n)$, and

$$\hat{R} = [\,R^T \quad 0\,]^T$$

is an $m \times n$ upper trapezoidal matrix with R upper triangular. It is well known that such a factorization exists for every real matrix A and that if A has full rank, then the diagonal entries of R may be chosen to be positive, in which case the factorization $A = QR$ is unique.

There are three well-known methods for computing a QR factorization. One is Gram–Schmidt orthogonalization (usually implemented in the modified Gram–Schmidt form), the second uses Householder transformations, and the third uses Givens rotations. For dense matrices the Householder method is more efficient than the others, and hence is usually the method of choice. However, for sparse matrices the method of Givens may be more efficient than the other two, particularly if the rotations are applied in an order that preserves sparsity in the matrix. Finding such an ordering has been the subject of much research [2], [4], [5], [7], [10], [12], [13]. These papers are primarily concerned with how best to order the Givens rotations to reduce the total work required by the factorizations for matrices with the strong Hall property or in block form. In this paper we present an ordering for the Givens

* Received by the editors August 4, 1993; accepted (in revised form) by H. Elman July 7, 1994.

†Department of Computer Science, University of Victoria, Victoria, British Columbia, Canada V8W 3P6. The work of the second author was partially supported by Natural Sciences and Engineering Research Council of Canada grant A-8214 and the University of Victoria President's Committee on Faculty Research and Travel.

rotations that, when A has the Hall property, is optimal with regard to storage for Q (a *tight* ordering) and that preserves sparsity by restricting fill to those locations in R that are necessarily nonzero.

2. Notation. Since the computation of a QR factorization by Givens rotations is stable regardless of the order in which the rotations are applied, it is possible to choose an ordering solely on the basis of efficiency considerations. Thus an ordering can be determined without knowing the actual values of the nonzero entries of the matrix. That is, we are only concerned here with the zero/nonzero structure of the matrix. For an $m \times n$ matrix $A = (a_{ij})$, we define the *structure* by

$$\text{struct}(A) = \{(i,j) | a_{ij} \neq 0\}.$$

When convenient, we depict struct(A) by an $m \times n$ array $\mathbf{A} = (\mathbf{a}_{ij})$ such that $\mathbf{a}_{ij} = \star$ if $a_{ij} \neq 0$ and $\mathbf{a}_{ij} = 0$ if $a_{ij} = 0$. The *transpose* of struct(A) is defined by $\{(i,j) | (j,i) \in \text{struct}(A)\}$, and we denote it by struct$(A)^T$ or $\mathbf{A}^T$.

For an $m \times n$ matrix A and an $n \times p$ matrix B, we define the *product of the structures* of A and B as follows:

$$\begin{aligned}\text{struct}(A)\ \text{struct}(B) = \{&(i,j) | 1 \leq i \leq m; 1 \leq j \leq p; \text{ and there is some } k,\\ &\text{where } 1 \leq k \leq n, \text{ such that } (i,k) \in \text{struct}(A)\\ &\text{and } (k,j) \in \text{struct}(B)\}.\end{aligned}$$

We depict this by $\mathbf{C} = \mathbf{AB}$, which is an $m \times p$ array with $\mathbf{c}_{ij} = \star$ if $(i,j) \in$ struct(A) struct(B) and $\mathbf{c}_{ij} = 0$ otherwise. The array $\mathbf{AB}$ is sometimes called the *symbolic product* of A and B [11].

The following lemmas follow readily from the definitions, and we state them without proof.

LEMMA 1. *Multiplication of structures is associative.*

LEMMA 2. *If*

$$\mathbf{C} = \text{struct}(A)\text{struct}(B),$$

then

$$\mathbf{C}^T = \text{struct}(B)^T\text{struct}(A)^T.$$

LEMMA 3. *Let A be an $m \times n$ matrix and B be an $n \times p$ matrix. Then*

$$\text{struct}(AB) \subseteq \text{struct}(A)\ \text{struct}(B).$$

As in [11], given an $m \times n$ matrix A with rank n, we denote by $\mathcal{M}(A)$ the set of all $m \times n$ full rank matrices whose structures are identical to struct(A). Let

$$\mathcal{Q}(A) = \bigcup_{B \in \mathcal{M}(A)} \{\text{struct}(Q) | B = QR\} \tag{1}$$

and

$$\mathcal{R}(A) = \bigcup_{B \in \mathcal{M}(A)} \{\text{struct}(R) | B = QR\}. \tag{2}$$

(Note that if B has full column rank, then struct(Q) and struct(R) are uniquely determined.) The structures $\mathcal{Q}(A)$ and $\mathcal{R}(A)$ are the smallest possible that can be

guaranteed to accommodate all the nonzeros of the factors Q and R, respectively, of any full rank $m \times n$ matrix with structure identical to struct(A).

It is often convenient to model a matrix structure with a bipartite graph [1], [5], [9], [12]. A bipartite graph corresponding to $\mathbf{A} \equiv$ *struct*(A) is defined as $H(\mathbf{A}) = (R(\mathbf{A}), C(\mathbf{A}), E(\mathbf{A}))$, where $R(\mathbf{A})$ is the vertex set $\{r_i = i | 1 \leq i \leq m\}$ corresponding to the rows of the matrix $A, C(\mathbf{A})$ is the vertex set $\{c_j = j | 1 \leq j \leq n\}$ corresponding to the columns of the matrix A, and $E(\mathbf{A})$ is the set of edges $r_i c_j$ such that $r_i c_j \in E(\mathbf{A})$ if and only if $a_{ij} \neq 0$. Note that $r_i c_j$ and $c_j r_i$ denote the same edge. If $a_{ij} \neq 0$, then r_i and c_j are said to be *adjacent.*

Let $k \geq 1$ and $v_i \in R(\mathbf{A})$ or $C(\mathbf{A}), 0 \leq i \leq k$. A *path* in $H(\mathbf{A})$ is denoted by $P = v_0 v_1 v_2 \cdots v_k$, where $k \geq 1$ and it is understood that $v_{i-1} v_i, i = 1, \ldots k$, is an edge in $E(\mathbf{A})$. P is called a path from v_0 to v_k, or a (v_0, v_k)-*path.* If P_1 is a (u, v)-path and P_2 is a (v, w)-path, then $P_1 P_2$ represents a (u, w)-path in $H(\mathbf{A})$. If P_3 is a (y, z)-path and if the edge vy exists in $E(\mathbf{A})$, then $P_1 P_3$ is a (u, z)-path in $H(\mathbf{A})$.

A bipartite graph with $|C(\mathbf{A})| \leq |R(\mathbf{A})|$ is said to have the *Hall property* (with respect to $C(\mathbf{A})$) [1] if every subset S of $C(\mathbf{A})$ is adjacent to at least $|S|$ vertices in $R(\mathbf{A})$. It is said to have the *strong Hall property* [9] if S is adjacent to more than $|S|$ vertices in $R(\mathbf{A})$ for all subsets S of $C(\mathbf{A})$ such that

(i) $1 \leq |S| \leq |C(\mathbf{A})| - 1$ if $|C(\mathbf{A})| = |R(\mathbf{A})| > 1$, or

(ii) $1 \leq |S| \leq |C(\mathbf{A})|$ if $|C(\mathbf{A})| < |R(\mathbf{A})|$.

Analogously, a matrix A with m rows and n columns, $m \geq n$, has the Hall property if every set of k columns, $1 \leq k \leq n$, has nonzeros in at least k rows, and has the strong Hall property if

(i) $m = n > 1$ and every set of k columns, $1 \leq k < n$, has nonzeros in more than k rows, or

(ii) $m > n$ and every set of k columns, $1 \leq k \leq n$, has nonzeros in more than k rows.

An $m \times n$ matrix with $m \geq n$ must have the Hall property if it has full rank.

Hare et al. [9] have characterized the structures $\mathcal{Q}(A)$ and $\mathcal{R}(A)$ in terms of path conditions in the bipartite graph. We now summarize a number of concepts that they introduced and that are of importance here.

Let A be an $m \times n$ matrix with the Hall property, and let $H(\mathbf{A})$ be the bipartite graph associated with A.

1. A *Hall set* of $\mathbf{A}$ (or A) is a subset S of $C(\mathbf{A})$ such that the corresponding columns of A have nonzeros in exactly $|S|$ rows of A. It is clear that the union of two Hall sets of $\mathbf{A}$ is itself a Hall set.

2. S_j is the (possibly empty) Hall set of maximum cardinality in $\mathbf{A}_j$, where $\mathbf{A}_j$ is the structure of the submatrix formed by the first j columns of A. Define $S_0 = \varnothing$. If $j \leq k$, then $S_j \subseteq S_k$.

3. s_j is the subset of $R(\mathbf{A})$ of all vertices adjacent to vertices in S_j, and $s_0 = \varnothing$. Note that if $\mathbf{a}_{ii} = \star$ for $1 \leq i \leq n$, then $S_j = s_j$ for $1 \leq j \leq n - 1$ (since $c_i = i$ and $r_i = i$). Thus, in this case, $c_i \in S_j$ if and only if $r_i \in s_j$.

4. For $1 \leq j \leq n$, the auxiliary bipartite graph $B_j(\mathbf{A}) = (R_j(\mathbf{A}), C_j(\mathbf{A}), E_j(\mathbf{A}))$ is the bipartite graph of $\mathbf{A}_j$ with the sets s_{j-1} and S_{j-1} removed. That is, $C_j(\mathbf{A}) \equiv \{c_k = k | 1 \leq k \leq j, c_k \notin S_{j-1}\}, R_j(\mathbf{A}) \equiv \{r_i = i | 1 \leq i \leq m, r_i \notin s_{j-1}$, and there exists $c_v \in C_j(\mathbf{A})$ such that $\mathbf{a}_{iv} = \star\}$, and there is an edge between $r_i \in R_j(\mathbf{A})$ and $c_k \in C_j(\mathbf{A})$ if and only if $\mathbf{a}_{ik} = \star$.

The main result of [9] is the following theorem, which characterizes $\mathcal{Q}(A)$ in terms of path conditions in $B_j(\mathbf{A}), 1 \leq j \leq n$.

FIG. 1. *Bipartite graphs* $H(\mathbf{A})$ *and* $B_4(\mathbf{A})$ *for Example* 1.

THEOREM 4 [9, Thm. 4.7]. *Let* $1 \le i \le m$ *and* $1 \le j \le n$. *Then* $(i,j) \in \mathcal{Q}(A)$ *if and only if there exists a* (c_j, r_i)*-path in* $B_j(\mathbf{A})$.

We also use the following result.

THEOREM 5 [9, Thm. 5.1]. *Let* $\mathbf{A}, \mathcal{Q}(A)$ *and* $\mathcal{R}(A)$ *be as defined above. Then* $\mathcal{R}(A)$ *is identical to the upper trapezoidal part of* $(\mathcal{Q}(A))^T\mathbf{A}$ *(i.e., the elements* (i,j) *of* $(\mathcal{Q}(A))^T\mathbf{A}$ *with* $i \le j$).

We end this section with an example to illustrate these concepts.

Example 1. Let

$$\mathbf{A} = \begin{pmatrix} \star & 0 & 0 & 0 \\ \star & \star & 0 & \star \\ 0 & \star & \star & \star \\ 0 & 0 & 0 & \star \\ \star & 0 & 0 & 0 \\ 0 & 0 & 0 & \star \end{pmatrix}.$$

The Hall sets S_j of $\mathbf{A}$ are

$$S_0 = S_1 = S_2 = \varnothing, \quad S_3 = S_4 = \{c_2, c_3\},$$

and since all $\mathbf{a}_{ii} = \star$,

$$s_0 = s_1 = s_2 = \varnothing, \quad s_3 = s_4 = \{r_2, r_3\}.$$

The bipartite graphs $H(\mathbf{A})$ and $B_4(\mathbf{A})$ are shown in Fig. 1. The structures $\mathcal{Q}(A)$ and $\mathcal{R}(A)$ are given by

$$\mathcal{Q}(A) = \begin{pmatrix} \star & \star & \star & 0 \\ \star & \star & \star & 0 \\ 0 & \star & \star & 0 \\ 0 & 0 & 0 & \star \\ \star & \star & \star & 0 \\ 0 & 0 & 0 & \star \end{pmatrix} \quad \text{and} \quad \mathcal{R}(A) = \begin{pmatrix} \star & \star & 0 & \star \\ 0 & \star & \star & \star \\ 0 & 0 & \star & \star \\ 0 & 0 & 0 & \star \end{pmatrix}.$$

3. Tight orderings and products of rotation matrices. Givens rotations G_{ij} can be used to compute the QR factorization by applying them sequentially to eliminate nonzeros from the lower trapezoidal part of A:

$$R = G_{i_r j_r} \cdots G_{i_2 j_2} G_{i_1 j_1} A$$

and

$$Q = G_{i_1 j_1}^T G_{i_2 j_2}^T \cdots G_{i_r j_r}^T.$$

If $i_k > j_k$ and the (i_k, j_k) entry of $G_{i_{k-1} j_{k-1}} \cdots G_{i_1 j_1} A$ is nonzero, the nonzero entries of a Givens rotation $G_{i_k j_k}$ are chosen so that the (i_k, j_k) entry of the product $G_{i_k j_k} \cdots G_{i_1 j_1} A$ is 0 (see, e.g., [8]). The entry $a_{j_k j_k}$ is called the pivot, and we call this *diagonal pivoting* (in contrast to *variable pivoting*); see, e.g., [2], [10]. If $a_{j_k j_k} \neq 0$, then the structure of an $m \times m$ rotation $G_{i_k j_k}$ is exactly

$$\mathbf{G}_{i_k j_k} = \{(i_k, j_k), (j_k, i_k)\} \bigcup \{(t,t) | t = 1, 2, \ldots, m\}. \tag{3}$$

Note that there may be more than one order in which the rotations can be applied to compute the (unique) QR factorization.

A corresponding *symbolic QR factorization* can be defined. Starting with $\bar{\mathbf{R}}_0 \equiv \text{struct}(A)$, define the sequence of structures

$$\bar{\mathbf{R}}_k = (\mathbf{G}_{i_k j_k} \bar{\mathbf{R}}_{k-1}) \setminus \{(i_k, j_k)\}, \quad k = 1, 2, \ldots, r,$$

where $(i_k, j_k) \in \bar{\mathbf{R}}_{k-1}$ with $i_k > j_k$, so that $\bar{\mathbf{R}}_r$ is an upper trapezoidal structure. Defining

$$\bar{\mathbf{Q}} = \left(\prod_{k=1}^{r} \mathbf{G}_{i_k j_k}^T \right) \setminus \{(i,j) | j > n\}, \tag{4}$$

Lemma 3 implies that $\text{struct}(Q) \subseteq \bar{\mathbf{Q}}$. Also, by definition, $\text{struct}(Q) \subseteq \mathcal{Q}(A)$. Furthermore, $\mathcal{Q}(A) \subseteq \bar{\mathbf{Q}}$ (since $\text{struct}(Q) \subseteq \bar{\mathbf{Q}}$ for every Q such that $B = QR$ and $B \in \mathcal{M}(A)$). However, it is not necessarily the case that $\bar{\mathbf{Q}} \subseteq \mathcal{Q}(A)$, as the following example illustrates.

Example 2. Let

$$\mathbf{A} = \begin{pmatrix} \star & 0 & 0 & 0 \\ 0 & \star & \star & 0 \\ \star & 0 & \star & 0 \\ \star & 0 & 0 & \star \end{pmatrix}. \tag{5}$$

There are two ways in which any matrix A with this structure can be reduced to triangular form by Givens rotations (using diagonal pivoting). One way is to apply the rotation G_{31} before G_{41}, in which case the $(4,3)$ entry will become nonzero (that is, there is fill at the $(4,3)$ position) and thus G_{43} must be applied as well. The resultant factorization is

$$A = G_{31}^T G_{41}^T G_{43}^T R, \tag{6}$$

with

$$\mathbf{G}_{31}^T \mathbf{G}_{41}^T \mathbf{G}_{43}^T = \begin{pmatrix} \star & 0 & \star & \star \\ 0 & \star & 0 & 0 \\ \star & 0 & \star & \star \\ \star & 0 & \star & \star \end{pmatrix} \equiv \bar{\mathbf{Q}}_1.$$

On the other hand, if G_{41} is applied before G_{31}, then there is no fill in positions of A below the main diagonal and the resultant factorization is

$$A = \tilde{G}_{41}^T \tilde{G}_{31}^T R, \tag{7}$$

so that

$$\mathbf{G}_{41}^T\mathbf{G}_{31}^T = \begin{pmatrix} \star & 0 & \star & \star \\ 0 & \star & 0 & 0 \\ \star & 0 & \star & 0 \\ \star & 0 & \star & \star \end{pmatrix} \equiv \bar{\mathbf{Q}}_2.$$

Both (6) and (7) determine the unique QR factorization of any full rank matrix A having the structure in (5) (if the diagonal entries of R are normalized); that is,

$$Q = G_{31}^T G_{41}^T G_{43}^T = \tilde{G}_{41}^T \tilde{G}_{31}^T.$$

Thus, $\mathcal{Q}(A) \subseteq \bar{\mathbf{Q}}_1$ and $\mathcal{Q}(A) \subseteq \bar{\mathbf{Q}}_2$; however, $\bar{\mathbf{Q}}_1 \not\subseteq \mathcal{Q}(A)$ since $(3,4) \notin \bar{\mathbf{Q}}_2$. (Note: By Theorem 4, $\mathcal{Q}(A) = \bar{\mathbf{Q}}_2$.)

We define an ordering of the Givens rotations for which $\mathcal{Q}(A) = \bar{\mathbf{Q}}$ to be a *tight ordering*. Algorithm 1 in §4 specifies a tight ordering for any full rank matrix A. In the terminology of [1], Algorithm 1 is "correct," or, in the terminology of [11], $\bar{\mathbf{Q}}$ is "tight." For applications in which the factor Q is explicitly required, a tight ordering minimizes the storage required for the computation.

It follows from Theorem 5 that for a tight ordering, $\mathcal{R}(A)$ is equal to the upper trapezoidal part of $\prod_{t=r}^{1} \mathbf{G}_{i_t j_t}\mathbf{A}$. Thus Algorithm 1 determines a tight structure for R when A is any full rank matrix.

In the next theorem, we determine the product of a sequence of structures (3). The proofs of Lemmas 6 and 7 follow directly from the definition of the product of structures.

LEMMA 6. *Let i, j, k, and l be distinct and let $\mathbf{H} = \mathbf{G}_{ij}\mathbf{G}_{kl}$. Then*

$$\mathbf{H} = \{(i,j),(j,i),(k,l),(l,k)\} \bigcup \{(t,t) | 1 \le t \le m\}.$$

LEMMA 7. *Let i, j, and k be distinct, and let $\mathbf{H}$ be any one of the products $\mathbf{G}_{ij}\mathbf{G}_{ik}, \mathbf{G}_{ji}\mathbf{G}_{ki}, \mathbf{G}_{ji}\mathbf{G}_{ik}$, or $\mathbf{G}_{ij}\mathbf{G}_{ki}$. Then*

$$\mathbf{H} = \{(i,j),(j,i),(i,k),(k,i),(j,k)\} \bigcup \{(t,t) | 1 \le t \le m\}.$$

We note that $(k,j) \notin \mathbf{H}$ in Lemma 7.

THEOREM 8. *Let $\mathbf{G} = \prod_{t=1}^{r} \mathbf{G}_{i_t j_t}$, where $\mathbf{G}_{i_t j_t}$ is defined by* (3). *Then $(x,y) \in \mathbf{G}$ if and only if*

(a) *$x = y$, or*

(b) *there exists s such that ($i_s = x$ and $j_s = y$) or ($i_s = y$ and $j_s = x$), or*

(c) *there is an ordered subsequence of the structures*

$$\mathbf{G}_{i_{s_1} j_{s_1}}, \mathbf{G}_{i_{s_2} j_{s_2}}, \ldots, \mathbf{G}_{i_{s_k} j_{s_k}}$$

with $2 \le k \le r$ and $s_1 < s_2 < s_3 < \cdots < s_k$ such that ($x = i_{s_1}$ or $x = j_{s_1}$), and ($y = i_{s_k}$ or $y = j_{s_k}$), and (at least) one of the equalities ($i_{s_t} = i_{s_{t+1}}$), ($i_{s_t} = j_{s_{t+1}}$), ($j_{s_t} = i_{s_{t+1}}$), ($j_{s_t} = j_{s_{t+1}}$) holds for each $t = 1, 2, \ldots, k-1$.

Proof. This follows from Lemmas 6 and 7 and induction on k (for the sufficiency of (a), (b), and (c)) and induction on r (for the necessity). □

Theorem 8, which we use to prove the correctness of Algorithm 1, characterizes the product (4) of structures of a sequence of Givens rotations (since $\mathbf{G}_{i_k j_k}^T = \mathbf{G}_{i_k j_k}$). Thus, if the ordering of the Givens rotations applied in a QR factorization is a tight ordering, then the three conditions of Theorem 8 determine the elements of $\mathcal{Q}(A)$.

4. An algorithm for symbolic factorization. In this section we present an algorithm for determining the structures of the upper triangular factor R and the factor Q of an $m \times n$ matrix A that has full rank.

Our strategy is to compute a QR factorization of A by applying Givens rotations in a specified order that depends upon struct(A). The ordering is restricted to the case where A has a zero-free diagonal and diagonal pivoting is used. The rows of any matrix with the Hall property can be reordered so that it has a zero-free diagonal (see [3], [6]). This reordering does not affect the structure of R and affects the structure of Q only by reordering its rows. Nonzeros below the diagonal are eliminated column by column; and, within each column, nonzeros in rows that are not in s_{n-1} are eliminated first, followed by those that are not in s_{n-2}, then those that are not in s_{n-3}, and so on. The order of elimination of nonzeros in rows not in some fixed s_k is arbitrary. This ordering of the Givens rotations is the main innovation of this paper.

From now on,

$$\bar{R} \equiv \left(\left(\prod_{k=r}^{1} \text{struct}(G_{i_k j_k}) \right) \text{struct}(A) \right) \setminus \{(i,j) | i > j\} \tag{8}$$

and

$$\bar{\mathbf{Q}} \equiv \left(\prod_{k=1}^{r} \text{struct}(G_{i_k j_k}^T) \right) \setminus \{(i,j) | j > n\}, \tag{9}$$

where the ordering for the Givens rotations is determined by Algorithm 1. In §5, we prove that $\bar{\mathbf{Q}} = \mathcal{Q}(A)$, that is, that the ordering of the Givens rotations in Algorithm 1 is a tight ordering.

ALGORITHM 1. *Determine* $\bar{\mathbf{R}}$ *and* $\bar{\mathbf{Q}}$.

Input. $m, n, \mathbf{A}$, the sets $s_1, s_2, \ldots, s_{n-1}$ (as defined in §2).

($m \geq n$ and $\mathbf{A}$ is the structure of an $m \times n$ matrix with a zero-free diagonal.)

Output. $\bar{\mathbf{Q}}, \bar{\mathbf{R}}$

Step 1. Initialize $\bar{\mathbf{Q}} = \{(i,i) | 1 \leq i \leq m\}$

Step 2. Initialize $\hat{\mathbf{A}} = \mathbf{A}$

Step 3. Iterate for $j = 1, 2, 3, \ldots, n-1$. (Elimination on c_j)

Initialize: rows $\leftarrow \{r_i | i > j$ and $(i,j) \in \hat{\mathbf{A}}\}$

For $k = n-1, n-2, \ldots, j$

for each $i \in$ rows$\backslash s_k$

set $\bar{\mathbf{Q}} \leftarrow \bar{\mathbf{Q}}$ struct(G_{ij}) (G_{ij} is an $m \times m$ Givens rotation that produces a zero entry at the (i,j) position.)

set $\hat{\mathbf{A}} \leftarrow$ struct$(G_{ij})\hat{\mathbf{A}}$

set rows $\leftarrow$ rows$\backslash\{r_i\}$

Step 4. For $j = n$ (Elimination on c_n)

Initialize: rows $\leftarrow \{r_i | i > j$ and $(i,j) \in \hat{\mathbf{A}}\}$

for each $i \in$ rows

set $\bar{\mathbf{Q}} \leftarrow \bar{\mathbf{Q}}$ struct(G_{ij})

set $\hat{\mathbf{A}} \leftarrow$ struct$(G_{ij})\hat{\mathbf{A}}$

set rows $\leftarrow$ rows$\backslash\{r_i\}$

Step 5. Set $\bar{\mathbf{R}} \leftarrow \hat{\mathbf{A}} \backslash \{(i,j) | i > j\}$
Set $\bar{\mathbf{Q}} \leftarrow \bar{\mathbf{Q}} \backslash \{(i,j) | j > n\}$
Step 6. Output $\bar{\mathbf{R}}$ and $\bar{\mathbf{Q}}$

Some notation follows that will be used throughout the rest of this section. For $1 \leq j \leq n$, let $\hat{\mathbf{A}}_j$ denote the structure $\hat{\mathbf{A}}$ after elimination on c_j in Algorithm 1 (that is, after the jth iteration of Step 3 or Step 4). By "fill" at the (i,k) position we mean that $(i,k) \notin \mathbf{A}$ and for some $j \geq 1, (i,k) \in \hat{\mathbf{A}}_j$. At any given point in Algorithm 1, we say that $\hat{\mathbf{a}}_{ik} = \star$ if $(i,k) \in \hat{\mathbf{A}}$ and $\hat{\mathbf{a}}_{ik} = 0$ if $(i,k) \notin \hat{\mathbf{A}}$. (Note that $\hat{\mathbf{A}}$ is a dynamic data structure, so that at different points in Algorithm 1, any $\hat{\mathbf{a}}_{ik}$ may be 0 or $\star$.) Similarly, we say that $\hat{\mathbf{a}}_{ik}^{(j)} = \star$ if $(i,k) \in \hat{\mathbf{A}}_j$ and $\hat{\mathbf{a}}_{ik}^{(j)} = 0$ if $(i,k) \notin \hat{\mathbf{A}}_j$.

We illustrate the application of this algorithm to the matrix in Example 1. In the array representation of the evolving structure of $\hat{\mathbf{A}}$, we use the symbol f to denote fill and the symbol $\odot$ to denote an eliminated entry. At the first iteration of Step 3, G_{51} is applied first since $r_5 \notin s_3$ but $r_2 \in s_3$. Thus

$$\hat{\mathbf{A}} \rightarrow \begin{pmatrix} \star & 0 & 0 & 0 \\ \star & \star & 0 & \star \\ 0 & \star & \star & \star \\ 0 & 0 & 0 & \star \\ \odot & 0 & 0 & 0 \\ 0 & 0 & 0 & \star \end{pmatrix} \rightarrow \begin{pmatrix} \star & f & 0 & f \\ \odot & \star & 0 & \star \\ 0 & \star & \star & \star \\ 0 & 0 & 0 & \star \\ \odot & 0 & 0 & 0 \\ 0 & 0 & 0 & \star \end{pmatrix}.$$

At the same time,

$$\bar{\mathbf{Q}} \rightarrow \begin{pmatrix} \star & 0 & 0 & 0 & \star & 0 \\ 0 & \star & 0 & 0 & 0 & 0 \\ 0 & 0 & \star & 0 & 0 & 0 \\ 0 & 0 & 0 & \star & 0 & 0 \\ \star & 0 & 0 & 0 & \star & 0 \\ 0 & 0 & 0 & 0 & 0 & \star \end{pmatrix} \rightarrow \begin{pmatrix} \star & \star & 0 & 0 & \star & 0 \\ \star & \star & 0 & 0 & 0 & 0 \\ 0 & 0 & \star & 0 & 0 & 0 \\ 0 & 0 & 0 & \star & 0 & 0 \\ \star & \star & 0 & 0 & \star & 0 \\ 0 & 0 & 0 & 0 & 0 & \star \end{pmatrix}.$$

In the second iteration of Step 3, only G_{32} is applied so that

$$\hat{\mathbf{A}} \rightarrow \begin{pmatrix} \star & f & 0 & f \\ \odot & \star & f & \star \\ 0 & \odot & \star & \star \\ 0 & 0 & 0 & \star \\ \odot & 0 & 0 & 0 \\ 0 & 0 & 0 & \star \end{pmatrix} \quad \text{and} \quad \bar{\mathbf{Q}} \rightarrow \begin{pmatrix} \star & \star & \star & 0 & \star & 0 \\ \star & \star & \star & 0 & 0 & 0 \\ 0 & \star & \star & 0 & 0 & 0 \\ 0 & 0 & 0 & \star & 0 & 0 \\ \star & \star & \star & 0 & \star & 0 \\ 0 & 0 & 0 & 0 & 0 & \star \end{pmatrix}.$$

No computation occurs during the third iteration of Step 3, and only G_{64} is applied at Step 4, so that

$$\hat{\mathbf{A}} \rightarrow \begin{pmatrix} \star & f & 0 & f \\ \odot & \star & f & \star \\ 0 & \odot & \star & \star \\ 0 & 0 & 0 & \star \\ \odot & 0 & 0 & 0 \\ 0 & 0 & 0 & \odot \end{pmatrix} \quad \text{and} \quad \bar{\mathbf{Q}} \rightarrow \begin{pmatrix} \star & \star & \star & 0 & \star & 0 \\ \star & \star & \star & 0 & 0 & 0 \\ 0 & \star & \star & 0 & 0 & 0 \\ 0 & 0 & 0 & \star & 0 & \star \\ \star & \star & \star & 0 & \star & 0 \\ 0 & 0 & 0 & \star & 0 & \star \end{pmatrix}.$$

Finally, at Step 5, the entries of $\hat{\mathbf{A}}$ denoted by $\odot$ and the last two columns of $\bar{\mathbf{Q}}$ are deleted, and the resulting structures returned are

$$\bar{\mathbf{R}} = \begin{pmatrix} \star & \star & 0 & \star \\ 0 & \star & \star & \star \\ 0 & 0 & \star & \star \\ 0 & 0 & 0 & \star \end{pmatrix} \quad \text{and} \quad \bar{\mathbf{Q}} = \begin{pmatrix} \star & \star & \star & 0 \\ \star & \star & \star & 0 \\ 0 & \star & \star & 0 \\ 0 & 0 & 0 & \star \\ \star & \star & \star & 0 \\ 0 & 0 & 0 & \star \end{pmatrix}.$$

5. Proof of correctness for Algorithm 1. In Theorems 20 and 21, respectively, we prove that the structures determined by Algorithm 1 are correct, i.e., $\bar{\mathbf{Q}} = \mathcal{Q}(A)$ and $\bar{\mathbf{R}} = \mathcal{R}(A)$. An overview of how this is accomplished follows.

Recall that elements of $\bar{\mathbf{Q}}$ (see (9)) are characterized by Theorem 8. For each of the three conditions in Theorem 8 that identifies an element (i,j) of $\bar{\mathbf{Q}}$, we show that $(i,j) \in \mathcal{Q}(A)$. This is done, respectively, in Lemma 9 (§5.1), Corollary 10.2 and Lemma 13 (§5.2), and Lemma 19 (§5.3), implying that $\bar{\mathbf{Q}} \subseteq \mathcal{Q}(\mathbf{A})$. However, it follows from Lemma 3 that $\mathcal{Q}(A) \subseteq \bar{\mathbf{Q}}$, and thus $\bar{\mathbf{Q}} = \mathcal{Q}(\mathbf{A})$. Having proven this, it follows from Theorem 5 that $\bar{\mathbf{R}} = \mathcal{R}(A)$.

5.1. Condition (a) of Theorem 8. This condition implies that $(i,i) \in \bar{\mathbf{Q}}$ for $i = 1, 2, \ldots, n$. The following result shows that these "diagonal" elements are also in $\mathcal{Q}(A)$.

LEMMA 9. *For $1 \leq i \leq n$, there is a (c_i, r_i)-path in $B_i(\mathbf{A})$ and hence $(i,i) \in \mathcal{Q}(A)$.*

Proof. This is clear since $\mathbf{A}$ has a zero-free diagonal and $r_i \notin s_{i-1}$. □

5.2. Condition (b) of Theorem 8. The next step in proving that $\bar{\mathbf{Q}} \subseteq \mathcal{Q}(A)$ is to show that if G_{ij} is applied in Algorithm 1, then $(i,j) \in \mathcal{Q}(A)$ and if $i \leq n$, then $(j,i) \in \mathcal{Q}(A)$. By condition (b) of Theorem 8, these are elements of $\bar{\mathbf{Q}}$.

The structure $\mathcal{Q}(\mathrm{A})$ is characterized in [9] by certain path conditions. In the following lemma, a path condition is obtained that corresponds to the occurrence of fill at certain positions (i,j) in $\hat{\mathbf{A}}$. To illustrate our notation, suppose

$$\mathbf{A} = \begin{pmatrix} \star & 0 & 0 & 0 & 0 & 0 \\ 0 & \star & 0 & 0 & 0 & 0 \\ \star & \star & \star & 0 & \star & 0 \\ 0 & 0 & 0 & \star & 0 & 0 \\ 0 & 0 & \star & 0 & \star & 0 \\ 0 & \star & 0 & \star & 0 & \star \end{pmatrix}.$$

Then on applying Algorithm 1, the nonzeros in positions $(3,1)$ and $(6,2)$ are eliminated, giving

$$\hat{\mathbf{A}} = \begin{pmatrix} \star & f & f & 0 & f & 0 \\ 0 & \star & 0 & f & 0 & f \\ \odot & \star & \star & 0 & \star & 0 \\ 0 & 0 & 0 & \star & 0 & 0 \\ 0 & 0 & \star & 0 & \star & 0 \\ 0 & \odot & 0 & \star & 0 & \star \end{pmatrix}.$$

Next the nonzero in the $(3,2)$ position is eliminated. In the notation of the following lemma, a nonzero in column $v = 2$ is eliminated creating fill at the (i,j) position (where $i = 3$ and $j = 4$), and $r_i \notin s_{k-1}$ with $k = 5$. The lemma shows that there is a (c_4, r_3)-path in $B_5(\mathbf{A})$.

LEMMA 10. *Let* $1 \le v \le n-1, v < j \le n, v < i \le m$, *and* $j \le k \le n$. *If* $r_i \notin s_{k-1}$ *and fill occurs at the* (i,j) *position during elimination of the nonzero entries of column* v *in Algorithm* 1, *then there is a* (c_j, r_i)*-path in* $B_k(\mathbf{A})$.

Proof. We use induction on v.

Base case. Consider elimination of the nonzero entries (below the diagonal) in column $v = 1$.

If fill occurs at the (i,j) position, then $\mathbf{a}_{i1} = \star$ and either

1. $\mathbf{a}_{1j} = \star$, or

2. there exists w such that G_{w1} is applied before $G_{i1}, \mathbf{a}_{w1} = \star$ and $\mathbf{a}_{wj} = \star$.

Now $\mathbf{a}_{i1} = \star$ implies that $c_1 \notin S_{k-1}$ and $r_1 \notin s_{k-1}$.

1. If $\mathbf{a}_{1j} = \star$, then $c_j \notin S_{k-1}$ and $r_j \notin s_{k-1}$. Therefore $B_k(\mathbf{A})$ contains the path $r_i c_1 r_1 c_j$ since the $(i,1), (1,1)$, and $(1,j)$ entries are all nonzero.

2. If there exists w such that G_{w1} is applied before $G_{i1}, \mathbf{a}_{w1} = \star$ and $\mathbf{a}_{wj} = \star$, then $r_w \notin s_{k-1}$ (by the order of elimination of entries in Algorithm 1), which implies that $c_j \notin S_{k-1}$. Thus we have the path $r_i c_1 r_w c_j$ in $B_k(\mathbf{A})$.

In either case there is a path in $B_k(\mathbf{A})$ from c_j to r_i.

Induction hypothesis. Suppose that during elimination on column p, where $1 \le p \le v$, if fill occurs at the (i,j) position, where $p < i \le m, p < j \le n, j \le k \le n$, and $r_i \notin s_{k-1}$, then there is a (c_j, r_i)-path in $B_k(\mathbf{A})$.

Induction step. We consider elimination on column $v+1$ and suppose that fill occurs at the (i,j) position, where $v+1 < i \le m, v+1 < j \le n, j \le k \le n$, and $r_i \notin s_{k-1}$. For this to happen, we must have $\hat{\mathbf{a}}^{(v)}_{i,v+1} = \star$, which means either $\mathbf{a}_{i,v+1} = \star$ or, by the induction hypothesis, there is a path in $B_k(\mathbf{A})$ from c_{v+1} to r_i. In either case, $c_{v+1} \notin S_{k-1}$, which implies that $r_{v+1} \notin s_{k-1}$. We must also have one of the following two cases.

1. $\hat{\mathbf{a}}^{(v)}_{v+1,j} = \star$, in which case there is a (c_j, r_{v+1})-path in $B_k(\mathbf{A})$. Call this path P_1 and denote by P_2 the (c_{v+1}, r_i)-path in $B_k(\mathbf{A})$. The zero-free diagonal of A implies that the $r_{v+1}c_{v+1}$ edge also exists, so that $P_1 P_2$ is a (c_j, r_i)-path in $B_k(\mathbf{A})$.

2. There exists w such that $G_{w,v+1}$ is applied before $G_{i,v+1}, \hat{\mathbf{a}}^{(v)}_{w,v+1} = \star$ and $\hat{\mathbf{a}}^{(v)}_{wj} = \star$. Since $G_{w,v+1}$ is applied before $G_{i,v+1}, r_w$ cannot be in s_{k-1} (by the order of elimination of entries in Algorithm 1). Thus, $\hat{\mathbf{a}}^{(v)}_{w,v+1} = \star$ and the induction hypothesis together imply that there is a (c_{v+1}, r_w)-path in $B_k(\mathbf{A})$. Similarly, $\hat{\mathbf{a}}^{(v)}_{wj} = \star$ implies that $B_k(\mathbf{A})$ contains a (c_j, r_w)-path. Since $B_k(\mathbf{A})$ also contains the (c_{v+1}, r_i)-path, together these paths give a (c_j, r_i)-path in $B_k(\mathbf{A})$. □

COROLLARY 10.1. *Let* $1 \le j \le n$ *and* $j < i \le m$, *and suppose that* G_{ij} *is applied in Algorithm* 1. *If there exists* k, *where* $j \le k \le n$, *such that* $r_i \notin s_{k-1}$, *then the* (c_j, r_i)*-path exists in* $B_k(\mathbf{A})$.

Proof. This is clear when $\mathbf{a}_{ij} = \star$ since the edge $r_i c_j$ exists in $B_k(\mathbf{A})$, and follows from Lemma 10 otherwise. □

COROLLARY 10.2. *If* G_{ij} *is applied in Algorithm* 1, *then* $(i,j) \in \mathcal{Q}(A)$.

Proof. Apply Corollary 10.1 with $k = j$ to obtain a (c_j, r_i)-path in $B_j(\mathbf{A})$. It follows from Theorem 4 that $(i,j) \in \mathcal{Q}(A)$. □

To prove that $(j,i) \in \mathcal{Q}(A)$ also (when $i \le n$), we use the following two lemmas.

LEMMA 11. *Let* j *be a fixed index with* $2 \le j \le n-1$ *and let* $v < j$, *and suppose* $c_j \in S_k$ *for some* k *such that* $j \le k \le n-1$. *Then during elimination on* c_v *using Algorithm* 1, *fill in* c_j *below* r_v *is restricted to rows in* s_k.

Proof. We use induction on v.

Base case. Considering elimination on column c_1, we have two cases.

Case 1. $r_1 \in s_k$. Then $c_1 \in S_k$ and hence only rows in s_k are involved in the elimination at this stage so that fill in c_j is necessarily restricted to rows in s_k.

Case 2. $r_1 \notin s_k$. Then c_j must have a zero in row 1, since $c_j \in S_k$, and in all other rows that are not in s_k. Since nonzeros in all rows not in s_k are eliminated *before* nonzeros in rows in s_k, it follows that $\hat{\mathbf{a}}_{1j} = 0$ until after elimination on rows in s_k begins. At that point, only rows in s_k remain to be processed, so that fill in c_j below r_1 is restricted to these rows.

Induction hypothesis. Suppose $v < j - 1$ and that during elimination of nonzeros in column $w, 1 \leq w \leq v$, fill below r_w in c_j is restricted to rows in s_k.

Induction step. We now consider elimination of nonzeros from column $v + 1$.

Case 1. $r_{v+1} \in s_k$. In this case $c_{v+1} \in S_k$ and by the induction hypothesis, c_{v+1} has nonzeros below the diagonal only in rows in s_k. Therefore, fill during this stage of the elimination is restricted to rows in s_k.

Case 2. $r_{v+1} \notin s_k$. In this case $\mathbf{a}_{v+1,j} = 0$ and by the induction hypothesis, $\hat{\mathbf{a}}_{v+1,j} = 0$ before elimination on c_{v+1} begins. Furthermore, all nonzeros in c_j below r_v are restricted to rows in s_k before processing of column c_{v+1} begins. The order of elimination guarantees that all the zeros in c_j below r_v are preserved until elimination of rows in s_k begins. At that point, $\hat{\mathbf{a}}_{v+1,j}$ may become $\star$, but fill in c_j below r_{v+1} is restricted to rows in s_k. Hence, during elimination on c_{v+1} fill in c_j below r_{v+1} is restricted to rows in s_k. □

LEMMA 12. *Let* $1 \leq j \leq n$ *and* $i > j$. *If* G_{ij} *is applied in Algorithm* 1, *then* $r_j \notin s_l$, *where* $l = \min\{i - 1, n - 1\}$.

Proof. Suppose $r_j \in s_l$. By Lemma 11, fill in c_j below r_j is restricted to rows in s_l. Therefore, since $r_i \notin s_l, \hat{\mathbf{a}}_{ij}^{(j-1)} = 0$. But then G_{ij} would not be applied in Algorithm 1. Therefore, if G_{ij} is applied, then $r_j \notin s_l$. □

We now prove the second of the two main results of this subsection.

LEMMA 13. *Let* $1 \leq j < i \leq n$. *If* G_{ij} *is applied in Algorithm* 1, *then there is a* (c_i, r_j)*-path in* $B_j(\mathbf{A})$ *and thus* $(j, i) \in \mathcal{Q}(A)$.

Proof. Since G_{ij} is applied, either $\mathbf{a}_{ij} = \star$ or fill occurs at the (i, j) position during elimination on columns $1, \ldots, j - 1$ (i.e., $\hat{\mathbf{a}}_{ij}^{(j-1)} = \star$). If $\mathbf{a}_{ij} = \star$, then $r_j \notin s_{i-1}$ by Lemma 12, so that the path $c_i r_i c_j r_j$ exists in $B_i(\mathbf{A})$. In the other case, there exists a (c_j, r_i)-path in $B_i(\mathbf{A})$ by Lemma 10, so this path and the $r_i c_i, r_j c_j$ edges imply the existence of a (c_i, r_j)-path in $B_i(\mathbf{A})$. Hence, in both cases, $(j, i) \in \mathcal{Q}(A)$ by Theorem 4. □

Corollary 10.2 and Lemma 13 together prove that if G_{ij} is applied in Algorithm 1, then $(i, j) \in \mathcal{Q}(A)$ and if $i \leq n$ then $(j, i) \in \mathcal{Q}(A)$. Thus entries of $\bar{\mathbf{Q}}$ that occur because of the second condition in Theorem 8 are also in $\mathcal{Q}(A)$.

5.3. Condition (c) of Theorem 8. Before considering entries in $\bar{\mathbf{Q}}$ that exist because of this condition, we first examine the orderings permitted by Algorithm 1.

DEFINITION 1. *A chain of length* $k \geq 2$ *that links* $G_{i_a j_a}$ *and* $G_{i_b j_b}$ *is an ordered subsequence*

$$G_{i_{s_1} j_{s_1}}, G_{i_{s_2} j_{s_2}}, \ldots, G_{i_{s_k} j_{s_k}} \tag{10}$$

of the rotations $G_{i_1 j_1}, \ldots, G_{i_r j_r}$ *with* $s_1 < s_2 < s_3 < \cdots < s_k$ *such that* $s_1 = a, s_k = b$, *and one of*

(i) $i_{s_t} = i_{s_{t+1}}$,
(ii) $i_{s_t} = j_{s_{t+1}}$,
(iii) $j_{s_t} = j_{s_{t+1}}$,
(iv) $j_{s_t} = i_{s_{t+1}}$

holds for each $t, 1 \le t \le k-1$.

We first show that (iv) is not possible in the ordering for the rotations imposed by Algorithm 1.

LEMMA 14. *Let* $1 \le i \le m, 1 \le j \le n$, *and* $1 \le k \le n$. *If* G_{ij} *and* G_{jk} *are both applied in Algorithm* 1, *then* G_{jk} *must be applied before* G_{ij}.

Proof. If G_{ij} is applied, then $i > j$. If G_{jk} is applied, then $j > k$, so that $k < j < i$. Since the columns are processed in increasing order, G_{jk} is applied before G_{ij}. □

Suppose that (i) holds for a fixed t in (10); that is, (10) contains an adjacent pair of rotations of the form

$$G_{i_{s_t} j_{s_t}}, G_{i_{s_t} j_{s_{t+1}}}.$$

If (i) also holds for $t+1$, then (10) contains

$$G_{i_{s_t} j_{s_t}}, G_{i_{s_t} j_{s_{t+1}}}, G_{i_{s_t} j_{s_{t+2}}}$$

and we can eliminate $G_{i_{s_t} j_{s_{t+1}}}$ from the chain (10) to get a chain of length $k-1$ linking $G_{i_a j_a}$ and $G_{i_b j_b}$. Similarly, if (ii) holds for $t+1$, then (10) contains

$$G_{i_{s_t} j_{s_t}}, G_{i_{s_t} j_{s_{t+1}}}, G_{i_{s_{t+2}} i_{s_t}}.$$

Once again, the rotation $G_{i_{s_t} j_{s_{t+1}}}$ can be eliminated to give a chain of length $k-1$ that still links $G_{i_a j_a}$ and $G_{i_b j_b}$.

If no rotations can be eliminated from a chain in such a manner, we say that the chain has *minimal* length. (We note that any chain of length 2 is necessarily minimal.) Let us suppose that

$$G_{i_1 j_1}, G_{i_2 j_2}, \ldots, G_{i_k j_k}$$

is a chain of minimal length that links $G_{i_1 j_1}$ and $G_{i_k j_k}$, and that the length is ≥ 3. Consideration of all cases as above gives the following result.

LEMMA 15. *The only possible sequences of rotations in a chain of minimal length* ≥ 3, *where the ordering for the rotations is determined by Algorithm* 1, *are the following:* (i) *followed by* (iii), (iii) *followed by* (i), (iii) *followed by* (ii), (ii) *followed by* (i), *and* (ii) *followed by* (ii), *where* (i), (ii), *and* (iii) *refer to the conditions in Definition* 1.

We now proceed to prove that the appropriate entry of $\mathcal{Q}(A)$ is $\star$ for every chain of length 2 that is permitted by Algorithm 1. In Lemmas 16, 17, and 18, we consider the cases (i), (ii), and (iii), respectively.

LEMMA 16. *Let* $1 \le i \le m, 1 \le j \le n$ *and* $1 \le k \le n$. *If* G_{ik} *is applied before* G_{ij} *in Algorithm* 1, *then there is a* (c_j, r_k)*-path in* $B_j(\mathbf{A})$.

Proof. Clearly $i > k$ and $i > j$. Since G_{ik} is applied before $G_{ij}, k < j$, so that $k < j < i$, which implies that $r_i \notin s_{j-1}$. By Corollary 10.1 there is a (c_k, r_i)-path and also a (c_j, r_i)-path in $B_j(\mathbf{A})$. Combining these two paths with the edge $r_k c_k$ gives a (c_j, r_k)-path in $B_j(\mathbf{A})$. □

LEMMA 17. *Let* $1 \le i \le m, 1 \le j \le n, 1 \le k \le n$, *and let* $l = \min\{i, n\}$. *If* G_{jk} *and* G_{ij} *are both applied in Algorithm* 1, *then there is an* (r_k, r_i)*-path in* $B_l(\mathbf{A})$. *Furthermore, if* $1 \le i \le n$, *then there is a* (c_i, r_k)*-path in* $B_i(\mathbf{A})$.

Proof. From the proof of Lemma 14, $k < j < i$. Lemma 12 implies that $r_j \notin s_{l-1}$ and since G_{jk} is applied, Corollary 10.1 gives a (c_k, r_j)-path in $B_l(\mathbf{A})$. Clearly $r_i \notin s_{l-1}$, so since G_{ij} is applied, Corollary 10.1 gives a (c_j, r_i)-path in $B_l(\mathbf{A})$. Combining these with the $r_k c_k$ and $c_j r_j$ edges gives an (r_k, r_i)-path in $B_l(\mathbf{A})$. Furthermore, if $i \le n$ then this path and the $r_i c_i$ edge give a (c_i, r_k)-path in $B_i(\mathbf{A}) \equiv B_l(\mathbf{A})$. □

LEMMA 18. *Let* $1 \leq j \leq n; 1 \leq i, k \leq m$, *with* $i \neq k$; *and* $l = \min\{n, k\}$. *If* G_{ij} *and* G_{kj} *are both applied in Algorithm* 1 *and if* G_{ij} *is applied before* G_{kj}, *then there is an* (r_i, r_k)*-path in* $B_l(\mathbf{A})$. *Furthermore, if* $k \leq n$, *then there is a* (c_k, r_i)*-path in* $B_k(\mathbf{A})$.

Proof. Since $l \leq k, r_k \notin s_{l-1}$. Applying Corollary 10.1 to G_{kj} gives a (c_j, r_k)-path in $B_l(\mathbf{A})$. Since G_{ij} is applied before G_{kj}, we know that $r_i \notin s_{l-1}$, so that applying Corollary 10.1 to G_{ij} gives a (c_j, r_i)-path in $B_l(\mathbf{A})$. Combining these paths gives an (r_i, r_k)-path in $B_l(\mathbf{A})$. If $k \leq n$, the $r_k c_k$ edge and this path imply that there is a (c_k, r_i)-path in $B_k(\mathbf{A})$. □

We are now ready to prove that if there is a chain of rotations as described in Theorem 8 so that $(i, j) \in \bar{\mathbf{Q}}$, then $(i, j) \in \mathcal{Q}(A)$ also.

LEMMA 19. *Let*

$$G_{i_1 j_1}, G_{i_2 j_2}, \ldots, G_{i_k j_k}, \tag{11}$$

denote a chain of minimal length $k \geq 2$ *of the rotations determined by the application of Algorithm* 1 *to a structure* $\mathbf{A}$.

(a) *If* $(i_1 = i_2$ *or* $i_1 = j_2)$ *and* $(i_{k-1} = j_k$ *or* $j_{k-1} = j_k)$, *then there is an* (r_{j_1}, r_{i_k})*-path in* $B_l(\mathbf{A})$ *where* $l = \min\{n, i_k\}$, *and if* $i_k \leq n$, *then* $(j_1, i_k) \in \mathcal{Q}(A)$.

(b) *If* $(i_1 = i_2$ *or* $i_1 = j_2)$ *and* $i_{k-1} = i_k$, *then* $(j_1, j_k) \in \mathcal{Q}(A)$.

(c) *If* $j_1 = j_2$ *and* $i_{k-1} = i_k$, *then* $(i_1, j_k) \in \mathcal{Q}(A)$.

(d) *If* $j_1 = j_2$ *and* $(i_{k-1} = j_k$ *or* $j_{k-1} = j_k)$, *then there is an* (r_{i_1}, r_{i_k})*-path in* $B_l(\mathbf{A})$ *where* $l = \min\{n, i_k\}$, *and if* $i_k \leq n$, *then* $(i_1, i_k) \in \mathcal{Q}(A)$.

Proof. The proof is by induction on k.

Base case. Let $k = 2$.

(a) becomes $G_{i_1 j_1}, G_{i_2 i_1}$ and the path exists by Lemma 17. If $i_2 \leq n$, then there is a (c_{i_2}, r_{j_1})-path in $B_{i_2}(\mathbf{A})$ also by Lemma 17 and $(j_1, i_2) \in \mathcal{Q}(A)$ by Theorem 4. All the other cases $(G_{i_1 j_1}, G_{i_1 i_1}; G_{i_1 j_1}, G_{i_1 j_1}; G_{i_1 i_1}, G_{i_2 i_1})$ are impossible.

(b) becomes $G_{i_1 j_1}, G_{i_1 j_2}$ and there is a (c_{j_2}, r_{j_1})-path in $B_{j_2}(\mathbf{A})$ by Lemma 16. Hence, $(j_1, j_2) \in \mathcal{Q}(A)$. The other case $G_{i_1 j_1}, G_{i_1 i_1}$ is impossible.

(c) becomes $G_{i_1 j_1}, G_{i_1 j_1}$, which never occurs since we do not eliminate the same nonzero twice.

(d) becomes $G_{i_1 j_1}, G_{i_2 j_1}$ and, by Lemma 18, the (r_{i_1}, r_{i_2})-path exists and if $i_2 \leq n$, then $(i_1, i_2) \in \mathcal{Q}(A)$. The other case $G_{j_1 j_1}, G_{i_2 j_1}$ is not possible.

Hence the lemma is true for the case $k = 2$.

Induction hypothesis. Suppose that the lemma is true for all chains of minimal length $\leq k - 1$.

Induction step. We consider a chain (11) of minimal length $k \geq 3$. Now

$$\text{chain 1: } G_{i_1 j_1}, G_{i_2 j_2}, \ldots, G_{i_{k-1} j_{k-1}}$$

and

$$\text{chain 2: } G_{i_2 j_2}, \ldots, G_{i_{k-1} j_{k-1}}, G_{i_k j_k}$$

are chains of length $k - 1$, so by the induction hypothesis, the lemma is true for these two subchains (and all other subchains) of (11).

There are nine cases to consider, one corresponding to each of the possible combinations of the equalities listed in the statement of the lemma.

Case 1. If $i_1 = i_2$ and $i_{k-1} = j_k$, then (11) may be written as

$$G_{i_1 j_1}, G_{i_1 j_2}, G_{i_3 j_2}, \ldots, G_{i_{k-1} j_{k-1}}, G_{i_k i_{k-1}}$$

with $k-1 \geq 3$ and $j_1 < j_2 < i_{k-1} \leq n$, since such a chain is impossible if $k-1=2$.

Case 2. If $i_1 = i_2$ and $j_{k-1} = j_k$, then (11) may be written as

$$G_{i_1 j_1}, G_{i_1 j_2}, G_{i_3 j_2}, \ldots, G_{i_{k-1} j_{k-2}}, G_{i_{k-1} j_{k-1}}, G_{i_k j_{k-1}},$$

where $k-1 \geq 2$.

The proofs of Cases 1 and 2 are identical. Chain 2 satisfies case (d) so that by the induction hypothesis, there is an (r_{i_1}, r_{i_k})-path in $B_l(\mathbf{A})$ where $l = \min\{n, i_k\}$. This implies that $r_{i_1} \notin s_{l-1}$. Since $G_{i_1 j_1}$ is applied, there is a (c_{j_1}, r_{i_1})-path in $B_l(\mathbf{A})$ by Corollary 10.1. These two paths combined with the $r_{j_1} c_{j_1}$ edge make an (r_{j_1}, r_{i_k})-path in $B_l(\mathbf{A})$. If $i_k \leq n$, then $l = i_k$ and the $r_{i_k} c_{i_k}$ edge also exists in $B_{i_k}(\mathbf{A})$ so that there is a (c_{i_k}, r_{j_1})-path as well. That is, $(j_1, i_k) \in \mathcal{Q}(A)$.

Case 3. If $i_1 = j_2$ and $i_{k-1} = j_k$, then (11) may be written as

$$G_{i_1 j_1}, G_{i_2 i_1}, \ldots, G_{i_{k-1} j_{k-1}}, G_{i_k i_{k-1}},$$

where $k-1 \geq 2$.

Case 4. If $i_1 = j_2$ and $j_{k-1} = j_k$, then (11) may be written as

$$G_{i_1 j_1}, G_{i_2 i_1}, \ldots, G_{i_{k-1} j_{k-2}}, G_{i_{k-1} j_{k-1}}, G_{i_k j_{k-1}}$$

with $k-1 \geq 3$, since such a chain is impossible if $k-1=2$.

The proofs of Cases 3 and 4 are identical. Either $i_2 = i_3$ or $i_2 = j_3$, so that chain 2 satisfies case (a) and by the induction hypothesis, there is an (r_{i_1}, r_{i_k})-path in $B_l(\mathbf{A})$ where $l = \min\{n, i_k\}$. Since $r_{i_1} \notin s_{l-1}$ and since $G_{i_1 j_1}$ is applied, there is a (c_{j_1}, r_{i_1})-path in $B_l(\mathbf{A})$ by Corollary 10.1. Combining these paths with the edge $r_{j_1} c_{j_1}$ gives an (r_{j_1}, r_{i_k})-path in $B_l(\mathbf{A})$, and if $i_k \leq n$, this path and the $r_{i_k} c_{i_k}$ edge give a (c_{i_k}, r_{j_1})-path, so that $(j_1, i_k) \in \mathcal{Q}(A)$.

Case 5. If $i_1 = i_2$ and $i_{k-1} = i_k$, then (11) may be written as

$$G_{i_1 j_1}, G_{i_1 j_2}, G_{i_3 j_2}, \ldots, G_{i_{k-1} j_{k-1}}, G_{i_{k-1} j_k}$$

with $k-1 \geq 3$, since such a chain is impossible if $k-1=2$. Now chain 2 satisfies case (c) so that there is a (c_{j_k}, r_{i_1})-path in $B_{j_k}(\mathbf{A})$. Thus, $r_{i_1} \notin s_{j_k-1}$, and since $G_{i_1 j_1}$ is applied, $B_{j_k}(\mathbf{A})$ contains a (c_{j_1}, r_{i_1})-path by Corollary 10.1. Combining these paths with the $r_{j_1} c_{j_1}$ edge gives a (c_{j_k}, r_{j_1})-path in $B_{j_k}(\mathbf{A})$. Hence, $(j_1, j_k) \in \mathcal{Q}(A)$.

Case 6. If $i_1 = j_2$ and $i_{k-1} = i_k$, then (11) may be written as

$$G_{i_1 j_1}, G_{i_2 i_1}, \ldots, G_{i_{k-1} j_{k-1}}, G_{i_{k-1} j_k},$$

where $k-1 \geq 2$. Since either $i_2 = i_3$ or $i_2 = j_3$, chain 2 satisfies case (b), and by the induction hypothesis, there is a (c_{j_k}, r_{i_1})-path in $B_{j_k}(\mathbf{A})$. Thus $r_{i_1} \notin s_{j_k-1}$, and since $G_{i_1 j_1}$ is applied, there exists a (c_{j_1}, r_{i_1})-path in $B_{j_k}(\mathbf{A})$ by Corollary 10.1. These paths combined with the $r_{j_1} c_{j_1}$ edge produce a (c_{j_k}, r_{j_1})-path in $B_{j_k}(\mathbf{A})$. Hence, $(j_1, j_k) \in \mathcal{Q}(A)$.

Case 7. If $j_1 = j_2$ and $i_{k-1} = i_k$, then (11) may be written as

$$G_{i_1 j_1}, G_{i_2 j_1}, \ldots, G_{i_{k-1} j_{k-1}}, G_{i_{k-1} j_k},$$

where $k-1 \geq 2$. Chain 1 satisfies case (d), so for $l = \min\{n, i_{k-1}\}$, $B_l(\mathbf{A})$ contains an $(r_{i_1}, r_{i_{k-1}})$-path. Thus, $r_{i_1} \notin s_{l-1}$, and, consequently, $r_{i_1} \notin s_{j_k-1}$ also (since $j_k \leq l$).

Either $i_2 = i_3$ or $i_2 = j_3$, so that chain 2 satisfies case (b) and hence there is a (c_{j_k}, r_{j_1})-path in $B_{j_k}(\mathbf{A})$ by the induction hypothesis. Since $G_{i_1 j_1}$ is applied and $r_{i_1} \notin s_{j_k - 1}$, by Corollary 10.1 there exists a (c_{j_1}, r_{i_1})-path in $B_{j_k}(\mathbf{A})$. The latter two paths combined with the $r_{j_1} c_{j_1}$ edge give a (c_{j_k}, r_{i_1})-path in $B_{j_k}(\mathbf{A})$, and hence $(i_1, j_k) \in \mathcal{Q}(A)$.

Case 8. If $j_1 = j_2$ and $i_{k-1} = j_k$, then (11) may be written as either

$$G_{i_1 j_1}, G_{i_2 j_1}, G_{i_3 i_2}, \ldots, G_{i_{k-1} j_{k-1}}, G_{i_k i_{k-1}}, \tag{12}$$

where $k - 1 \geq 2$, or

$$G_{i_1 j_1}, G_{i_2 j_1}, G_{i_2 j_3}, \ldots, G_{i_{k-1} j_{k-1}}, G_{i_k i_{k-1}}, \tag{13}$$

where $k - 1 \geq 4$. Consider the subchain

$$G_{i_3 i_2}, \ldots, G_{i_{k-1} j_{k-1}}, G_{i_k i_{k-1}}$$

of (12). If $k - 1 = 2$, then this subchain reduces to the single rotation $G_{i_3 i_2}$. In this case, with $l = \min\{n, i_3\}$, we have $r_{i_2} \notin s_{l-1}$ (by Lemma 12) and clearly $r_{i_3} \notin s_{l-1}$, so Corollary 10.1 gives a (c_{i_2}, r_{i_3})-path in $B_l(\mathbf{A})$, which together with the $c_{i_2} r_{i_2}$ edge gives an (r_{i_2}, r_{i_3})-path in $B_l(\mathbf{A})$. If $k - 1 > 2$, the subchain is of length at least 2 and satisfies case (a), and thus an (r_{i_2}, r_{i_k})-path exists in $B_l(\mathbf{A})$ by the induction hypothesis.

Consider now the subchain

$$G_{i_2 j_3}, \ldots, G_{i_{k-1} j_{k-1}}, G_{i_k i_{k-1}}$$

of (13), which is of length at least 2 and satisfies case (d). Thus the induction hypothesis implies that there is an (r_{i_2}, r_{i_k})-path in $B_l(\mathbf{A})$.

So in every instance, there is an (r_{i_2}, r_{i_k})-path in $B_l(\mathbf{A})$, which implies $r_{i_2} \notin s_{l-1}$. Since $G_{i_1 j_1}$ is applied before $G_{i_2 j_1}, r_{i_1} \notin s_{l-1}$. Corollary 10.1 applied to $G_{i_1 j_1}$ gives a (c_{j_1}, r_{i_1})-path in $B_l(\mathbf{A})$, and applying it to $G_{i_2 j_1}$ gives a (c_{j_1}, r_{i_2})-path in $B_l(\mathbf{A})$. Combining these three paths produces an (r_{i_1}, r_{i_k})-path. Furthermore, if $i_k \leq n$, then $B_l(\mathbf{A})$ also contains the $r_{i_k} c_{i_k}$ edge, which gives a (c_{i_k}, r_{i_1})-path, so that $(i_1, i_k) \in \mathcal{Q}(A)$.

Case 9. If $j_1 = j_2$ and $j_{k-1} = j_k$, then (11) may be written as either

$$G_{i_1 j_1}, G_{i_2 j_1}, G_{i_2 j_3}, \ldots, G_{i_{k-1} j_{k-1}}, G_{i_k j_{k-1}}, \tag{14}$$

where $k - 1 \geq 3$, or

$$G_{i_1 j_1}, G_{i_2 j_1}, G_{i_3 i_2}, \ldots, G_{i_{k-1} j_{k-1}}, G_{i_k j_{k-1}}, \tag{15}$$

where $k-1 \geq 4$. The subchain $G_{i_2 j_3}, \ldots, G_{i_{k-1} j_{k-1}}, G_{i_k j_{k-1}}$ of (14) is of length at least 2 and satisfies case (d). By the induction hypothesis, there is an (r_{i_2}, r_{i_k})-path in $B_l(\mathbf{A})$, where $l = \min\{i_k, n\}$. Similarly, the subchain $G_{i_3 i_2}, \ldots, G_{i_{k-1} j_{k-1}}, G_{i_k j_{k-1}}$ of (15) satisfies case (a), and the induction hypothesis implies the existence of an (r_{i_2}, r_{i_k})-path in $B_l(\mathbf{A})$. Thus, $r_{i_2} \notin s_{l-1}$, and since $G_{i_1 j_1}$ is applied before $G_{i_2 j_1}, r_{i_1} \notin s_{l-1}$. So Corollary 10.1 applied to $G_{i_1 j_1}$ gives a (c_{j_1}, r_{i_1})-path in $B_l(\mathbf{A})$, and applying it to $G_{i_2 j_1}$ gives a (c_{j_1}, r_{i_2})-path in $B_l(\mathbf{A})$. Combining these paths with the (r_{i_2}, r_{i_k})-path gives an (r_{i_1}, r_{i_k})-path in $B_l(\mathbf{A})$. Furthermore, if $i_k \leq n$, then $l = i_k$ and the $r_{i_k} c_{i_k}$ edge is in $B_{i_k}(\mathbf{A})$, so there is a (c_{i_k}, r_{i_1})-path as well. That is, $(i_1, i_k) \in \mathcal{Q}(A)$. □

5.4. Main results.

THEOREM 20. *Let* $\mathbf{A}$ *be the structure of an* $m \times n$ *matrix with the Hall property and a zero-free diagonal. Then the structure* $\bar{\mathbf{Q}}$ *in* (9) *that results from application of Algorithm* 1 *to* $\mathbf{A}$ *is identical to the structure* $\mathcal{Q}(A)$ *in* (1).

Proof. The elements of $\bar{\mathbf{Q}}$ are characterized by Theorem 8. Let $(x, y) \in \bar{\mathbf{Q}}$. Then $y \leq n$. If $x = y$, then $(x, y) \in \mathcal{Q}(A)$ by Lemma 9. If $x \neq y$ and G_{xy} is one of the applied rotations, then $(x, y) \in \mathcal{Q}(A)$ by Corollary 10.2, whereas if G_{yx} (with $x \leq n$) is one of the applied rotations, then $(x, y) \in \mathcal{Q}(A)$ by Lemma 13. Finally, if there is a chain of rotations

$$G_{i_{s_1} j_{s_1}}, G_{i_{s_2} j_{s_2}}, \ldots, G_{i_{s_k} j_{s_k}}$$

with $x = i_{s_1}$ or $x = j_{s_1}$ and $y = i_{s_k}$ or $y = j_{s_k}$, so that $(x, y) \in \bar{\mathbf{Q}}$, then $(x, y) \in \mathcal{Q}(A)$ by Lemma 19. Given Lemma 14, these exhaust all possibilities in Theorem 8. Thus $\bar{\mathbf{Q}} \subseteq \mathcal{Q}(A)$.

To show the reverse inclusion, suppose $(x, y) \in \mathcal{Q}(A)$. For any full rank $m \times n$ matrix A such that $\text{struct}(A) = \mathbf{A}$, suppose its QR factorization is computed using a sequence of Givens transformations

$$G_{i_1 j_1}, G_{i_2 j_2}, \ldots, G_{i_r j_r}$$

as determined by Algorithm 1. By Lemma 3, $\text{struct}(Q) \subseteq \bar{\mathbf{Q}}$. Thus $\mathcal{Q}(A) \subseteq \bar{\mathbf{Q}}$, completing the proof. □

THEOREM 21. *Let* $\mathbf{A}$ *be the structure of an* $m \times n$ *matrix with the Hall property and a zero-free diagonal. Then the structure* $\bar{\mathbf{R}}$ *in* (8) *that results from application of Algorithm* 1 *to* $\mathbf{A}$ *is identical to the structure* $\mathcal{R}(A)$ *in* (2).

Proof. Algorithm 1 computes

$$\begin{aligned}
\bar{\mathbf{R}} &= \left(\left(\prod_{k=r}^{1} \text{struct}(G_{i_k j_k}) \right) \mathbf{A} \right) \setminus \{(i,j) | i > j\} \\
&= \left(\left(\prod_{k=1}^{r} \text{struct}(G_{i_k j_k}^T) \right)^T \mathbf{A} \right) \setminus \{(i,j) | i > j\} \\
&= \bar{\mathbf{Q}}^T \mathbf{A} \setminus \{(i,j) | i > j\}.
\end{aligned}$$

But since $\bar{\mathbf{Q}} = \mathcal{Q}(A)$ by Theorem 20, it follows that $\bar{\mathbf{R}} = \mathcal{R}(A)$ by Theorem 5.

6. Conclusions. We have shown that Algorithm 1 determines a tight ordering for any matrix that has the Hall property and a zero-free diagonal. We have not considered any particular row or column ordering schemes. However, our algorithm could be used with any fixed-pivot row ordering scheme (i.e., one that uses only one pivot in each column) since row interchanges have no substantive effect on Hall sets (provided a zero-free diagonal is maintained). Algorithm 1 can also be used with any a priori column ordering scheme, but not with a scheme that reorders columns at each step, since this changes the Hall sets.

Algorithm 1 can be adapted to perform a numerical QR factorization by replacing the structures with numeric matrices and the products of structures with matrix multiplication. Such a numeric computation $A = QR$ is optimal with regard to storage because the ordering is tight; that is, barring accidental cancellation, $\text{struct}(Q) = \bar{\mathbf{Q}} = \mathcal{Q}(A)$ and $\text{struct}(R) = \bar{\mathbf{R}} = \mathcal{R}(A)$. If Algorithm 1 is first used to determine $\hat{\mathbf{A}}$ and $\bar{\mathbf{Q}}$, then the entire numerical computation may be performed within

the (static) structures $\hat{\mathbf{A}}$ and $\bar{\mathbf{Q}}$. If there is no need to compute Q explicitly, it may be saved in factored form in the lower trapezoidal part of $\hat{\mathbf{A}}$, using a single parameter to characterize each rotation as described in [8].

We can find an upper bound on the time complexity of Algorithm 1 (or an analogous algorithm for computing a numeric QR factorization) for an $m \times n$ full rank matrix. Corollary 10.2 restricts the number of rotations required to the number of nonzeros below the diagonal in $\mathcal{Q}(A)$. The number of operations required for each rotation is linear in the number of nonzeros in the two rows involved, and this number is certainly not more than $2n$. This gives a bound on the operation count of $\mathcal{O}(n\tau(\mathcal{Q}))$, where $\tau(\mathcal{Q})$ is the number of nonzeros in the lower trapezoidal part of $\mathcal{Q}(A)$.

For a tight ordering, there is no unnecessary intermediate storage for $\mathcal{Q}(A)$ and $\mathcal{R}(A)$; that is, all storage allocated in Algorithm 1 is ultimately required for these structures. However, the amount of computation (i.e., the number of Givens rotations required) is not necessarily minimized. As the following example illustrates, for a given structure Algorithm 1 may produce different tight orderings that use a different number of rotations.

Example 3. For

$$\mathbf{A} = \begin{pmatrix} \star & 0 & 0 \\ 0 & \star & 0 \\ \star & \star & \star \\ \star & 0 & 0 \end{pmatrix},$$

$s_0 = s_1 = s_2 = \varnothing$ and $s_3 = \{r_2, r_3\}$. The orderings of the rotations in the QR factorizations

$$A = G_{31}^T G_{41}^T G_{32}^T G_{42}^T G_{43}^T R$$

and

$$A = \tilde{G}_{41}^T \tilde{G}_{31}^T \tilde{G}_{32}^T R$$

are both tight.

A consequence of Theorems 20 and 21 is that if A has the strong Hall property or if A has the Hall property and is in Dulmage–Mendelsohn form (see, e.g., [1]), then all orderings that compute the QR factorization column-by-column using diagonal pivoting are tight. In the first case, there are no nonempty Hall sets. The second case follows similarly since the diagonal blocks have the strong Hall property and the QR factorization of A is easily obtained from the QR factorizations of the diagonal blocks of A. These results correspond to those of Coleman, Edenbrandt, and Gilbert [1].

REFERENCES

[1] T. F. Coleman, A. Edenbrandt, and J. R. Gilbert, *Predicting fill for sparse orthogonal factorization*, J. Assoc. Comput. Mach., 33 (1986), pp. 517–532.

[2] I. S. Duff, *Pivot selection and row ordering in Givens reduction on sparse matrices*, Computing, 13 (1974), pp. 239–248.

[3] ———, *Algorithm* 575. *Permutations for a zero-free diagonal*, ACM Trans. Math. Software, 7 (1981), pp. 387–390.

[4] A. George and M. T. Heath, *Solution of sparse linear least squares problems using Givens rotations*, Linear Algebra Appl., 34 (1980), pp. 69–83.

[5] A. George, J. Liu, and E. Ng, *Row-ordering schemes for sparse Givens transformations. 1. Bipartite graph model*, Linear Algebra Appl., 61 (1984), pp. 55–81.

[6] ———, *A data structure for sparse* QR *and* LU *factorizations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 100–121.

[7] A. George and E. Ng, *On row and column orderings for sparse least squares problems*, SIAM J. Numer. Anal., 20 (1983), pp. 326–344.

[8] G. H. GOLUB AND C. F. VAN LOAN , *Matrix Computations*, The Johns Hopkins University Press, Baltimore, Maryland, 1989.

[9] D. HARE, C. R. JOHNSON, D. D. OLESKY, AND P. VAN DEN DRIESSCHE, *Sparsity analysis of the* QR *factorization*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 655–669.

[10] J. W. H. LIU, *On general row merging schemes for sparse Givens transformations*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1190–1211.

[11] E. G. NG AND B. W. PEYTON, *A Tight and Explicit Representation of* Q *in Sparse* QR *Factorization*, Oak Ridge National Laboratory, Report TM-12059, Oak Ridge, Tennessee, 1992.

[12] G. OSTROUCHOV, *Symbolic Givens reduction and row-ordering in large, sparse, least squares problems*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 248–264.

[13] T. H. ROBEY AND D. L. SULSKY, *Row ordering for a sparse* QR *decomposition*, preprint (1992).

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, pp. 1043–1055, October 1995

ON A STURM SEQUENCE OF POLYNOMIALS FOR UNITARY HESSENBERG MATRICES*

ANGELIKA BUNSE-GERSTNER† AND CHUNYANG HE‡

Abstract. Unitary matrices have a rich mathematical structure that is closely analogous to real symmetric matrices. For real symmetric matrices this structure can be exploited to develop very efficient numerical algorithms and for some of these algorithms unitary analogues are known. Here we present a unitary analogue of the bisection method for symmetric tridiagonal matrices. Recently Delsarte and Genin introduced a sequence of so-called γ_n-symmetric polynomials that can be used to replace the classical Szegö polynomials in several signal processing problems. These polynomials satisfy a three-term recurrence relation and their roots interlace on the unit circle. Here we explain this sequence of polynomials in matrix terms. For an $n \times n$ unitary Hessenberg matrix, we introduce, motivated by the Cayley transformation, a sequence of modified unitary submatrices. The characteristic polynomials of the modified unitary submatrices $p_k(z), k = 1, 2, \ldots, n$ are exactly the γ_n-symmetric polynomials up to a constant. These polynomials can be considered as a sort of Sturm sequence and can serve as a basis for a bisection method for computing the eigenvalues of the unitary Hessenberg matrix. The Sturm sequence properties allow identification of the number of roots of $p_n(z)$, the characteristic polynomial of the unitary Hessenberg matrix itself, on any arc of the unit circle by computing the sign agreements of certain related real polynomials at a given point.

Key words. unitary matrices, Sturm sequence, root interlacing, bisection method

AMS subject classifications. 65F15, 15A18, 42C05

1. Introduction.

Numerical methods especially developed for unitary eigenvalue problems have been developed in [1], [4], [6], [7], [12], [15], [18], [19], [25]. Such eigenvalue problems arise for example in signal processing [2], [5], [8], [14], [20]–[23], [26], or more generally in trigonometric approximation problems [9], [17], [24]. These special methods make use of the fact that any unitary Hessenberg matrix with positive subdiagonal elements can be parameterized by n parameters $\gamma_1, \ldots, \gamma_n$, called reflection coefficients or Schur parameters, and essentially only these parameters must be considered in the numerical process.

The mathematical structure of the unitary eigenvalue problem is closely analogous to the structure of the real symmetric eigenvalue problem. Thus one can hope to find unitary analogues for the good numerical methods that exist for the symmetric eigenvalue problem. Some such unitary analogues have been developed. There are unitary QR methods [6], [18], a divide-and-conquer method [4], [19], and a method to solve the inverse unitary eigenvalue problem [3]. In addition there have been special unitary developments like methods for the real orthogonal eigenvalue problem [1], [7] and a pencil method [12].

For symmetric matrices we also have the bisection method, which computes the eigenvalues or only the number of the eigenvalues in a given interval. If for a unitary matrix only the eigenvalues or the number of eigenvalues on a certain arc of the unit circle are of interest, then a unitary analogue for this bisection method would be helpful. The basis for the bisection method for symmetric tridiagonal $n \times n$ matrices is the fact that the characteristic polynomials of the leading principal submatrices,

* Received by the editors May 4, 1992; accepted for publication (in revised form) by P. Van Dooren September 27, 1994.

† Fachbereich Mathematik und Informatik, Universität Bremen, Postfach 33 04 40, 28334 Bremen, Deutschland.

‡ Universität Chemnitz, Fachbereich Mathematik, PSF 964, 09009 Chemnitz, Deutschland (he@mathematik.tu-chemitz.de). The work of this author was supported by the Alexander von Humboldt Research Foundation and the SFB 343 of Bielefeld University.

$d_1(x), \ldots, d_n(x)$ say, form a Sturm sequence. They can be evaluated by a three-term recurrence and the roots of consecutive d's interlace. The number of sign agreements in consecutive terms of the numerical sequence $\{d_j(\hat{x}), j \doteq 1, \ldots, n\}$ is the number of the eigenvalues that are smaller than $\hat{x}$.

Recently a new family of polynomials, $q_0(z), q_1(z), \ldots, q_n(z)$, called γ_n-symmetric polynomials, has been introduced in a series of papers by Delsarte and Genin [10]–[13]. Given reflection coefficient $\gamma_1, \ldots, \gamma_n$ with $|\gamma_k| < 1$ for $k = 1, \ldots, n-1$ and $|\gamma_n| = 1$, and η_0 with $|\eta_0| = 1$, called the circle parameter, these polynomials can be constructed from Szegö polynomials and satisfy the following three-term recurrence relations:

$$q_{-1}(z) = 0, q_0(z) = q_0,$$

$$q_{k+1}(z) = (\beta_k + \bar{\beta}_k z) q_k(z) - z q_{k-1}(z), \tag{1}$$

for $k = 0, 1, \ldots, n-1$. Here q_0 and β_0 with $|\beta_0| > 1/2$ are nonzero real numbers and the β_k are obtainable by the recurrence

$$\beta_k = \eta_0 \beta_{k-1}^{-1} (1 + \eta_0 \rho_k \bar{\gamma}_{k-1})^{-1} (1 - \bar{\rho}_k \gamma_k)^{-1},$$

for $k = 1, 2, \ldots, n-1$ and the ρ_k by

$$\rho_k = \frac{\gamma_k + \eta_0 \rho_{k+1}}{1 + \eta_0 \bar{\gamma}_k \rho_{k+1}} \tag{2}$$

for $k = n-1, n-2, \ldots, 1$ with initial value $\rho_n = \gamma_n$. The $\{\rho_k\}$ are of modulus one and are called pseudo reflection coefficients [13], [12]. The roots of these polynomials are all on the unit circle and interlace each other [11]. Delsarte and Genin also showed that the new family of polynomials could be used to replace the classical Szegö polynomials in several signal processing problems and thus provide new techniques for the interpolation problem [11], the retrieval of harmonics problem [13], [12], and Toeplitz systems [10].

In this paper we explain the γ_n-symmetric polynomials in terms of unitary Hessenberg matrices. For a unitary $n \times n$ Hessenberg matrix H, the $k \times k$ leading principal submatrix can be modified in a simple way to a $k \times k$ unitary matrix $\tilde{H}_k$ by replacing the kth reflection coefficient γ_k by $\rho_k = \frac{\gamma_k + \bar{\xi}_0 \rho_{k+1}}{1 + \bar{\xi}_0 \bar{\gamma}_k \rho_{k+1}}$ for $k = 1, \ldots, n-1$ and $\rho_n = \gamma_n$. We call ξ_0 the cutting point. Note by comparing this definition with (2) that $\bar{\xi}_0$ is the circle parameter in the notation of Delsarte and Genin. This definition of modified unitary submatrices is motivated by the Cayley transformation: Choosing the cutting point ξ_0 we get a Hermitian matrix A via the (generalized) Cayley transformation as $A = i(\xi_0 I - H)^{-1}(\xi_0 I + H)$. Then A_k, the kth leading principal submatrix of A, can be shown to be exactly the Cayley transform of $\tilde{H}_k$, i.e., $A_k = i(\xi_0 I - \tilde{H}_k)^{-1}(\xi_0 I + \tilde{H}_k)$. The monotonicity of the Cayley transformation then implies that the eigenvalues of the $\tilde{H}_k$-s interlace on the unit circle with respect to the cutting point ξ_0. The characteristic polynomials of modified unitary submatrices are, up to a constant, the sequence of polynomials given by Delsarte and Genin. We show that for any ξ on the unit circle the number of sign agreements in consecutive terms of the sequence of real numbers $\{(\frac{2i}{\xi - \xi_0})^k \frac{p_k(\xi)}{p_k(\xi_0)}, k = 1, \ldots, n\}$ is the number of eigenvalues of H that are located on the arc between ξ_0 and ξ when moving counterclockwise along the unit

circle. Thus we can consider these polynomials as a sort of Sturm sequence for unitary Hessenberg matrices.

The paper is organized as follows. In §2 we introduce the modified unitary submatrices $\tilde{H}_k$ and show that their Cayley transform is the kth leading principal submatrix of the Cayley transform of H. In §3 we give a three-term recurrence relation for the characteristic polynomials of $\tilde{H}_k, k = 1, \ldots, n$ and have a closer look at the cutting point. In §4 we introduce related polynomials, $\{d_k(\xi), k = 0, 1, \ldots, n\}$ given by $d_k(\xi) = (\frac{2i}{1+\xi})^k \frac{p_k(\xi)}{p_k(-1)}$ and prove that the number of sign agreements of consecutive elements of the sequence $\{d_k(\xi), k = 0, 1, \ldots, n\}$ is the number of the roots of $p_n(z)$ which lie on the arc between -1 and ξ when moving counterclockwise along the unit circle. This can serve as a basis for a bisection method for the unitary eigenvalue problem. In §5, a Christoffel–Darboux-type formula is derived analogous to the formula given by Delsarte and Genin.

2. Modified unitary submatrices. Let H be an $n \times n$ unitary Hessenberg matrix with real positive subdiagonal elements. Then it is well known that H can be written as $H = G_1 G_2 \ldots G_n$, where

$$G_k = \operatorname{diag}(I_{k-1}, \begin{pmatrix} -\gamma_k & \sigma_k \\ \sigma_k & \bar{\gamma}_k \end{pmatrix}, I_{n-k-1}), \quad k = 1, \ldots, n-1$$

and

$$G_n = \operatorname{diag}(I_{n-1}, -\gamma_n).$$

The parameters γ_k, $k = 1, \ldots, n$, are called reflection coefficients in signal processing and satisfy $|\gamma_k|^2 + \sigma_k^2 = 1$, $\sigma_k > 0$ and $|\gamma_n| = 1$. H is of the explicit form

$$\begin{pmatrix} -\gamma_1 & -\sigma_1\gamma_2 & \cdots & -\sigma_1 \ldots \sigma_{n-1}\gamma_n \\ \sigma_1 & -\bar{\gamma}_1\gamma_2 & \cdots & -\bar{\gamma}_1\sigma_2 \ldots \sigma_{n-1}\gamma_n \\ & \ddots & \ddots & \vdots \\ & & \sigma_{n-1} & -\bar{\gamma}_{n-1}\gamma_n \end{pmatrix}$$

and is uniquely determined by $\gamma_1, \ldots, \gamma_n$. We denote this $n \times n$ unitary Hessenberg matrix by $H(\gamma_1, \ldots, \gamma_n)$. This representation was introduced by Gragg and is the basic condensed form for the development of unitary eigenvalue algorithms analogous to the symmetric tridiagonal matrix in symmetric eigenvalue algorithms [17], [18]. Let H_k be the kth leading principal submatrix of H. Then $\chi_k(z) = \det(zI - H_k)$, $k = 1, \ldots, n$, the monic characteristic polynomials of the H_k-s, are the well-known Szegö polynomials [20]. They satisfy the following recurrence relations:

$$\chi_0 = 1, \tilde{\chi}_0 = 1,$$

$$\chi_k(z) = z\chi_{k-1}(z) + \gamma_k\tilde{\chi}_{k-1}(z),$$

$$\tilde{\chi}_k(z) = \tilde{\chi}_{k-1}(z) + z\bar{\gamma}_k\chi_{k-1}(z), \quad k = 1, \ldots, n.$$

The $\tilde{\chi}_k(z)$, $k = 1, \ldots, n$, are auxiliary polynomials and it follows by induction that $\tilde{\chi}_k(z) = z^k \bar{\chi}_k(\frac{1}{z})$.

The matrix H_k is not unitary for $k < n$ and the roots of $\chi_k(z)$ are inside the unit circle. It will however become unitary if γ_k is replaced by any number on the unit circle. Assume for the following that -1 is not an eigenvalue of H. Motivated by (2), we introduce the following sequence of modified unitary submatrices:

$$\tilde{H}_k = \begin{pmatrix} -\gamma_1 & -\sigma_1\gamma_2 & \cdots & -\sigma_1 \cdots \sigma_{k-1}\rho_k \\ \sigma_1 & -\bar{\gamma}_1\gamma_2 & \cdots & -\bar{\gamma}_1\sigma_2 \cdots \sigma_{k-1}\rho_k \\ & \ddots & \ddots & \vdots \\ & & \sigma_{k-1} & -\bar{\gamma}_{k-1}\rho_k \end{pmatrix}, \qquad k = 1, 2, \ldots, n, \tag{3}$$

where $\rho_n = \gamma_n$ and

$$\rho_k = \frac{\gamma_k - \rho_{k+1}}{1 - \bar{\gamma}_k\rho_{k+1}}, \quad k = n-1, n-2, \ldots, 1.$$

Here we have chosen $\xi_0 = -1$ in (2) for the definition of the ρ_k-s. We call the point -1 cutting point, because the unit circle under the Cayley transformation is "cut" at this point and "stretched to the real axis." The choice -1 as cutting point is for convenience. In fact any point ξ_0 can serve as cutting point, as discussed in the next section. Because all ρ_k are of modulus one, the modified submatrices $\tilde{H}_k$ are unitary and $\tilde{H}_k = H(\gamma_1, \ldots, \gamma_{k-1}, \rho_k)$. The Cayley transformation can shed some light on this special definition of modified submatrices with the following theorem.

THEOREM 2.1. *Let $A = i(I + H)^{-1}(I - H)$ and A_k be the kth leading principal submatrix of A. Then*

$$A_k = i(I + \tilde{H}_k)^{-1}(I - \tilde{H}_k), \tag{4}$$

i.e., the Cayley transform of $\tilde{H}_k$ is the kth leading principle submatrix A_k of A, the Cayley transform of H.

To prove this result we need the following two lemmas. The first one shows that A_k is related to the Schur complement of the $(n-k)$th trailing principal submatrix of $H + I$. It is proved in [16].

LEMMA 2.2 ([16]). *Let H be partitioned as*

$$\begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}, \tag{5}$$

where H_{11} is a $k \times k$ matrix and H_{22} is an $(n-k) \times (n-k)$ matrix. Let $A = i(I + H)^{-1}(I - H)$ be the Cayley transform of H and A_k its kth leading principal submatrix.

Then A_k is the Cayley transform of $H_{11} - H_{12}(H_{22} + I)^{-1}H_{21}$.

The second lemma we need, expresses ρ_k in terms of $\gamma_k, \gamma_{k+1}, \ldots, \gamma_n$. Note here that for the kth leading principal submatrix H_k of H we get $H_k = H(\gamma_1, \ldots, \gamma_{k-1}, 1)\tilde{D}_k$, where $\tilde{D}_k = \operatorname{diag}(1, \ldots, 1, \gamma_k)$ and the $(n-k)$th trailing principal submatrix of H is given by $D_{n-k}H(\gamma_{k+1}, \ldots, \gamma_n)$, where $D_{n-k} = \operatorname{diag}(\bar{\gamma}_k, 1, \ldots, 1)$.

LEMMA 2.3. *Let $\rho_n = \gamma_n$ and*

$$\rho_k = \frac{\gamma_k - \rho_{k+1}}{1 - \bar{\gamma}_k \rho_{k+1}} \quad \text{for } k = n-1, n-2, \ldots, 1, \tag{6}$$

then ρ_k *can be expressed in terms of* $\gamma_k, \gamma_{k+1}, \ldots, \gamma_n$ *as*

$$\rho_k = \gamma_k + \sigma_k^2((-1)^{n-k-2}\gamma_n)\frac{\bar{w}_{k+1}}{w_k}, \tag{7}$$

where $w_k = \det(D_{n-k}H(\gamma_{k+1}, \ldots, \gamma_n) + I)$, $D_{n-k} = \operatorname{diag}(\bar{\gamma}_k, 1, \ldots, 1)$ *and* $w_n = 1$.

Proof. The proof is by induction. For $k = n-1$ (6) leads to

$$\rho_{n-1} = \gamma_{n-1} + \sigma_{n-1}^2 \frac{-\gamma_n}{1 - \bar{\gamma}_{n-1}\gamma_n} = \gamma_{n-1} + \sigma_{n-1}^2(-\gamma_n)\frac{\bar{w}_n}{w_{n-1}},$$

where $w_n = 1$ and $w_{n-1} = -\bar{\gamma}_{n-1}\gamma_n + 1$.

Assume that (7) holds for $k = n-1, \ldots, \ell+1$. The essential step in proving (7) for $k = \ell$ is the observation that the ω_k-s satisfy a three-term recurrence relation. It is easy to see that

$$\operatorname{diag}(1, H(\gamma_{\ell+2}, \ldots, \gamma_n))[D_{n-\ell}H(\gamma_{\ell+1}, \ldots, \gamma_n) + I]H^H(\gamma_{\ell+1}, \ldots, \gamma_n)$$

$$= \begin{pmatrix} \bar{\gamma}_\ell & \\ & H(\gamma_{\ell+2}, \ldots, \gamma_n) \end{pmatrix} + \begin{pmatrix} \begin{pmatrix} -\bar{\gamma}_{\ell+1} & \sigma_{\ell+1} \\ \sigma_{\ell+1} & \gamma_{\ell+1} \end{pmatrix} & \\ & I_{n-\ell-2} \end{pmatrix},$$

the determinant of which is equal to $-w_\ell$, because $\det(\operatorname{diag}(1, H(\gamma_{\ell+2}, \ldots, \gamma_n)) = (-1)^{n-\ell-3}\gamma_n$ and $\det(H^H(\gamma_{\ell+1}, \ldots, \gamma_n)) = (-1)^{n-\ell-2}\bar{\gamma}_n$. The Laplace determinant theorem then implies that

$$-w_\ell = (\bar{\gamma}_\ell - \bar{\gamma}_{\ell+1})((-1)^{n-\ell-3}\gamma_n)\bar{w}_{\ell+1} - \sigma_{\ell+1}^2 w_{\ell+2},$$

or equivalently

$$(-1)^{n-\ell-2}\bar{\gamma}_n \frac{w_\ell}{\bar{w}_{\ell+1}} = (\bar{\gamma}_\ell - \bar{\gamma}_{\ell+1}) - \sigma_{\ell+1}^2((-1)^{n-\ell-3}\bar{\gamma}_n)\frac{w_{\ell+2}}{\bar{w}_{\ell+1}}.$$

According to (7) for $k = \ell+1$ we have $\sigma_{\ell+1}^2((-1)^{n-\ell-3}\gamma_n)\bar{w}_{\ell+2}/w_{\ell+1} = \rho_{\ell+1} - \gamma_{\ell+1}$ and therefore

$$(-1)^{n-\ell-2}\bar{\gamma}_n \frac{w_\ell}{\bar{w}_{\ell+1}} = (\bar{\gamma}_\ell - \bar{\gamma}_{\ell+1}) - (\bar{\rho}_{\ell+1} - \bar{\gamma}_{\ell+1}) = \bar{\gamma}_\ell - \bar{\rho}_{\ell+1}. \tag{8}$$

From (6) we get

$$\bar{\rho}_{\ell+1} = \frac{1 - \bar{\gamma}_\ell \rho_\ell}{\gamma_\ell - \rho_\ell}$$

and thus it follows from (8) that

$$((-1)^{n-\ell-2}\bar{\gamma}_n)\frac{w_\ell}{\bar{w}_{\ell+1}} = \frac{-\sigma_\ell^2}{\gamma_\ell - \rho_\ell},$$

or equivalently

$$\rho_\ell = \gamma_\ell + \sigma_\ell^2((-1)^{n-\ell-2}\gamma_n)\frac{\bar{w}_{\ell+1}}{w_\ell}. \qquad \square$$

Proof of Theorem 2.1. Suppose that H is partitioned as in (5). We show that

$$\tilde{H}_k = H_{11} - H_{12}(H_{22} + I)^{-1}H_{21}. \tag{9}$$

Recall that

$$H_{11} = H(\gamma_1, \ldots, \gamma_{k-1}, 1)\tilde{D}_k,$$

$$H_{12} = -\sigma_k H(\gamma_1, \ldots, \gamma_{k-1}, 1)e_k e_1^T H(\gamma_{k+1}, \ldots, \gamma_n),$$

$$H_{21} = \sigma_k e_1 e_k^T,$$

$$H_{22} = D_{n-k}H(\gamma_{k+1}, \ldots, \gamma_n).$$

A simple evaluation leads to

$$H_{11} - H_{12}(H_{22} + I)^{-1}H_{21} = H(\gamma_1, \ldots, \gamma_{k-1}, 1)\tilde{D}_k + l_k H(\gamma_1, \ldots, \gamma_{k-1}, 1)e_k e_k^T,$$

where

$$l_k = \sigma_k^2(e_1^H H(\gamma_{k+1}, \ldots, \gamma_n)(D_{n-k}H(\gamma_{k+1}, \ldots, \gamma_n) + I_{n-k})^{-1}e_1).$$

By Lemma 2.3 we obtain $l_k = \rho_k - \gamma_k$ and

$$H_{11} - H_{12}(H_{22} + I)^{-1}H_{21} = \tilde{H}_k,$$

and Lemma 2.2 then shows that $\tilde{H}_k$, our kth modified leading principal submatrix of H, is the Cayley transform of A_k. $\square$

The one-dimensional Cayley transformation $x = i\frac{1-\xi}{1+\xi}$ mapping the unit circle onto the real axis is known to be a strictly monotone function of $\theta = \arctan\frac{\mathrm{im}(\xi)}{1+\mathrm{re}(\xi)}$.

Because the eigenvalues of A_k interlace those of A_{k+1} on the real line, the strict monotonicity of the Cayley transformation assures that the eigenvalues of $\tilde{H}_k$ interlace those of $\tilde{H}_{k+1}$ on the unit circle. The cutting point -1 corresponds to both $-\infty$ and $+\infty$ under this transformation and we thus get the following corollary.

COROLLARY 2.4. *Let $\tilde{H}_k, k = 1, \ldots, n$, be the modified unitary submatrices defined by* (3). *If we number the eigenvalues of $\tilde{H}_k$ starting from -1 moving counterclockwise along the unit circle, then the eigenvalues of $\tilde{H}_k$ interlace those of $\tilde{H}_{k+1}$ in the following sense: the jth eigenvalue of $\tilde{H}_k$ lies on the arc between the jth and the $j+1$st eigenvalue of $\tilde{H}_{k+1}$.*

Example 1. Let us consider a 4×4 unitary Hessenberg matrix with

$\gamma_1 = -0.5 + i0.3,$
$\gamma_2 = 0.4 + i0.21,$
$\gamma_3 = 0.8 - i0.1,$
$\gamma_4 = i.$

Setting $\rho_4 = \gamma_4$, we calculate $\rho_k, k = 1, 2, 3$ by (6). The eigenvalues of $\tilde{H}_k$ are denoted by $\lambda_1^k, \ldots, \lambda_k^k$. Figure 1 illustrates the interlacing property.

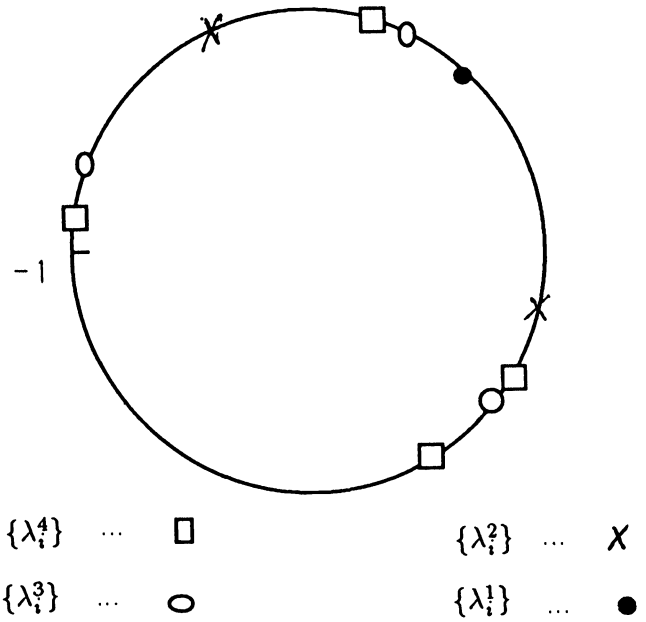


FIG. 1.

3. Three-term recurrence and cutting points. We denote the characteristic polynomials of $\tilde{H}_k$ by $p_k(z)$, i.e., $p_k(z) = \det(zI - \tilde{H}_k)$. It is easy to show that the $p_k(z)$ satisfy the following three-term recurrence relations:

$$p_0(z) = 1, p_1(z) = z + \rho_1,$$

$$p_{k+1}(z) = (z + \rho_{k+1}\bar{\rho}_k)p_k(z) - \alpha_k z p_{k-1}(z), k = 1, \ldots, n-1, \tag{10}$$

where

$$\alpha_k = \bar{\gamma}_{k-1}(\rho_k - \gamma_k) + \rho_{k+1}(\bar{\rho}_k - \bar{\gamma}_k) \tag{11}$$

with $\gamma_0 = 1$. It can be shown that the $p_k(z)$ differs from $q_k(z)$ given by (1) by a constant, namely, $p_k(z) = \frac{1}{\beta_0\beta_1\ldots\beta_{k-1}} q_k(z)$. As seen in Corollary 2.4 the roots of $p_k(z)$ interlace those of $p_{k+1}(z)$. We refer to $p_k(z)$ as the Sturm sequence of polynomials corresponding to H, because they are the analogues to the Sturm sequence of characteristic polynomials of symmetric tridiagonal matrices.

As an immmediate consequence of (10) we make the following observation.

LEMMA 3.1. *Let $\tilde{H}_k, k = 1, \ldots, n$, be the modified unitary submatrices defined by* (3) *and let all $\sigma_1, \ldots, \sigma_{n-1}$ be nonzero. Then $\tilde{H}_k$ and $\tilde{H}_{k+1}$ have no common eigenvalues.*

Proof. All $\sigma_1, \ldots, \sigma_{n-1}$ being nonzero implies that $\alpha_k \neq 0$ for $k = 2, \ldots, n-1$. Thus from (10) we see that if $\tilde{H}_\ell$ and $\tilde{H}_{\ell+1}$ have a common eigenvalue ξ then ξ is an eigenvalue of all $\tilde{H}_k$. It is then easily seen that this can only hold if ξ is the cutting point -1. □

Figure 1 shows that the cutting point -1 is a common starting point to number the eigenvalues for all $p_k(z)$ such that we get our root interlacing for every two consecutive polynomials in the sequence. The choice -1 as a cutting point is for convenience. Any point ξ_0 on the unit circle that is not an eigenvalue of H may serve as a cutting point and we must then use the generalized Cayley transformation with respect to ξ_0

$$A = i(\xi_0 I - H)^{-1}(\xi_0 I + H)$$

and the definition of the $\rho_n, \ldots, \rho_1$ must be adapted to: $\rho_n = \gamma_n$, and

$$\rho_k = \frac{\gamma_k + \bar{\xi}_0 \rho_{k+1}}{1 + \bar{\xi}_0 \bar{\gamma}_k \rho_{k+1}} \quad k = n-1, \ldots, 1. \tag{12}$$

The $\rho_k, k = 1, \ldots, n$, are well defined, because $|\gamma_k| < 1$. The generalized Cayley transformation then maps the modified unitary submatrices of H defined with this choice of $\rho_n, \ldots, \rho_1$ to A_k, the kth leading principal submatrices of $A = i(\xi_0 I - H)^{-1}(\xi_0 I + H)$ and Lemmas 2.2, 2.3, and 3.1 and Theorem 2.1 still hold when -1 is replaced in the suitable way by ξ_0.

The three-term recurrence relation for the Sturm sequence of polynomials, $p_k(z) = \det(zI - \tilde{H}_k)$, is exactly the one given by (10) and (11), but with ρ_k from (12). The roots of $p_k(z)$ now interlace those of $p_{k+1}(z)$ on the unit circle in the following sense. If we number the roots of $p_k(z)$ starting from ξ_0 moving counterclockwise along the unit circle, then the jth root of $p_k(z)$ lies on the arc between the jth root and $(j+1)$st root of $p_{k+1}(z)$, $j = 1, \ldots, k$.

So far we have had to assume that the cutting point is not an eigenvalue of H. An easy way to check whether ξ_0 is an eigenvalue of H is by looking at the sequence $\{\rho_k\}$ defined in (12). In fact, ξ_0 is an eigenvalue of H if and only if

$$\xi_0 + \rho_1 = 0.$$

In our matrix terms this can be seen in what follows.

Let all σ_k be nonzero. Then $H - \xi_0 I$ is singular if and only if the Schur complement of the $(n-1)$st trailing principal submatrix of $H - \xi_0 I$ is zero. With the notation of (5) for $k = 1$ this Schur complement is

$$\Delta = H_{11} - \xi_0 - H_{12}(H_{22} - \xi_0 I)^{-1} H_{21}.$$

Theorem 2.1, adapted to the general cutting point ξ_0, then gives an adaptation of (9), which for $k = 1$ reads

$$\tilde{H}_1 = \Delta + \xi_0.$$

Thus we have $\Delta = 0$ if and only if $-\rho_1 = \tilde{H}_1 = \xi_0$.

4. Counting eigenvalues. For a Sturm sequence of real polynomials $\{d_k(x)\}$, $\alpha(\mu)$, the number of sign agreements between consecutive terms of the numerical sequence $\{d_k(\mu), k = 0, 1, \ldots, n\}$, is the number of roots of $d_n(x)$, which are smaller than μ (see [27]). An analogous result can be derived for the Sturm sequence $\{p_k(z)\}$ of the unitary Hessenberg matrix H.

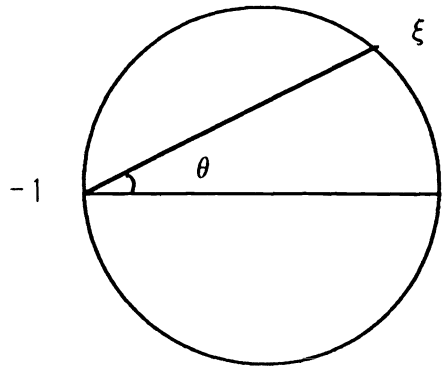


FIG. 2.

Consider the one-dimensional Cayley transformation on the unit circle

$$x = i\frac{1-\xi}{1+\xi}, \tag{13}$$

where $|\xi| = 1$. If $\xi = 1$, then $x = 0$. If ξ is on the upper semicircle then $x > 0$ and if ξ is on the lower semicircle then $x < 0$. A geometrical interpretation for ξ is given by the following formula and Fig. 2.

$$x = \tan(\theta) = \frac{\mathrm{im}\xi}{1 + \mathrm{re}\xi}.$$

The Cayley transformation is a strictly monotone function of $\theta = \arctan \frac{\mathrm{im}(\xi)}{1+\mathrm{re}(\xi)}$ (see Fig. 2). Thus for two points ξ and η on the unit circle it is reasonable to define $\xi \leq \eta$ if $i\frac{1-\xi}{1+\xi} \leq i\frac{1-\eta}{1+\eta}$. This gives a complete ordering of the points on the unit circle. Note that the complete ordering excludes the cutting point -1. Let $\{\lambda_j^k\}_{j=1}^k$ denote the k roots of $p_k(z)$ with

$$\lambda_1^k \leq \lambda_2^k \leq \cdots \leq \lambda_k^k.$$

The Cayley transformation (13) preserves the order and therefore the interlacing is preserved under the Cayley transformation. For the transformed points on the real line we get

$$i\frac{1-\lambda_1^k}{1+\lambda_1^k} \leq i\frac{1-\lambda_2^k}{1+\lambda_2^k} \leq \cdots \leq i\frac{1-\lambda_k^k}{1+\lambda_k^k}.$$

TABLE 1

ξ	$\alpha(\xi)$	$d_0(\xi)$	$d_1(\xi)$	$d_2(\xi)$	$d_3(\xi)$	$d_4(\xi)$
$-i$	0	+	-	+	-	+
1	2	+	-	-	+	+
$-0.9900 + 0.1411i$	4	+	+	+	+	+

For any ξ on the unit circle, we define $\alpha(\xi)$ as the number of sign agreements of the numerical sequence $\{d_k(\xi)\}$, $k = 0, 1, \ldots, n$, where

$$d_0(\xi) = 1,$$

$$d_k(\xi) = \left(i\frac{1-\xi}{1+\xi} - i\frac{1-\lambda_1^k}{1+\lambda_1^k}\right)\left(i\frac{1-\xi}{1+\xi} - i\frac{1-\lambda_2^k}{1+\lambda_2^k}\right)\cdots\left(i\frac{1-\xi}{1+\xi} - i\frac{1-\lambda_k^k}{1+\lambda_k^k}\right).$$

Via the Cayley transformation (13) the sequence $\{d_k(\xi)\}$ can be considered as a sequence of real polynomials with interlacing roots. Thus $\alpha(\xi)$ is the number of $i\frac{1-\lambda_1^n}{1+\lambda_1^n}, \ldots, i\frac{1-\lambda_n^n}{1+\lambda_n^n}$, which are less than $i\frac{1-\xi}{1+\xi}$ or the number of $\lambda_1^n, \ldots, \lambda_n^n$, which are smaller than ξ.

The evaluation of the functions $\{d_k(\xi)\}$ seems to be difficult on first sight. But noting that

$$i\frac{1-\xi}{1+\xi} - i\frac{1-\lambda_j^k}{1+\lambda_j^k} = \left(\frac{2i}{1+\xi}\right)\left(\frac{\xi-\lambda_j^k}{-1-\lambda_j^k}\right),$$

we fortunately find that

$$d_k(\xi) = \left(\frac{2i}{1+\xi}\right)^k \frac{p_k(\xi)}{p_k(-1)}. \tag{14}$$

Thus via (14) $d_k(\xi)$ can easily be computed by the three-term recurrence relation (10) for $p_k(z)$.

Summarizing the considerations above, we get the following theorem.

THEOREM 4.1. *Let $p_n(-1) \neq 0$. The number $\alpha(\xi)$ of sign agreements for consecutive terms of the numerical sequence $d_k(\xi)$, $k = 0, 1, \ldots, n$, is the number of roots of $p_n(z)$ which are smaller than ξ.*

For Example 1 the signs of $\{d_k(\xi),\ k = 0, 1, \ldots, n\}$ are given in Table 1.

From Theorem 4.1 it follows that for given two points ξ and η on the unit circle the number of roots of $p_n(z)$ on the arc between ξ and η is $\alpha(\eta) - \alpha(\xi)$. Thus we can compute the eigenvalues of the unitary Hessenberg matrix H by a bisection method based on the evaluation of the sequence $d_k(\xi)$.

5. A Christoffel–Darboux-type formula. A Christoffel–Darboux-type formula for the γ_n-symmetric polynomials has been derived by Delsarte and Genin [12], [20]. The purpose this section is to show how a Christoffel–Darboux-type formula for our unitary Sturm sequence of polynomials can be derived in matrix terms. We also show that H is similar to a Hermitian matrix, which is a product of a lower bidiagonal matrix and the inverse of an upper bidiagonal matrix and will thus shed light on the

relation between the unitary QR Hessenberg method [18] and the pencil method in [12].

According to the three-term recurrence relation (10), we have

$$(p_0(z),\dots,p_{n-1}(z))(Z_1 z - Z_2) = (0,\dots,0,p_n(z)), \tag{15}$$

where

$$Z_1 = \begin{pmatrix} 1 & -\alpha_1 & & \\ & 1 & \ddots & \\ & & \ddots & -\alpha_{n-1} \\ & & & 1 \end{pmatrix}, \quad Z_2 = \begin{pmatrix} -\rho_1 & & & \\ 1 & -\rho_2\bar{\rho}_1 & & \\ & \ddots & \ddots & \\ & & 1 & -\rho_n\bar{\rho}_{n-1} \end{pmatrix}.$$

The next theorem shows that the eigenvalue problem for H is equivalent to the eigenvalue problem for the matrix pencil $\lambda Z_1 - Z_2$, i.e., H is similar to $H_0 = Z_2 Z_1^{-1}$. H_0 is a Hessenberg matrix of the form

$$H_0 = \begin{pmatrix} -\rho_1 & -\rho_1\alpha_1 & \dots & -\rho_1\alpha_1\dots\alpha_{n-1} \\ 1 & (\alpha_1 - \rho_2\bar{\rho}_1) & \dots & (\alpha_1 - \rho_2\bar{\rho}_1)\alpha_2\dots\alpha_{n-1} \\ & \ddots & \ddots & \vdots \\ & & 1 & (\alpha_{n-1} - \rho_n\bar{\rho}_{n-1}) \end{pmatrix}$$

and the following can easily be proved.

THEOREM 5.1. *Let the matrix R be given by*

$$R = \begin{pmatrix} 1 & \frac{\rho_2\beta_1}{\sigma_1} & \frac{\rho_3\beta_1\beta_2}{\sigma_1\sigma_2} & \dots & \frac{\rho_n\beta_1\dots\beta_{n-1}}{\sigma_1\dots\sigma_{n-1}} \\ & \frac{1}{\sigma_1} & \frac{\bar{\gamma}_1\rho_3\beta_2}{\sigma_1\sigma_2} & \dots & \frac{\bar{\gamma}_1\rho_n\beta_2\dots\beta_{n-1}}{\sigma_1\dots\sigma_{n-1}} \\ & & \frac{1}{\sigma_1\sigma_2} & \dots & \frac{\bar{\gamma}_2\rho_n\beta_3\dots\beta_{n-1}}{\sigma_1\dots\sigma_{n-1}} \\ & & & \ddots & \vdots \\ & & & & \frac{1}{\sigma_1\dots\sigma_{n-1}} \end{pmatrix}$$

with $\beta_j = 1 - \bar{\gamma}_j\rho_j, j = 1,\dots,n-1$. Then $H = R^{-1}H_0R$.

We define dual polynomials $\{\tilde{p}_k(z)\}$ by

$$\tilde{p}_0(z) = z^{n-1}, \tilde{p}_1(z) = z^{n-2}(z + \rho_1),$$

$$\tilde{p}_{k+1}(z) = \bar{\alpha}_k z^{-1}((z + \rho_{k+1}\bar{\rho}_k)\tilde{p}_k(z) - \tilde{p}_{k-1}(z)), \quad k = 1,\dots,n-1. \tag{16}$$

Here we can show that $\tilde{p}_k(z) = z^{n-1-k}\bar{\alpha}_1\dots\bar{\alpha}_k p_k(z)$ for $k = 0,1,\dots,n-1$ and $\tilde{p}_n(z) = \bar{\alpha}_1\dots\bar{\alpha}_{n-1}p_n(z)$. These properties are called γ_n-symmetric properties by Delsarte and Genin. From (16) we get

$$(Z_1 w - Z_2)\begin{pmatrix} \tilde{p}_0(w) \\ \tilde{p}_1(w) \\ \vdots \\ \tilde{p}_{n-1}(w) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \tilde{p}_n(w) \end{pmatrix}. \tag{17}$$

Postmultiplying (15) by $(\tilde{p}_0(w), \ldots, \tilde{p}_{n-1}(w))^T$, premultiplying (17) by $(p_0(z), \ldots, p_{n-1}(z))$, and subtracting the results, we obtain the following Christoffel–Darboux-type formula.

$$(z-w)(p_0(z),\ldots,p_{n-1}(z)) \begin{pmatrix} 1 & -\alpha_1 & & \\ & 1 & \ddots & \\ & & \ddots & -\alpha_{n-1} \\ & & & 1 \end{pmatrix} \begin{pmatrix} \tilde{p}_0(w) \\ \tilde{p}_1(w) \\ \vdots \\ \tilde{p}_{n-1}(w) \end{pmatrix}$$

$$= p_n(z)\tilde{p}_{n-1}(w) - \tilde{p}_n(w)p_{n-1}(z).$$

6. Conclusions. For a unitary Hessenberg matrix H we defined a sequence of modified unitary submatrices, which are slight modifications of the kth leading principal submatrix of H. The characteristic polynomials of the modified unitary submatrices, $p_k(z)$, $k = 1, \ldots, n$, can be evaluated by a three-term recurrence relation and, up to a constant, they coincide with the γ_n-symmetric polynomials of Delsarte and Genin, which in turn are modifications of the Szegö polynomials.

We showed that the Cayley transforms of the modified unitary submatrices are the leading principal submatrices of the Cayley transform of H itself. This allowed us to view their characteristic polynomials as a Sturm sequence of polynomials for the unitary Hessenberg matrix H. For a given ξ on the unit circle we derived from $p_1(\xi), \ldots, p_n(\xi)$ a sequence of real numbers $d_1(\xi), \ldots, d_n(\xi)$, such that the number of sign agreements of consecutive terms in the sequence is the number of eigenvalues of H that lie between ξ_0 and ξ when moving counterclockwise along the unit circle.

We can therefore determine the number of eigenvalues between any ξ and η on the unit circle by computing the corresponding sequence of d's and have thus developed a basis for a bisection method for the unitary eigenvalue problem.

REFERENCES

[1] G. S. AMMAR, W. B. GRAGG, AND L. REICHEL, *On the Eigenproblem for Orthogonal Matrices*, in Proc. 25th IEEE Conference on Decision and Control, Athens, Greece, 1986, pp. 1963–1966.

[2] ———, *Determination of Pisarenko frequency estimates as eigenvalues of an orthogonal matrix*, in Advanced Algorithms and Architectures for Signal Processing II, VOL. 826, F.T. Luk, ed., Proc. SPIE, Internat. Soc. for Optical Engineering, 1987, pp. 143–145.

[3] ———, *Constructing a Unitary Hessenberg Matrix from Spectral Data*, in Numerical Linear Algebra, Digital Signal Processing, and Parallel Algorithms, G. Golub and P. V. Dooren, eds., Springer-Verlag, Berlin, 1991, pp. 385–396.

[4] G. S. AMMAR, L. REICHEL, AND D. C. SORENSEN, *An implementaion of a divide and conquer algorithm for the unitary eigenproblem*, ACM Trans. Math. Software, 1992, to appear.

[5] A. BRUCKSTEIN AND T. KAILATH, *Some matrix factorization identities for discrete inverse scattering*, Linear Algebra Appl., 74 (1986), pp. 157–172.

[6] A. BUNSE-GERSTNER AND L. ELSNER, *Schur parameter pencils for the solution of the unitary eigenproblem*, Linear Algebra Appl., 154 (1991), pp. 741–778.

[7] G. CYBENKO, *Computing Pisarenko Frequency Estimates*, in Proc. Information Systems and Sciences, Princeton, NJ, 1984, pp. 587–591.

[8] ———, *Fast approximation of dominate harmonics*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 317–331.

[9] ———, *Restrictions of normal operators, Padé approximation and autogressive time series*, SIAM J. Math. Anal., 15 (1984), pp. 753–767.

[10] P. DELSARTE AND Y. GENIN, *The Split Levinson Algorithm*, IEEE Trans. ASSP, 34 (1986), pp. 470–478.

[11] ———, *The tridiagonal approach to Szegö's orthogonal polynomials, Toeplitz linear systems, and related interpolation problems*, SIAM J. Math. Anal., 19 (1988), pp. 718–735.

[12] ———, *Tridiagonal approach to the algebraic environment of Toeplitz matrices, part* II*: zero and eigenvalue problems*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 432–448.

[13] ———, *Tridiagonal approach to the algebraic environment of Toeplitz matrices, part* I*: basic results*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 220–238.

[14] P. DELSARTE, Y. GENIN, Y. KAMP, AND P. VAN DOOREN, *Speech modelling and the trigonometric moment problem*, Philips J. Res., 37 (1982), pp. 277–292.

[15] P. EBERLEIN AND C. P. HUANG, *Global convergence of the QR algorithm for unitary matrices with some results for normal matrices*, SIAM J. Numer. Anal., 12 (1975), pp. 97–104.

[16] L. ELSNER AND C. HE, *Perturbation and interlace theorems for the unitary eigenvalue problem*, Linear Algebra Appl., 188/189 (1993), pp. 207 – 230.

[17] W. B. GRAGG, *Positive Definite Toeplitz Matrices, the Arnoldi Process for Isometric Operators, and Gaussian Quadrature on the Unit Circle*, in Numerical Methods in Linear Algebra, E. S. Nikolaev, ed., Moscow University Press, Moscow, 1982, pp. 16–32. (In Russian.)

[18] ———, *The QR algorithm for unitary Hessenberg matrices*, J. Comput. Appl. Math., 16 (1986), pp. 1–8.

[19] W. B. GRAGG AND L. REICHEL, *A divide and conquer method for the unitary and orthogonal eigenproblem*, Numer. Math., 57 (1990), pp. 695–718.

[20] U. GRENANDER AND G. SZEGÖ, *Toeplitz Forms and their Applications*, Cambridge University Press, Cambridge, 1958.

[21] S. Y. KUNG, *A Toeplitz Approximation Method and Some Applications*, in International Symposium on Mathematical Theory of Networks and Systems, Santa Monica, CA, 1981.

[22] K. MAKHOUL, *Linear Prediction: A Tutorial Review*, Proc. IEEE, 63 (1975), pp. 561–580.

[23] V. F. PISARENKO, *The retrieval of harmonics from a covariance function*, Geophys. J. R. Astr. Soc., 33 (1973), pp. 347–366.

[24] L. REICHEL AND G. S. AMMAR, *Fast Approximation of Dominant Harmonics by Solving an Orthogonal Eigenvalue Problem*, in Proc. Second IMA Conference on Mathematics in Signal Processing, J. McWhirter et al., ed., Oxford University Press, Oxford, 1990.

[25] H. RUTISHAUER, *Bestimmung der Eigenwerte orthogonaler Matrizen*, Numer. Math., 9 (1966), pp. 104–108.

[26] S. SAGAYAMA AND F. ITAKURA, *Duality theory of composite sinusoidal modelling and linear prediction*, in Proc. Internat. Acoustics Speech Signal Processing, Tokyo, 1986, pp. 1261–1264.

[27] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, pp. 1056–1073, October 1995

LEAST SQUARES SIGN-SOLVABILITY*

BRYAN L. SHADER†

This paper is dedicated to John Maybee in honor of his 65th birthday and his numerous contributions to mathematics.

Abstract. Let $Ax = b$ be a linear system. We study the relationship between the sign pattern of the least squares solution to $Ax = b$ and the sign patterns of A and b. The system $Ax = b$ is least squares sign-solvable if the signs of the entries in its least square solution can be determined solely from the signs of the entries of A and b. We construct a family of least squares sign-solvable linear systems from the vertex-incidence matrices of trees. General properties of least squares sign-solvable linear systems are developed and the structure of a least squares sign-solvable system is shown to be analogous to that of sign-solvable linear systems. Square matrices whose sign pattern determines the sign pattern of its inverse have been extensively studied. We study matrices whose sign pattern determines the sign pattern of its generalized inverse.

Key words. least squares solution, sign-solvability, generalized inverse

AMS subject classifications. 05B20, 05C20, 15A09

1. Introduction.

Let $p, q, \ldots, v$ be positive numbers and consider the linear system $Ax = b$ where

$$A = \begin{bmatrix} p & r & t \\ q & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & u \end{bmatrix} \text{ and } b = \begin{bmatrix} 0 \\ 0 \\ 0 \\ v \end{bmatrix}.$$

Since the matrix of order 3 obtained from A by deleting its last row is invertible, $Ax = b$ has no solution. However the least squares solution to $Ax = b$ has an interesting property. First note that the columns of A are linearly independent regardless of the magnitudes of the numbers $p, q, \ldots, u$. Hence, the least squares solution to $Ax = b$ is the solution to the normal equation $A^TAx = A^Tb$. Computing the normal equation we have

$$\begin{bmatrix} p^2+q^2 & pr & pt \\ pr & r^2+s^2 & rt \\ pt & rt & t^2+u^2 \end{bmatrix} x = \begin{bmatrix} 0 \\ 0 \\ uv \end{bmatrix}. \tag{1}$$

The solution to the linear system (1)

$$\frac{uv}{\det A^TA} \begin{bmatrix} -pts^2 \\ -q^2rt \\ p^2s^2+q^2r^2+q^2s^2 \end{bmatrix}.$$

Since the columns of A are linearly independent, A^TA is a positive definite matrix and in particular $\det A^TA > 0$. It follows that regardless of the magnitudes of $p, q, \ldots, v$, the first two entries of the least squares solution to $Ax = b$ are negative and the last entry of the least squares solution is positive. Thus the signs of the entries of the least

* Received by the editors November 1993; accepted for publication (in revised form) by J. Maybee August 1, 1994. This work was partially supported by National Security Agency grant MDA904-94-H-2051.

† Department of Mathematics, University of Wyoming, Laramie, Wyoming 82071 (bshader@uwyo.edu).

squares solution to $Ax = b$ depend only on the signs of the entries of A and of b. We study such linear systems in this paper.

To be more precise, we make the following definitions. The *sign* of a real number a is 1 if a is positive, 0 if a is zero and -1 if a is negative. Throughout we consider only matrices with real entries. Let $A = [a_{ij}]$ be an m by n matrix. The *qualitative class*, $\mathcal{Q}(A)$, of A is the set of all m by n real matrices $\widetilde{A} = [\tilde{a}_{ij}]$ such that the sign of $\tilde{a}_{ij}$ and the sign of a_{ij} are equal for all i and j. The *sign pattern* of A is the unique $(0, 1, -1)$-matrix in $\mathcal{Q}(A)$. Let b be an n by 1 column vector. The linear system $Ax = b$ is a *least squares sign-solvable* provided the vectors in

$$\{u : \text{there exist } \widetilde{A} \in \mathcal{Q}(A) \text{ and } \tilde{b} \in \mathcal{Q}(b) \text{ such that } \|\widetilde{A}u - \tilde{b}\| = \min_{x \in R^m} \|\widetilde{A}x - \tilde{b}\|\}$$

are contained in a single qualitative class. The above example shows that

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} x = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \tag{2}$$

is a least squares sign-solvable system. Other examples of least squares sign-solvable linear systems are given in §2.

The notion of least squares sign-solvability generalizes the previously studied notion of sign-solvability (see [6], [11]). The linear system $Ax = b$ is *sign-solvable system* provided each of the systems

$$\widetilde{A}x = \tilde{b} \qquad (A \in \mathcal{Q}(A), b \in \mathcal{Q}(b))$$

has a solution and the vectors in

$$\{u : \text{ there exists } \widetilde{A} \in \mathcal{Q}(A) \text{ and } \tilde{b} \in \mathcal{Q}(b) \text{ such that } \widetilde{A}u = \tilde{b} \}$$

belong to a single qualitative class. Clearly a sign-solvable linear system is least squares sign-solvable. The linear system in (2) is least squares sign-solvable but is not sign-solvable. The notion of sign-solvability has been generalized in other ways (see [1], [3], [9]).

In [11] the structure of sign-solvable linear systems is described in terms of two special classes of matrices, called *L-matrices* and *S^*-matrices*. An m by n matrix is an *L-matrix* provided every matrix in its qualitative class has linearly independent rows. If $m = n$, an L-matrix is called a *sign nonsingular* matrix, which we abbreviate to SNS-matrix. A matrix is an SNS-matrix if and only if it has a nonzero term in its determinant expansion and each of the nonzero terms in its determinant expansion have the same sign. An S^*-matrix is an m by $m+1$ matrix such that every submatrix of order m is an SNS-matrix. An m by $m+1$ matrix is an S^*-matrix if and only if there exists a $(1, -1)$-vector u such that for each $\widetilde{A} \in \mathcal{Q}(A)$ the null space of $\widetilde{A}$ is contained in $\mathcal{Q}(u) \cup \mathcal{Q}(-u) \cup \mathcal{Q}(0)$. If $u = (1, 1, \ldots, 1)^T$, then A is an *S-matrix*. The structure of least squares sign-solvable linear systems is described in §2.

Let A be an SNS-matrix. Then every matrix $\widetilde{A}$ in the qualitative class of A has an inverse. However, in general, $\widetilde{A}^{-1}$ need not be in the qualitative class of A^{-1}. The matrix A is a *strong* SNS-*matrix* (S^2NS-matrix) if $\widetilde{A}^{-1} \in \mathcal{Q}(A^{-1})$ for all $\widetilde{A} \in \mathcal{Q}(A)$. Thus if A is an S^2NS-matrix, then the signs of the entries of its inverse depend only

on the signs of the entries of A. Irreducible[1] S^2NS-matrices are characterized in [5], [14] and S^2NS-matrices have also been studied in [7], [12], [13].

In this paper we generalize S^2NS-matrices to the least squares setting (see [1], [6] for other generalizations of strong sign-nonsingularity). Let M be an m by n matrix of rank n. The *generalized inverse* of M, denoted by $M^\dagger$, is the matrix $(M^TM)^{-1}M^T$, and that the least squares solution to $Mx = b$ is $M^\dagger b$. Now let A be an m by n matrix such that A^T is an L-matrix. Then A has a *signed generalized inverse* if the matrices in

$$\{\widetilde{A}^\dagger : \widetilde{A} \in \mathcal{Q}(A)\}$$

belong to the same qualitative class. Note that if $m = n$, then A has a signed generalized inverse if and only if A is an S^2NS-matrix. It is easy to verify that if $a, b, c, d > 0$ then

$$\begin{bmatrix} a & 0 \\ b & c \\ 0 & d \end{bmatrix}^\dagger = \frac{1}{a^2c+a^2d^2+b^2d^2} \begin{bmatrix} a(c^2+d^2) & bd^2 & -bcd \\ -abc & ca^2 & d(a^2+b^2) \end{bmatrix}.$$

Hence

$$\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$$

has a signed generalized inverse. Matrices with sign generalized inverses and their relationship to vertex-edge incidence matrices of trees are studied in §3.

2. Least squares sign-solvability. Recall that the linear system $Ax = b$ is least squares sign-solvable if the signs of the entries of its least squares solution can be determined solely from the signs of the entries of A and of b. Clearly, if P and Q are permutation matrices, and D and E are invertible diagonal matrices, then $Ax = b$ is least squares sign-solvable if and only if $(PDAQE)x = PDb$ is. It is known that if $Ax = b$ is sign-solvable [11], then A^T is an L-matrix. We extend this result to least squares sign-solvability.

LEMMA 2.1. *If $Ax = b$ is a least squares sign-solvable linear system, then A^T is an L-matrix.*

Proof. Suppose that A^T is not an L-matrix. Then there exists a matrix $\widetilde{A} \in \mathcal{Q}(A)$ and a nonzero vector y such that $\widetilde{A}y = 0$. Let z be a least squares solution to $\widetilde{A}x = b$. Then since $\widetilde{A}(z + \lambda y) = \widetilde{A}z$, the vector $z + \lambda y$ is also a least squares solution to $\widetilde{A}x = b$ for all real numbers λ. For some choice of λ the vectors z and $z + \lambda y$ belong to different qualitative classes, therefore $Ax = b$ is not least squares sign-solvable. □

In [3] the notion of sign-solvability is generalized by relaxing the requirement that each system $\widetilde{A}x = \tilde{b}$ has a solution. The linear system $Ax = b$ is *conditionally sign-solvable* provided there is a matrix $\widetilde{A}$ in $\mathcal{Q}(A)$ and a vector $\tilde{b}$ in $\mathcal{Q}(b)$ such that $\widetilde{A}x = \tilde{b}$

[1] A square matrix X is *irreducible* provided there does not exist a permutation matrix P such that PXP^T has the form

$$\begin{bmatrix} X_1 & O \\ X_3 & X_2 \end{bmatrix},$$

where X_1 and X_2 are square nonvacuous matrices.

has a solution and each of the vectors in

$$\{u : \text{there exists } \tilde{A} \in \mathcal{Q}(A) \text{ and } \tilde{b} \in \mathcal{Q}(b) \text{ with } \tilde{A}u = \tilde{b}\}$$

belong to the same qualitative class. Thus if $Ax = b$ is conditionally sign-solvable then the signs of the entries of a solution to $\tilde{A}x = \tilde{b}$, provided a solution exists, are determined by the signs of the entries of A and of b. Clearly, any sign-solvable linear system is also conditionally sign-solvable. Many of the results for sign-solvability can be generalized to conditional sign-solvability. In particular, in [3] it is shown that if $Ax = b$ is conditionally sign-solvable then A^T is an L-matrix, and that the structure of conditionally sign-solvable linear systems can be described in terms of L-matrices and generalizations of S-matrices known as conditionally S-matrices. The following corollary relates least squares sign-solvable linear systems and conditionally sign-solvable systems.

COROLLARY 2.2. *If $Ax = b$ is a least squares sign-solvable linear system, then either $Ax = b$ is conditionally sign-solvable or $\begin{bmatrix} A & -b \end{bmatrix}^T$ is a an L-matrix.*

Proof. Suppose that $Ax = b$ is least squares sign-solvable and that $[A\ -b]^T$ is not an L-matrix. Then there exists a matrix $\hat{A} \in \mathcal{Q}(A)$ and a vector $\hat{b} \in \mathcal{Q}(A)$ such that the columns of $[\hat{A}\ -\hat{b}]$ are linearly dependent. By Lemma 2.1, A^T is an L-matrix, and hence the columns of $\hat{A}$ are linearly independent. It follows that $\hat{A}x = \hat{b}$ has a solution. Since

$$\{u : \text{there exists } \tilde{A} \in \mathcal{Q}(A) \text{ and } \tilde{b} \in \mathcal{Q}(b) \text{ such that } \tilde{A}u = \tilde{b}\}$$

is a subset of

$$\Big\{u : \text{there exists } \tilde{A} \in \mathcal{Q}(A) \text{ and } \tilde{b} \in \mathcal{Q}(b) \text{ such that } u \text{ is a least squares solution to } \tilde{A}u = \tilde{b}\Big\},$$

it follows that $Ax = b$ is conditionally sign-solvable. □

If the columns of A are linearly independent, then the zero vector is the least squares solution to $Ax = b$ if and only if each column of A is orthogonal to b. Two vectors $u = (u_1, u_2, \ldots, u_n)^T$ and $v = (v_1, v_2, \ldots, v_n)^T$ are *combinatorially orthogonal* if $\tilde{u}^T\tilde{v} = 0$ for all $\tilde{u} \in \mathcal{Q}(u)$ and all $\tilde{v} \in \mathcal{Q}(v)$. Clearly, u and v are combinatorially orthogonal if and only if $u_iv_i = 0$ for $i = 1, 2, \ldots, n$.

COROLLARY 2.3. *The linear system $Ax = b$ is least squares sign-solvable and its solution is the zero vector if and only if A^T is an L-matrix and each column of A is combinatorially orthogonal to b.*

Proof. Suppose that A^T is an L-matrix and that each column of A is combinatorially orthogonal to b. Let $\tilde{A} \in \mathcal{Q}(A)$ and $\tilde{b} \in \mathcal{Q}(b)$. Then the least squares solution to $\tilde{A}x = \tilde{b}$ is the solution to $\tilde{A}^T\tilde{A}x = \tilde{A}^T\tilde{b}$. Since A^T is an L-matrix $\tilde{A}^T\tilde{A}$ is nonsingular and since $\tilde{A}^T\tilde{b} = 0$, the only solution to $\tilde{A}^T\tilde{A}x = \tilde{A}^T\tilde{b}$ is the zero vector. Hence $Ax = b$ is least squares sign-solvable and its solution is the zero vector.

Conversely, suppose that $Ax = b$ is least squares sign-solvable and its solution is the zero vector. By Lemma 2.1, A^T is an L-matrix. If $\tilde{A} \in \mathcal{Q}(A)$ and $\tilde{b} \in \mathcal{Q}(b)$, then since the zero vector is the least squares solution to $\tilde{A}x = \tilde{b}$ we have $\tilde{A}^T\tilde{b} = \tilde{A}^T\tilde{A}(0) = 0$. It follows that each column of A is combinatorially orthogonal to column b. □

COROLLARY 2.4. *Let*

$$A = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \text{ and } b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

be m by 1 matrices. Then $Ax = b$ is least squares sign-solvable if and only if some entry of A is nonzero and the numbers $a_i b_i$ $(i = 1, 2, \ldots, m)$ are all nonpositive or all nonnegative.

Proof. Since A is an m by 1 matrix, A^T is an L-matrix if and only if A has a nonzero entry. Lemma 2.1 implies that if $Ax = b$ is least squares sign-solvable, then A has a nonzero entry. Thus, we may assume that A has a nonzero entry. If $\tilde{A} \in \mathcal{Q}(A)$ and $\tilde{b} \in \mathcal{Q}(b)$, then the least squares solution to $\tilde{A}x = \tilde{b}$ is $(1/\ell)\tilde{A}^T\tilde{b}$ where ℓ is the sum of the squares of the entries of $\tilde{A}$. Hence, $Ax = b$ is least squares sign-solvable if and only if the sign of $\tilde{A}^T\tilde{b}$ is independent of the choice of $\tilde{A}$ and $\tilde{b}$. The corollary now follows. □

Note that there do exist least squares sign-solvable linear systems $Ax = b$ such that the sign pattern of $\tilde{A}^T\tilde{b}$ depends on the choice of $\tilde{A} \in \mathcal{Q}(A)$ and $\tilde{b} \in \mathcal{Q}(b)$. For example, let

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \text{ and } b = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Then $Ax = b$ is sign-solvable, and hence least squares sign-solvable. But the sign pattern of $\tilde{A}^T\tilde{b}$ is not determined by the sign pattern of A and of b.

COROLLARY 2.5. *Let A_1 and A_2 be m_1 by n_1 and m_2 by n_2 matrices, respectively, and let b_1 and b_2 be m_1 by 1 and m_2 by 1 column vectors, respectively. Then*

$$\begin{bmatrix} A_1 & O \\ O & A_2 \end{bmatrix} x = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \tag{3}$$

is least squares sign-solvable if and only if both $A_1 u = b_1$ and $A_2 v = b_2$ are least squares sign-solvable.

Proof. Let

$$A = \begin{bmatrix} A_1 & O \\ O & A_2 \end{bmatrix}.$$

Clearly, A^T is an L-matrix if and only if both A_1^T and A_2^T are L-matrices. By Lemma 2.1 we may assume that A^T, and hence A_1^T and A_2^T are L-matrices. The normal equation of (3) is

$$\begin{bmatrix} A_1^T A_1 & O \\ O & A_2^T A_2 \end{bmatrix} x = \begin{bmatrix} A_1^T b_1 \\ A_2^T b_2 \end{bmatrix},$$

which is equivalent to

$$A_1^T A_1 u = A_1^T b_1 \text{ and } A_2^T A_2 v = A_2^T b_2,$$

where $x = [u\ v]^T$. It now follows that the sign pattern of the solution to the normal equation of (3) is determined by the sign patterns of A, b_1, and b_2 if and only if the

sign pattern of the solution to the normal equation of $A_1u = b_1$ is determined by the sign patterns of A_1 and b_1, and the sign pattern of the solution to the normal equation of $A_2v = b_2$ is determined by the sign patterns of A_2 and b_2. □

Before further developing the theory of least squares sign-solvability we give a family of examples. A matrix A of order n has an *identically zero determinant* if the determinant of each matrix in $\mathcal{Q}(A)$ equals 0. By the Frobenius–König theorem, A has an identically zero determinant if and only if A contains a k by ℓ zero submatrix for some positive integers k and ℓ with $k+\ell > n$. Let G be a graph with vertices 1, 2, ..., m and edges e_1, e_2, ..., e_n. The *vertex-edge incidence matrix* of G is the m by n $(0,1)$-matrix $A = [a_{ij}]$ such that $a_{ij} = 1$ if and only if vertex i belongs to the edge e_j. A *tree* is an acyclic connected graph. Thus a tree with m vertices has $m-1$ edges. Let A be the vertex-edge incidence matrix of a tree. It is easy to verify that A^T is an L-matrix, and that the rows and columns of each square submatrix of A can be permuted to obtain a lower triangular matrix. Thus, each square submatrix of A is either an SNS-matrix or has an identically zero determinant. Moreover, this implies that each square submatrix of A is either an S^2NS-matrix or has an identically zero determinant. We now construct m least squares sign-solvable linear systems for each tree with m vertices.

Let X be an m by n matrix. If α is a subset of $\{1,2,\ldots,m\}$ and β is a subset of $\{1,2,\ldots,n\}$ then $X[\alpha,\beta]$ denotes the submatrix of X whose rows have index in α and whose columns have index in β. By $\overline{\alpha}$, respectively $\overline{\beta}$, we mean the complement of α in $\{1,2,\ldots,m\}$, respectively of β in $\{1,2,\ldots,n\}$. Thus $X[\overline{\alpha},\overline{\beta}]$ is the submatrix of X obtained by deleting the rows with index in α and the columns with index in β. If $\alpha = \{1,2,\ldots,m\}$, we abbreviate $X[\alpha,\beta]$ to $X[:,\beta]$. Similarly, if $\beta = \{1,2,\ldots,n\}$ we abbreviate $X[\alpha,\beta]$ to $X[:,\beta]$. If X is a column vector, then we abbreviate $X[\alpha,:]$ to simply $X[\alpha]$. If b is an m by 1 column vector, and j is an integer with $1 \le j \le n$ then $X_{j\leftarrow b}$ denotes the m by n matrix obtained from X by replacing its jth column by b.

THEOREM 2.6. *Let $A = [a_{ij}]$ be an m by $m-1$ $(0,1)$-matrix which is the vertex-edge incidence matrix of a tree with $m \ge 2$ vertices. Let v be a vertex of the tree and let b be the n by 1 column vector with a 1 in row v and 0's elsewhere. Then the linear system $Ax = b$ is least squares sign-solvable.*

Proof. Let T be the tree with vertices 1, 2, ..., m and edges e_1, e_2, ..., e_{m-1} such that A is the vertex edge-incidence matrix of T. By replacing A by QAP, x by P^Tx, and b by Qb where P and Q are appropriate permutation matrices, we may assume without loss of generality that $v = m$. Thus $b_m = 1$ is the only nonzero entry in b. Let $\widetilde{A} \in \mathcal{Q}(A)$ and let $\tilde{b} \in \mathcal{Q}(b)$. Since A^T is an L-matrix, the columns of $\widetilde{A}$ are linearly independent and hence the least squares solution to $\widetilde{A}x = \tilde{b}$ is the solution to the normal equation

$$\widetilde{A}^T\widetilde{A}x = \widetilde{A}^T\tilde{b}. \tag{4}$$

Let $u = (u_1, u_2, \ldots, u_{m-1})^T$ be the solution to (4) and let j be a fixed integer with $1 \le j \le m-1$. Since $\widetilde{A}^T\widetilde{A}$ is positive definite, it follows from Cramer's rule that the sign of u_j equals the sign of

$$\det\,(\widetilde{A}^T\widetilde{A})_{j\leftarrow\widetilde{A}^T\tilde{b}}.$$

Note that

$$(\widetilde{A}^T\widetilde{A})_{j\leftarrow\widetilde{A}^T\tilde{b}} = \widetilde{A}^T(\widetilde{A}_{j\leftarrow\tilde{b}}).$$

Thus it suffices to show that the sign of $\det \widetilde{A}^T(\widetilde{A}_{j\leftarrow\tilde{b}})$ is independent of the choice of $\widetilde{A}$ in $\mathcal{Q}(A)$ and $\tilde{b} \in \mathcal{Q}(b)$.

By the Cauchy–Binet determinantal formula we have

$$\det \widetilde{A}^T(\widetilde{A}_{j\leftarrow\tilde{b}}) = \sum_{i=1}^{m} \det \widetilde{A}[\overline{\{i\}}, :] \det \widetilde{A}_{j\leftarrow\tilde{b}}[\overline{\{i\}}, :]. \tag{5}$$

The graph T' obtained from T by removing the edge e_j has two connected components. Let α be the set of vertices of the connected component of T' which does not contain vertex m, and let β be the set consisting of the remaining vertices. Let γ be the set of all k such that e_k is an edge of the connected component of T' which does not contain vertex m, and let δ be the set of all k such that e_k is an edge of the connected component of T' which contains m. Thus, $\alpha \cup \beta = \{1, 2, \ldots, m\}$, $\gamma \cup \delta = \{1, 2, \ldots, m-1\} \setminus \{j\}$, $|\alpha| = |\gamma| + 1$ and $|\beta| = |\delta| + 1$. In addition, $A[\alpha, \delta] = O$ and $A[\beta, \gamma] = O$. By permuting the first $m-1$ rows and the columns of A we may assume without loss of generality that $j = m-1$, the elements of α come before those of β, and the elements of γ come before those of δ. This may change the actual sign of u_j but will not change the fact the sign is determined by the signs of the entries of A and of b. If $i \in \beta$ then $A_{j\leftarrow b}[\alpha, \delta \cup j]$ is a zero submatrix of $A_{j\leftarrow b}[\overline{\{i\}}, :]$, and since $|\alpha| + |\delta| = m - 1$ we conclude that $\det A_{j\leftarrow\tilde{b}}[\overline{\{i\}}, :] = 0$. Suppose that $i \in \alpha$. Since $A[\beta, \gamma]$ is a zero submatrix of $A[\overline{\{i\}}, :]$ the matrix $A[\overline{\{i\}}, :]$ is block upper triangular and hence

$$\det \widetilde{A}[\overline{\{i\}}, :] = \det \widetilde{A}[\alpha \setminus \{i\}, \gamma] \det \widetilde{A}[\beta, \delta \cup \{j\}].$$

Because $A_{j\leftarrow b}[\alpha \setminus \{i\}, \delta \cup \{j\}]$ and $A_{j\leftarrow b}[\beta, \gamma]$ are zero submatrices of $A_{j\leftarrow b}$, the matrix $A_{j\leftarrow b}$ is the direct sum of $A_{j\leftarrow b}[\alpha \setminus \{i\}, \gamma]$ and $A_{j\leftarrow b}[\beta, \gamma \cup \{j\}]$. Since the only nonzero entry in $\tilde{b}$ is the entry $\tilde{b}_m$ in its last row, it follows that

$$\det \widetilde{A}_{j\leftarrow b}[\overline{\{i\}}, :] = \tilde{b}_m \det \widetilde{A}[\alpha \setminus \{i\}, \gamma] \det \widetilde{A}[\beta \setminus \{m\}, \gamma].$$

Therefore, we have

$$\det \widetilde{A}[\overline{\{i\}}, :] \det \widetilde{A}_{j\leftarrow\tilde{b}}[\overline{\{i\}}, :] = \tilde{b}_m (\det \widetilde{A}[\alpha \setminus \{i\}, \gamma])^2 \det \widetilde{A}[\beta, \delta \cup \{j\}] \det \widetilde{A}[\beta \setminus \{m\}, \delta].$$

Since A is the vertex-edge incidence matrix of a tree each of $A[\alpha \setminus \{i\}, \gamma]$, $A[\beta, \delta \cup \{j\}]$ and $A[\beta \setminus \{m\}, \delta]$ is an SNS-matrix or has identically zero determinant. Thus if $A[\alpha \setminus \{i\}, \gamma]$ does not have an identically zero determinant, then the sign of

$$\det \widetilde{A}[\overline{\{i\}}, :] \det \widetilde{A}_{j\leftarrow\tilde{b}}[\overline{\{i\}}, :]$$

is independent of i and of the choice of $\widetilde{A}$ and $\tilde{b}$. Thus, by (5), the sign of $\det \widetilde{A}^T(\widetilde{A}_{j\leftarrow\tilde{b}})$, and hence the sign of u_j, is independent of the choice of $\widetilde{A}$ and $\tilde{b}$. We conclude that $Ax = b$ is least squares sign-solvable. □

Let A be an m by $m-1$ matrix which is the vertex-edge incidence matrix of a tree and let b be an m by 1 vector with exactly one nonzero entry. It is easy to verify that $[A\ b]^T$ is an L-matrix and hence it follows that the least squares sign-solvable linear systems described in Theorem 2.6 are not conditionally sign-solvable. If b is an

m by 1 nonzero vector and $A = b$, then $Ax = b$ is a least squares sign-solvable linear system which is also conditionally sign-solvable. It can be shown that

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix} x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

is conditionally sign-solvable but is not least squares sign-solvable.

Let B be an m by $m-1$ matrix such that the matrix A obtained from B by replacing each nonzero entry by a 1 is the vertex-edge incidence matrix of a tree. Then it is easy to show that there exist invertible diagonal matrices D and E such that $DBE = A$. It now follows from Theorem 2.6 that if b has exactly one nonzero entry, then $Bx = b$ is least squares sign-solvable.

COROLLARY 2.7. *Let A be an m by $m-1$ matrix which is the vertex-edge incidence matrix of a tree, and let b be a column vector with exactly one nonzero entry. Then for $j = 1, 2, \ldots, m-1$, the linear system*

$$A_{j \leftarrow b} x = A_j$$

is least squares sign-solvable, where A_j is the jth column of A.

Proof. The corollary follows easily from Theorem 2.6 and Corollary 2.5 by induction on m. □

We now study the structure of least squares sign-solvable linear systems. An m by n matrix A is *balanceable* provided there exists a diagonal matrix D of order m each of whose diagonal entries is 1 or -1 such that each nonzero column of DA has both a positive entry and a negative entry. If A is balanceable, then for each row j there exists a matrix $\widetilde{A} \in \mathcal{Q}(A)$ such that the jth row of $\widetilde{A}$ is a linear combination of the other rows of $\widetilde{A}$. If $Ax = b$ is a least squares sign-solvable linear system, then the ith-entry of the system is *exact* if for each $\widetilde{A} \in \mathcal{Q}(A)$ and each $\tilde{b} \in \mathcal{Q}(b)$ the ith-entry of $\tilde{b} - \widetilde{A}u$ is zero where u is the least squares solution to $\widetilde{A}x = \tilde{b}$. For example, it can be verified that only the third entry of the least squares sign-solvable linear system

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} x = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

is exact. Suppose that $Ax = b$ is a least squares sign-solvable linear system whose ith entry is exact. Let u be a least squares solution to $Ax = b$. Then $Au - b$ has a 0 in its ith entry and is orthogonal to the columns of A. Thus $A[\overline{\{i\}}, :]u - b[\overline{\{i\}}]$ is orthogonal to the columns of $A[\overline{\{i\}}, :]$. This implies that u is a least squares solution to $A[\overline{\{i\}}, :]x = b[\overline{\{i\}}]$.

LEMMA 2.8. *Let $Ax = b$ be a least squares sign-solvable linear system such that the least squares solution to $Ax = b$ has no zero entries. If A has no rows of all 0's and A is balanceable, then no entry of the system $Ax = b$ is exact.*

Proof. Assume that A has no rows of all 0's and that A is balanceable. Suppose to the contrary that the last entry of $Ax = b$ is exact. Let B be the matrix obtained from A by removing its last row and let w^T be the last row of A. Since A is balanceable, there exists a matrix $\widetilde{B} \in \mathcal{Q}(B)$ and a vector $\tilde{w} \in \mathcal{Q}(w)$ such that $\tilde{w}$ belongs to the row space of $\widetilde{B}$. Let u be the least squares solution to

$$\begin{bmatrix} \widetilde{B} \\ \tilde{w}^T \end{bmatrix} x = b. \tag{6}$$

Some entry, say the first, of $\tilde{w}^T$ is nonzero and without loss of generality is positive. Let v be the least squares solution to

$$\begin{bmatrix} \widetilde{B} \\ \tilde{w}^T + e_1^T \end{bmatrix} x = b, \tag{7}$$

where $e_1^T = (1, 0, 0, 0, \ldots, 0)$. Since the last entry of $Ax = b$ is exact $\tilde{w}^T u = (\tilde{w}^T + e_1^T)v$, and both u and v are least squares solutions to $\widetilde{B}x = b'$ where b' is the vector obtained from b by deleting its last row. Thus $\widetilde{B}(u - v) = 0$. Since $\tilde{w}^T$ belongs to the row space of $\widetilde{B}$, we have $\tilde{w}^T(u - v) = 0$. It follows that ${e_1}^T v = 0$ and hence that v has a zero entry, contrary to assumption. Therefore, each entry of $Ax = b$ is exact. □

THEOREM 2.9. *Let*

$$\begin{bmatrix} A & O & O \\ B & C & D \\ O & O & E \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \tag{8}$$

be a linear system where the vectors and matrices are conformally parititioned, and the entries of b_1, b_2, and b_3 are nonnegative. Assume that:

(i) *A has no zero rows and is balanceable*;

(ii) *$Ax_1 = b_1$ is least squares sign-solvable and its least squares solution has only positive entries*;

(iii) *each row of $[B \; -b_2]$ is nonnegative or nonpositive*;

(iv) *the matrix $[C \; b_2']$ is an S-matrix where b_2' is the row sum vector of $[B \; -b_2]$*;

(v) *E^T is an L-matrix or E^T has no rows and columns*; *and*

(vi) *the columns of E are combinatorially orthogonal to b_3.*

Then (8) *is a least squares sign-solvable linear system.*

Proof. Let M be the coeffecient matrix of (8) and let

$$b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}.$$

Consider a linear system

$$\begin{bmatrix} \widetilde{A} & O & O \\ \widetilde{B} & \widetilde{C} & \widetilde{D} \\ O & O & \widetilde{E} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \tilde{b}_1 \\ \tilde{b}_2 \\ \tilde{b}_3 \end{bmatrix}, \tag{9}$$

where the vectors and matrices belong to the appropriate qualitative classes. Let $\widetilde{M}$ and $\tilde{b}$ be the coeffecient matrix and the vector on the right-hand side of (9), respectively. By Lemma 2.1, A^T is an L-matrix. Since $[C \; b_2']$ is an S-matrix, C is an SNS-matrix, $Cx = -b_2'$ is sign-solvable, and each entry of the solution to $Cx = -b_2'$ is positive. Since E^T is an L-matrix or E has no rows and columns it now follows that M^T is an L-matrix and hence that there is a unique least squares solution u to (9). Let u_1 be the least squares solution to $\widetilde{A}x = \tilde{b}_1$. Then the distance between the column space of $\widetilde{A}$ and $\tilde{b}_1$ equals $||\widetilde{A}u_1 - \tilde{b}_1||$. Since E^T is an L-matrix and the columns of E are combinatorially orthogonal to b_3, it follows from Corollary 2.3 that the distance between the column space of $\widetilde{E}$ and $\tilde{b}_3$ equals $||\tilde{b}_3||$. Thus, the distance between the column space of $\widetilde{M}$ and $\tilde{b}$ is at least $\sqrt{||\widetilde{A}u_1 + \tilde{b}_1||^2 + ||\tilde{b}_3||^2}$. By (iii) and

(iv), $Cx = \tilde{b}_2 - \tilde{B}u_1$ is sign-solvable. Let u_2 be the solution to to $\tilde{C}x_2 = \tilde{b}_2 - \tilde{B}u_1$, and let

$$u = \begin{bmatrix} u_1 \\ u_2 \\ 0 \end{bmatrix}.$$

Then the distance between $\widetilde{M}u$ and $\tilde{b}$ equals $\sqrt{||\tilde{A}u_1 + \tilde{b}_1||^2 + ||\tilde{b}_3||^2}$. Therefore, u is the least squares solution to $\widetilde{M}x = \tilde{b}$. Since the sign patterns of u_1 and of u_2 are determined, it follows that (8) is a least squares sign-solvable linear system. □

We now prove a converse to Theorem 2.9. Note that if $Ax = b$ is least squares sign-solvable with least squares solution u then there exist diagonal matrices D and E each of whose main diagonal entries is 1 or -1 such that Db and Eu are nonnegative, and $(DAE)x = Db$ is least squares sign-solvable. Hence there is no loss in generality in assuming that both b and u are nonnegative. We use the following facts in the proof of the next theorem. For each matrix X there exist unique sets (possibly empty) γ and δ such that $X[\gamma, \overline{\delta}] = O$, $A[\overline{\gamma}, \overline{\delta}]$ is an L-matrix, and $X[\gamma, \delta]$ is balanceable [6]. Note that if X^T is an L-matrix then $X[\overline{\gamma}, \overline{\delta}]$ is an SNS-matrix. In [8] it is shown that if $X = [x_{ij}]$ and $Y = [y_{ij}]$ are distinct m by $m+1$ S-matrices such that $x_{ij} = 0$ if and only if $y_{ij} = 0$, then at least two corresponding entries of X and Y are not equal.

THEOREM 2.10. *Let $M = [m_{ij}]$ be an m by n matrix and let b be an m by n vector such that $Mx = b$ is a least squares sign-solvable linear system, b is nonnegative and the least squares solution $u = (u_1, u_2, \ldots, u_n)^T$ to $Mx = b$ is nonnegative. Let*

$$\beta = \{j : u_j \neq 0\} \text{ and } \alpha = \{i : m_{ij} \neq 0 \text{ for some } j \in \beta\}.$$

Then $M[\alpha, \beta]x = b[\alpha]$ is least squares sign-solvable and in particular $M[\alpha, \beta]^T$ is an L-matrix. Let γ and δ be the unique subsets of α and β, respectively, such that $M[\gamma, \beta \setminus \delta] = O$, $M[\alpha \setminus \gamma, \beta \setminus \delta]$ is an SNS*-matrix and $M[\gamma, \delta]$ is balanceable. Then $Mx = b$ has the form* (8) *and satisfies* (i)–(vi) *of Theorem* 2.10 *where $A = M[\gamma, \delta]$, $B = M[\alpha \setminus \gamma, \delta]$, $C = M[\alpha \setminus \gamma, \beta \setminus \delta]$, $D = M[\alpha \setminus \gamma, \overline{\beta}]$, $E = M[\overline{\alpha}, \overline{\beta}]$, $b_1 = b[\gamma]$, $b_2 = b[\alpha \setminus \gamma]$, and $b_3 = b[\overline{\alpha}]$.*

Proof. Since $Mx = b$ is least squares sign-solvable and $u[\overline{\beta}] = 0$, if $\widetilde{N} \in \mathcal{Q}(M[\alpha, \beta])$, $\tilde{c} \in \mathcal{Q}(b[\alpha])$ and v is a least squares solution to $\widetilde{N}x = \tilde{c}$, then $\begin{bmatrix} v \\ 0 \end{bmatrix}$ is the least squares solution to $\widetilde{M}x = \tilde{b}$ for each $\widetilde{M} \in \mathcal{Q}(M)$ and $\tilde{b} \in \mathcal{Q}(b)$ such that $\widetilde{M}[\alpha, \beta] = \widetilde{N}$ and $\tilde{b}[\alpha] = \tilde{c}$. Thus $M[\alpha, \beta]x = b[\alpha]$ is a least squares sign-solvable linear system. By Lemma 2.1, $M[\alpha, \beta]^T$ is an L-matrix. Thus the sets γ and β exist. Let A, B, C, D, E, b_1, b_2, and b_3 be defined as in the statement of the theorem.

Statement (i) holds by definition, and since $M[\alpha, \beta]^T$ is an L-matrix, C is an SNS-matrix. Thus, if $\widetilde{A} \in \mathcal{Q}(A)$, $\widetilde{B} \in \mathcal{Q}(B)$, $\widetilde{C} \in \mathcal{Q}(C)$, and $\tilde{b} \in \mathcal{Q}(b)$ and if v_1 is a least squares solution to

$$\begin{bmatrix} \widetilde{A} \\ \widetilde{B} \end{bmatrix} x = \begin{bmatrix} \tilde{b}_1 \\ \tilde{b}_2 \end{bmatrix}$$

and v_2 is the solution to $\widetilde{C}x = \tilde{b}_2 - \tilde{B}v_1$, then

$$\begin{bmatrix} v_1 \\ v_2 \\ 0 \end{bmatrix}$$

is the least squares solution to any system $\widetilde{M}x - \tilde{b}$ such that $\widetilde{M} \in \mathcal{Q}(M)$ and

$$\widetilde{M}[\alpha, \beta] = \begin{bmatrix} \widetilde{A} & O \\ \widetilde{B} & \widetilde{C} \end{bmatrix}.$$

Since C is an SNS-matrix, it follows that $Ax = b[\gamma]$ is a least squares sign-solvable system and that (i) holds. In addition, each linear system $Cx = \tilde{b}_2 - \tilde{B}v_1$ is sign-solvable and each entry of its solution is positive. It follows that each of the matrices $[\ \widetilde{C} \quad \tilde{b}_2 - \widetilde{B}v_1\ ]$ is an S-matrix. By the remark about S-matrices preceding the statement of the theorem we conclude that (iii) and (iv) hold.

Suppose that $M[\gamma, \overline{\beta}] \neq 0$. Let $i \in \gamma$, $j \in \overline{\beta}$ be integers such that $m_{ij} \neq 0$. By Lemma 2.8 there exists a matrix $\widetilde{A} \in \mathcal{Q}(A)$ and a vector $\tilde{b}_1 \in \mathcal{Q}(b_1)$ such that the ith entry of $\widetilde{A}v_1 - \tilde{b}_1$ is nonzero, where v is the least squares solution to $\widetilde{A}x = \tilde{b}_1$. Let $\widetilde{M}$ be the matrix in $\mathcal{Q}(M)$ such that $\widetilde{M}[\gamma, \delta] = \widetilde{A}$, the (i,j)-entry of $\widetilde{M}$ equals 1 and all other entries of $\widetilde{M}$ are 0, ϵ or $-\epsilon$. Let $\tilde{b}$ be a vector in $\mathcal{Q}(b)$ such that $\tilde{b}[\gamma] = \tilde{b}_1$. Let v_2 be the solution to $\widetilde{A}x = \tilde{b}_2 - \widetilde{B}v_1$. Then

$$\begin{bmatrix} v_1 \\ v_2 \\ 0 \end{bmatrix}$$

is the least squares solution to $\widetilde{M}x = \tilde{b}$. But for ϵ sufficiently small the jth column of $\widetilde{M}$ is not orthogonal to $\widetilde{M}v - \tilde{b}$, a contradiction. Thus, $M[\gamma, \overline{\beta}] = O$.

Since M^T is an L-matrix and C is an SNS-matrix, it follows that E^T is an L-matrix or E^T has no rows and columns. Hence (v) holds. If $b_3 = 0$, then clearly (vi) holds. Assume that $b_3 \neq 0$. By Corollary 2.3, to show that (vi) holds, it suffices to show that $Ex = b_3$ is least squares sign-solvable and its solution is the zero vector. Since E^T is an L-matrix, each $\widetilde{E}x = \tilde{b}_3$ $(\widetilde{E} \in \mathcal{Q}(E), \tilde{b}_3 \in \mathcal{Q}(b_3))$ has exactly one least squares solution. Suppose that the least squares solution v_3' to $\widetilde{E}x = \tilde{b}_3$ is nonzero. Let v_1 be the least squares solution to $Ax = b_1$ and let v_2' be the solution to $Cx = b_2 - Bv_1 - Dv_3'$. Then

$$||\widetilde{M} \begin{bmatrix} v_1 \\ v_2' \\ v_3' \end{bmatrix} - \tilde{b}|| < ||\widetilde{M} \begin{bmatrix} v_1 \\ v_2 \\ 0 \end{bmatrix} - \tilde{b}||,$$

where v_2 is the solution to $Cx = b_2 - Bx_1$,

$$\widetilde{M} = \begin{bmatrix} A & O & O \\ B & C & D \\ O & O & \widetilde{E} \end{bmatrix} \text{ and } \tilde{b} = \begin{bmatrix} b_1 \\ b_2 \\ \tilde{b}_3 \end{bmatrix}.$$

This contradicts the fact that

$$\begin{bmatrix} v_1 \\ v_2 \\ 0 \end{bmatrix}$$

is the least squares solution to $\widetilde{M}x = \tilde{b}$. Thus $Ex = b_3$ is least squares sign-solvable and has the zero vector as its least squares solution. Therefore (vi) holds and the proof is complete. □

Note that if in Theorem 2.10 the system $Mx = b$ is sign-solvable, then each row is exact and hence Lemma 2.8 implies that the matrix A has no rows and no columns. Theorems 2.9 and 2.10 show that the study of least squares sign-solvable linear systems reduces to the study of L-matrices and to least squares sign-solvable linear systems for which no entry is exact and for which the least squares solution has no zero entries. This reduction is quite similar to those for sign-solvable linear systems and conditionally sign-solvable linear systems presented in [11] and [3], respectively.

Consider a linear system $Ax = b$. Among all least squares solutions to $Ax = b$ there is a unique solution u such that $||u||$ is minimal. The solution u is the *minimal least squares solution* to $Ax = b$. If $Ax = b$ is least squares sign-solvable, then by Lemma 2.1 the minimal least squares solution is the unique least squares solution for each linear system $\widetilde{A}x = \tilde{b}$ ($\widetilde{A} \in \mathcal{Q}(A), \tilde{b} \in \mathcal{Q}(b)$). As suggested by the referee, it may be of interest to study linear systems $Ax = b$ for which the sign pattern of the minimal least squares solution is completely determined by the sign patterns of A and b. Any linear system $Ax = b$ for which the columns of A are combinatorially orthogonal to b is an example of a linear system whose minimal least squares solution is completely determined. If A^T is not an L-matrix, then $Ax = b$ will not be least-squares sign-solvable.

3. Signed generalized inverses. Let $A = [a_{ij}]$ be an m by n matrix such that A^T is an L-matrix, and let p and q be integers such that $1 \leq p \leq n$ and $1 \leq q \leq m$. Then the (p, q)-entry of $A^\dagger$ is *signed* provided the (p, q)-entries of the matrices in

$$\{\widetilde{A}^\dagger : \widetilde{A} \in \mathcal{Q}(A)\}$$

all have the same sign. For example let

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}.$$

Then A^T is an L-matrix and for $\widetilde{A} \in \mathcal{Q}(A)$ there exist positive numbers a, b, c, d, e such that

$$\widetilde{A} = \begin{bmatrix} a & b \\ c & d \\ e & 0 \end{bmatrix}.$$

Then

$$\widetilde{A}^\dagger = \frac{1}{\det \widetilde{A}^T \widetilde{A}} \begin{bmatrix} ad^2 - cbd & cb^2 - dab & e(b^2 + d^2) \\ -acd + bc^2 + be^2 & -cab + da^2 + de^2 & -e(ab + cd) \end{bmatrix}.$$

Thus, it follows that the $(1, 3)$ and $(2, 3)$-entry of $A^\dagger$ are signed, and no other entry of $A^\dagger$ is signed. More generally each entry in the jth column is of $A^\dagger$ is signed if and only if $Ax = e_j$ is least squares sign-solvable where e_j is the $(0, 1)$-vector whose only nonzero entry is a 1 in its jth row. The next theorem establishes necessary and sufficient conditions for the (p, q)-entry of A^T to be signed. If T is a subset of $\{1, 2, \ldots, m\}$ and q is a positive integer, then $\text{inv}(q, T)$ is the number of elements in T which are less than q.

THEOREM 3.1. *Let $A = [a_{ij}]$ be an m by n matrix such that A^T is an L-matrix, and let p and q be integers with $1 \leq p \leq n$ and $1 \leq q \leq m$. Define α to be the*

collection of subsets T of $\{1,2,\dots,m\}$ of cardinality n which contain q such that neither $A[T \setminus \{q\}, \{p\}]$ nor $A[T,:]$ has an identically zero determinant. Then the (p,q)-entry of $A^\dagger$ is signed if and only if one of the following hold:

(i) *α is the empty set.*

(ii) *For each $\widetilde{A} \in \mathcal{Q}(A)$ the numbers*

$$(-1)^{\mathrm{inv}(q,T)} \det \widetilde{A}[T \setminus \{q\}, \overline{\{p\}}] \det \widetilde{A}[T,:] \qquad (T \in \alpha)$$

are nonnegative and at least one is positive.

(iii) *For each $\widetilde{A} \in \mathcal{Q}(A)$ the numbers*

$$(-1)^{\mathrm{inv}(q,T)} \det \widetilde{A}[T \setminus \{q\}, \overline{\{p\}}] \det \widetilde{A}[T,:] \qquad (T \in \alpha)$$

are nonpositive and at least one is negative.

Proof. Let $\widetilde{A} = [\tilde{a}_{ij}]$ be a matrix in $\mathcal{Q}(A)$. Since $\widetilde{A}^\dagger = (\widetilde{A}^T \widetilde{A})^{-1} \widetilde{A}^T$ the (p,q)-entry of $\widetilde{A}^\dagger$ is given by

$$\sum_{k=1}^{n} \frac{(-1)^{p+k} \det(\widetilde{A}^T \widetilde{A})[\overline{\{p\}}, \overline{\{k\}}] \tilde{a}_{qk}}{\det(\widetilde{A}^T \widetilde{A})}$$

(10)

$$= \frac{(-1)^p}{\det(\widetilde{A}^T \widetilde{A})} \sum_{k=1}^{n} (-1)^k \tilde{a}_{qk} \det(\widetilde{A}^T \widetilde{A})[\overline{\{p\}}, \overline{\{k\}}].$$

By the Cauchy–Binet determinantal formula,

$$\det(\widetilde{A}^T \widetilde{A})[\overline{\{p\}}, \overline{\{k\}}] = \sum_{S} \det \widetilde{A}[S, \overline{\{p\}}] \det \widetilde{A}[S, \overline{\{k\}}], \tag{11}$$

where the summation is over all subsets S of $\{1,2,\dots,m\}$ of cardinality $n-1$. Thus by (10) and (11), the (p,q)-entry of $\widetilde{A}^T$ is given by

(12)

$$\frac{(-1)^p}{\det(\widetilde{A}^T \widetilde{A})} \sum_{k=1}^{n} \sum_{S} (-1)^k \tilde{a}_{qk} \det \widetilde{A}[S, \overline{\{p\}}] \det \widetilde{A}[S, \overline{\{k\}}]$$

$$= \frac{(-1)^p}{\det(\widetilde{A}^T \widetilde{A})} \sum_{S} \det \widetilde{A}[S, \overline{\{p\}}] \left(\sum_{k=1}^{n} (-1)^k \tilde{a}_{qk} \det \widetilde{A}[S, \overline{\{k\}}] \right),$$

where the summation involving S is over all subsets S of $\{1,2,\dots,m\}$ of cardinality $n-1$.

Let S be a subset of $\{1,2,\dots,m\}$ of cardinality $n-1$. Then

$$\sum_{k=1}^{n} (-1)^k \tilde{a}_{qk} \det A[S, \overline{\{p\}}] \det A[S, \overline{\{k\}}]$$

$$= \det A[S, \overline{\{p\}}] \sum_{k=1}^{n} (-1)^k \tilde{a}_{qk} \det A[S, \overline{\{k\}}]$$

$$= -\det A[S, \overline{\{p\}}] \left(\det \begin{bmatrix} A[\{q\},:] \\ A[S,:] \end{bmatrix} \right).$$

If q belongs to S, then clearly

$$\det \begin{bmatrix} A[\{q\},:] \\ A[S,:] \end{bmatrix} = 0.$$

Otherwise,

$$\det \begin{bmatrix} A[\{q\},:] \\ A[S,:] \end{bmatrix} = (-1)^{\mathrm{inv}(q,S)} \det A[S \cup \{q\},:].$$

It now follows from (12) that the (p,q)-entry of $\widetilde{A}^{\dagger}$ equals

$$\frac{(-1)^{p+1}}{\det(\widetilde{A}^T \widetilde{A})} \sum_{T \in \alpha} (-1)^{\mathrm{inv}(q,T)} \det \widetilde{A}[T \setminus \{q\}, \overline{\{p\}}] \det \widetilde{A}[T,:]. \tag{13}$$

A^T is an L-matrix, $\widetilde{A}^T \widetilde{A}$ is a positive definite matrix and in particular $\det(\widetilde{A}^T \widetilde{A}) > 0$ for all $\widetilde{A} \in \mathcal{Q}(A)$. Therefore, if (i), (ii), or (iii) holds then the sign of (13) is independent of the choice of $\widetilde{A}$, and hence the (p,q)-entry of $A^{\dagger}$ is signed.

Conversely, assume that the sign of the (p,q)-entry of $A^{\dagger}$ is signed. If α is empty, then (i) holds. Assume that α is nonempty, and consider a subset T belonging to α. It is easy to verify that there exists a matrix $\widehat{A} \in \mathcal{Q}(A)$ such that

$$\det \widehat{A}[T \setminus \{q\}, \overline{\{p\}}] \det \widehat{A}[T,:] \neq 0.$$

For any such matrix $\widehat{A}$, let $\widetilde{A}$ be the matrix in $\mathcal{Q}(A)$ such that $\widetilde{A}[T,:] = \widehat{A}[T,:]$ and each nonzero entry in $\widetilde{A}[\overline{T},:]$ is $\pm\epsilon$ for some real number ϵ. For ϵ sufficiently small it follows from (13) that the sign of the (p,q)-entry of $\widetilde{A}$ equals the sign of

$$\frac{(-1)^{p+1}}{\det(\widetilde{A}^T \widetilde{A})} \det \widehat{A}[T \setminus \{q\}, \overline{\{p\}}] \det \widehat{A}[T,:].$$

Since the (p,q)-entry of $A^{\dagger}$ is signed this implies that either (i) or (ii) holds. □

As noted by the referee, the fact that the (p,q)-entry of $\widetilde{A}^{\dagger}$ is given by (13) is a special case of the formula in [2] for the (p,q)-entry of the generalized inverse of a matrix whose entries belong to an integral domain.

We now study the structure of matrices whose generalized inverse is signed. We use the fact that if X is a square matrix which does not have an identically zero determinant, then X is an SNS-matrix if and only if the determinant of each matrix in $\mathcal{Q}(A)$ has the same sign (see [6]).

LEMMA 3.2. *Let A be matrix of order n such that A does not have an identically zero determinant. Suppose that for each r and s the numbers*

$$(-1)^{r+s} \widetilde{A}[\overline{\{r\}}, \overline{\{s\}}] \det \widetilde{A} \quad (\widetilde{A} \in \mathcal{Q}(A)) \tag{14}$$

are all nonnegative or all nonpositive. Then A is an S^2NS*-matrix.*

Proof. We show first show that A is an SNS-matrix. For suppose not. Then there exists matrices $\widetilde{A}$ and $\widehat{A}$ in $\mathcal{Q}(A)$ such that $\widetilde{A}$ has rank n and $\widehat{A}$ has rank less than n. Since changing a single entry of a matrix can only change the rank by 1, we may assume without loss of generality that $\widehat{A}$ has rank $n-1$. Furthermore, we may assume that all but one nonzero entry $\widehat{A}$ is equal to the corresponding entry of

$\widetilde{A}$. Without loss of generality assume that $(1,1)$-entries of $\widehat{A}$ and $\widetilde{A}$ are not equal. Let c and d be the $(1,1)$-entries of $\widehat{A}$ and $\widetilde{A}$, respectively. Since $\widehat{A}$ is not invertible and $\widetilde{A}$ is invertible, $\det \widehat{A}[\overline{\{1\}}, \overline{\{1\}}] \neq 0$. Let A_ϵ be the matrix obtained from $\widehat{A}$ by adding ϵ to the $(1,1)$-entry. Then for ϵ with $|\epsilon| < |d-c|$, A_ϵ is a matrix in $\mathcal{Q}(A)$ with $\det A_\epsilon = \epsilon \det \widehat{A}[\overline{\{1\}}, \overline{\{1\}}]$, and $\det A_\epsilon[\overline{\{1\}}, \overline{\{1\}}] = \det \widehat{A}[\overline{\{1\}}, \overline{\{1\}}]$. It follows that there are numbers in (14) of different sign, contrary to assumption. Therefore, A is an SNS-matrix.

If $\widetilde{A} \in \mathcal{Q}(A)$, then the (s,r)-entry of $\widetilde{A}^{-1}$ is

$$\frac{(-1)^{r+s} \det \widetilde{A}[\overline{\{r\}}, \overline{\{s\}}]}{\det \widetilde{A}}. \tag{15}$$

Since A is an SNS-matrix the assumptions on the numbers (14) imply that determinant of each matrix in $\mathcal{Q}(A[\overline{\{r\}}, \overline{\{s\}}])$ is nonnegative or the determinant of each such matrix is nonpositive. By the comment preceding the statement of the lemma, we conclude that $A[\overline{\{r\}}, \overline{\{s\}}]$ is either an SNS-matrix or has an identically zero determinant. Thus by (15), the sign of the (s,r)-entry of $\widetilde{A}^{-1}$ is determined. Therefore, A is an S^2NS-matrix. □

THEOREM 3.3. *Let A be an m by n matrix such that A^T is an L-matrix and A has a signed generalized inverse. Then each submatrix of A of order n either has an identically zero determinant or is an* S^2NS*-matrix.*

Proof. Let T be a subset of $\{1, 2, \ldots, m\}$ of cardinality m such that $A[T, :]$ does not have an identically zero determinant. Theorem 3.1 implies that if p and q are integers with $1 \leq p \leq m$ and $q \in T$ then the numbers

$$(-1)^{\mathrm{inv}(q,T)} \det \widetilde{A}[T \setminus \{q\}, \{p\}] \det \widetilde{A}[T, :] \quad (\widetilde{A} \in \mathcal{Q}(A))$$

are all nonnegative or all nonpositive. The theorem follows from Lemma 3.2. □

Clearly an S^2NS-matrix of order m has a generalized inverse and each of its submatrices of order m is an S^2NS-matrix. Let A be an m by $m-1$ matrix such that the matrix obtained from A by replacing its nonzero entries by 1's is the vertex-edge incidence matrix of a tree. By Corollary 2.7 each system $Ax = b$ where b is a vector with exactly one nonzero entry is least squares sign-solvable. Hence, A has a generalized inverse. Since no submatrix of A of order $m-1$ has an identically zero determinant, Theorem 3.3 implies that each submatrix of A of order $m-1$ is an S^2NS-matrix. We now show that an m by n matrix with $m \geq 2$ which has a signed generalized inverse and each of whose submatrices of order n is an S^2NS-matrix is either an S^2NS-matrix, or the matrix obtained by replacing its nonzeros by 1's is the vertex-edge incidence matrix of a tree.

LEMMA 3.4. *Let A be an m by n matrix with $m > n$ such that A^T is an L-matrix and A has a signed generalized inverse. Suppose no submatrix of $B = A[\{1, 2, \ldots, n+1\}, :]$ of order n has an identically zero determinant. Then the matrix*

$$M = \begin{bmatrix} & \det B[\overline{\{1\}}, :] \\ & -\det B[\overline{\{2\}}, :] \\ B & \vdots \\ & (-1)^n \det B[\overline{\{n+1\}}, :] \end{bmatrix}$$

is an S^2NS*-matrix of order $n+1$.*

Proof. It follows from Lemma 3.3 that M is an SNS-matrix and that each submatrix of order n of M which does not contain the last column is an SNS-matrix. Consider the submatrix $M[\overline{\{q\}}, \overline{\{p\}}]$ of order where $p \neq n+1$. By Laplace expansion of the determinant along the last column we have

$$\begin{aligned}\det M[\overline{\{q\}}, \overline{\{p\}}] &= \sum_{k=1}^{q-1}(-1)^{n+k}(-1)^{k+1} \det B[\overline{\{k\}}, :] \det B[\overline{\{k,q\}}, \overline{\{p\}}] \\ &\quad + \sum_{k=q+1}^{n+1} (-1)^{n+k-1}(-1)^{k+1} \det B[\overline{\{k\}}, :] \det B[\overline{\{k,q\}}, \overline{\{p\}}]. \end{aligned} \tag{16}$$

For $k = 1, 2, \ldots, n+1$ and $k \neq q$, let T_k be the set $\{1, 2, \ldots, n+1\} \setminus \{k\}$. Then $\operatorname{inv}(q, T_k)$ equals $q-1$ if $k > q$ and equals $q-2$ if $k < q$. Thus by (16) we have

$$\begin{aligned}&\det M[\overline{\{q\}}, \overline{\{p\}}] \\ &\quad = (-1)^{n+q+k+1} \sum_{\substack{k=1 \\ k\neq q}}^{n} (-1)^{\operatorname{inv}(q,T_k)} \det B[T_k, :] \det B[T_k \setminus \{k,q\}, \overline{\{p\}}]. \end{aligned} \tag{17}$$

Since the (p,q)-entry of $A^\dagger$ is signed, Theorem 3.1 now implies that the sign of $\det M[\overline{\{q\}}, \overline{\{p\}}]$ does not depend upon the magnitudes of the entries in of M. Hence $M[\overline{\{q\}}, \overline{\{p\}}]$ either has an identically zero determinant or is an SNS-matrix. Therefore, M is an S^2NS-matrix. □

In [6], it is shown that if C is an S^2NS-matrix of order $n+1$ such that the last column of C contains no zero entries and no submatrix of order n which does not intersect column $n+1$ has an identically zero determinant, then the matrix obtained from $C[\{1, 2, \ldots, n+1\}, \overline{\{n+1\}}]$ by replacing its nonzero entries by 1's is the vertex-edge incidence matrix of a tree. In [4] it is shown that if A is an m by n matrix with $m \geq 2$ such that each submatrix of order n is an SNS-matrix then $m \leq n+2$, and if $m = n+2$ then each column of A contains exactly three nonzero entries. An $n+2$ by n matrix each of whose submatrices of order n is an SNS-matrix is a *totally L-matrix.* These results, along with Lemma 3.4 imply the following.

THEOREM 3.5. *Let A be an m by n matrix with $n \geq 2$ such that A^T is an L-matrix, A has a signed generalized inverse, and no submatrix of A of order n has an identically zero determinant. Then either $m = n$ and A is an S^2NS-matrix, or $m = n+1$ and the matrix obtained from A by replacing its nonzero entries by 1's is the vertex-edge incidence matrix of a tree.*

There do exist m by n matrices with signed generalized inverses some of whose submatrices of order m have identically zero determinant. For example, the matrix

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

has a signed generalized inverse, and $A[\{1,2\}, \{1,2\}]$ has an identically zero determinant. We now show that matrices with a signed generalized inverse have a very specific structure. The following lemma is essentially contained in [6].

LEMMA 3.6. *Let A be an m by n matrix such that A^T is an L-matrix and each submatrix of A of order n either has an identically zero determinant or is an* SNS-*matrix. Then there exist permutation matrices P and Q and an integer k such that PAQ has the form*

$$\begin{bmatrix} A_1 & O & \cdots & O \\ A_{21} & A_2 & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ A_{k1} & A_{k2} & \cdots & A_k \end{bmatrix},$$

where each A_i^T is either an S^2NS-*matrix, an* S^*-*matrix a totally L-matrix, or a matrix with one column and no nonzero entries.*

LEMMA 3.7. *Let B be an m by n matrix and let D be an r by s matrix such that B^T and D^T are L-matrices. Let A be a matrix of the form*

$$\begin{bmatrix} B & O \\ C & D \end{bmatrix}.$$

If the generalized inverse of A is signed, then the generalized inverses of B and D are signed.

Proof. Since B^T and D^T are L-matrices, A^T is an L-matrix. Assume that A has a signed generalized inverse. Since A^T is an L-matrix, there exists a submatrix $B[\gamma, :]$ of B and a submatrix $D[\delta, :]$ of D neither of which has an identically zero determinant. By Theorem 3.3, each submatrix of A of order $n+s$ is either an S^2NS-matrix or has an identically zero determinant. Let T be a subset of $\{1, 2, \ldots, m\}$ of cardinality n. Then $A[T \cup \delta, :]$ is a submatrix of A of order $n+s$, and is lower triangular. It follows that $B[T, :]$ is either an S^2NS-matrix or has an identically zero determinant. Hence each submatrix of B of order n is either an S^2NS-matrix or has an identically zero determinant.

We show that B has a signed generalized inverse by showing that each entry of its generalized inverse is signed. Let p and q be integers such that $1 \le p \le m$ and $1 \le q \le n$. Let α be the set of T such that neither $B[T \setminus \{q\}, \overline{\{p\}}]$ nor $B[T,]$ has an identically zero determinant. Then it follows that if $T \in \alpha$ both $B[T \setminus \{q\}, \overline{\{p\}}]$ and $B[T, :]$ are SNS-matrices. Hence by Theorem 3.1, it suffices to show that either α is empty or the signs of the nonzero numbers

$$(-1)^{\mathrm{inv}(q,T)} \det B[T \setminus \{q\}, \overline{\{p\}}] \det B[T, :] \quad (T \in \alpha) \tag{18}$$

are all the same.

Assume that α is nonempty. Since the (p, q) entry of $A^\dagger$ is signed, it follows from Theorem 3.1 that the numbers

$$(-1)^{\mathrm{inv}(q,T\cup\delta)} \det A[(T \cup \alpha) \setminus \{q\}, \overline{\{p\}}] \det A[T \cup \delta, :] \quad (T \in \alpha) \tag{19}$$

are all nonnegative or all nonpositive. Clearly, $\mathrm{inv}(q, T \cup \delta) = \mathrm{inv}(q, T)$, $\det A[(T \cup \delta) \setminus \{q\}, \overline{\{p\}}] = \det B[T \setminus \{q\}, \overline{\{p\}}] \det D[\delta, ;]$ and $\det A[T \cup \delta, :] = \det B[T, :] \det D[\delta, :]$. Since the numbers in (19) are all nonpositive or all nonnegative, it now follows that the nonzero numbers (18) are all of the same sign. Hence the (p, q)-entry of $B^\dagger$ is signed. Therefore, B has a signed generalized inverse. A similar argument shows that D has a signed generalized inverse. □

The following theorem describes the structure of a matrix with a signed generalized inverse, and is an immediate consequence of Lemmas 3.6 and 3.7 and Theorem 3.5.

THEOREM 3.8. *Let A be an m by n matrix with no zero rows such that A^T is an L-matrix. If A has a signed generalized inverse, then there exist permutation matrices P and Q, diagonal matrices D and E, and an integer k such that PAQ has the form*

$$\begin{bmatrix} A_1 & O & \cdots & O \\ A_{21} & A_2 & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ A_{k1} & A_{k2} & \cdots & A_k \end{bmatrix}, \tag{20}$$

where each A_i is a matrix of one column each of whose entries is 1, *an* S^2NS*-matrix, or the vertex-edge incidence matrix of a tree.*

Let A be an m by n matrix of the form (20) such that each A_i is a matrix of one column and each entry is 1, an S^2NS-matrix, or the vertex-edge incidence matrix of a tree. Necessary and sufficient conditions on the matrices A_{ij} in order that A have a signed generalized inverse are not presently known, and are a topic of continuing research.

Let A be an m by n matrix. Even if the columns of A are linearly dependent, the Moore–Penrose inverse $A^\dagger$ exists. As suggested by the referee, it might be interesting to study those matrices A for which the sign-pattern of the Moore–Penrose inverse of A is completely determined by the sign-pattern of A. In this paper we have only studied the case in which A^T is an L-matrix.

REFERENCES

[1] L. BASSETT, J. MAYBEE, AND J. QUIRK, *Qualitative economics and the scope of the correspondence principle*, Econometrica, 36(1968), pp. 544–563.

[2] R. B. BAPAT, K. P. S. BHASKARA ROA, AND K. M. PRASAD, *Generalized inverses over integral domains*, Linear Algebra Appl., 140 (1990), pp. 181–196.

[3] R. A. BRUALDI, K. L. CHAVEY, AND B. L. SHADER, *Conditional sign-solvability*, Math. Comp. Model., 17 (993), pp. 141–148.

[4] ———, *Rectangular L-matrices*, Linear Algebra Appl., 196 (1994), pp. 37–62.

[5] ———, *Bipartite graphs and inverse sign patterns of strong sign-nonsingular matrices*, J. Comb. Theory Ser B., 62 (1994), pp. 133–150.

[6] R. A. BRUALDI AND B. L. SHADER, *The matrices of sign-solvable linear systems*, Cambridge University Press, Cambridge, to appear.

[7] C. ESCHENBACH, F. HALL, AND C. R. JOHNSON, *Self-inverse sign patterns*, IMA Preprint Series 1005, 1992.

[8] V. KLEE, *Recursive structure of S-matrices and an $O(m^2)$ algorithm for recognizing strong sign solvability*, Linear Algebra Appl., 96 (1987), pp. 233–242.

[9] V. KLEE, B. VON HOHENBALKEN, AND T. LEWIS, *Cone-systems and sign-solvability*, Linear Algebra Appl., 192(1993), pp. 187–204.

[10] ———, *S-systems, L-systems, and an extension of sign-solvability*, preprint.

[11] V. KLEE, R. LADNER, AND R. MANBER, *Sign solvability revisited*, Linear Algebra Appl., 59 (1984), pp. 131–157.

[12] G. LADY AND J. MAYBEE, *Qualitatively invertible matrices*, J. Math. Social Sciences, 6 (1983), pp. 397–407.

[13] T. J. LUNDY AND J. S. MAYBEE, *Inverses of sign nonsingular matrices*, preprint.

[14] C. THOMASSEN, *When the sign pattern of a square matrix determines uniquely the sign pattern of its inverse*, Linear Algebra Appl., 119 (1989), pp. 27–34.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, pp. 1074–1085, October 1995

003

ON EIGENVALUE ESTIMATES FOR BLOCK INCOMPLETE FACTORIZATION METHODS*

O. AXELSSON† AND H. LU†

Abstract. Eigenvalue estimates of block incomplete preconditioners are considered. We investigate how the block diagonal entries and off-block diagonal entries influence the bounds of all eigenvalues. The results presented here improve and unify some previous results. We generalize the well-known inequality that the spectral radius is bounded by the trace for symmetric positive semidefinite matrices to block form. Some of the methods can also be useful to estimate lower bounds of block incomplete preconditioners.

Key words. eigenvalue estimates, incomplete factorization, preconditioners

AMS subject classifications. 65F10, 65F15, 65F50

1. Introduction.

To estimate the rate of convergence of preconditioned iterative methods such as the Chebyshev iterative method and the conjugate gradient iterative method, one needs to know the extreme eigenvalues and the distribution of eigenvalues of the preconditioned matrix, respectively; see [1], [2], [8], [7], [5], [12]. Naturally, this problem by itself is difficult, especially for the distribution of all eigenvalues. Fortunately, it has been shown (see [4]) that under certain conditions lower and upper bounds of some eigenvalues can be derived and they provide the information necessary to compare modified and unmodified incomplete factorization methods for symmetric positive definite matrices, for instance.

Consider the implicit preconditioner on factorized form

$$C = (X + L)X^{-1}(X + L^T)$$

of a symmetric matrix A. Let $A = D_A + L_A + L_A^T$, where D_A is a block diagonal matrix. If A is a Stieltjes matrix and $L = L_A$ in some cases, some methods to estimate upper bounds of eigenvalues of $C^{-1}A$ were derived in [6], [4], [10], [9]. However, the assumptions limit the applicability of the results because for incomplete factorization methods they do not hold in general. In this paper, we discuss upper bounds and distribution of eigenvalues of block incomplete preconditioners for the general case of A being only a symmetric matrix. All of the results allow that L_A differs from L. As we will see, even when the assumption of A is weakened, we can have strong results. The results here also unify some of the previous results on upper bounds of eigenvalues of incomplete preconditioners.

The paper is organized as follows. Under the assumption of A being a symmetric matrix, in §2 we focus our attention on both estimates of upper bounds and distribution of eigenvalues of block incomplete preconditioners. The result presented in this section can also be useful to estimate lower bounds of block incomplete preconditioners. In §3, some further useful methods to estimate upper bounds and distribution of eigenvalues are presented based on the fundamental result in the previous section. We generalize the well-known inequality $\rho(A) \leq \mathrm{tr}(A)$ for A symmetric positive semidef-

* Received by the editors August 6, 1993; accepted for publication (in revised form) by A. Greenbaum July 9, 1994.

† Department of Mathematics, University of Nijmegen, Toernooiveld, 6525 ED Nijmegen, The Netherlands. The second author was supported by The Netherlands Organization for Scientific Research, grant 611-302-025.

inite to block form, which with the result in §2 yields a new upper bound depending only on the block order of matrices for the largest eigenvalue.

For convenience, $\lambda_i(A)$ denotes the ith eigenvalue of matrix A and it is assumed that all eigenvalues of a matrix are ordered in a nonincreasing order. For any pair of matrices A, B of the same order, $A \geq B$ means that the same inequality holds elementwise. The notation s.p.d. means symmetric positive definite while s.p.s.d. means symmetric positive semidefinite.

2. Upper and lower bounds of eigenvalues. Let A be a symmetric matrix partitioned in a block form

$$A = D_A + L_A + L_A^T,$$

where D_A, L_A is the block diagonal part and strictly lower block triangular part of A, respectively. Consider a preconditioner C in the form

$$C = (X + L)X^{-1}(X + L^T),$$

i.e., a so-called implicit preconditioner, where X is a block diagonal and s.p.d. matrix and L is a block lower triangular matrix. X and L are partitioned in blocks consistently with D_A and L_A, respectively. We present first a result for upper bounds of eigenvalues, which extends some results in [4], [10].

THEOREM 2.1. *Let A be symmetric and assume that X is* s.p.d. *and $\sigma X - K$ and $K - \beta X$ are* s.p.s.d. *for some constants σ, β. Then*

$$\lambda_i(M(\beta)) \leq \lambda_i(C^{-1}A) \leq \lambda_i(M(\sigma)), \tag{1}$$

where $K = A - L - L^T$, $C = (X + L)X^{-1}(X + L^T)$, and

$$M(\alpha) = (I + \tilde{L})^{-1} + (I + \tilde{L}^T)^{-1} + (\alpha - 2)(I + \tilde{L})^{-1}(I + \tilde{L}^T)^{-1}, \tag{2}$$

$$\tilde{L} = X^{-\frac{1}{2}} L X^{-\frac{1}{2}}.$$

Proof. We have

$$A = K - \sigma X + (X + L) + (X + L^T) + (\sigma - 2)X.$$

A computation with a similarity transformation of $C^{-1}A$ shows that

$$\begin{aligned}
&X^{-\frac{1}{2}}(X + L^T)C^{-1}A(X + L^T)^{-1}X^{\frac{1}{2}} \\
&= X^{\frac{1}{2}}(X + L)^{-1}A(X + L^T)^{-1}X^{\frac{1}{2}} \\
&= X^{\frac{1}{2}}(X + L)^{-1}(K - \sigma X)(X + L^T)^{-1}X^{\frac{1}{2}} \\
&\quad + X^{\frac{1}{2}}(X + L)^{-1}X^{\frac{1}{2}} + X^{\frac{1}{2}}(X + L^T)^{-1}X^{\frac{1}{2}} \\
&\quad + (\sigma - 2)X^{\frac{1}{2}}(X + L)^{-1}X(X + L^T)^{-1}X^{\frac{1}{2}} \\
&= X^{\frac{1}{2}}(X + L)^{-1}(K - \sigma X)(X + L^T)^{-1}X^{\frac{1}{2}} + M(\sigma).
\end{aligned}$$

Since, by assumption, $K - \sigma X$ is negative semidefinite, this shows that

$$\lambda_i(C^{-1}A) \leq \lambda_i(M(\sigma)).$$

Similarly, using that $K - \beta X$ is s.p.s.d., we prove the first inequality in (1). □

If $L = L_A$ is nonpositive and X is a block diagonal Stieltjes matrix, the special case of Theorem 2.1 for the maximum eigenvalue of $C^{-1}A$ can be found in [10].

The following five propositions give situations in which the theorem is applicable.

PROPOSITION 2.2. *X is* s.p.s.d. *if X is a symmetric Z-matrix and $X\mathbf{v} \geq 0$ for some vector $\mathbf{v} > 0$.*

Proof. Let $\mathbf{v} = (v_1, v_2, \ldots, v_k)^T$ and $D = \operatorname{diag}(v_1, v_2, \ldots, v_k)$. $DXD + \varepsilon I$ is a diagonally dominant Z-matrix for any $\varepsilon > 0$, which implies, in particular, that X is s.p.s.d. □

PROPOSITION 2.3. *Let X be symmetric. If $\sigma X - D_A$ is a Z-matrix and the entries of $L + L^T$ are not larger than the corresponding entries of $L_A + L_A^T$, then $\sigma X - K$ is a Z-matrix. If, in addition, $\sigma X\mathbf{v} - K\mathbf{v} \geq 0$ for some positive vector $\mathbf{v}$, then $\sigma X - K$ is* s.p.s.d.

Proof. A direct calculation shows that $\sigma X - K = (\sigma X - D_A) + (L + L^T - L_A - L_A^T)$, which shows that $\sigma X - K$ is a Z-matrix. An application of Proposition 2.2 completes the proof. □

PROPOSITION 2.4. *$\lambda_i(M(\sigma)) \leq \min(\lambda_i(M(2)), 1/(2-\sigma))$ if $\sigma \in [0, 2]$. The inequality is strict if $\sigma < 2$.*

Proof. $\lambda_i(M(\sigma)) \leq \lambda_i(M(2))$ is straightforward. By a simple computation

$$\begin{aligned} M(\sigma) &= (I + \tilde{L})^{-1} + (I + \tilde{L}^T)^{-1} + (\sigma - 2)(I + \tilde{L})^{-1}(I + \tilde{L}^T)^{-1} \\ &= \frac{1}{2-\sigma} I - (2 - \sigma)\left((I + \tilde{L})^{-1} - \frac{1}{2-\sigma} I\right)\left((I + \tilde{L}^T)^{-1} - \frac{1}{2-\sigma} I\right), \end{aligned}$$

we finish the proof. □

If $L = L_A$, the upper bound $1/(2-\sigma)$ can be found in [4].

PROPOSITION 2.5. *Suppose matrix X is* s.p.d. *and $\sigma X - K + \gamma I$ is* s.p.s.d. *Then $(\sigma + \gamma/\lambda_{\min}(X))X - K$ is* s.p.s.d. *if $\gamma \geq 0$ and $(\sigma + \gamma/\lambda_{\max}(X))X - K$ is* s.p.s.d. *if $\gamma \leq 0$.*

PROPOSITION 2.6. *$\lambda_{\max}(X^{-1}K)X - K$ and $\lambda_{\min}(X^{-1}K)X + K$ are* s.p.s.d. *if X is* s.p.d.

3. Some alternative upper bounds. As we have seen in the previous section, the maximum eigenvalue of $C^{-1}A$ can be bounded by $\frac{1}{2-\sigma}$ if $\sigma < 2$, but the situation is not so fortunate if $\sigma > 2$. It is impossible to derive a bound by involving σ alone. The bound of the eigenvalues must depend on both σ and the lower triangular matrix $\tilde{L}$. In this section, first, we discuss how to estimate the eigenvalue bound of $C^{-1}A$ if $\sigma > 2$. Though $\frac{1}{2-\sigma}$ is an upper bound provided $\sigma < 2$, it is still very large if σ is close to 2. In the second half of this section, we reconsider the bound in this case. The discussion is based on our generalization of the well-known inequality $\rho(A) \leq \operatorname{tr}(A)$ for A s.p.s.d. to block form. It is shown that $2 - \sigma + 2(\sigma - 1)m$ is another upper bound of $\lambda(C^{-1}A)$ if $1 < \sigma \leq 2$ and A is an $m \times m$ block s.p.s.d. matrix.

Let $\tilde{M} = (I + \tilde{L})(I + \tilde{L}^T)$, where $\tilde{L}$ stands for the same matrix as in Theorem 2.1. The following result gives a method to estimate upper bounds of eigenvalues

$$\lambda_i(C^{-1}A) \quad \text{if} \quad \sigma \geq 2 - \lambda_{n-i+1}^{\frac{1}{2}}(\tilde{M}) = 2 - \lambda_i^{-\frac{1}{2}}(\tilde{M}^{-1}).$$

THEOREM 3.1. *Let matrices A and C satisfy the conditions of* Theorem 2.1. *If $\kappa_i \geq \lambda_i(\tilde{M}^{-1})$, where $\tilde{M} = (I + \tilde{L})(I + \tilde{L}^T)$, and $\sigma \geq 2 - \kappa_i^{-\frac{1}{2}}$, then*

$$\lambda_i(C^{-1}A) \leq (\sigma - 2)\kappa_i + \kappa_i^{\frac{1}{2}}. \tag{3}$$

Proof. Using Weyl's theorem (cf. Parlett [11, p. 192]), we find for any $\mu < 2$ and $\mu \leq \sigma$ that

$$\begin{aligned}\lambda_i(M(\sigma)) &= \lambda_i((I+\tilde{L})^{-1} + (I+\tilde{L}^T)^{-1} + (\sigma-2)(I+\tilde{L})^{-1}(I+\tilde{L}^T)^{-1}) \\ &\leq \lambda_{\max}((I+\tilde{L})^{-1} + (I+\tilde{L}^T)^{-1} + (\mu-2)(I+\tilde{L})^{-1}(I+\tilde{L}^T)^{-1}) \\ &\quad +(\sigma-\mu)\lambda_i(I+\tilde{L})^{-1}(I+\tilde{L}^T)^{-1} \\ &\leq \frac{1}{2-\mu} + (\sigma-\mu)\kappa_i.\end{aligned}$$

Therefore

$$\begin{aligned}\lambda_i(M(\sigma)) &\leq \min_{\mu<2}\left((2-\mu)^{-1} + (\sigma-\mu)\kappa_i\right) \\ &= 2\kappa_i^{\frac{1}{2}} + (\sigma-2)\kappa_i.\end{aligned}$$

The minimum is taken for $\mu = 2 - \kappa_i^{-\frac{1}{2}}$. An application of Theorem 2.1 ends the proof of inequality (3). □

The bound given by (3) is clearly an improvement of $1/(2-\sigma)$ if $\sigma \geq 2 - \lambda_{n-i+1}^{-\frac{1}{2}}(\tilde{M})$.

Let now

$$I+\tilde{L} = \begin{pmatrix} I_{n_1} & & & \\ L_{21} & I_{n_2} & & \\ \vdots & \ddots & \ddots & \\ L_{m1} & \cdots & L_{m,m-1} & I_{n_m} \end{pmatrix},$$

$\tilde{L}_1$ and $\tilde{L}_2$ be the lower triangular submatrices of $I+\tilde{L}$ of the form

$$\tilde{L}_1 = \begin{pmatrix} I_{n_1} & & & \\ L_{21} & \ddots & & \\ \vdots & \ddots & \ddots & \\ L_{k1} & \cdots & L_{k,k-1} & I_{n_k} \end{pmatrix},$$

$$\tilde{L}_2 = \begin{pmatrix} I_{n_{k+1}} & & & \\ L_{k+2,k+1} & \ddots & & \\ \vdots & \ddots & \ddots & \\ L_{m,k+1} & \cdots & L_{m,m-1} & I_{n_m} \end{pmatrix},$$

and let $\tilde{L}_{21}$ be $(m-k) \times k$ block submatrix of $I+\tilde{L}$ at the southwest corner, i.e.,

$$\tilde{L}_{21} = \begin{pmatrix} L_{k+1,1} & \cdots & L_{k+1,k} \\ \vdots & & \vdots \\ L_{m1} & \cdots & L_{mk} \end{pmatrix}.$$

Hence

$$I+\tilde{L} = \begin{pmatrix} \tilde{L}_1 & 0 \\ \tilde{L}_{21} & \tilde{L}_2 \end{pmatrix}.$$

Let $\Lambda(D)$ denote the set of all nonzero eigenvalues of matrix D. Let G_1 be an $n \times m$ matrix and G_2 is an $m \times n$ matrix. It is well known that $\Lambda(G_1G_2)=\Lambda(G_2G_1)$. Set $\tilde{M}_i = \tilde{L}_i\tilde{L}_i^T$, $i = 1, 2$, and denote $\tilde{k} = n_1 + \cdots + n_k$, $\tilde{p} = n_{k+1} + \cdots + n_m$. We now consider how to estimate $\lambda_i(\tilde{M})$. To this end, we need the following lemmas.

LEMMA 3.2. *Let α_1 and α_2 be positive numbers, B be a $p \times k$ real matrix, and D be a matrix of the form*

$$D = \begin{pmatrix} \alpha_1^{\frac{1}{2}} I_k & 0 \\ B & \alpha_2^{\frac{1}{2}} I_p \end{pmatrix} \begin{pmatrix} \alpha_1^{\frac{1}{2}} I_k & B^T \\ 0 & \alpha_2^{\frac{1}{2}} I_p \end{pmatrix}, \tag{4}$$

where I_k denotes the unit matrix of order k. Then the eigenvalues of D are given by

$$\lambda_i(D) = \begin{cases} f_+(\alpha_1, \alpha_2, \mu_i), & \text{if } 1 \le i \le \min(p, k), \\ \alpha_1, & \text{if } p < i \le k, \\ \alpha_2, & \text{if } k < i <\le p, \\ f_-(\alpha_1, \alpha_2, \mu_{k+p+1-i}), & \text{if } \max(p, k) < i \le k + p, \end{cases}$$

where $\mu_1, \ldots, \mu_{\min(p,k)}$ are the first $\min(p, k)$ eigenvalues of BB^T, $f_+(\alpha, \beta, \mu)$ and $f_-(\alpha, \beta, \mu)$ are the largest, respectively, the smallest zero of the function

$$t \rightsquigarrow (\alpha - t)(\beta - t) - \mu t.$$

Proof. If $k \ge p$, a computation, using a block decomposition of $\lambda I_{k+p} - D$ shows the characteristic polynomial of D

$$\begin{aligned} f_D(\lambda) &= \det(\lambda I_{k+p} - D) \\ &= (\lambda - \alpha_1)^{k-p} \det((\lambda - \alpha_1)(\lambda - \alpha_2) I_p - \lambda BB^T) \\ &= (\lambda - \alpha_1)^{k-p} \prod_{i=1}^{p} ((\lambda - \alpha_1)(\lambda - \alpha_2) - \mu_i \lambda). \end{aligned}$$

Thus, $f_+(\alpha_1, \alpha_2, \mu_i)$, $f_-(\alpha_1, \alpha_2, \mu_i)$, and α_1 are eigenvalues of D. Since $f_+(\alpha_1, \alpha_2, x)$ and $f_-(\alpha_1, \alpha_2, x)$ are monotonously increasing and monotonously decreasing, respectively, the lemma follows immediately due to the fact that $f_+(\alpha_1, \alpha_2, x) \ge \max(\alpha_1, \alpha_2)$ and $f_-(\alpha_1, \alpha_2, x) \le \min(\alpha_1, \alpha_2)$ for $x \ge 0$. Note that $\Lambda(B^TB) = \Lambda(BB^T)$. Similarly, one can prove the case $k < p$. □

LEMMA 3.3. *With $L = \begin{pmatrix} L_1 & 0 \\ L_{21} & L_2 \end{pmatrix}$, where L_i are nonsingular, and $\tilde{L} = \begin{pmatrix} \sigma_1 & 0 \\ L_{21} & \sigma_2 \end{pmatrix}$, where $\sigma_i^2 = \lambda_{\min}(L_i^T L_i)$, we have that $\lambda_i(LL^T) \ge \lambda_i(\tilde{L}\tilde{L}^T)$.*

Proof. Note that it is readily seen that $\lambda_i(BG) \ge \lambda_i(DG)$ if B, D, G and $B - D$ are s.p.s.d. Hence, we have

$$\begin{aligned} \lambda_i(LL^T) &= \lambda_i\left(\begin{pmatrix} L_1 & 0 \\ 0 & I_p \end{pmatrix} \begin{pmatrix} I_k & 0 \\ L_{21} & L_2 \end{pmatrix} \begin{pmatrix} I_k & L_{21}^T \\ 0 & L_2^T \end{pmatrix} \begin{pmatrix} L_1^T & 0 \\ 0 & I_p \end{pmatrix} \right) \\ &= \lambda_i\left(\begin{pmatrix} L_1^T L_1 & 0 \\ 0 & I_p \end{pmatrix} \begin{pmatrix} I_k & 0 \\ L_{21} & L_2 \end{pmatrix} \begin{pmatrix} I_k & L_{21}^T \\ 0 & L_2^T \end{pmatrix} \right) \\ &\ge \lambda_i\left(\begin{pmatrix} \sigma_1^2 I_k & 0 \\ 0 & I_p \end{pmatrix} \begin{pmatrix} I_k & 0 \\ L_{21} & L_2 \end{pmatrix} \begin{pmatrix} I_k & L_{21}^T \\ 0 & L_2^T \end{pmatrix} \right) \\ &= \lambda_i\left(\begin{pmatrix} \sigma_1 I_k & 0 \\ 0 & I_p \end{pmatrix} \begin{pmatrix} I_k & 0 \\ L_{21} & L_2 \end{pmatrix} \begin{pmatrix} I_k & L_{21}^T \\ 0 & L_2^T \end{pmatrix} \begin{pmatrix} \sigma_1 I_k & 0 \\ 0 & I_p \end{pmatrix} \right) \end{aligned}$$

$$= \lambda_i \left(\begin{pmatrix} \sigma_1 I_k & 0 \\ L_{21} & L_2 \end{pmatrix} \begin{pmatrix} \sigma_1 I_k & L_{21}^T \\ 0 & L_2^T \end{pmatrix} \right) \equiv \tilde{\lambda}_i.$$

Similarly, it follows that $\tilde{\lambda}_i \geq \lambda_i(\tilde{L}\tilde{L}^T)$. □

Since $\Lambda(\tilde{L}_{21}\tilde{L}_{21}^T) = \Lambda(\tilde{L}_{21}^T\tilde{L}_{21})$, it follows from the proof of Theorem 3.1 with using Lemmas 3.2 and 3.3 that

$$\lambda_i(\tilde{M}) \geq \begin{cases} f_+(\lambda_{\min}(\tilde{M}_1), \lambda_{\min}(\tilde{M}_2), \mu_i), & \text{if } 1 \leq i \leq \min(\tilde{p}, \tilde{k}), \\ \lambda_{\min}(\tilde{M}_1), & \text{if } \tilde{p} < i \leq \tilde{k}, \\ \lambda_{\min}(\tilde{M}_2), & \text{if } \tilde{k} < i \leq \tilde{p}, \\ f_-(\lambda_{\min}(\tilde{M}_1), \lambda_{\min}(\tilde{M}_2), \mu_{n+1-i}), & \text{if } \max(\tilde{p}, \tilde{k}) < i \leq n, \end{cases}$$

where $\mu_1, \mu_2, \ldots, \mu_{\min(\tilde{p},\tilde{k})}$ are the first $\min(\tilde{p}, \tilde{k})$ eigenvalues of $L_{21}L_{21}^T$ numbered in a nonincreasing order.

LEMMA 3.4. *Let B be a $p \times k$ real matrix and D be a matrix of the form*

$$D = \begin{pmatrix} \alpha_1 I_k & B^T \\ B & \alpha_2 I_p \end{pmatrix}.$$

Then the eigenvalues of D are given by

$$\lambda_i(D) = \begin{cases} g_+(\alpha_1, \alpha_2, \beta_i), & \text{if } 1 \leq i \leq \min(p, k), \\ \alpha_1, & \text{if } p < i \leq k, \\ \alpha_2, & \text{if } k < i \leq p, \\ g_-(\alpha_1, \alpha_2, \beta_{k+p+1-i}), & \text{if } \max(p, k) < i \leq k + p, \end{cases}$$

where $\beta_1, \beta_2, \ldots,$ and $\beta_{\min(p,k)}$ are the first $\min(p, k)$ eigenvalues of BB^T,

$$g_+(\alpha_1, \alpha_2, x) = \frac{1}{2}(\alpha_1 + \alpha_2 + ((\alpha_1 - \alpha_2)^2 + 4x)^{\frac{1}{2}}),$$

$$g_-(\alpha_1, \alpha_2, x) = \frac{1}{2}(\alpha_1 + \alpha_2 - ((\alpha_1 - \alpha_2)^2 + 4x)^{\frac{1}{2}}).$$

Proof. The proof is similar to that of Lemma 3.2. □

THEOREM 3.5. *Let $A = (A_{ij})_{i,j=1}^m$ be a block matrix partitioning of an* s.p.s.d. *matrix. Then*

$$\rho(A) \leq \sum_{i=1}^{m} \rho(A_{ii}). \tag{5}$$

Proof. Consider first the case $m = 2$. Let

$$B = \begin{pmatrix} \rho(A_{11})I & A_{12} \\ A_{21} & \rho(A_{22})I \end{pmatrix},$$

which is clearly s.p.s.d. and $\rho(A) \leq \rho(B)$. Lemma 3.4 shows that

$$\rho(A_{11}) + \rho(A_{22}) - ((\rho(A_{11}) - \rho(A_{22})^2 + \|A_{12}\|_2^2)^{\frac{1}{2}}) = 2\lambda_{\min}(B) \geq 0$$

and, hence

$$\rho(B) \leq \frac{1}{2}(\rho(A_{11}) + \rho(A_{22}) + ((\rho(A_{11}) - \rho(A_{22})^2 + \|A_{12}\|_2^2)^{\frac{1}{2}})) \leq \rho(A_{11}) + \rho(A_{22}).$$

By induction, we find for any s.p.s.d. matrix, $A = (A_{ij})_{i,j=1}^m$ (5) holds. □

If A has scalar form, (5) reduces to the well-known inequality

$$\rho(A) \leq \mathrm{tr}(A).$$

If $1 < \sigma \leq 2$, we give now an alternative method to estimate an upper bound of the eigenvalues of $C^{-1}A$, which yields $2 - \sigma + 2(\sigma - 1)m$ as an upper bound if A is s.p.s.d.

THEOREM 3.6. *Let A and C satisfy all conditions of* Theorem 2.1. *If A is* s.p.s.d. *and $1 < \sigma \leq 2$, then*

$$\lambda_{\max}(C^{-1}A) \leq 2 - \sigma + 2(\sigma - 1)m.$$

Proof. It holds that

$$\begin{aligned}
&(I + \tilde{L})^{-1}(I + \tilde{L}^T)^{-1} \\
&= \left(\sum_{i=0}^{m-1}(-1)^i \tilde{L}^i\right)\left(\sum_{i=0}^{m-1}(-1)^i (\tilde{L}^T)^i\right) \\
&= I + \sum_{i=1}^{m-1}(-1)^i \tilde{L}^i + \sum_{i=1}^{m-1}(-1)^i (\tilde{L}^T)^i + \left(\sum_{i=1}^{m-1}(-1)^i \tilde{L}^i\right)\left(\sum_{i=1}^{m-1}(-1)^i (\tilde{L}^T)^i\right) \\
&= M(2) - I + \left(\sum_{i=1}^{m-1}(-1)^i \tilde{L}^i\right)\left(\sum_{i=1}^{m-1}(-1)^i (\tilde{L}^T)^i\right).
\end{aligned}$$

Since $(\sum_{i=1}^{m-1}(-1)^i \tilde{L}^i)(\sum_{i=1}^{m-1}(-1)^i (\tilde{L}^T)^i)$ is s.p.s.d. and $1 < \sigma \leq 2$, Theorem 2.1 and the above yield

$$\begin{aligned}
&\lambda_i(C^{-1}A) \\
&\leq \lambda_i(M(2) + (\sigma - 2)(I + \tilde{L})^{-1}(I + \tilde{L}^T)^{-1}) \\
&= \lambda_i\left((2 - \sigma)I + (\sigma - 1)M(2) + (\sigma - 2)\left(\sum_{i=1}^{m-1}(-1)^i \tilde{L}^i\right)\left(\sum_{i=1}^{m-1}(-1)^i (\tilde{L}^T)^i\right)\right) \\
&\leq 2 - \sigma + (\sigma - 1)\lambda_i(M(2)).
\end{aligned}$$

Since A is s.p.s.d. and C is s.p.d., Proposition 2.4 shows that $M(2)$ is s.p.s.d. Since the diagonal part of $M(2)$ is $2I$, (5) finishes the proof. □

Under some additional assumptions, the bound of Theorem 3.6 can be reduced. For example, if

1. A is a Stieltjes matrix,
2. $L = L_A$,
3. X is a Stieltjes matrix such that offdiag$(X) \leq$offdiag(D_A), and
4. there is a positive vector $\mathbf{v}$ such that

$$C\mathbf{v} \geq 0, \tag{6}$$

$$(L^T - L)\mathbf{v} \leq C\mathbf{v}, \tag{7}$$

$$(X + L^T)\mathbf{v} \geq A\mathbf{v}, \tag{8}$$

$$L_p^T \mathbf{v} > 0,$$

where $X = L_p P L_p^T$ denotes the point LU decomposition of X, then it has been shown in [9] that

$$\lambda_{\max}(C^{-1}A) \leq m + 1.$$

These assumptions imply actually that $2X + L + L^T - A$ is s.p.s.d. In this case, $2X + L + L^T - A$ is a Stieltjes matrix. This follows because, as has been shown in [9], (6) and (7) imply $(X + L)\mathbf{v} \geq 0$. Hence using (8) shows that

$$(2X + L + L^T - A)\mathbf{v} \geq 0,$$

which implies that $2X + L + L^T - A$ is s.p.s.d.

The matrix $M(2)$ acts as a key for estimating the eigenvalues $\lambda_i(C^{-1}A)$ as we have seen. In general, $\lambda_i(M(2))$ can be estimated by using Lemma 3.4. Denote $\bar{M}_i = \tilde{L}_i^{-1} + (\tilde{L}_i^T)^{-1}$, $i = 1, 2$, and $\eta_i = \lambda_{\max}(\bar{M}_i)$. According to the partitioning of $I + \tilde{L}$, it is easy to check that

$$M(2) = \begin{pmatrix} \bar{M}_1 & \bar{L}_{21}^T \\ \bar{L}_{21} & \bar{M}_2 \end{pmatrix}$$

and hence

$$\lambda_i(M(2)) \leq \lambda_i \begin{pmatrix} \eta_1 I_{\tilde{k}} & \bar{L}_{21}^T \\ \bar{L}_{21} & \eta_2 I_{\tilde{p}} \end{pmatrix}.$$

Again using $\Lambda(\bar{L}_{21}\bar{L}_{21}^T) = \Lambda(\bar{L}_{21}^T\bar{L}_{21})$ and Lemma 3.4, we have that

$$\lambda_i(M(2)) \leq \begin{cases} g_+(\eta_1, \eta_2, \gamma_i), & \text{if } 1 \leq i \leq \min(\tilde{p}, \tilde{k}), \\ \eta_1, & \text{if } \tilde{p} < i \leq \tilde{k}, \\ \eta_2, & \text{if } \tilde{k} < i \leq \tilde{p}, \\ g_-(\eta_1, \eta_2, \gamma_{n+1-i}), & \text{if } \max(\tilde{p}, \tilde{k}) < i \leq n, \end{cases}$$

where $\bar{L}_{21} = -\tilde{L}_1^{-1}\tilde{L}_{21}\tilde{L}_2^{-1}$, $\gamma_1, \ldots, \gamma_{\min(\tilde{p},\tilde{k})}$ are the first $\min(\tilde{p}, \tilde{k})$ eigenvalues of $\bar{L}_{21}\bar{L}_{21}^T$ numbered in a nonincreasing order.

4. Application to generalized SSOR preconditioned matrices. As an application of the results presented in §§2 and 3, we now consider upper bounds of the condition number of the preconditioned matrix when the generalized symmetric successive overrelaxation (SSOR) method is applied to symmetric block tridiagonal matrices.

Let A be a block tridiagonal matrix of the form

$$A = \text{blocktridiag}(A_{i,i-1}, A_{ii}, A_{i,i+1}),$$

where $A_{ii} = \text{tridiag}(-b, a, -b)$ and $A_{i,i-1} = A_{i,i+1} = -cI$, $i = 1, 2, \ldots, m$. All blocks have order $n \times n$. In addition, we assume that $b, c \geq 0$ and $a \geq 2(b + c)$. Consider

$$A = D_A - L - L,$$

a splitting of A, and the generalized SSOR preconditioned matrix

$$C = (D - L)D^{-1}(D - L^T),$$

where $D_A = \text{blockdiag}(A_{11}, A_{22}, \ldots, A_{mm})$, L is the lower block tridiagonal part of A, $D = \text{blockdiag}(D_1, D_2, \ldots, D_m)$ partitioned as D_A.

We compute a preconditioner C for A in a common way as follows (see [3]):

$$D_1 = A_{11},$$
$$D_i = A_{ii} - A_{i,i-1}X_{i-1}A_{i-1,i} + D_i', \qquad i = 2, 3, \ldots, m,$$

where X_i, $i \geq 1$, is a sparse approximation to D_i^{-1} and D_i' is a diagonal matrix such that

$$D_i'\mathbf{v} = A_{i,i-1}(X_{i-1} - D_{i-1}^{-1})A_{i-1,i}\mathbf{v}, \qquad i = 2, 3, \ldots, m$$

for some positive vector v. Hence we have

$$D_1\mathbf{v} = A_{11}v, \tag{9}$$
$$D_i\mathbf{v} = (A_{ii} - A_{i,i-1}D_{i-1}^{-1}A_{i-1,i})\mathbf{v}, \qquad i = 2, 3, \ldots, m. \tag{10}$$

Since $A_{ii} = (a - 2b)I + b\,\text{tridiag}(-1, 2, -1)$, the smallest eigenvalue of A_{ii} is

$$a - 2b + b\left(2\sin\frac{\pi}{2(n+1)}\right)^2,$$

denoted by λ, where n is the order of A_{ii}. Let $\mathbf{v}$ be the eigenvector of A_{ii} corresponding to λ. Equations (9) and (10) imply that $\mathbf{v}$ is also an eigenvector of D_i and the corresponding smallest eigenvalue of D_i becomes

$$\lambda_1 = \lambda, \qquad \lambda_i = \lambda - c^2\lambda_{i-1}^{-1}, \qquad i = 2, 3, \ldots, m.$$

It is readily seen that λ_i converges monotonically to the lower bound

$$\tilde{\lambda} = \frac{1}{2}(\lambda + (\lambda^2 - 4c^2)^{\frac{1}{2}}).$$

Let $\sigma = \frac{\lambda}{\lambda - c^2\tilde{\lambda}^{-1}}$. We have $\sigma = 1 + \frac{c^2}{\lambda\tilde{\lambda} - c^2} < 2$. A computation shows that

$$\sigma D_1\mathbf{v} - A_{11}\mathbf{v} \geq 0,$$

$$\sigma D_i\mathbf{v} - A_{ii}\mathbf{v} = \lambda\frac{\lambda - c^2\lambda_{i-1}^{-1}}{\lambda - c^2\tilde{\lambda}^{-1}}\mathbf{v} - \lambda\mathbf{v} \geq 0, \qquad i = 2, 3, \ldots, m,$$

which implies that $\sigma X - L - L^T - A$ is s.p.s.d. Now Proposition 2.4 shows that

$$\rho(C^{-1}A) \leq \frac{1}{2 - \sigma} = \frac{\lambda\tilde{\lambda} - c^2}{\lambda\tilde{\lambda} - 2c^2}.$$

On the other hand, using $\lambda \geq 2c$, it is also seen that $\lambda_i \geq \hat{\lambda}_i$, which satisfy

$$\hat{\lambda}_1 = 2c, \ \hat{\lambda}_i = 2c - c^2\hat{\lambda}_i^{-1}, \ i = 2, 3, \ldots, m.$$

Hence $\hat{\lambda}_i = \frac{i+1}{i}c$. Let

$$\alpha = \frac{\hat{\lambda}_1}{\hat{\lambda}_1 - c^2\hat{\lambda}_m^{-1}} = 2 - \frac{2}{m+2}.$$

Similarly, we find that $\alpha D - L - L^T - A$ is also s.p.s.d.

Consider the lower block tridiagonal matrix $T = I - D^{1/2}LD^{-1/2}$. We have

$$T^{-1} = (T_{ij}) = \sum_{t=0}^{m-1} (D^{-\frac{1}{2}}LD^{-\frac{1}{2}})^t,$$

where $T_{ii} = I_n$, $T_{ij} = c^{i-j}D_i^{-\frac{1}{2}}D_{i-1}^{-1}\dots D_{j+1}^{-1}D_j^{-\frac{1}{2}}$, $i > j$, $T_{ij} = 0$, $i < j$.

Partition $(T^T)^{-1}T^{-1}$ into an $m \times m$ block matrix (B_{ij}) consistently with the partitioning of A. Clearly, $(T^T)^{-1}T^{-1}$ is a nonnegative matrix. Applying $D_i^{-1}\mathbf{v} \le \frac{i}{(i+1)c}\mathbf{v}$ shows that

$$T_{ij}\mathbf{v} \le \left(\frac{j(j+1)}{i(i+1)}\right)^{\frac{1}{2}}\mathbf{v} \le \frac{j+1}{i+1}\mathbf{v} \quad \text{and} \quad T_{ij}^T\mathbf{v} \le \frac{j+1}{i+1}\mathbf{v}, \quad i \ge j.$$

Hence,

$$B_{ij}\mathbf{v} = \sum_{k=1}^{m} T_{ki}^T T_{kj}\mathbf{v} \le \sum_{k\ge\max(i,j)}^{m} \frac{(i+1)(j+1)}{(k+1)^2}\mathbf{v},$$

$$\sum_{j=1}^{m} B_{ij}\mathbf{v} = \sum_{j=1}^{i-1}\sum_{k=i}^{m} \frac{(i+1)(j+1)}{(k+1)^2}\mathbf{v} + \sum_{j=i}^{m}\sum_{k=j}^{m} \frac{(i+1)(j+1)}{(k+1)^2}\mathbf{v}$$

$$= \sum_{k=i}^{m}\sum_{j=1}^{k} \frac{(i+1)(j+1)}{(k+1)^2}\mathbf{v} = \frac{i+1}{2}\sum_{k=i}^{m}\left(1 + \frac{1}{k+1} - \frac{2}{(k+1)^2}\right)\mathbf{v}$$

$$< \frac{i+1}{2}\sum_{k=i}^{m}\left(1 + \frac{1}{k+1} - \frac{2}{(k+1)(k+2)}\right)\mathbf{v}$$

$$\le \frac{i+1}{2}\left(m - i + 1 + \int_{i+1}^{m+1}\frac{dx}{x} + \frac{1}{i+1} - 2\sum_{k=i}^{m}\left(\frac{1}{k+1} - \frac{1}{k+2}\right)\right)\mathbf{v}$$

$$= \frac{1}{2}\left((i+1)\left(m - i + 1 + \log(m+1) - \log(i+1) + \frac{2}{m+2}\right) - 1\right)\mathbf{v}$$

$$< \frac{1}{8}\left(\left(m + 1 + \log(2) + \frac{2}{m+2}\right)\left(m + 3 + \log(2) + \frac{2}{m+2}\right) - 4\right)\mathbf{v}$$

$$< \frac{1}{8}(m + 2 + \log(2))^2\mathbf{v},$$

which implies that

$$\rho(\tilde{M}^{-1}) = \rho((T^T)^{-1}T^{-1}) \le \frac{1}{8}(m + 2 + \log(2))^2 \equiv \kappa,$$

where $\tilde{M}$ stands for the same matrix as in Theorem 3.1. Since $\kappa \ge \rho(\tilde{M}^{-1})$ and $\alpha > 2 - \kappa^{-1/2}$, applying Theorem 3.1 shows that

$$\rho(C^{-1}A) \le (\alpha - 2)\kappa + 2\kappa^{\frac{1}{2}} \le \frac{2\sqrt{2} - 1}{4}(m + 2 + \log(2)).$$

This bound is approximately $0.4571(m + 2 + \log(2))$. The result can be further improved if we can estimate $\rho(\tilde{M}^{-1})$ more accurately. Application of the result in [9]

(Theorem 4.3) shows only an $m/2$ upper bound for this example, although the result requires that A is a Stieltjes matrix, $L = L_A$ and some other additional conditions.

Furthermore, because $A-C$ is a Z-matrix and $(A-C)\mathbf{v} = 0$, we have $\lambda_{\min}(C^{-1}A) = 1$ and, therefore,

$$\operatorname{cond}(C^{-1}A) \leq \min\left(\frac{\lambda\tilde{\lambda} - c^2}{\lambda\tilde{\lambda} - 2c^2}, \frac{2\sqrt{2}-1}{4}(m + 2 + \log(2))\right).$$

For the model second order elliptic difference equation on a rectangular $n \times m$ mesh with uniform meshwidth $h = \frac{1}{n+1}$, we have $A_{ii} = \operatorname{tridiag}(-1, 4, -1)$, $c = 1$. In this case, using the previously given bound on λ, we find

$$\frac{\lambda\tilde{\lambda} - 1}{\lambda\tilde{\lambda} - 2} \simeq \frac{n+1}{2\pi}.$$

Therefore,

$$\operatorname{cond}(C^{-1}A) \leq \min\left(\frac{n+1}{2\pi}, \frac{2\sqrt{2}-1}{4}(m + 2 + \log(2))\right).$$

It turns out that the second part holds also for the more common choice of the vector $\mathbf{e} = (1, 1, \ldots, 1)^T$, because we have $\tilde{D}_i\mathbf{e} \geq \frac{i+1}{i}\mathbf{e}$, where $\tilde{D}_i$ are the corresponding matrices of D_i obtained by using $\mathbf{e}$.

The general bound $2m$ of the condition number is hence not very accurate for the model type problem. Bounds involving only m, the number of blocks, are of particular interest when an elliptic second order difference equation is solved on an oblong rectangular domain with number of nodepoints $N_1 \times N_2$ where we assume that $N_1 \gg N_2$. If we number the points such that the order of the matrix blocks is N_1, i.e., there are $m = N_2$ blocks in the main diagonal, then applying Theorem 3.6 shows that

$$\operatorname{cond}(C^{-1}A) \leq 2N_2,$$

or $0.4571(N_2 + 2 + \log(2))$ for the model problem, both of which hence do not depend on N_1. It is therefore efficient to choose big blocks for such domains.

Acknowledgment. Careful comments by two anonymous referees regarding the presentation of some results are gratefully acknowledged.

REFERENCES

[1] O. AXELSSON, *A class of iterative methods for finite element equations*, Comput. Methods Appl. Mech. Engrg., 9 (1976), pp. 123–137.

[2] ———, *Solution of linear systems of equations: iterative methods*, in Sparse Matrix Techniques, Lecture Notes in Mathematics 572, V. A. Barker, ed., Springer-Verlag, Berlin, Heidelberg, New York, 1977, pp. 1–50.

[3] ———, *Incomplete block-matrix factorization preconditioning methods. the ultimate answer?*, J. Comput. Appl. Math., 12 & 13 (1985), pp. 3–18.

[4] ———, *Bounds of eigenvalues of preconditioned matrices*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 847–862.

[5] O. AXELSSON AND G. LINDSKOG, *On the rate of convergence of the preconditioned conjugate gradient method*, Numer. Math., 48 (1986), pp. 499–523.

[6] R. BEAUWENS, *Upper eigenvalue bounds for pencils of matrices*, Linear Algebra Appl., 62 (1984), pp. 87–104.

[7] A. GREENBAUM, *Comparision of splittings used with the conjugate gradient algorithm*, Numer. Math., 33 (1979), pp. 181–194.

[8] A. JENNINGS, *Influence of the eigenvalues spectrum of the convergence rate of the conjugate gradient method*, IMA J. of Numer. Anal., 20 (1977), pp. 61–72.

[9] M. M. MAGOLU, *Analytical bounds for block approximate factorization methods*, Linear Algebra Appl., 179 (1993), pp. 33–57.

[10] M. M. MAGOLU AND Y. NOTAY, *On the conditioning analysis of block approximate factorization methods*, Linear Algebra Appl., 154-156 (1991), pp. 583–599.

[11] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[12] A. VAN DER SLUIS AND H. A. VAN DER VORST, *The rate of convergence conjugate gradients*, Numer. Math. 48 (1986), pp. 543–560.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, pp. 1086–1099, October 1995

004

DIAGONAL DOMINANCE IN THE PARALLEL PARTITION METHOD FOR TRIDIAGONAL SYSTEMS*

CHRIS H. WALSHAW†

Abstract. The partition method for the parallel solution of tridiagonal linear systems is discussed and the coefficients of the reduced global system derived. It is shown that if the full system is diagonally dominant then the reduced system retains this property. This has important implications for the stability of calculations in this reduced system and eliminates the need for global pivoting with its expensive communication overhead.

Key words. diagonal dominance, parallel partition method, tridiagonal linear systems

AMS subject classifications. 65F05, 65Y05, 65F50, 15A06

1. Introduction.

This paper deals with the solution of a tridiagonal linear system of equations,

$$T\mathbf{x} \equiv \begin{bmatrix} \beta_1 & \gamma_1 & & & & \\ \alpha_2 & \beta_2 & \gamma_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & & & \gamma_{n-1} & \\ & & & \alpha_n & \beta_n & \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_n \end{bmatrix} \equiv \mathbf{d}, \tag{1}$$

on a network of P distributed memory processors. It is assumed that $P \ll n$ and each node has approximately $m \approx n/P$ unknowns assigned to it.

The system is said to be strictly diagonally dominant if

$$|\beta_i| > |\alpha_i| + |\gamma_i| \qquad i = 1, \ldots, n, \tag{2}$$

where $\alpha_1 = \gamma_n = 0$. This is a useful property as it guarantees that a (sequential) method of solution such as Gaussian elimination can be performed without row or column interchanges and that the computations are stable with respect to the growth of rounding errors. It is even more valuable for parallel algorithms as parallel pivoting with its inherent global communication would account for a much higher overhead than a sequential version. Throughout the following we deal with diagonally dominant systems and consider the implications of this property.

1.1. The parallel solution of tridiagonal systems.

Many nearest neighbour problems (e.g., those involving spatial finite difference approximations) have at their heart a tridiagonal matrix. Their sparsity pattern suggests that, for large values of n, such systems are ideal candidates for parallelisation. However the common methods for solving tridiagonal systems, such as Gaussian elimination or matrix decomposition, tend to be inherently sequential in nature and, as a result, this topic was one of the earliest subjects investigated in the field of parallel linear algebra. Amongst the first such schemes were those introduced by Stone in the early seventies [8], [9] employing a recursive doubling algorithm. Odd-even cyclic reduction was another early method developed by Golub and stabilized by Buneman [1], [2]. Since its first introduction for

* Received by the editors March 11, 1993; accepted for publication (in revised form) by B. Kågström September 22, 1994.

† School of Mathematics, Statistics and Computing, University of Greenwich, Wellington Street, Woolwich, London, SE18 6PF, United Kingdom (c.walshaw@gre.ac.uk).

symmetric constant coefficient matrices it has been extended to general nonsymmetric tridiagonal systems [10].

In 1981 Wang [12] introduced what he called a new (partition) method. This has since also been referred to as the spike algorithm (e.g., see spikes in Fig. 1) as it proceeds by a completely parallel local Gaussian elimination to transform the system from a tridiagonal one into a diagonal one with spikes or fill-ins at the interprocessor boundaries. With one communication at the end of this local reduction phase, the result is a global order $P-1$ tridiagonal system in terms of the boundary variables (the unknowns at the interprocessor boundaries). Figure 1 shows an example matrix before and after the elimination phase, the dotted lines representing the interprocessor boundaries and the bold $\times$ symbols the coefficients of the $O(P-1)$ system. This *reduced system* can then be solved globally with, for example, cyclic reduction or two-way Gaussian elimination and finally the internal solutions calculated each as a linear combination of up to two boundary variables.

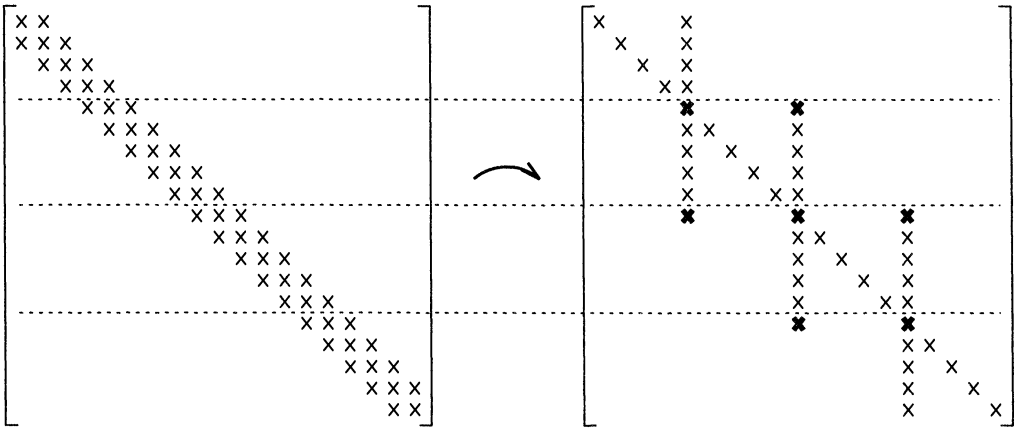


FIG. 1. *Matrix transformation for the partition algorithm*

A slightly different variant, the method of Sameh and Kuck [7] and Lawrie and Sameh [6], uses no communication in the reduction phase but results in a pentadiagonal matrix. However, in a generalisation to narrow banded systems, Johnsson [4] has shown that the extra communication is valuable and preserves positive definiteness and a form of diagonal dominance (albeit not in the classical sense; see §3 for further discussion).

The competing methods and their implementation on different architectures have all been comprehensively reviewed by Johnsson in [5]. He shows that a hybrid GECR (Gaussian elimination locally, cyclic reduction globally) algorithm has similar arithmetic complexity to the full cyclic reduction algorithm and that the best method therefore depends on communication considerations.

Having established that the partition method is competitive this paper proceeds in the following manner. In §2 the partition algorithm is described in full and values for the coefficients of the reduced matrix are derived. It is important for stability purposes to know which properties of the original system are retained by this reduced system and some previous results are discussed in §3. Finally, in §4, a proof that diagonal dominance is indeed retained by the reduced matrix is given.

2. The partition algorithm. Consider the solution, on a ring of P processors, of the order n tridiagonal system (1) which is assumed to be diagonally dominant (2). For convenience the system will be relabelled and throughout the following section, superscripts denote the partition a particular value is assigned to and subscripts the position in a vector.

2.1. Assignment of equations. The variables are distributed over the processors by dividing up the system into P tridiagonal subsystems of order m, interspersed with $P-1$ single equations. For simplicity it is assumed that n and P are such that m is the same for each processor and so

$$n = mP + P - 1.$$

An example of the relabelling of $\mathbf{x}$ is shown in (3); $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, and $\mathbf{d}$ are relabelled and partitioned in the same way.

(3)
$$\left.\begin{bmatrix} \vdots \\ \hline x_1^{i-1} \\ \vdots \\ x_m^{i-1} \\ \hline x^i \\ \hline x_1^i \\ \vdots \\ x_m^i \\ \hline x^{i+1} \\ \hline x_1^{i+1} \\ \vdots \\ x_m^{i+1} \\ \hline \vdots \end{bmatrix}\right\} \mathbf{x}^i.$$

The subsystems are assigned to partitions in a natural way; the first m equations to partition 1, equations $m+2$ to $2m+1$ to partition 2, etc., ... Thus each partition i contains the system

$$\begin{bmatrix} a_1^i \\ 0 \\ \vdots \\ 0 \end{bmatrix} x^i + \begin{bmatrix} b_1^i & c_1^i & & & \\ a_2^i & & & & \\ & \ddots & \ddots & \ddots & \\ & & & & c_{m-1}^i \\ & & & a_m^i & b_m^i \end{bmatrix} \begin{bmatrix} x_1^i \\ \vdots \\ x_m^i \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ c_m^i \end{bmatrix} x^{i+1} = \begin{bmatrix} d_1^i \\ \vdots \\ d_m^i \end{bmatrix},$$

where $a_1^1 = c_m^P = 0$. For brevity these systems will be denoted by

$$a_1^i \mathbf{e}_1 x^i + T^i \mathbf{x}^i + c_m^i \mathbf{e}_m x^{i+1} = \mathbf{d}^i, \tag{4}$$

where $\mathbf{e}_1$, $\mathbf{e}_m$ denote the standard unit basis vectors with $[e_j]_k = \delta_{jk}$.

Bordering each tridiagonal system (4) are the boundary equations

$$a^i x_m^{i-1} + b^i x^i + c^i x_1^i = d^i \tag{5}$$

and

$$a^{i+1} x_m^i + b^{i+1} x^{i+1} + c^{i+1} x_1^{i+1} = d^{i+1}. \tag{6}$$

2.2. The algorithm. The method can be broken into three phases. A completely parallel reduction phase (Phase 1), the solution of the reduced matrix (Phase 2), and the completely parallel back-substitution (Phase 3).

Phase 1. Reduction. The tridiagonal systems are manipulated to give $\mathbf{x}^i$ as a linear combination of the two local boundary variables x^i and x^{i+1},

$$\mathbf{x}^i = \mathbf{w}^i - \mathbf{r}^i x^i - \mathbf{s}^i x^{i+1}, \tag{7}$$

where

$$\mathbf{w}^i \overset{\text{def}}{=} (T^i)^{-1}\mathbf{d}^i, \quad \mathbf{r}^i \overset{\text{def}}{=} a_1^i (T^i)^{-1}\mathbf{e}_1, \quad \mathbf{s}^i \overset{\text{def}}{=} c_m^i (T^i)^{-1}\mathbf{e}_m.$$

Note that although $(T^i)^{-1}$ is written here there is no need to explicitly calculate the inverse of T^i and that $\mathbf{w}^i$, $\mathbf{r}^i$, and $\mathbf{s}^i$ can be found by Gaussian elimination or a matrix factorization algorithm such as LU-decomposition. Diagonal dominance guarantees that local pivoting is not required and that the decomposition is stable.

The expressions for x_1^i, x_1^{i+1}, x_m^{i-1}, and x_m^i are now substituted into the boundary equations to yield

$$-a^i r_m^{i-1} x^{i-1} + (b^i - a^i s_m^{i-1} - c^i r_1^i)x^i - c^i s_1^i x^{i+1} = d^i - a^i w_m^{i-1} - c^i w_1^i. \tag{8}$$

These can now be written as a reduced $O(P-1)$ tridiagonal system whose unknowns are the boundary variables:

$$R\mathbf{x} = \begin{bmatrix} \tilde{b}^2 & \tilde{c}^2 & & & \\ \tilde{a}^3 & & & & \\ & \ddots & \ddots & \ddots & \\ & & & & \tilde{c}^{P-1} \\ & & & \tilde{a}^P & \tilde{b}^P \end{bmatrix} \begin{bmatrix} x^2 \\ \\ \vdots \\ \\ x^P \end{bmatrix} = \begin{bmatrix} \tilde{d}^2 \\ \\ \vdots \\ \\ \tilde{d}^P \end{bmatrix}. \tag{9}$$

The coefficients are given by:

$$\begin{aligned} \tilde{a}^i &= -a^i r_m^{i-1} && i = 3, \ldots, P, \\ \tilde{b}^i &= b^i - a^i s_m^{i-1} - c^i r_1^i && i = 2, \ldots, P, \\ \tilde{c}^i &= -c^i s_1^i && i = 2, \ldots, P-1, \\ \tilde{d}^i &= d^i - a^i w_m^{i-1} - c^i w_1^i && i = 2, \ldots, P. \end{aligned} \tag{10}$$

Phase 2. Solution of the reduced system. The second phase is the solution of this reduced system for the boundary variables. Again this can be accomplished in a number of ways, for example, via cyclic reduction [3] or two-way Gaussian elimination/matrix decomposition, e.g., see [11]. This involves a simultaneous sweep in from both ends of the chain of processors followed by a simultaneous sweep out.

For any of these methods it is important to know if the diagonal dominance of the full system (1) is retained by the reduced system. Were it not the case the scheme might require pivoting, a costly operation globally. This question is discussed below.

Phase 3. Back-substitution. Once the reduced system is solved the values of the two boundary variables can be substituted into (7) to construct the solution. This operation is local and can be executed completely in parallel.

3. Properties of the reduced matrix. An important question arising in partition type methods is what properties possessed by the full matrix A are retained by the reduced matrix R after Phase 1. In his original description of the algorithm, [12, p. 182], Wang points to a paper by Wilkinson, [13, §8], as demonstrating that diagonal dominance is conserved. This comprehensive study of error analysis of matrix inversion can be used to show that the fill-ins are still dominated by the diagonals. However, published long before the partition method was of interest it does not (and had no reason to) address the elements of the reduced matrix—elements that do not arise as standard fill-ins.

Johnsson [4], in an extension of the partition method to narrow banded systems, gives a proof that a matrix which is diagonally dominant *in a matrix sense* gives rise to a diagonally dominant reduced matrix. This proof is given for narrow banded systems, but when restricted to the tridiagonal case requires the condition on the full system that, for each i, the boundary equations be diagonally dominant, i.e.,

$$|a^i| + |c^i| < |b^i| \tag{11}$$

and that the matrix systems satisfy the following dominance condition

$$\left\|(T^i)^{-1}a_1^i\mathbf{e}_1\right\| + \left\|(T^i)^{-1}c_m^i\mathbf{e}_m\right\| < 1. \tag{12}$$

This is, of course, not the same as the classical sense of diagonal dominance, (2), which in the new notation requires the boundary equation condition, (11), plus conditions on all the other equations in the system, i.e.,

$$|a_j^i| + |c_j^i| < |b_j^i| \tag{13}$$

for each $i = 1, \ldots, P$ and for each $j = 1, \ldots, m$.

Theorem 4.1 (below) assumes the classical characterisation of diagonal dominance for the full matrix and then goes on to show that, as a result, the reduced matrix is also diagonally dominant. This is a useful result as conditions (11) and (13), which henceforth shall be collectively referred to as the *classical diagonal dominance conditions*, are a much more natural request to make of a system than (11) and (12)—henceforth *matrix diagonal dominance conditions*. First, the classical conditions are much easier to check and second, they arise naturally in many applications.

Another important distinction is that the classical diagonal dominance conditions are completely independent of the number of processors and of the way the variables are distributed among the processors. The same is not true of the matrix conditions. As a counterexample, consider the (somewhat trivial) matrix

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 2\epsilon & 1 & 1-\epsilon & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1-\epsilon & 1 & 2\epsilon \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{14}$$

with ϵ small. If this is solved on two processors with the third equation as the boundary equation (which is diagonally dominant), then the first matrix condition,

$$\begin{aligned} \left\|(T^1)^{-1}c_2^1\mathbf{e}_2\right\| &= \left\| \begin{bmatrix} 1 & 0 \\ 2\epsilon & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1-\epsilon \end{bmatrix} \right\| \\ &= \left\| \begin{bmatrix} 1 & 0 \\ -2\epsilon & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1-\epsilon \end{bmatrix} \right\| = \left\| \begin{bmatrix} 0 \\ 1-\epsilon \end{bmatrix} \right\| < 1, \end{aligned}$$

(and by symmetry the second one also) is satisfied and hence the reduced matrix is diagonally dominant. However, if solved on three processors then neither of the boundary equations, two or four, are diagonally dominant and hence the proof can make no claims about the reduced matrix. Thus it can be seen that the matrix conditions are dependent on both P and on the way the system is distributed.

The above example also shows that a matrix satisfying the matrix conditions is not necessarily classically diagonally dominant. In the next example it can be seen that classical diagonal dominance does not imply the matrix condition. Consider, then,

$$T^i = I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad a_1^i = c_2^i = 1 - \epsilon.$$

This would stem from a tridiagonal system containing the entries

$$\begin{array}{cccccc} * & * & * & & & \\ & 1-\epsilon & 1 & 0 & & \\ & & 0 & 1 & 1-\epsilon & \\ & & & * & * & * \end{array}$$

which is diagonally dominant in the classical sense. However the matrix condition is that

$$\left\| I_2^{-1} \begin{bmatrix} 1-\epsilon \\ 0 \end{bmatrix} \right\| + \left\| I_2^{-1} \begin{bmatrix} 0 \\ 1-\epsilon \end{bmatrix} \right\| < 1,$$

which is certainly not true.

Thus it can be seen that neither of these two characterisations of diagonal dominance is implied by the other. Hence a matrix can fail either one of the tests but still render a diagonally dominant matrix.

Note that the property of *positive definiteness* has been addressed by Johnsson [4], who presents a proof employing permutation matrices that a *positive definite symmetric* matrix retains those qualities in reduction.

4. A diagonal dominance proof.

THEOREM 4.1. *Consider a tridiagonal matrix T as proposed in* (1) *and a reduced matrix R as derived in* (9). *If T is strictly diagonally dominant,* (2), *then so is R.*

The proof is deferred to §4.3. Some preliminary results are established in §4.1 and the main lemma is given in §4.2.

4.1. Preliminaries. It is convenient to express the entries of the reduced matrix in a different form and to do this first consider the tridiagonal systems (4). Denote the minors of each matrix T^i by $T_{11}^i, T_{12}^i, \ldots$ so that the inverse

$$(15) \qquad (T^i)^{-1} = \frac{1}{\det(T^i)} \begin{bmatrix} T_{11}^i & -T_{21}^i & \cdots & (-1)^{m+1} T_{m1}^i \\ -T_{12}^i & T_{22}^i & \cdots & (-1)^m T_{m2}^i \\ \vdots & \vdots & & \vdots \\ (-1)^{m+1} T_{1m}^i & (-1)^m T_{2m}^i & \cdots & T_{mm}^i \end{bmatrix},$$

the transposed matrix of cofactors, divided by the determinant of T^i.

In particular the four corner minors, $T^i_{11}, T^i_{m1}, T^i_{1m}$, and T^i_{mm}, will be required. Whence,

$$T^i_{m1} = \det \begin{bmatrix} c^i_1 & & & & \\ b^i_2 & c^i_2 & & & \\ a^i_3 & b^i_3 & c^i_3 & & \\ & \ddots & \ddots & \ddots & \\ & & a^i_{m-1} & b^i_{m-1} & c^i_{m-1} \end{bmatrix} = \prod_{j=1}^{m-1} c^i_j, \tag{16}$$

and similarly,

$$T^i_{1m} = \prod_{j=2}^{m} a^i_j. \tag{17}$$

Also

$$T^i_{11} = \det \begin{bmatrix} b^i_2 & c^i_2 & & & \\ a^i_3 & & & & \\ & \ddots & \ddots & \ddots & \\ & & & & c^i_{m-1} \\ & & & a^i_m & b^i_m \end{bmatrix} = \det(T^i_{\text{lower}}), \quad \text{say}, \tag{18}$$

and

$$T^i_{mm} = \det \begin{bmatrix} b^i_1 & c^i_1 & & & \\ a^i_2 & & & & \\ & \ddots & \ddots & \ddots & \\ & & & & c^i_{m-2} \\ & & & a^i_{m-1} & b^i_{m-1} \end{bmatrix} = \det(T^i_{\text{upper}}), \quad \text{say}. \tag{19}$$

Now consider the entries of the reduced matrix R as given in (10). For $i = 2, \ldots, P$, the main diagonal

$$\begin{aligned} \tilde{b}^i &= b^i - a^i s^{i-1}_m - c^i r^i_1 \\ &= b^i - a^i c^{i-1}_m [(T^{i-1})^{-1} \mathbf{e}_m]_m - c^i a^i_1 [(T^i)^{-1} \mathbf{e}_1]_1 \\ &= b^i - a^i c^{i-1}_m [(T^{i-1})^{-1}]_{mm} - c^i a^i_1 [(T^i)^{-1}]_{11}. \end{aligned}$$

So from (15), (19), and (18),

$$\tilde{b}^i = b^i - a^i c^{i-1}_m \frac{\det(T^{i-1}_{\text{upper}})}{\det(T^{i-1})} - c^i a^i_1 \frac{\det(T^i_{\text{lower}})}{\det(T^i)}.$$

But consider $\det(M^i)$ defined for $i = 2, \ldots, P$ by

$$M^i \overset{\text{def}}{=} \begin{bmatrix} b^{i-1}_1 & c^{i-1}_1 & & & & & & & \\ a^{i-1}_2 & & & & & & & & \\ & \ddots & \ddots & \ddots & & & & & \\ & & a^{i-1}_m & b^{i-1}_m & c^{i-1}_m & & & & \\ & & & a^i & b^i & c^i & & & \\ & & & & a^i_1 & b^i_1 & c^i_1 & & \\ & & & & & \ddots & \ddots & \ddots & \\ & & & & & & & & c^i_{m-1} \\ & & & & & & & a^i_m & b^i_m \end{bmatrix}.$$

Then, on expansion by the row containing a^i, b^i, and c^i,

$$\det(M^i) = b^i \det(T^i) \det(T^{i-1}) - a^i c_m^{i-1} \det(T^{i-1}_{\text{upper}}) \det(T^i) - c^i a_1^i \det(T^i_{\text{lower}}) \det(T^{i-1}).$$

Thus, for $i = 2, \ldots, P$,

$$\tilde{b}^i = \frac{\det(M^i)}{\det(T^{i-1}) \det(T^i)}. \tag{20}$$

For $i = 2, \ldots, P-1$, the lower diagonal

$$\begin{aligned} \tilde{a}^i &= -a^i r_m^{i-1} \\ &= -a^i a_1^{i-1} [(T^{i-1})^{-1} \mathbf{e}_1]_m \\ &= -a^i a_1^{i-1} [(T^{i-1})^{-1}]_{m1} \\ &= -a^i a_1^{i-1} (-1)^{m+1} T^{i-1}_{1m} / \det(T^{i-1}). \end{aligned}$$

Then from (17),

$$\tilde{a}^i = (-1)^m \frac{a^i a_1^{i-1} \prod_{j=2}^{m} a_j^{i-1}}{\det(T^{i-1})}$$

and writing $a^i \equiv a_{m+1}^{i-1}$

$$\tilde{a}^i = (-1)^m \frac{\prod_{j=1}^{m+1} a_j^{i-1}}{\det(T^{i-1})}. \tag{21}$$

Similarly, for $i = 3, \ldots, P$, the upper diagonal

$$\tilde{c}^i = -c^i s_1^i = -c^i c_m^i [(T^i)^{-1} \mathbf{e}_m]_1$$

and so from (16) with $c_0^i \equiv c^i$,

$$\tilde{c}^i = (-1)^m \frac{\prod_{j=0}^{m} c_j^i}{\det(T^i)}. \tag{22}$$

Some simple technical results, the corollaries of the following lemma, will also be required.

LEMMA 4.2. *Consider the series* $\{a_i\}, \{b_i\}$, *and* $\{c_i\}$ *and let*

$$b_i = |a_i| + |c_i| + \delta_i \quad \forall i \geq 1,$$

where $\delta_i > 0$ *is a series of strictly positive numbers. Now if*

$$\begin{aligned} x_1 &> 0, \\ x_2 - |c_2| x_1 &> 0, \end{aligned}$$

and

$$x_i = b_i x_{i-1} - a_i c_{i-1} x_{i-2} \quad i \geq 3. \tag{23}$$

Then $x_i > 0$ *for all* $i \geq 1$ *and* $x_i - |c_i| x_{i-1} > 0$ *for all* $i \geq 2$.

Proof. Suppose, for induction,

$$x_{i-2} > 0 \qquad \text{and} \quad x_{i-1} - |c_{i-1}| x_{i-2} > 0 \qquad (\Rightarrow \quad x_{i-1} > 0).$$

Then,

$$\begin{aligned} x_i &= b_i x_{i-1} - a_i c_{i-1} x_{i-2} \\ &\geq b_i x_{i-1} - |a_i c_{i-1}| x_{i-2} \\ &= (|a_i| + |c_i| + \delta_i) x_{i-1} - |a_i c_{i-1}| x_{i-2} \\ &= |a_i| (x_{i-1} - |c_{i-1}| x_{i-2}) + (|c_i| + \delta_i) x_{i-1} > 0. \end{aligned}$$

Also rearranging the last line

$$x_i - |c_i| x_{i-1} \geq |a_i| (x_{i-1} - |c_{i-1}| x_{i-2}) + \delta_i x_{i-1} > 0.$$

But, by hypothesis,

$$x_1 > 0 \quad \text{and} \quad x_2 - |c_2| x_1 > 0.$$

Hence, by induction,

$$x_i > 0 \quad \forall i \geq 1 \quad \text{and} \quad x_i - |c_i| x_{i-1} > 0 \quad \forall i \geq 2. \qquad \square$$

COROLLARY 4.3. *The determinant of a strictly diagonally dominant tridiagonal matrix with strictly positive main diagonal is positive.*

Proof. Let T be the tridiagonal matrix (of order n) with $\{a_i\}_{i=2}^n$, $\{b_i\}_{i=1}^n$, and $\{c_i\}_{i=1}^{n-1}$ on the lower, main, and upper diagonals, respectively. Then, in the above lemma, define x_i as the determinants of the leading principal submatrices of T. So

$$\begin{aligned} x_1 &= b_1 > 0, \\ x_2 &= b_2 x_1 - a_2 c_1 \\ &\geq b_2 b_1 - |a_2 c_1| \\ &= (|a_2| + |c_2| + \delta_2) b_1 - |a_2 c_1| \\ &= |a_2| (|a_1| + |c_1| + \delta_1) + (|c_2| + \delta_2) b_1 - |a_2 c_1| \\ &= |a_2| (|a_1| + \delta_1) + (|c_2| + \delta_2) x_1 > 0, \end{aligned}$$

and rearranging the last line,

$$x_2 - |c_2| x_1 = |a_2| (|a_1| + \delta_1) + \delta_2 x_1 \ > \ 0.$$

Now expanding x_i by the bottom row,

$$x_i = b_i x_{i-1} - a_i c_{i-1} x_{i-2} \quad 3 \leq i \leq n.$$

Hence, by the lemma,

$$x_i > 0 \quad 1 \leq i \leq n$$

and, in particular,

$$x_n = \det(T) > 0. \qquad \square$$

COROLLARY 4.4. *If* $x_1 > 0$, $x_2 - |c_2| x_1 > 0$ *and this time* $x_i \geq b_i x_{i-1} - a_i c_{i-1} x_{i-2}$ *in the above lemma. Then again* $x_i > 0$ *for all* $i \geq 1$.

Proof. An obvious modification of the proof of the lemma. $\square$

4.2. The main lemma. From the results in (20), (21), and (22) it can be seen that the elements of each row of the reduced matrix can be represented solely in terms of the corresponding boundary equation and the two tridiagonal systems it lies between. Since the proof seeks to derive a row condition on this matrix it is simpler to treat one such system. It is convenient therefore to abandon the notation conventions used up to now and use superscripts to denote rows above the "central" or "boundary" row, and subscripts to denote rows below. Consider then the system

$$
(24)\qquad \begin{bmatrix}
a^n & b^n & c^n & & & & & & & & \\
 & a^{n-1} & b^{n-1} & c^{n-1} & & & & & & & \\
 & & \ddots & \ddots & \ddots & & & & & & \\
 & & & a^1 & b^1 & c^1 & & & & & \\
 & & & & a & b & c & & & & \\
 & & & & & a_1 & b_1 & c_1 & & & \\
 & & & & & & \ddots & \ddots & \ddots & & \\
 & & & & & & & a_{m-1} & b_{m-1} & c_{m-1} & \\
 & & & & & & & & a_m & b_m & c_m
\end{bmatrix}.
$$

Now define the series of determinants to be used in the following proof by

$$
M_0^0 \overset{\text{def}}{=} b,
$$

$$
M_m^n \overset{\text{def}}{=} \det \begin{bmatrix}
b^n & c^n & & & & & & & \\
a^{n-1} & & & & & & & & \\
 & \ddots & \ddots & \ddots & & & & & \\
 & & a^1 & b^1 & c^1 & & & & \\
 & & & a & b & c & & & \\
 & & & & a_1 & b_1 & c_1 & & \\
 & & & & & \ddots & \ddots & \ddots & \\
 & & & & & & & c_{m-1} & \\
 & & & & & & a_m & b_m &
\end{bmatrix} \qquad n, m \geq 0,
$$

$$
U^0 \overset{\text{def}}{=} 1,
$$

$$
U^n \overset{\text{def}}{=} \det \begin{bmatrix}
b^n & c^n & & & & \\
a^{n-1} & & & & & \\
 & \ddots & \ddots & \ddots & & \\
 & & & & c^2 & \\
 & & & a^1 & b^1 &
\end{bmatrix} \qquad n \geq 1,
$$

$$
L_0 \overset{\text{def}}{=} 1,
$$

$$
L_m \overset{\text{def}}{=} \det \begin{bmatrix}
b_1 & c_1 & & & & \\
a_2 & & & & & \\
 & \ddots & \ddots & \ddots & & \\
 & & & & c_{m-1} & \\
 & & & a_m & b_m &
\end{bmatrix} \qquad m \geq 1,
$$

$$A^0 \stackrel{\text{def}}{=} 1,$$
$$A^n \stackrel{\text{def}}{=} \prod_{j=1}^{n} a^j \quad n \geq 1,$$

$$C_0 \stackrel{\text{def}}{=} 1,$$
$$C_m \stackrel{\text{def}}{=} \prod_{j=1}^{m} c_j \quad m \geq 1.$$

The bulk of the proof now is presented as a lemma (because the notation has changed) but M_m^n, U^n, L_m, A^n, and C_m can all be readily identified with elements of the results in (20), (21), and (22). It is worth remarking that $|.|$ will denote modulus and not determinant.

LEMMA 4.5. *Consider the tridiagonal system* (24) *and suppose that it has a strictly positive main diagonal and is strictly diagonally dominant, i.e.,*

$$b^i = |a^i| + |c^i| + \delta^i \qquad i = 1, \ldots, n,$$
$$b = |a| + |c| + \delta,$$
$$b_j = |a_j| + |c_j| + \delta_j \qquad i = 1, \ldots, m,$$

where

$$\delta, \delta^i, \delta_j > 0 \quad \forall i, j \geq 1.$$

Then, if M_m^n, U^n, L_m, A^n, and C_m are as defined above,

$$M_m^n > |aA^nL_m| + |cC_mU^n| \quad \forall m, n \geq 0.$$

Proof. Corollary 4.3 $\Rightarrow M_m^n, U^n, L_m > 0$ for all $m, n \geq 0$.
Now define

$$\epsilon_m^n \stackrel{\text{def}}{=} M_m^n - |aA^nL_m| - |cC_mU^n| \quad \forall m, n \geq 0. \tag{25}$$

The aim is now to show that $\epsilon_m^n > 0$.

First expand M_m^n by the top row and substitute for M_m^{n-1} and M_m^{n-2} from (25) to get

$$\begin{aligned} M_m^n &= b^n M_m^{n-1} - c^n a^{n-1} M_m^{n-2} \\ &= b^n |aA^{n-1}L_m| - c^n a^{n-1} |aA^{n-2}L_m| \\ &\quad + b^n |cC_mU^{n-1}| - c^n a^{n-1} |cC_mU^{n-2}| \\ &\quad + b^n \epsilon_m^{n-1} - c^n a^{n-1} \epsilon_m^{n-2}. \end{aligned} \tag{26}$$

Now considering the constituent parts of the right-hand side in (26)

$$\begin{aligned} b^n |aA^{n-1}L_m| - c^n a^{n-1} |aA^{n-2}L_m| &\geq b^n |aA^{n-1}L_m| - |c^n a^{n-1}||aA^{n-2}L_m| \\ &= (b^n - |c^n|)|aA^{n-1}L_m| \\ &= (\delta^n + |a^n|)|aA^{n-1}L_m| \\ &= \delta^n |aA^{n-1}L_m| + |aA^nL_m|. \end{aligned} \tag{27}$$

Also, since $U^n > 0$ for all n,

$$b^n|cC_mU^{n-1}| - c^n a^{n-1}|cC_mU^{n-2}| = (b^nU^{n-1} - c^n a^{n-1}U^{n-2})|cC_m| \tag{28}$$
$$= |cC_mU^n|.$$

So substituting (27) and (28) into (26) gives

$$M_m^n \geq \delta^n|aA^{n-1}L_m| + |aA^nL_m| + |cC_mU^n| + b^n\epsilon_m^{n-1} - c^n a^{n-1}\epsilon_m^{n-2},$$

and hence from the definition of ϵ_m^n and dropping the δ^n term

$$\epsilon_m^n \geq b^n\epsilon_m^{n-1} - c^n a^{n-1}\epsilon_m^{n-2}. \tag{29}$$

Similarly, expansion by the bottom row of M_m^n gives

$$\epsilon_m^n \geq b_m\epsilon_{m-1}^n - a_m c_{m-1}\epsilon_{m-2}^n. \tag{30}$$

The proof now goes as follows.

(a) First it is shown $\epsilon_0^n > 0$ for all n by induction on n.

(b) Next it is shown $\epsilon_1^n - |c_1|\epsilon_0^n > 0$ for all n.

(c) Finally it is shown $\epsilon_m^n > 0$ for all n, m by induction on m.

(a) Consider

$$M_0^0 = b = |a| + |c| + \delta = |aA^0L_0| + |cC_0U^0| + \delta.$$

Thus

$$\epsilon_0^0 = \delta > 0. \tag{31}$$

Also

$$\begin{aligned} M_0^1 &= b^1b - c^1a \\ &\geq b^1b - |c^1a| \\ &= (|a^1| + |c^1| + \delta^1)|a| + b^1(|c| + \delta) - |c^1a| \\ &= |a^1a| + b^1|c| + \delta^1|a| + b^1\delta \\ &= |aA^1L_0| + |cC_0U^1| + \delta^1|a| + b^1\delta. \end{aligned}$$

So

$$\begin{aligned} \epsilon_0^1 &= \delta^1|a| + b^1\delta \\ &= \delta^1|a| + |a^1|\delta + |c^1|\delta + \delta^1\delta. \end{aligned}$$

Hence, since $\epsilon_0^0 = \delta$,

$$\epsilon_0^1 - |a^1|\epsilon_0^0 = \delta^1|a| + |c^1|\delta + \delta^1\delta > 0. \tag{32}$$

Thus, using (31), (32), and (29) together with Corollary 4.4, it can be seen that

$$\epsilon_0^n > 0 \quad \forall n \geq 0. \tag{33}$$

N.B. The a's and c's swap roles from the statement of the corollary as the determinant is expanded from the top and not the bottom.

(b) Now consider

$$\begin{aligned}\epsilon_1^n - |c_1|\epsilon_0^n &= M_1^n - |aA^nL_1| - |cC_1U^n| \\ &\quad -|c_1|(M_0^n - |aA^nL_0| - |cC_0U^n|) \\ &= M_1^n - |aA^nb_1| - |cc_1U^n| \\ &\quad -|c_1|M_0^n + |aA^nc_1| + |cc_1U^n| \\ &= M_1^n - |c_1|M_0^n - |aA^n|(b_1 - |c_1|) \\ &= M_1^n - |c_1|M_0^n - |aA^n|(|a_1| + \delta_1).\end{aligned}$$

Then expanding M_1^n by the bottom row,

$$\begin{aligned}\epsilon_1^n - |c_1|\epsilon_0^n &= b_1M_0^n - a_1cU^n - |c_1|M_0^n - |aA^n|(|a_1| + \delta_1) \\ &= (b_1 - |c_1|)M_0^n - a_1cU^n - |aA^n|(|a_1| + \delta_1) \\ &= (|a_1| + \delta_1)M_0^n - a_1cU^n - |aA^n|(|a_1| + \delta_1) \\ &= (|a_1| + \delta_1)(M_0^n - |aA^n|) - a_1cU^n \\ &= (|a_1| + \delta_1)(M_0^n - |aA^nL_0|) - a_1cU^n \\ &= (|a_1| + \delta_1)(|cC_0U^n| + \epsilon_0^n) - a_1cU^n \\ &= |a_1cU^n| + |a_1|\epsilon_0^n + \delta_1(|cC_0U^n| + \epsilon_0^n) - a_1cU^n \\ &\geq |a_1|\epsilon_0^n + \delta_1(|cC_0U^n| + \epsilon_0^n) \\ &> 0, \qquad (34)\end{aligned}$$

and this is true for all $n \geq 0$ by (33).

(c) Finally, using (33), (34), and (30) together with Corollary 4.4, it can be seen that

$$\epsilon_m^n > 0 \quad \forall n, m \geq 0. \qquad (35)$$

Hence,

$$M_m^n > |aA^nL_m| + |cC_mU^n| \quad \forall m, n \geq 0. \qquad (36)$$

□

4.3. Proof of the theorem.

Proof. This is a direct consequence of Lemma 4.5. First, without loss of generality, each row of the full matrix A can be multiplied by $\operatorname{sign}(b_i)$ to ensure that the main diagonal is positive. Now for each $i = 2, \ldots, P$ identify U^n with $\det(T^{i-1})$, L_m with $\det(T^i)$ and hence M_m^n with $\det(M^i)$. Then from (20), (21) and (22), and Lemma 4.5

$$\begin{aligned}|\tilde{b}^i| &= \left|\frac{\det(M^i)}{\det(T^{i-1})\det(T^i)}\right| \\ &= \frac{M_m^n}{U^nL_m} \\ &> \frac{|aA^n|}{U^n} + \frac{|cC_m|}{L_m} \\ &= \frac{\left|\prod_{j=1}^{m+1} a_j^{i-1}\right|}{|\det(T^{i-1})|} + \frac{\left|\prod_{j=0}^{m} c_j^i\right|}{|\det(T^i)|} \\ &= |\tilde{a}^i| + |\tilde{c}^i|.\end{aligned}$$

N.B. This proof allows m, the order of T^i, to be different for each processor.

5. Conclusion. It has been explicitly proven that classical diagonal dominance is retained by the reduced system required in the partition algorithm for tridiagonal systems. It follows that methods such as two-way Gaussian elimination, two-way matrix decomposition, or cyclic reduction are stable to growth in round-off errors when used to solve the reduced system. It would be desirable to extend this result to cover narrow banded systems, but it is not clear that the same method of proof will suffice.

Acknowledgments. This work was carried out at Heriot–Watt University, Edinburgh, and benefitted greatly from the help of Dr. D. B. Duncan, my supervisor there. I would also like to thank the referees for their help in making the paper more clear.

REFERENCES

[1] O. BUNEMAN, *A compact non-iterative Poisson solver*, Tech. Report 294, Stanford University Institute for Plasma Research, Stanford, CA, 1969.

[2] B. BUZBEE, G. GOLUB, AND C. NEILSON, *On direct methods for solving Poisson's equations*, SIAM J. Numer. Anal., 7 (1970), pp. 627–656.

[3] R. W. HOCKNEY AND C. R. JESSHOPE, *Parallel Computers* 2, Adam Hilger, Bristol, 1988.

[4] S. L. JOHNSSON, *Solving narrow banded systems on ensemble architectures*, ACM Trans. Math. Software, 11 (1985), pp. 271–288.

[5] ———, *Solving tridiagonal systems on ensemble architectures*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 354–392.

[6] D. H. LAWRIE AND A. H. SAMEH, *The computation and communication complexity of a parallel banded system solver*, ACM Trans. Math. Software, 10 (1984), pp. 185–195.

[7] A. H. SAMEH AND D. J. KUCK, *On stable parallel linear system solvers*, J. ACM, 25 (1978), pp. 81–91.

[8] H. S. STONE, *An efficient parallel algorithm for the solution of a tridiagonal linear system of equations*, J. ACM, 20 (1973), pp. 27–38.

[9] ———, *Parallel tridiagonal equation solvers*, ACM Trans. Math. Software, 1 (1975), pp. 289–307.

[10] P. N. SWARZTRAUBER, *A parallel algorithm for solving general tridiagonal equations*, Math. Comp., 33 (1979), pp. 185–199.

[11] C. H. WALSHAW, *Parallel Algorithms for Large Sparse Systems of Differential Equations*, Ph.D. thesis, Heriot–Watt University, Edinburgh, 1991.

[12] H. H. WANG, *A parallel method for tridiagonal equations*, ACM Trans. Math. Software, 7 (1981), pp. 170–183.

[13] J. H. WILKINSON, *Error analysis of direct methods of matrix inversion*, J. ACM, 8 (1961), pp. 281–330.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, pp. 1100–1106, October 1995

005

MATRICES WITH SIGN CONSISTENCY OF A GIVEN ORDER*

J.M. PEÑA†

Abstract. In this paper, the matrices whose minors of a given order have the same sign are characterized in several ways. In particular, given an $n \times d$ matrix A (with $n > d$), a criterion involving $(n-d)d+1$ minors to determine if all $d \times d$ minors of A have the same strict sign is obtained. A test of $O(m^3)$ elementary operations (with $m = \max\{n-d, d\}$) to check if a given matrix satisfies these properties is also provided. Finally, these results are applied to improve the characterizations of alternating polytopes.

Key words. sign consistency, total positivity, alternating polytopes

AMS subject classifications. 15A15, 15A48, 15A57, 65F40, 52B40

1. Introduction.

Following [9], a matrix A is said to be *sign consistent of order* k if all nonzero $k \times k$ minors of A have the same sign. If all the $k \times k$ minors of A are nonzero and have the same sign, then A is called *strictly sign consistent of order* k. When A is strictly sign consistent of order k, we show in Theorem 2.2 that the number of minors of A to check can be considerably reduced.

It is particularly interesting to know when an $n \times d$ matrix (with $n > d$) is strictly sign consistent of order d because this class of matrices has important applications. It can be applied to characterize alternating polytopes (see [12], [13]). Besides, strictly sign consistent matrices of maximal order are very closely related with generalized convexity preserving transformations (see [9, Chap. 6, §3], [5]). On the other hand, inspired by Ostrowsky, Karlin in [9, Chap. 6, §5] showed that strictly sign consistent matrices of maximal order characterize some variation diminishing properties of certain systems of polynomials. In this last application, the variation diminishing properties of these matrices play an important role. Let us recall that, given a vector $x = (x_1, \ldots, x_m)^T \in \mathbf{R}^m$, $S^+(x)$ denotes the *maximum* number of sign changes of the sequence $x_1, \ldots, x_m$ that can be obtained by counting zeros as either $+$ or $-$. The following result (cf. [9, Chap. 5, Thm. 1.1]) characterizes the $n \times d$ $(n > d)$ matrices which are strictly sign consistent of order d by a variation diminishing property.

THEOREM 1.1. *Let A be a real $n \times d$ matrix with $n > d$. Then A is strictly sign consistent of order d if and only if $S^+(Ax) \leq d-1$ for any nonzero $x \in \mathbf{R}^d$.*

An $n \times d$ $(n > d)$ matrix A which is strictly sign consistent of order d has all its $\binom{n}{d}$ minors of order d nonzero and with the same sign. In Theorem 2.2(i) we show that it is sufficient to consider $(n-d)d+1$ minors of A. Besides, in §3 we provide a test of $O(m^3)$ elementary operations (with $m = \max\{n-d, d\}$) to determine if an $n \times d$ $(n > d)$ matrix is strictly sign consistent of order d or if it is sign consistent of order d with rank d.

A $p \times q$ matrix A which is (strictly) sign consistent of order k for all $k \leq \min\{p, q\}$ is called (*strictly*) *sign-regular*. If this sign is (strictly) positive for all $k \leq \min\{p, q\}$ then A is called (*strictly*) *totally positive*. Some of the main tools used to obtain our results are the criteria and tests for totally positive and strictly totally positive

*Received by the editors March 9, 1994; accepted for publication (in revised form) by R. Brualdi September 23, 1994.

†Departamento de Matemática Aplicada, Universidad de Zaragoza, 50009 Zaragoza, Spain (jmpena@cc.unizar.es). This work was supported by research grant DGICYT PB93-0310, Spain.

matrices given in [6]. Totally positive matrices have an increasing importance in approximation theory, theoretical economics, probability theory, and other fields (see [1], [9]). New applications to computer aided geometric design can be found in [4].

Finally, in §4 we give an application of our results to characterize alternating polytopes using a very reduced number of simplices, improving previous well-known characterizations.

2. Characterizations with a reduced number of minors. Our notation follows, in essence, that of [1]. Given k, $n \in \mathbf{N}$, $1 \leq k \leq n$, $Q_{k,n}$ will denote the set of all increasing sequences of k natural numbers less than or equal to n. Given $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k) \in Q_{k,n}$, the complement of the k-tuple α is the unique $(n-k)$-tuple $\alpha' \in Q_{n-k,n}$ such that $\alpha \cup \alpha' = \{1, 2, \ldots, n\}$. Let A be an $m \times n$ real matrix. For $k \leq m$, $l \leq n$, and for any $\alpha \in Q_{k,m}$ and $\beta \in Q_{l,n}$, we denote by $A[\alpha|\beta]$ the $k \times l$ submatrix of A containing rows numbered by α and columns numbered by β. Finally, when $\alpha = \beta$, $A[\alpha|\alpha]$ will be denoted by $A[\alpha]$. On the other hand, if M is an $m \times q$ matrix and N is a $p \times q$ matrix, $\begin{pmatrix} M \\ N \end{pmatrix}$ will denote the $(m+p) \times q$ matrix whose first m rows are the rows of M and whose p last rows are the rows of N.

Although the next result is very close to the proof of Lemma 4 of [13], we include the proof for the sake of completeness.

PROPOSITION 2.1. *Let A be a real $n \times d$ matrix, $n > d$, such that* $\det A[1, \ldots, d] \neq 0$, *and let $K = (k_{ij})_{1 \leq i,j \leq d}$ be the matrix whose nonzero entries are: $k_{ij} = (-1)^{j-1}$ if $i + j = d + 1$. Then the matrix $B := A(A[1, \ldots, d])^{-1}K$ is of the form*

$$B = \begin{pmatrix} K \\ C \end{pmatrix}$$

and there is a bijection between the $d \times d$ submatrices $A[\lambda|1, \ldots, d]$, $\lambda \in Q_{d,n} \setminus \{(1, \ldots, d)\}$, and all the submatrices of C. In this bijection, the corresponding determinants coincide up to the sign of $\det A[1, \ldots, d]$.

Proof. Let ε be the sign of $\det A[1, \ldots, d]$. Obviously, there is a natural bijection between the submatrices $B[\lambda|1, \ldots, d]$ and the submatrices $A[\lambda|1, \ldots, d]$, where $\lambda \in Q_{d,n} \setminus \{(1, \ldots, d)\}$, such that $\det B[\lambda|1, \ldots, d] = \varepsilon \det A[\lambda|1, \ldots, d]$ because $\det K = 1$. A submatrix of C is of the form $C[\alpha|\beta]$ with $\alpha \in Q_{r,n-d}$, $\beta \in Q_{r,d}$ and $1 \leq r \leq \min\{d, n-d\}$. Then we may define a submatrix $B[\lambda|1, \ldots, d]$, $\lambda \in Q_{d,n} \setminus \{(1, \ldots, d)\}$, such that

$$\begin{aligned} &\lambda_{d-r+1-j} = d + 1 - \beta'_j \quad \text{for } j = 1, \ldots, d-r, \\ &\lambda_{d-r+i} = d + \alpha_i \qquad\qquad \text{for } i = 1, \ldots, r. \end{aligned} \tag{2.1}$$

Conversely, given any submatrix $B[\lambda|1, \ldots, d]$ with $\lambda \in Q_{d,n} \setminus \{(1, \ldots, d)\}$, choose $r \in \{1, \ldots, d\}$ such that $\lambda_{d-r} \leq d < \lambda_{d-r+1}$ and define $\alpha \in Q_{r,n-d}$, $\beta \in Q_{r,d}$ so that they satisfy (2.1). Thus there is a bijection between the submatrices $B[\lambda|1, \ldots, d]$ with $\lambda \in Q_{d,n} \setminus \{(1, \ldots, d)\}$ and all the submatrices $C[\alpha|\beta]$ of C.

Now, using the Laplace expansion of $\det B[\lambda|1, \ldots, d]$ with respect to the first $d - r$ rows, we may obtain that $\det B[\lambda|1, \ldots, d] = \det C[\alpha|\beta]$ and so

$$\det A[\lambda|1, \ldots, d] = \varepsilon \det C[\alpha|\beta]. \qquad \square \tag{2.2}$$

Now we shall characterize the strictly sign consistent matrices of a given order, obtaining a particularly simplified characterization in the case of maximal order.

THEOREM 2.2. *Let A be an $n \times d$ matrix with $n > d$. Then we have what follows.*

(i) *A is strictly sign consistent of order d if and only if the following $(n-d)d+1$ submatrices of A have determinant with the same strict sign: $A[k, k+1, \ldots, k+d-1|1, \ldots, d]$ for any $k \in \{1, \ldots, n-d+1\}$, $A[1, 2, \ldots, d-r, j, j+1, \ldots, j+r-1|1, \ldots, d]$ for any r such that $1 \leq r < d$ and for any $j \geq d-r+2$ such that $j+r-1 \leq n$.*

(ii) *A is strictly sign consistent of order k for a given $k \in \{1, \ldots, d-1\}$ if and only if the following submatrices of A have determinant with the same strict sign: $A[\lambda|\mu]$ with λ, μ of the form $(t, t+1, \ldots, t+k-1)$ for any $t \in \{1, \ldots, n-k+1\}$ or of the form $(1, 2, \ldots, k-r, j, j+1, \ldots, j+r-1)$ for any r such that $1 \leq r < k$ and for any $j \geq k-r+2$ such that $j+r-1 \leq n$.*

Proof. (i) We must prove that, if the given minors have the same strict sign, then all the $\binom{n}{d}$ minors $\det A[i_1, i_2, \ldots, i_d|1, \ldots, d]$ $(1 \leq i_1 < i_2 < \cdots < i_d \leq n)$ have the same strict sign. Let K and

$$B = A(A[1, \ldots, d])^{-1}K = \begin{pmatrix} K \\ C \end{pmatrix}$$

be the matrices given in Proposition 2.1. By the same proposition it is sufficient to prove the equivalence of our hypotheses with the fact that the matrix C is strictly totally positive, that is, $\det C[\alpha|\beta] > 0$ for any $\alpha \in Q_{r,n-d}$, $\beta \in Q_{r,d}$, $1 \leq r \leq \min\{d, n-d\}$. In Theorem 4.1 of [6] it is shown that, in order to prove the strict total positivity of a matrix, it is necessary and sufficient to prove that it has strictly positive row-initial minors (the minors formed by consecutive initial rows and consecutive columns) and strictly positive column-initial minors (the minors formed by consecutive initial columns and consecutive rows). Thus we have only to check that $\det C[\alpha|\beta] > 0$ for $\alpha = (1, \ldots, r)$ and β formed by r consecutive columns among columns of C, and for $\beta = (1, \ldots, r)$ and α formed by r consecutive rows among rows $1, \ldots, n-d$ of C. Since the bijection of submatrices given in Proposition 2.1 satisfies (2.1) and (2.2), let us identify the corresponding $d \times d$ minors of A.

If $\alpha = (1, \ldots, r)$ then we have that $\lambda_{d-r+i} = d+i$ for $i = 1, \ldots, r$. If $\beta_1, \ldots, \beta_r$ are consecutive, since $\beta \cup \beta' = \{1, \ldots, d\}$, we deduce that $(\beta'_1, \ldots, \beta'_{d-r})$ is formed by $(r+1, r+2, \ldots, d)$ or $(1, 2, \ldots, d-r)$ or $(1, 2, \ldots, j+1, d+1-i, d+2-i, \ldots, d)$ (where $i \geq 0$, $j \geq 1$ and $i+j = d-r-1$). Thus $(\lambda_1, \lambda_2, \ldots, \lambda_{d-r})$ is $(1, 2, \ldots, d-r)$ or $(r+1, r+2, \ldots, d)$ or $(1, 2, \ldots, i, d-j, d-j+1, \ldots, d)$, respectively. By our hypotheses and (2.2), $\det C[\alpha|\beta] = \varepsilon \det A[\lambda|1, \ldots, d] > 0$ for all these d-tuples λ.

Now, if $\beta = (1, 2, \ldots, r)$, we have that $\beta' = (r+1, r+2, \ldots, d)$ and so we have that $(\lambda_1, \lambda_2, \ldots, \lambda_{d-r}) = (1, 2, \ldots, d-r)$. If $\alpha_1, \ldots, \alpha_r$ are consecutive, we obtain that $(\lambda_{d-r+1}, \ldots, \lambda_d)$ is formed by consecutive numbers among $d+1, \ldots, n$. Again by our hypotheses and (2.2), $\det C[\alpha|\beta] = \varepsilon \det A[\lambda|1, \ldots, d] > 0$ for all these d-tuples λ.

Finally let us count how many minors we have used. Analogously to the bijection between the set of $\binom{n}{d} - 1$ submatrices of C and the set of submatrices $A[i_1, i_2, \ldots, i_d|1, \ldots, d]$ (except $A[1, 2, \ldots, d]$), there exists a bijection between the set of submatrices that we have used in the statement (except $A[1, 2, \ldots, d]$) and the set of column-initial and row-initial submatrices of C. In order to count this number of submatrices, we must sum the following arithmetic progressions: $(n-d)+(n-d-1)+\cdots+(n-2d+1)$ and $d+(d-1)+\cdots+1$ if $n-d \geq d$, and $d+(d-1)+\cdots+2d-n+1$ and $(n-d)+(n-d-1)+\cdots+1$ if $d > n-d$. In the first case we obtain $(n-d)d+d$ and in the second case we obtain $(n-d)d+n-d$. Now we must substract the leading principal submatrices of C because they have been counted twice: their number is $\min\{d, n-d\}$. Thus (i) holds.

(ii) Let us consider any submatrix of A of the form $A[1,\dots,n|\mu]$ with $\mu=(t,t+1,\dots,t+k-1)$ for any $t\in\{1,\dots,n-k+1\}$ or $\mu=(1,2,\dots,k-r,j,j+1,\dots,j+r-1)$ for any r such that $1\le r<k$ and for any $j\ge k-r+2$ such that $j+r-1\le n$. Then this submatrix satisfies the hypotheses of (i). Thus this submatrix is strictly sign consistent of order k and so all minors of the form $\det A[\lambda|\mu]$ with $\lambda\in Q_{k,n}$ and μ as above have the same nonzero sign.

Now let us consider any minor $\det A[\lambda|\mu]$ ($\lambda\in Q_{k,n}$, $\mu\in Q_{k,d}$) of A. Let us consider the matrix $(A[\lambda|1,\dots,d])^T$ (which is the transpose of a submatrix of A). Let us observe that now this matrix satisfies the hypotheses of (i) and so $\det A[\lambda|\mu]$ has also the same nonzero sign, and (ii) follows. □

The next matrix shows that we cannot improve Theorem 2.2 (i) by checking only the signs of minors with consecutive rows:

$$A=\begin{pmatrix}2&-4\\1&-1\\-2&3\end{pmatrix}.$$

In [9, Chap. 2, Thm. 3.2] there is a sufficient condition (due to Fekete) to prove that a matrix is strictly sign consistent (or sign consistent) of maximal order d, which needs that all minors of order $d-1$ formed with the first $d-1$ columns and all minors of order d with consecutive rows have the same sign (always strict for the $(d-1)\times(d-1)$ minors). The next proposition gives an application of this result (in fact of [9, Chap. 2, Thm. 3.1], which is a consequence of the other one) in terms of only $d\times d$ minors.

PROPOSITION 2.3. *Let A be an $n\times d$ ($n\ge 2d$) matrix such that the following submatrices have determinant with the same strict sign: $A[1,\dots,d]$ and $A[1,2,\dots,d-r,j,j+1,\dots,j+r-1|1,\dots,d]$ for any r such that $1\le r<d$ and for any $j\ge d+1$ such that $j+r-1\le n$. If the nonzero minors among (respectively, if all the minors)* $\det A[k,k+1,\dots,k+d-1|1,\dots,d]$ *for any $k\in\{d,\dots,n-d+1\}$ have the same strict sign then $A[d+1,d+2,\dots,n|1,\dots,d]$ is sign consistent of order d (respectively, strict sign consistent of order d).*

Proof. Let K,B,C be the matrices of Proposition 2.1. Taking into account the bijection between $d\times d$ minors of A and minors of C (given in (2.1)) and (2.2), we deduce from our hypotheses the positivity of the minors of C formed by its first k columns and k consecutive rows (for $k=1,\dots,d-1$) and the nonnegativity (respectively, the positivity) of the $d\times d$ minors of C formed by consecutive rows. Then [9, Chap. 2, Thm. 3.1] implies that all the $d\times d$ minors of C are nonnegative (respectively, positive). Finally, applying again (2.2), the proposition holds. □

The previous results used the positivity of some minors of the constructed matrix C. In the next curious result, whose hypotheses will involve $(n-d)d+2$ minors of A, the matrix C will be strictly totally negative. A matrix is said to be *strictly totally negative* if all its minors are negative. These matrices were characterized in [8].

PROPOSITION 2.4. *Let A be an $n\times d$ ($n=2d$) matrix such that* $\det A[1,\dots,d]$ *has a strict sign ε. Then the $d\times d$ submatrices $A[\lambda|1,\dots,d]$, $\lambda\in Q_{d,n}\setminus\{(1,\dots,d)\}$, have determinant with strict sign $-\varepsilon$ if and only if the following submatrices of A have determinant with strict sign $-\varepsilon$: $A[2,3,\dots,d,n|1,\dots,d]$, $A[k,k+1,\dots,k+d-1|1,\dots,d]$ for any $k\in\{1,\dots,n-d+1\}$, $A[1,2,\dots,d-r,j,j+1,\dots,j+r-1]$ for any r such that $1\le r<d$ and for any $j\ge d-r+2$ such that $j+r-1\le n$.*

Proof. Let K,B,C be again the matrices of Proposition 2.1. By the same proposition (taking into account (2.1) and (2.2)), it is sufficient to prove the equivalence of our hypotheses with the fact that the matrix C is strictly totally negative. In Remark

3.6 of [8] it is shown that, in order to prove the strict total negativity of a matrix, it is necessary and sufficient to prove that it has negative row-initial minors, negative column-initial minors, and the element of its last row and column is also negative. But, by (2.1) and (2.2), these properties are equivalent with our hypotheses. □

Finally, let us remark that, taking into account the best criteria (using minors) to check that a matrix is totally positive (cf. [1, Thm. 2.1], [7, Thm. 3.1]), we cannot obtain a similar result to Theorem 2.2 to check the (not strict) sign consistency of a matrix. However, in the next section we provide a test to check that a matrix is sign consistent of the maximal order.

3. A test to check (strict) sign consistency of maximal order. Here we shall obtain a test requiring $O(m^3)$ elementary operations (with $m = \max\{n-d, d\}$) to determine if an $n \times d$ $(n > d)$ matrix is strict sign consistent of order d or if it is sign consistent of order d with rank d. Our main tool will be the use of the so-called Neville elimination. This elimination process was described in detail in [6] and we shall briefly recall it for the reader's convenience.

Neville elimination is a procedure to create zeros in a matrix by means of adding to a given row a suitable multiple of the previous one. For an $m \times n$ matrix $A = (a_{ij})_{1\le i\le m, 1\le j\le n}$ $(m \ge n)$, it consists of $n-1$ major steps resulting in a sequence of matrices as follows:

$$A := A_1 \to A_2 \to \cdots \to A_n,$$

where $A_t = (a_{ij}^{(t)})_{1\le i\le m, 1\le j\le n}$ has zeros below its main diagonal in the $t-1$ first columns. The matrix A_{t+1} is obtained from A_t $(t = 1, \ldots, n)$ according to the formula

$$a_{ij}^{(t+1)} := \begin{cases} a_{ij}^{(t)} & \text{if } i \le t, \\ a_{ij}^{(t)} - (a_{it}^{(t)}/a_{i-1,t}^{(t)})a_{i-1,j}^{(t)} & \text{if } i \ge t+1 \text{ and } j \ge t+1, \\ 0 & \text{otherwise.} \end{cases}$$

In this process the element

$$p_{ij} := a_{ij}^{(j)}, \quad 1 \le j \le n, \quad j \le i \le m, \tag{3.1}$$

is called the (i,j) *pivot* of the Neville elimination of A. The process would break down if any one of the pivots p_{ij} $(j \le i < m)$ is zero. In that case we can move the corresponding rows to the bottom and proceed with the new matrix, as described in [6].

If A has maximal column rank, the matrix $U := A_n$ is an $m \times n$ matrix with zeros below its main diagonal. The *complete Neville elimination* of a matrix A consists in performing the Neville elimination of A until getting the matrix U and, afterwards, proceeding with the Neville elimination of U^T (the transpose of U). When we say that the complete Neville elimination of A is possible without row or column exchanges, we mean that there have not been any row exchanges in the Neville elimination of A or U^T. Finally, the (i,j) pivot of the complete Neville elimination of A is the (i,j) pivot of the Neville elimination of A if $i \ge j$ and the (j,i) pivot of the Neville elimination of U^T if $i \le j$.

The following result, which will be very useful for our purposes, is a reformulation of the equivalence of the conditions (1) and (2) of [6, Thms. 4.1, 5.4].

PROPOSITION 3.1. *A matrix A is strictly totally positive if and only if the complete Neville elimination of A can be carried out without row or column exchanges and all the pivots are strictly positive. A matrix A is totally positive if and only if the*

complete Neville elimination of A can be carried out without row or column exchanges up to null rows or null columns and all the pivots are nonnegative.

Now we shall use Neville elimination to obtain the announced test. Let $\mathcal{S}$ denote the class of $n \times d$ $(n > d)$ matrices A which are strictly sign consistent of order d and let $\mathcal{N}$ denote the class of $n \times d$ $(n > d)$ matrices A which are sign consistent of order d and with $\det A[1, \ldots, d] \neq 0$. We shall obtain the matrix

$$B = \begin{pmatrix} K \\ C \end{pmatrix}$$

as in Proposition 2.1. Then, by the same proposition, one must check (using Proposition 3.1) if C is strictly totally positive (respectively, totally positive) to determine if $A \in \mathcal{S}$ (respectively, if $A \in \mathcal{N}$). Thus we have the following steps of the test.

STEPS OF THE TEST

I. Check that $\det A[1, \ldots, d] \neq 0$.
II. Obtain $B := A(A[1, \ldots, d])^{-1}K$.
III. Perform the complete Neville elimination of C.
IV. $A \in \mathcal{S}$ if and only if the complete Neville elimination of C can be carried out without row or column exchanges and all the pivots are strictly positive. $A \in \mathcal{N}$ if and only if the complete Neville elimination of C can be carried out without row or column exchanges up to null rows or null columns and all the pivots are nonnegative.

Finally, taking into account that the computational cost of Neville elimination coincides with the computational cost of Gaussian elimination (so that we may approximate the number of elementary operations in III by about $\frac{(n-d)^3}{2}$), the number of elementary operations used in this test can be approximated by about

$$d^3 + (n-d)d^2 + \frac{(n-d)^3}{2}.$$

4. An application to characterize alternating polytopes. Given an oriented matroid M it is often interesting (and difficult) to find a small subset of bases whose orientations completely determine M. In this section we shall apply the previous results to provide a solution to this problem for the alternating matroid, that is, the oriented matroid associated with the standard cyclic polytope. Let us start by introducing the main definitions.

Given the $(d-1)$-dimensional Euclidean space $\mathbf{E}^{d-1}$, a polytope $P \in \mathbf{E}^{d-1}$ is *cyclic* if it is isomorphic to the convex hull of a finite subset of the *moment curve* $\{(t, t^2, \ldots, t^{d-1}) \in \mathbf{E}^{d-1} | t \in \mathbf{R}\}$. Among all $(d-1)$-polytopes with n vertices $(n > d \geq 1)$, the cyclic polytopes have the maximum number of j-dimensional faces for all $j = 1, \ldots, d-2$ (McMullen's upper bound theorem [10]). The vertices of cyclic polytopes are usually labeled with respect to succession on the moment curve. A cyclic polytope is *alternating* if all subpolytopes of P are cyclic, and if the corresponding isomorphisms to cyclic polytopes are all induced by the above canonical labeling of P. This terminology comes from oriented matroid theory. Shemer proved that every even-dimensional cyclic polytope is alternating [11, Thm. 2.12], but this result does not hold for odd-dimensional cyclic polytopes.

As usual, we consider $(d-1)$-dimensional Euclidean space $\mathbf{E}^{d-1}$ embedded as an affine hyperplane in the vector space $\mathbf{R}^d$. We start by recalling the following well-known result (cf. [13, Lem. 2], [12, Rem. 2.3]). In fact, it is a geometric reformulation

of Remark 2.3 of [12], where it is shown that *cyclic chirotopes* as introduced in [3] are equivalent to *alternating matroids* as introduced in [2].

THEOREM 4.1. *A set* $\{x_1, x_2, \ldots, x_n\} \subset \mathbf{E}^{d-1}$ *is the (canonically labeled) set of vertices of an alternating* $(d-1)$*-polytope if and only all* $\binom{n}{d}$ *oriented simplices* $[x_{i_1}, x_{i_2}, \ldots, x_{i_d}]$ $(1 \leq i_1 < i_2 < \cdots < i_d \leq n)$*, have the same nonzero orientation.*

In the next theorem, which is a reformulation of Theorem 2.2(i), we shall considerably reduce the number of oriented simplices. They will be formed by, at most, two subsets of consecutive vertices, one of them formed by initial vertices.

THEOREM 4.2. *A set* $\{x_1, x_2, \ldots, x_n\} \subset \mathbf{E}^{d-1}$ *is the (canonically labeled) set of vertices of an alternating* $(d-1)$*-polytope if and only the following* $(n-d)d+1$ *oriented simplices have the same nonzero orientation:*

(i) $[x_k, x_{k+1}, \ldots, x_{k+d-1}]$ *for any* $k \in \{1, \ldots, n-d+1\}$,

(ii) $[x_1, x_2, \ldots, x_{d-r}, x_j, x_{j+1}, \ldots, x_{j+r-1}]$ *for any* r *such that* $1 \leq r < d$ *and for any* $j \geq d-r+2$ *such that* $j+r-1 \leq n$.

Finally, let us observe that the test given in §3 (using $O(m^3)$ elementary operations with $m = \max\{n-d, d\}$) to recognize that an $n \times d$ $(n > d)$ matrix is strictly sign consistent of order d allows us to check if a $(d-1)$-polytope of n vertices is alternating.

REFERENCES

[1] T. ANDO, *Totally positive matrices*, Linear Algebra Appl., 90 (1987), pp. 165–219.

[2] R.G. BLAND AND M. LAS VERGNAS, *Orientability of matroids*, J. Combin. Theory Ser. B, 24 (1978), pp. 94–123.

[3] J. BOKOWSKI AND B. STURMFELS, *On the coordinatization of oriented matroids*, Discrete Comput. Geom., 1 (1986), pp. 293–306.

[4] J.M. CARNICER AND J.M. PEÑA, *Shape preserving representations and optimality of the Bernstein basis*, Adv. Comput. Math., 1 (1993), pp. 173–196.

[5] J.M. CARNICER, M. GARCÍA, AND J.M. PEÑA, *Generalized convexity preserving transformations*, Computer Aided Geometric Design, to appear.

[6] M. GASCA AND J.M. PEÑA, *Total positivity and Neville elimination*, Linear Algebra Appl., 165 (1992), pp. 25–44.

[7] M. GASCA AND J.M. PEÑA, *Total positivity, QR factorization and Neville elimination*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1132–1140.

[8] M. GASCA AND J.M. PEÑA, *A test for strict sign-regularity*, Linear Algebra Appl., 197-198 (1994), pp. 133–142.

[9] S. KARLIN, *Total positivity*, Stanford University, Stanford, CA, 1968.

[10] P. MCMULLEN AND G.C. SHEPHARD, *Convex polytopes and the upper bound conjecture*, London Math. Soc. Lecture Notes 3, Cambridge University, Cambridge, 1971.

[11] I. SHEMER, *Neighborly polytopes*, Israel J. Math., 43 (1982), pp. 291–314.

[12] B. STURMFELS, *Neighborly polytopes and oriented matroids*, European J. combin., 9 (1988), pp. 537–546.

[13] ———, *Totally positive matrices and cyclic polytopes*, Linear Algebra Appl., 107 (1988), pp. 275–281.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, pp. 1107–1126, October 1995

006

ON A QR-LIKE ALGORITHM FOR SOME STRUCTURED EIGENVALUE PROBLEMS *

A. GEORGE†, Kh. D. IKRAMOV‡, E. V. MATUSHKINA‡, AND W.-P. TANG†

Abstract. In this paper, a QR-like algorithm, called the con-QR algorithm, for computing the Youla form of a general complex matrix is presented. The Youla form is an analog of the Schur form where unitary congruences instead of unitary similarities are employed. We introduce a set of invariants of a unitary congruence transformation which are called coneigenvalues, and discuss their condition. Finally, the practical value of the Youla form is discussed.

Key words. consimilarity, unitary congruence, coneigenvalue, Youla form, con-QR algorithm

AMS subject classification. 65F10

1. Introduction.

A number of popular numerical techniques are based, explicitly or implicitly, on using the matrix commutativity relation and the following important theorem [11, p. 166]:

THEOREM 1.1. *Assume that a matrix* $G \in \mathbf{C}^{n\times n}$ *is upper block triangular*

$$G = \begin{bmatrix} G_{11} & G_{12} & \cdots & G_{1u} \\ 0 & G_{22} & \cdots & G_{2u} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & G_{uu} \end{bmatrix} \tag{1}$$

and the spectra of different diagonal blocks are disjoint:

$$\sigma(G_{ii}) \bigcap \sigma(G_{jj}) = \emptyset, \qquad i \neq j.$$

Then, every matrix $B \in \mathbf{C}^{n\times n}$ *commuting with* G *has block triangular form*

$$B = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1u} \\ 0 & B_{22} & \cdots & B_{2u} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & B_{uu} \end{bmatrix} \tag{2}$$

conformal to (1).

It follows from Theorem 1.1 that if a dense matrix $H \in \mathrm{C}^{n\times n}$ is reduced to form (1) by some similarity transformation

$$H \quad \longrightarrow \quad G = Q^{-1}HQ, \tag{3}$$

then the same similarity applied to *any* matrix A commuting with H reduces A to form (2):

$$A \quad \longrightarrow \quad B = Q^{-1}AQ. \tag{4}$$

* Received by the editors September 9, 1993; accepted for publication (in revised form) by A. Bunse-Gerstner September 29, 1994. This work was supported by the Natural Sciences and Engineering Research Council of Canada, and by the Information Technology Research Centre, a Centre of Excellence funded by the Province of Ontario.

† Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1 (jageorge@sparse1.uwaterloo.ca).

‡ Moscow State University, Faculty of Numerical Mathematics and Cybernetics, Moscow, Russia 119899.

Now, we would like to substantiate our assertion in the very beginning by the following examples.

Example 1. It is well known (see, for instance, [18]) that any $n \times n$ centrosymmetric matrix A, i.e., a matrix with the property

$$a_{ij} = a_{n+1-i,n+1-j} \quad \forall i, j$$

could be transformed to block diagonal form

$$B = B_1 \oplus B_2$$

by similarity (4) where diagonal blocks B_1 and B_2 are of order $\lfloor n/2 \rfloor$ and $\lceil n/2 \rceil$, respectively. When $n = 2m$ is even, we have

$$Q = \frac{1}{\sqrt{2}} \begin{bmatrix} I_m & I_m \\ P_m & -P_m \end{bmatrix}, \tag{5}$$

and P_m is the special permutation matrix

$$P_m = \begin{bmatrix} & & & 1 \\ & 0 & 1 & \\ & \cdot & & \\ & \cdot & 0 & \\ 1 & & & \end{bmatrix}.$$

When n is odd, the matrix Q is a little more complicated.

From the point of view adopted in this paper, centrosymmetric matrices are a matrix class Δ_H generated by the matrix $H = P_n$ via the commutativity relation

$$\Delta_H = \{A \in \mathrm{C}^{n \times n} \mid AH = HA\}. \tag{6}$$

Such a class is sometimes called a *centralizer*, and H a *generator* of Δ_H. The matrix (5) reduces P_n to block diagonal form

$$G = I_m \oplus -I_m$$

by unitary similarity.

Example 2. The discrete Fourier transformation (DFT) is a standard method for solving linear equations with a circulant matrix. Circulant matrices can be written as polynomials of the cycle permutation matrix

$$\Pi_n = \begin{bmatrix} 0 & 1 & & & & \\ & 0 & 1 & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & & \\ & & & & 0 & 1 \\ 1 & & & & & 0 \end{bmatrix}$$

and the columns of an $n \times n$ DFT matrix are the eigenvectors of Π_n.

The main motivation of this paper is to justify the consideration of matrix classes which are defined, similar to (6), by

$$\Gamma_H = \{A \in \mathrm{C}^{n \times n} \mid \overline{A}H = HA\}, \tag{7}$$

where H is a given $n \times n$ matrix.

DEFINITION 1. *The matrix relation in* (7) *is called* concommutativity, *the matrix class* Γ_H *a* concentralizer, *and* H *a* generator *of* Γ_H.

Again, we wish to point out some examples where Definition 1 is useful.

Example 3. *Centrohermitian* matrices are defined in [13] as $n \times n$ matrices A satisfying the relationship

$$a_{ij} = \overline{a}_{n+1-i,n+1-j} \quad \forall i, j. \tag{8}$$

It was shown in [13] that centrohermitian matrices constitute a real algebra isomorphic to the algebra $\mathbf{R}^{n\times n}$ of real $n \times n$ matrices. The isomorphism is established by similarity in (4), where A is any centrohermitian matrix, B is a real matrix, and Q (for even n) is the unitary matrix

$$Q = \frac{1}{\sqrt{2}} \begin{bmatrix} I_m & iI_m \\ P_m & -iP_m \end{bmatrix}. \tag{9}$$

This fact has an important computational implication which was not mentioned in [13]. Solving a real algebraic problem requires, in a typical case, from three to four times less computational work than solving a complex one of the same size. In fact, only $O(n^2)$ additions/subtractions and division by 2 are needed for the similarity transformation (4). After that, we will have a real matrix to work on.

Once more, we prefer to replace the original definition (8) of centrohermitian matrices with their description as a concentralizer where P_n is a generator. The fact the centrohermitian matrices could be simultaneously converted to real ones is simply a particular case of the general rule in [12]. According to that rule, a concentralizer Γ_H can be transformed to $\mathbf{R}^{n\times n}$ by (4) if and only if the generator H satisfies the relation

$$H\overline{H} = \alpha I_n, \qquad \alpha > 0. \tag{10}$$

If (10) is fulfilled then H could be represented (nonuniquely) as a product

$$H = \overline{Q}Q^{-1}, \tag{11}$$

and any such matrix Q is appropriate for similarity (4). In particular, (11) holds for $H = P_n$ and the matrix in (9).

Example 4. A matrix A of even dimension $n = 2m$ with the special block form

$$A = \begin{bmatrix} A_{11} & A_{12} \\ \overline{A}_{12} & \overline{A}_{11} \end{bmatrix} \tag{12}$$

will be called *crosshermitian.* Such matrices are found in some quantum theory problems [5]. We consider these matrices as a concentralizer with a generator

$$H = \begin{bmatrix} 0 & I_m \\ I_m & 0 \end{bmatrix}.$$

With this matrix H, (10) is satisfied, and for representation (11) the unitary matrix

$$Q = \frac{1}{\sqrt{2}} \begin{bmatrix} I_m & -iI_m \\ I_m & iI_m \end{bmatrix} \tag{13}$$

could be taken. It follows that similarity (4) with matrix (13) converts all crosshermitian matrices to real ones.

The problem we address in this paper can be expressed as follows. Suppose we are to solve a sequence of linear algebra problems with $n \times n$ matrices $A_1, A_2, \ldots$. What kind of computational advantages could be extracted from a priori knowledge that all the matrices A_i, $i = 1, 2, \ldots$ belong to a concentralizer where a generator H is known? We have already answered this question for the case when (10) is satisfied. But, what if it does not?

Our solution to this problem is based on a theory which we will discuss in §3. This theory is parallel to the commutativity theory sketched at the beginning of this section. However, it uses consimilarity transformations rather than the usual similarities. Therefore, we include in §2 a review of some basic facts from the theory of consimilarity. We relate a set of invariants called *coneigenvalues* to a consimilarity transformation. The discussion of the numerical condition of coneigenvalues is also provided in §2.

From the computational point of view, the most appealing type of consimilarity transformations are unitary congruences. The condensed form of a complex matrix with respect to unitary congruences is the so-called Youla form. A QR-like technique, called the *con*-QR *algorithm*, for computing the Youla form and the corresponding unitary transformation is presented in §4. For a discussion of our computer implementation of the con-QR algorithm and numerical experiments we refer the reader to the report [6]. In the last section, applications of our results are shown and concluding remarks are given.

2. Consimilarity and coneigenvalues. We recall the definition of consimilarity [10, p. 244]:

DEFINITION 2. *Complex matrices G and H are said to be* consimilar *if*

$$G = \overline{Q}^{-1} H Q. \tag{14}$$

The relation (14) *itself is called a* consimilarity.

It is mentioned in [10] that matrices

$$H_R = H\overline{H}, \qquad H_L = \overline{H}H \tag{15}$$

play a very important role in the theory of consimilarity. The reason is that the usual similarity transformations

$$G_R = \overline{Q}^{-1} H_R \overline{Q}, \qquad G_L = Q^{-1} H_L Q \tag{16}$$

correspond to consimilarity (14). This means that the eigenvalues of either of the matrices H_R and H_L can be considered as the invariants of consimilarity transformations.

The matrices H_R and H_L have an identical Jordan structure, so it will be enough to refer to one of them, say H_L. The spectrum of H_L has two remarkable properties [10, pp. 252–253].

1. It is symmetric with respect to the real axis. Moreover, eigenvalues λ and $\overline{\lambda}$ are of the same multiplicity.
2. The negative real eigenvalues of H_L (if any) are necessarily of even algebraic multiplicity.

For reasons which will be evident later, we prefer to deal with the square roots of eigenvalues of H_L rather than with the eigenvalues themselves, and it is these roots that we consider as the invariants of consimilarities.

DEFINITION 3. *If*

$$\sigma(H_L) = \{\lambda_1, \ldots, \lambda_n\} \tag{17}$$

is the spectrum of H_L, then the square roots with nonnegative real part of the numbers $\lambda_1, \ldots, \lambda_n$ are called coneigenvalues of the matrix H:

$$\mu_i = \lambda_i^{1/2}, \qquad \mathrm{Re}\mu_i \geq 0. \tag{18}$$

If μ is not purely imaginary then its multiplicity is defined as the multiplicity of the corresponding eigenvalue $\lambda = \mu^2$. For a purely imaginary μ, we define the multiplicity of μ as one-half of the multiplicity of $\lambda = \mu^2$. The set

$$c\sigma(H) = \{\mu_1, \ldots, \mu_n\} \tag{19}$$

is called the conspectrum *of the matrix H.*

We should emphasize that our definitions of coneigenvalues and conspectrum are different from the definitions in [10, p. 245]. Coneigenvalues as defined in [10] may not exist; when they do exist, there are always infinitely many of them. Our definition guarantees that every complex $n \times n$ matrix has exactly n coneigenvalues if the multiplicities are taken into account.

Next, we mention some interesting analogies between the theory of coneigenvalues and that of singular values. It is well known that, for an $n \times n$ matrix A with the singular values $\sigma_1, \ldots, \sigma_n$, the eigenvalues of the $2n \times 2n$ Hermitian matrix

$$\begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \tag{20}$$

are $\sigma_1, \ldots, \sigma_n, -\sigma_1, \ldots, -\sigma_n$.

By analogy with (20), we define, for a given $n \times n$ matrix A, the matrix

$$B_A = \begin{bmatrix} 0 & A \\ \overline{A} & 0 \end{bmatrix}. \tag{21}$$

THEOREM 2.1. *If $\mu_1, \ldots, \mu_n$ are the coneigenvalues of an $n \times n$ matrix A then*

$$\sigma(B_A) = \{\mu_1, \ldots, \mu_n, -\mu_1, \ldots, -\mu_n\}. \tag{22}$$

Proof. The assertion of the theorem follows easily from the equality

$$B_A^2 = A_R \oplus A_L. \qquad \square$$

The con-QR algorithm that we describe in §4 amounts to a sequence of special unitary similarity transformations of the matrix B_A (precisely as the singular value decomposition (SVD) algorithm of Golub and Kahan is equivalent to a sequence of special unitary similarities of matrix (20)). So, it would be useful to relate the condition of eigenvalues of B_A with that of A_R and A_L. This will be done in Theorems 2.5–2.7 below. Here, we first mention some simple connections between the eigenvectors of A_R and A_L.

LEMMA 2.2. *If λ is an eigenvalue of A_R, with u and p the corresponding right and left eigenvectors, then*

(a) *$\overline{u}$ and $\overline{p}$ are the right and left eigenvector corresponding to the eigenvalue $\overline{\lambda}$ of A_L, respectively.*

(b) *If $\overline{A}u$ ($A^T p$) is nonzero, then it is the right (left) eigenvector for the eigenvalue λ of A_L.*

Proof. The first assertion follows from the relation $A_L = \overline{A}_R$. To prove the first part of (b), premultiply $\overline{A}$ on both sides of $A_R u = \lambda u$, yielding

$$(\overline{A}A)(\overline{A}u) = \lambda(\overline{A}u).$$

To prove the second part of (b), postmultiply A on both sides of $p^T A_R = \lambda p^T$. □

Let $\mathrm{cond}(\mu, H)$ denote the condition number of a simple eigenvalue μ of the matrix H. Recall that if u and p are the corresponding right and left eigenvectors, respectively, then

$$\mathrm{cond}(\mu, H) = \frac{||u||_2 ||p||_2}{|p^T u|}. \tag{23}$$

COROLLARY 2.3. $\mathrm{cond}(\lambda, A_R) = \mathrm{cond}(\overline{\lambda}, A_L)$.

Let $\lambda = \mu^2$ be a simple eigenvalue of the matrix A_L. Then, we have two possibilities for λ:

(a) λ is a nonnegative real number;

(b) λ is nonreal.

Since a generator H of a concentralizer Γ_H is usually a nonsingular matrix, we assume henceforth the matrix A in B_A to be nonsingular as well. This assumption excludes the case $\lambda = 0$.

Consider the eigenvalue μ of matrix (21), which corresponds to a simple eigenvalue μ^2 of A_L. Let

$$w = \begin{bmatrix} u \\ v \end{bmatrix}, \qquad r = \begin{bmatrix} p \\ q \end{bmatrix}, \qquad u, v, p, q \in \mathbf{C}^n, \tag{24}$$

be the right and left eigenvector associated with μ, respectively. From $B_A w = \mu w$, and $r^T B_A = \mu r^T$, we have

$$Av = \mu u, \tag{25}$$

$$\overline{A}u = \mu v, \tag{26}$$

$$q^T \overline{A} = \mu p^T, \tag{27}$$

$$p^T A = \mu q^T. \tag{28}$$

It is clear from (25)–(28) that none of the vectors u, v, p, q could be zero. Indeed, if, for example, $v = 0$, then we will have $u = 0$ (recall $\mu \neq 0$). Therefore, $w = 0$, which is impossible.

The second consequence of (25)–(28) is that the vectors

$$\hat{w} = \begin{bmatrix} u \\ -v \end{bmatrix} \quad \text{and} \quad \hat{r} = \begin{bmatrix} p \\ -q \end{bmatrix} \tag{29}$$

are the right and left eigenvectors corresponding to the eigenvalue $-\mu$ of the matrix B_A, respectively. Therefore, the following lemma holds.

LEMMA 2.4. *If μ is a simple eigenvalue of B_A, then*

$$\mathrm{cond}(\mu, B_A) = \mathrm{cond}(-\mu, B_A).$$

The next implication of the formulas (25)–(28) is that u and p (v and q) are the right and left eigenvector of the matrix A_R (A_L) associated with its eigenvalue μ^2, respectively. To show this, for example, for the vector u we premultiply (26) by A and use (25):

$$A\overline{A}u = \mu Av = \mu^2 u. \tag{30}$$

The last and very important consequence is the equality

$$p^T u = q^T v. \tag{31}$$

Using (25), (26), and (30), we have

$$p^T A\overline{A}u = \mu^2 p^T u = \mu^2 q^T v.$$

Now we are ready to prove the first main result of this section.

THEOREM 2.5. *If μ is a simple positive eigenvalue of B_A, then*

$$\operatorname{cond}(\mu, B_A) = \operatorname{cond}(\mu^2, A_R) = \operatorname{cond}(\mu^2, A_L). \tag{32}$$

Proof. First, we point out that μ^2 is a simple eigenvalue for both matrices A_L and A_R. Now, if we take eigenvector (24) of the matrix B_A, then both vectors v and $\overline{u}$ are the eigenvectors of A_L corresponding to the eigenvalue μ^2 (see Lemma 2.2, part (a) and the discussion preceding this theorem). Therefore, there exists a nonzero number $\alpha \in \mathbf{C}$ such that

$$v = \alpha\overline{u}. \tag{33}$$

Analogously,

$$q = \beta\overline{p}$$

for some nonzero $\beta \in \mathbf{C}$.

Next, we show that

$$|\alpha| = |\beta| = 1. \tag{34}$$

Indeed, combining (25), (26), and (33), we get

$$A\overline{u} = \frac{\mu}{\alpha}u$$

and

$$A\overline{u} = \mu\overline{\alpha}u.$$

For the last equality, we have used the assumption that μ is real. Thus, $\alpha^{-1} = \overline{\alpha}$, yielding part of (34). The second part of (34) can be shown in the same way. Now, we have

$$\operatorname{cond}(\mu, B_A) = \frac{||w||_2||r||_2}{|r^T w|} \tag{35}$$

$$\begin{aligned} &= \frac{(||u||_2^2 + ||v||_2^2)^{1/2}(||p||_2^2 + ||q||_2^2)^{1/2}}{|p^T u + q^T v|} \\ &= \frac{\sqrt{2}\ \ ||u||_2\sqrt{2}\ \ ||p||_2}{2|p^t u|} \\ &= \operatorname{cond}(\mu^2, A_R) = \operatorname{cond}(\mu^2, A_L). \end{aligned} \tag{36}$$

We have used (31), and (33)–(34) for the middle equality, and Corollary 2.3 for the last equality. □

It is impossible to obtain such a simple and nice result for a nonreal eigenvalue μ. One of the reasons is that the condition numbers $\mathrm{cond}(\mu^2, A_R)$ and $\mathrm{cond}(\mu^2, A_L)$ are generally different if $\mathrm{Im}\mu^2 \neq 0$. (Notice that it is the numbers $\mathrm{cond}(\mu^2, A_R)$ and $\mathrm{cond}(\overline{\mu}^2, A_L)$ that are equal according to Corollary 2.3.)

It might be conjectured that, for a nonreal simple eigenvalue μ of B_A, the inequality

$$\mathrm{cond}(\mu, B_A) \leq \max\{\mathrm{cond}(\mu^2, A_R), \mathrm{cond}(\mu^2, A_L)\} \tag{37}$$

should be valid. However, we are only able to show the following weaker result at this moment.

THEOREM 2.6. *Suppose that an $n \times n$ matrix A is nonsingular, and μ is a simple nonreal eigenvalue of B_A. Then*

$$\mathrm{cond}(\mu, B_A) \leq \frac{(1 + \mathrm{cond}_2(A))^{1/2}}{\sqrt{2}} \max\{\mathrm{cond}(\mu^2, A_R), \mathrm{cond}(\mu^2, A_L)\}. \tag{38}$$

The proof of the theorem above follows along the same lines as that of Theorem 2.5 and is omitted.

The extension of the two theorems above to the purely imaginary eigenvalue μ of B_A is not possible. The reason is that μ^2 is a double eigenvalue for either of matrices A_R and A_L, and definition (23) for the condition number is not applicable. However, there exist more general definitions that hold for any simple invariant subspace. The definition in [16, Chap. 5, §2.2] is one of them. For the present situation, it could be stated as follows.

Let λ be a semisimple eigenvalue of multiplicity 2 of the matrix A_R, and let u_1, u_2 and p_1, p_2 be two biorthogonal systems of eigenvectors associated with λ, the right and left ones, respectively. Let

$$U = [u_1, u_2], \qquad P = [p_1, p_2]. \tag{39}$$

The biorthogonality implies

$$P^T U = I_2.$$

Now we can define the condition number of λ by the formula

$$\mathrm{cond}(\lambda, A_R) = ||U||_F ||P||_F. \tag{40}$$

Since the biorthogonal systems above can be chosen in infinitely many ways, it makes more sense to replace (40) by

$$\mathrm{cond}(\lambda, A_R) = \inf_{\substack{U \in R(\lambda), P \in L(\lambda) \\ P^T U = I_2}} ||U||_F ||P||_F. \tag{41}$$

Here $R(\lambda)$ and $L(\lambda)$ are, respectively, the right and left eigenspace of A_R associated with λ. In [16], a choice of orthonormal u_1 and u_2 is suggested:

$$\mathrm{cond}(\lambda, A_R) = ||U||_F ||P||_F, \tag{42}$$

where

$$U \in R(\lambda), P \in L(\lambda),$$
$$U^*U = I_2, P^TU = I_2.$$

Notice that $\overline{u}_1$, $\overline{u}_2$ and $\overline{p}_1$, $\overline{p}_2$ are the right and left eigenvectors of A_L for the eigenvalue $\overline{\lambda}$. Therefore, whichever of the definitions (40)–(42) is employed, we have

$$\text{cond}(\lambda, A_R) = \text{cond}(\overline{\lambda}, A_L),$$

exactly as in Corollary 2.3. In particular, for a negative real λ,

$$\text{cond}(\lambda, A_R) = \text{cond}(\lambda, A_L).$$

THEOREM 2.7. *Let $\lambda = \mu^2$ be a negative real eigenvalue of multiplicity 2 for either of matrices A_R and A_L. If λ is semisimple, then*

$$\text{cond}(\mu, B_A) = \text{cond}(\lambda, A_R). \tag{43}$$

Proof. We have seen above that an eigenvector w of B_A is compiled of eigenvectors of the matrices A_R and A_L. Now, we show that from any (right) eigenvector u of A_R, belonging to $\lambda = \mu^2$, an eigenvector of B_A associated with μ can be constructed. We do not need the assumption of $\lambda < 0$ for this part of the proof; the condition $\lambda \neq 0$ is all that is required.

Define the vector v as

$$v = \frac{1}{\mu}\overline{A}u. \tag{44}$$

Since

$$Av = \mu^{-1}A\overline{A}u = \mu u, \tag{45}$$

we have

(a) the vector v is a nonzero vector;
(b) equations (25), (26) are satisfied which means that

$$w = \begin{bmatrix} u \\ v \end{bmatrix}$$

is an eigenvector of B_A for the eigenvalue μ.

In the same way, we can reconstruct a left eigenvector r of B_A for the eigenvalue μ from a left eigenvector p of A_R belonging to $\lambda = \mu^2$. The required vector q is given by

$$q = \frac{1}{\mu}A^Tp. \tag{46}$$

Now suppose that u_1, u_2 and p_1, p_2 are two biorthogonal systems of eigenvectors of A_R for eigenvalue $\lambda = \mu^2$. Let w_1, w_2 and r_1, r_2 be the corresponding eigenvectors of B_A constructed from (44) and (46). Note that (31) holds for any eigenvectors w and r. Therefore,

$$r_i^Tw_j = p_i^Tu_j + q_i^Tv_j = 2p_i^Tu_j = \begin{cases} 2, & i = j, \\ 0, & i \neq j. \end{cases}$$

So, the systems w_1, w_2 and r_1, r_2 are *almost* biorthogonal.

We consider the matrices

$$V = [v_1, v_2], \qquad Q = [q_1, q_2]$$

along with the matrices (39). Since $\overline{u_1}$, $\overline{u_2}$ constitute a basis in $R(\overline{\lambda})$, we have

$$V = \overline{U}S$$

for some 2×2 nonsingular matrix S. Using (44) and (45), we obtain

$$A\overline{U} = \mu U\overline{S} = \mu U S^{-1}. \tag{47}$$

In deriving the left equality, we used for the first time the fact that μ is real. Now, (47) gives

$$\overline{S}S = I_2. \tag{48}$$

In a similar way, we show that in equation $Q = \overline{P}T$ the matrix T is such that $\overline{T}T = I_2$. Moreover, since $Q^T V = P^T U = I_2$, the relation

$$T = S^{-1} \tag{49}$$

holds.

Assume that from initial systems u_1, u_2 and p_1, p_2 we moved to new, and also biorthogonal systems $\hat{u}_1$, $\hat{u}_2$ and $\hat{p}_1$, $\hat{p}_2$. Then for the new matrices

$$\hat{U} = [\hat{u}_1, \hat{u}_2], \qquad \hat{P} = [\hat{p}_1, \hat{p}_2],$$

we have

$$\hat{U} = UZ, \qquad \hat{P} = PZ^{-T}, \tag{50}$$

where Z is some 2×2 nonsingular matrix.

Let $\hat{w}_1$, $\hat{w}_2$ and $\hat{r}_1$, $\hat{r}_2$ be the right and left eigenvectors, respectively, of B_A for the eigenvalue λ, which are constructed from $\hat{u}_1$, $\hat{u}_2$ and $\hat{p}_1$, $\hat{p}_2$, as described above. Again, we obtain the relations

$$\hat{V} = \overline{\hat{U}}\hat{S}, \qquad \hat{Q} = \overline{\hat{P}}\hat{T} \tag{51}$$

for the corresponding matrices $\hat{U}$, $\hat{V}$, $\hat{P}$, $\hat{Q}$ with a nonsingular 2×2 matrix $\hat{S}$, and $\hat{T} = \hat{S}^{-T}$. Now, from

$$\hat{U} = UZ, \quad \hat{V} = VZ, \quad \text{and} \qquad V = \overline{U}S,$$

we obtain

$$\hat{V} = \overline{\hat{U}}\,\overline{Z}^{-1}SZ,$$

i.e.,

$$\hat{S} = \overline{Z}^{-1}SZ.$$

So, the change of a basis in $R(\lambda)$ leads to the matrix S transformed according to the consimilarity rule!

It is shown in [10, p. 247] that a 2×2 matrix S could be transformed by a consimilarity into the identity matrix if and only if S satisfies (48). We, therefore, come to the following conclusion: there exists in $R(\lambda)$ basis $\hat{u}_1$, $\hat{u}_2$, such that the corresponding eigenvectors $\hat{w}_1$, $\hat{w}_2$ have a form

$$\hat{w}_i = \begin{bmatrix} \hat{u}_i \\ \overline{\hat{u}_i} \end{bmatrix}, \qquad i = 1, 2. \tag{52}$$

Since $\hat{S} = I_2$, then $\hat{T} = I_2$ as well. Hence, for the left eigenvectors a representation analogous to (52) is valid:

$$\hat{r}_i = \begin{bmatrix} \hat{p}_i \\ \overline{\hat{p}_i} \end{bmatrix}, \qquad i = 1, 2. \tag{53}$$

It is the vectors $\hat{w}_1$, $\hat{w}_2$, $\hat{r}_1$, $\hat{r}_2$ that are used for computing the condition number in (40). Letting

$$\hat{W} = \begin{bmatrix} \hat{U} \\ \overline{\hat{U}} \end{bmatrix}, \qquad \hat{R} = \begin{bmatrix} \hat{P} \\ \overline{\hat{P}} \end{bmatrix},$$

and noting that

$$\hat{R}^T \hat{W} = 2I_2,$$

we obtain

$$\operatorname{cond}(\mu, B_A) = \frac{1}{2} \, ||\hat{W}||_F \, ||\hat{R}||_F \; = ||\hat{U}||_F \, ||\hat{P}||_F = \operatorname{cond}(\lambda, A_R). \qquad \square$$

It should be noted that the proof of Theorem 2.7 was based on the condition number definition (40). We believe that equality (43) holds with the stronger definitions (41) or (42). We will deal with this question in a later publication.

The key message contained in the previous three theorems is that the sensitivity of coneigenvalues is much the same whether we compute them from the matrices A_R and A_L or from the matrix B_A. At the same time, equivalent perturbations with these two approaches could be quite different. For the former, the size of elements in the equivalent perturbation matrix is proportional to the size of elements in A_R (or A_L), while for the latter, it is proportional to the elements in B_A (or, essentially, the matrix A itself).

Notice that if the second way is chosen, we need not apply a general eigenvalue technique to B_A. In fact, consimilarity

$$A \longrightarrow \tilde{A} = \overline{Q}^{-1} A Q$$

induces the similarity transformation

$$B_A \longrightarrow Y^{-1} B_A Y = B_{\tilde{A}},$$

where $Y = \overline{Q} \oplus Q$.

So, it is sufficient to work with A, reducing it by consimilarities to a form that permits an easy computation of the eigenvalues of B_A. One such form is the so-called Youla form, discussed in the following section.

3. Concentralizers and unitary congruences. It is well known that the matrix Sylvester equation

$$AX - XB = C \tag{54}$$

has a unique solution X, if $A \in \mathbf{C}^{m\times m}$, $B \in \mathbf{C}^{n\times n}$, and $C \in \mathbf{C}^{m\times n}$ are known matrices, and the spectra of A and B are disjoint. In [2], an analogous statement concerning semilinear Sylvester-type matrix equation

$$AX - \overline{X}B = C \tag{55}$$

is proved.

THEOREM 3.1. *Let $A \in \mathbf{C}^{m\times m}$ and $B \in \mathbf{C}^{n\times n}$ be given matrices. Then the following assertions are equivalent.*

1. *Equation* (55) *has a unique solution for every $C \in \mathbf{C}^{m\times n}$.*
2. *The homogeneous equation*

$$AX - \overline{X}B = 0 \tag{56}$$

has the unique solution $X = 0$.

3. *The set $c\sigma(A) \bigcap c\sigma(B)$ is empty.*

Although our definition of the conspectrum is somewhat different from the one in [2], this theorem is still valid with our definition. We are now in a position to state the "con" version of Theorem 1.1. Its proof and the following corollary are straightforward consequences of Theorem 3.1, so are omitted.

THEOREM 3.2. *Assume that a matrix $G \in \mathbf{C}^{n\times n}$ has upper block triangular form* (1) *and the conspectra of its different diagonal blocks are disjoint:*

$$c\sigma(G_{ii}) \bigcap c\sigma(G_{jj}) = \emptyset, \qquad i \neq j. \tag{57}$$

Then every matrix $B \in \mathbf{C}^{n\times n}$ concommuting with G,

$$\overline{B}G = GB \tag{58}$$

has block tridiagonal form (2) *conformal to* (1).

COROLLARY 3.3. *If the matrix G is block diagonal rather than block triangular in Theorem* 3.1

$$G = G_{11} \oplus G_{22} \oplus \cdots \oplus G_{uu}, \tag{59}$$

then every matrix $B \in \Gamma_G$ must also have block diagonal form $B = B_{11} \oplus B_{22} \oplus \cdots \oplus B_{uu}$ conformal with (59).

The proof of this corollary is obtained by applying Theorem 3.2 to G and G^T.

Consider a concentralizer Γ_H with the generator H (see (7)). Apply to all matrices $A \in \Gamma_H$ the simultaneous similarity transformation

$$A \longrightarrow B = Q^{-1}AQ. \tag{60}$$

The image of Γ_H under similarity (60) is described in the following theorem.

THEOREM 3.4. *Under similarity* (60), *a concentralizer Γ_H transforms onto the concentralizer Γ_G where*

$$G = \overline{Q}^{-1}HQ. \tag{61}$$

The proof is straightforward and is omitted.

COROLLARY 3.5. *If, as a result of consimilarity, G has a block triangular form* (1), *and the conspectra of its different diagonal blocks are pairwise disjoint* (see (57)), *then every matrix B in* (60) *has the conformal block triangular form* (2).

OBSERVATION 1. *The assertion analogous to Corollary* 3.5 *holds where G is block diagonal* (59) *rather than block triangular.*

In numerical linear algebra, the preferred type of similarity transformations are unitary similarities. In the case of consimilarities, we also wish to deal with unitary ones. When Q is unitary, consimilarity (61) becomes just a unitary congruence

$$G = Q^T H Q, \qquad Q^* Q = I_n. \tag{62}$$

Now, the question is whether there exists a condensed form of a complex matrix under unitary congruences and if so, the nature of that form. The answer is given in [20]; see also [9].

THEOREM 3.6. *Every matrix $A \in \mathbf{C}^{n\times n}$ is unitarily congruent to a block triangular matrix*

$$R = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1t} \\ 0 & R_{22} & \cdots & R_{2t} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & R_{tt} \end{bmatrix} \tag{63}$$

with 1×1 and 2×2 diagonal blocks R_{ii}. The 1×1 blocks correspond to real coneigenvalues of A, and the 2×2 blocks correspond to pairs of conjugate complex coneigenvalues.

The matrix R in (63) constitutes an analog to the Schur form (and to the real Schur form) in the theory of unitary (orthogonal) similarity. It is called the *Youla form* of the matrix A.

Remark 1. (a) It would be more exact to call R the upper Youla form, since a lower block triangular matrix with analogous properties can be constructed from A by a unitary congruence. In this paper, only the upper block triangular case is studied.

(b) Even if a particular type—upper or lower—of the Youla form is chosen, this form is still not uniquely determined. We have the same kind of nonuniqueness as with the classical Schur form. For example, we can assign an arbitrary ordering of the coneigenvalues, and then find the Youla form with this same ordering for the coneigenvalues on the block diagonal of R. In particular, there always exist Youla forms with the property below.

DEFINITION 4. *The Youla form of a matrix $A \in \mathbf{C}^{n\times n}$ is called a good Youla form if for every multiple coneigenvalue of A all the corresponding diagonal blocks in R are in consecutive positions.*

The con-QR algorithm presented in §4 is a numerical approach for constructing a Youla form of a general complex matrix and the corresponding unitary transformation. If a given matrix A has some multiple coneigenvalues, then the computed Youla form will not necessarily be a good Youla form. However, effective techniques can be developed for reordering diagonal blocks in the Youla form. They are similar to the well-known approaches for reordering the Schur form [15].

After this discussion, we are now ready to answer the question posed in the introduction, namely, what kind of computational advantages could be deduced from a priori knowledge that a set of matrices belongs to a concentralizer with a known generator H. The answer is given by the following strategy.

1. Compute the unitary matrix Q which transforms H into its good Youla form. We can use the con-QR algorithm, with a possible application of a procedure for reordering the diagonal blocks, for this purpose.
2. For $k = 1, 2, \ldots$
   (a) apply the unitary similarity

$$A_k \longrightarrow B_k = Q^* A_k Q \tag{64}$$

   to A_k. As a result, the initial problem with the matrix A_k is replaced by a new one with the block triangular matrix

$$B_k = \begin{bmatrix} B_{11}^{(k)} & B_{12}^{(k)} & \cdots & B_{1u}^{(k)} \\ 0 & B_{22}^{(k)} & \cdots & B_{2u}^{(k)} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & B_{uu}^{(k)} \end{bmatrix}. \tag{65}$$

   (b) Solve the sequence of problems with smaller matrices $B_{uu}^{(k)}, \ldots, B_{11}^{(k)}$. This last stage may be organized as a successive process (if the problem is to solve the linear equations $A_k = B_k$, and B_k is indeed block triangular rather than block diagonal) or as a concurrent one (if only eigenvalues of A_k need to be computed, and B_k is block diagonal).

This approach is similar to the well-known technique for the direct solution of a sequence of sparse linear systems with an identical sparsity pattern (see, for example, [7]). Step 1 corresponds to a symbolic factorization of a sparse matrix, step 2(a) to a numerical factorization, and step 2(b) to forward and backward solvers.

Whether this technique is cost effective depends on a number of considerations: the number of problems to be solved, the cost for computing the matrix Q, the cost of similarity (64), coneigenvalue multiplicities of the matrix H, and so on. It is certainly more promising when matrices (1) and (65) are block diagonal. We conclude this section by presenting a matrix class that does have a block diagonal Youla form.

DEFINITION 5. *A matrix* $H \in \mathbf{C}^{n \times n}$ *is called* conjugate normal *if*

$$HH^* = \overline{H^* H}. \tag{66}$$

It is shown in [17] that the Youla form of a matrix H is block diagonal if H is a conjugate normal matrix. In particular, any complex symmetric matrix H is conjugate normal since

$$H = H^T \quad \longrightarrow \quad \overline{H} = H^*.$$

All the coneigenvalues of a symmetric matrix H are nonnegative real numbers; in fact, they coincide with H's singular values. The Youla form of H is a nonnegative diagonal matrix Λ, and step 1 of the procedure above is equivalent to computing the decomposition

$$H = P^T \Lambda P, \qquad P^* P = I_n.$$

It is called the Takagi factorization of a symmetric matrix H.

The computation of the Takagi factorization is the theme of [4], and we gratefully acknowledge the influence of this paper on this work.

4. The con-QR algorithm. In this section, a unitary method that we call the con-QR algorithm, is described. This technique transforms a given matrix A into its Youla form. If A is a complex symmetric matrix, the con-QR algorithm becomes the symmetric singular value decomposition (SSVD) algorithm of Bunse-Gerstner and Gragg.

The reduction of a given matrix to the Hessenberg form is not, strictly speaking, required as part of the usual QR-algorithm (which can be used for a dense matrix equally well). But for practical reasons of efficiency, an application of the iterative QR procedure is almost always preceded by this reduction. Therefore, it would be useful to have an analogous reduction in the case of unitary congruence.

For the reduction we can use transformations with elementary unitary matrices as well as with complex Householder matrices. Then analogues for the Givens procedure and the Householder procedure, respectively, will be obtained. Let us, for example, consider the last one.

The description of the con-Householder procedure is obtained from that of the usual Householder algorithm simply by replacing unitary similarities by unitary congruences, the final Hessenberg matrix H being the product of the form

$$H = \mathcal{H}_{n-2} \dots \mathcal{H}_2 \mathcal{H}_1 A \mathcal{H}_1^T \mathcal{H}_2^T \dots \mathcal{H}_{n-2}^T . \tag{67}$$

Here the $\mathcal{H}_i$ are the usual Householder matrices.

Corresponding to Lemma 3.1 in [4], we have the following result that describes the freedom in the Hessenberg form of a given matrix.

THEOREM 4.1. *Let Q_1 and Q_2 be unitary matrices such that both*

$$H_1 = Q_1^T A Q_1 \quad \text{and} \quad H_2 = Q_2^T A Q_2$$

are upper Hessenberg matrices. Assume that at least one of the matrices H_1, H_2 is unreduced (i.e., its elements on the subdiagonal (2, 1), (3, 2), ..., $(n, n-1)$ are all nonzero). If the first column of Q_2 is a multiple of the first column of Q_1 then there exists a unitary diagonal matrix D such that

$$Q_2 = Q_1 D$$

and

$$H_2 = D H_1 D . \tag{68}$$

(Note that (68) implies that in reality both matrices H_1 and H_2 are unreduced.)

The proof of this theorem can be obtained by a slight modification of the proof for the case of unitary similarity (see, for example, [19, p. 352]).

From this point on we may deal only with the Hessenberg matrix H instead of the initial dense matrix A. Our goal is to construct an iterative procedure that ultimately reduces H to its Youla form.

Recall from §2 that every consimilarity executed for a matrix A is accompanied by corresponding similarities for matrices $\overline{A}A$ and $A\overline{A}$. Now assume that we have constructed a sequence of matrices H_0, $H_1, \dots, H_k, \dots$, where $H_0 = H$ and the following conditions are fulfilled:

1. The matrices H_0, H_1 ..., $H_k, \dots$ are unitarily congruent;
2. The matrices H_0, H_1 ..., $H_k, \dots$ are upper Hessenberg;
3. Every transformation

$$H_k \longrightarrow H_{k+1} \tag{69}$$

is accompanied by similarity

$$F_k^{(L)} \longrightarrow F_{k+1}^{(L)}, \tag{70}$$

where we let $F_k^{(L)} = \overline{H_k} H_k$ and (70) is equivalent to one step of some version of the usual QR algorithm.

The same is true of similarity

$$F_k^{(R)} \longrightarrow F_{k+1}^{(R)},$$

where $F_k^{(R)} = H_k \overline{H}_k$.

So, the construction of a matrix sequence H_k is nothing more than the implicitly implemented QR algorithm for the matrices $F_0^{(L)} = \overline{H} H$ and $F_0^{(R)} = H \overline{H}$.

Consider now the limiting matrix $F_\infty^{(L)}$ of the sequence $\{F_k^{(L)}\}$ (understanding the limits as is customary for the QR algorithm where one sometimes could not speak of limits in the rigorous sense). In the typical case, the matrix $F_\infty^{(L)}$ is upper triangular. If H_∞ is a corresponding matrix for the sequence $\{H_k\}$, then

$$F_\infty^{(L)} = \overline{H}_\infty H_\infty.$$

In particular,

$$0 = \{F_\infty^{(L)}\}_{i+2,i} = \{\overline{H}_\infty\}_{i+2,i+1} \{H_\infty\}_{i+1,i}, \qquad i = 1, 2, \ldots, n-2.$$

We conclude that the upper Hessenberg matrix H_∞ has the following property: at least one of every two consecutive elements of its first subdiagonal is zero. This property is characteristic for the Youla form.

Now we describe how the transformation (69) is implemented in the con-QR algorithm. In doing so we will differentiate between two kinds of the usual QR steps: a single step with a real shift τ and a double step with nonreal conjugate shifts τ and $\overline{\tau}$. The idea of the analysis followed is borrowed from [4].

We may omit indices and consider, for definiteness, the transition from the matrix H to H_1. If τ is real then a single step of the QR algorithm with τ as a shift is described for the matrix $F^{(L)}$ by the formulas

$$F^{(L)} - \tau I = QR, \qquad F_1^{(L)} = RQ + \tau I. \tag{71}$$

In the unusual event that τ is an exact eigenvalue, it can be shown that $F_1^{(L)}$ has in this case a zero in one (or both) of the positions $(n-1, n-3)$ and $(n, n-2)$. Thus, H_1 is reduced and $h_{n,n-1}^{(1)} = 0$, so deflation will occur naturally after one step of the algorithm. Otherwise, we deduce from (71) that

$$F^{(R)} - \tau I = \overline{QR}$$

and

$$Q^T = \overline{R}(F^{(R)} - \tau I)^{-1}. \tag{72}$$

Now, using (71) and (72), we get

$$H_1 = Q^T H Q = \overline{R}(F^{(R)} - \tau I)^{-1} H (F^{(L)} - \tau I) R^{-1} = \overline{R} H R^{-1}.$$

We see that the matrix $H_1 = Q^T H Q$ is again upper Hessenberg. Moreover, it is unreduced.

We propose to construct the matrix Q by an implicit procedure akin to that employed in the usual QR algorithm. Suppose we have a unitary matrix P for which the following conditions are fulfilled: (i) the first column of P is a multiple of the first column of Q; (ii) the matrix $P^T H P$ is unreduced upper Hessenberg. Then, in view of Theorem 4.1, P essentially coincides with Q and $P^T H P$ is practically the same matrix as H_1.

The matrix P can be constructed by the following procedure.

1. Find the nonzero elements of the first column f_1 of $F^L - \tau I$. Note there are only three such elements: f_{11}, f_{21}, f_{31}.
2. Define $\mathcal{H}_0$ as the Householder matrix that eliminates from f_1 its second and third elements.
3. Implement the congruence

$$H \longrightarrow \hat{H} = \mathcal{H}_0^T H \mathcal{H}_0.$$

   The matrix $\hat{H}$ is not Hessenberg anymore because of the nonzero elements in positions (3,1), (4,1), and (4,2).
4. Reduce $\hat{H}$ to the upper Hessenberg form by the con-Householder procedure (67):

$$\hat{H}_1 = \mathcal{H}_{n-2} \dots \mathcal{H}_1 \hat{H} \mathcal{H}_1^T \dots \mathcal{H}_{n-2}^T = \mathcal{H}_{n-2} \dots \mathcal{H}_1 \mathcal{H}_0^T H \mathcal{H}_0 \mathcal{H}_1^T \dots \mathcal{H}_{n-2}^T.$$

The first column of the matrix

$$P = \mathcal{H}_0 \mathcal{H}_1^T \dots \mathcal{H}_{n-2}^T$$

coincides with the first column of $\mathcal{H}_0$ and, therefore, also with the first column of Q.

Note that in each of the matrices

$$\mathcal{H}_i = I - 2 w_i w_i^*,$$

the vector w_i has no more than three nonzero elements.

Consider now the case when τ is a complex number, not equal to any of eigenvalues of F^L. A double step of the QR algorithm with τ and $\overline{\tau}$ as shifts is described for the matrix $F^{(L)}$ by formulas

$$F^{(L)} - \tau I = Q_1 R_1, \qquad \tilde{F}^{(L)} = R_1 Q_1 + \tau I, \tag{73}$$

and

$$\tilde{F}^{(L)} - \overline{\tau} I = Q_2 R_2, \qquad F_1^{(L)} = R_2 Q_2 + \overline{\tau} I. \tag{74}$$

The following equality can be easily derived from (73) and (74):

$$F^{(L)2} - 2\alpha F^{(L)} + \gamma I = QR. \tag{75}$$

Here

$$\alpha = \mathrm{Re}\tau, \quad \beta = \mathrm{Im}\tau, \quad \gamma = |\tau|^2 = \alpha^2 + \beta^2, \quad Q = Q_1 Q_2, \quad R = R_2 R_1.$$

So, the matrices Q and R give us a unitary-triangular decomposition of the matrix in the left-hand side of (75). Replacing all the matrices in (75) by their conjugates we obtain

$$F^{(R)2} - 2\alpha F^{(R)} + \gamma I = \overline{QR}. \tag{76}$$

Note that

$$H(F^{(L)2} - 2F^{(L)} + \gamma I) = (F^{(R)2} - 2F^{(R)} + \gamma I)H.$$

Using (75) and (76) we have

$$\begin{aligned} H_1 &= Q^T H Q \\ &= \overline{R}(F^{(R)2} - 2\alpha F^{(R)} + \gamma I)^{-1} H (F^{(L)2} - 2\alpha F^{(L)} + \gamma I) R^{-1} \\ &= \overline{R} H R^{-1}. \end{aligned}$$

Therefore, H_1 is again an unreduced upper Hessenberg matrix.

As for constructing the matrix Q we can repeat almost word for word all that was said above in the case of real τ. The only substantial difference relates to the number of nonzero elements in the vectors w_i. There are now five nonzero elements in the first column of $F^{(L)2}$ and, therefore, five nonzero elements in the first column of Q. Every vector w_i has also no more than five nonzero elements.

5. Concluding remarks. In the final section of this paper, we point out two situations where the Youla form and the con-QR algorithm for computing it are advantageous.

Numerical solution of semilinear matrix equations. Let $A \in \mathbf{C}^{m\times m}$, $B \in \mathbf{C}^{n\times n}$ and $C \in \mathbf{C}^{m\times n}$. Suppose that

$$c\sigma(A) \bigcap c\sigma(B) = \emptyset.$$

Then, according to Theorem 3.1, the matrix Sylvester-type equation

$$AX - \overline{X}B = C \tag{77}$$

has a unique solution $X \in \mathbf{C}^{m\times n}$.

Now assume that $U \in \mathbf{C}^{m\times m}$ and $V \in \mathbf{C}^{n\times n}$ are unitary matrices. Then from (77) we have

$$(U^T A U)(U^* X V) - (U^T \overline{X} \overline{V})(V^T B V) = U^T C V. \tag{78}$$

Letting

$$R = U^T A U, \quad S = V^T B V, \quad D = U^T C V, \tag{79}$$

and

$$Y = U^* X V, \tag{80}$$

we may rewrite (78) as a new Sylvester-type equation

$$RY - \overline{Y}S = D. \tag{81}$$

Suppose the unitary matrices U, V above transform A and B to their Youla forms R, S, respectively. Then the computation of the matrix Y, the unique solution of (81), is reduced to solving a number of equations

$$R_\alpha Y_{\alpha\,\beta} - \overline{Y}_{\alpha\,\beta} S_\beta = D_{\alpha\,\beta}, \tag{82}$$

where 1×1 or 2×2 matrices R_α, S_β are diagonal blocks in R and S, respectively. Every small matrix equation of the (82) type can be considered as a system of real linear equations consisting of two, four, or eight equations. After Y is computed, the solution X of the initial equation (77) is obtained by (80).

Matrices U, V in the approach outlined can be computed by the application of con-QR algorithm to A and B, respectively. This technique can be viewed as an exact analogue of the Bartels–Stewart algorithm for the classical Sylvester equation $AX - XB = C$ [1]. We may transform (via unitary congruences) one of the matrices A, B to its Youla form and the other only to the Householder form. Then we obtain an analogue of the second algorithm for Sylvester equations, namely, the Golub–Nash–Van Loan algorithm [8].

Computation of functions of a matrix $B = \overline{A}A$. It was indicated in [14] that, for a triangular matrix T with different diagonal elements, any functions $F = f(T)$ could be restored from the commutativity relation $FT = TF$, if the diagonal elements $f_{ii} = f(t_{ii})$, $i = 1, \ldots, n$, are known. Fewer than $n^3/3$ multiplications are required to form F from its diagonal. The confluent case when some of the elements t_{ii} are equal is more complicated. However, the commutativity could still be used with advantage. The same is true for a block triangular matrix T. For a dense matrix B, the computation of $f(B)$ could be performed by first reducing B to its Schur form.

A similar idea is used in [3] where the computation of square roots of a matrix B (which are not necessarily polynomials of B) is considered.

If B is a product of the type $B = \overline{A}A$ (or $B = A\overline{A}$), and A is reduced to its Youla form, then B also acquires a block triangular form very similar to the Schur form. This implies that for such a matrix B, we could replace, using Parlett's technique, the reduction of B to the Schur form by the reduction of A to the Youla form. This avoids the explicit computation of B, if A is known.

REFERENCES

[1] R. H. Bartels and G. W. Stewart, *Solution of the equation* $AX + XB = C$, Comm. ACM, 15 (1972), pp. 820–826.

[2] J. H. Bevis, F. J. Hall, and R. E. Hartwig, *Consimilarity and the matrix equation* $A\overline{X} - XB = C$, in Proceedings of the Third Auburn Matrix Theory Conference, F. Uhlig and R. Grone, eds., pp. 51–64, North Holland, Amsterdam, 1987.

[3] Å. Björck and S. Hammarling, *A Schur method for the square root of a matrix*, Linear Algebra Appl., 52 (1983), pp. 127–140.

[4] A. Bunse-Gerstner and W. B. Gragg, *Singular value decompositions of complex symmetric matrices*, J. Comput. Math. Physics, 21 (1988), pp. 41–54.

[5] J. H. P. Colpa, *Diagonalization of the quadratic boson hamiltonian with zero modes*, I. Math. Physica 134A, 2 (1986), pp. 377–419.

[6] A. George, K. Ikramov, L. Matushkina, and W.-P. Tang, *On a* QR*-like algorithm for some structured eigenvalue problems*, Tech. Report CS-9405, University of Waterloo, Waterloo, Ontario, Canada, 1994.

[7] A. George and J. Liu, *Computer Solution of Large Sparse Positive Definite Linear Equations*, Prentice-Hall, New York, 1981.

[8] G. H. Golub, S. Nash, and C. Van Loan, *A Hessenberg–Schur method for the matrix problem* $AX + XB = C$, IEEE Trans. Automat. Control, AC-24 (1979), pp. 909–913.

[9] Y. P. HONG AND R. A. HORN, *A characterization of unitary congruence*, Linear Multilinear Algebra, 25 (1989), pp. 105–119.
[10] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, London, 1988.
[11] H. D. IKRAMOV, *Linear Algebra: Problem Book*, Mir, Moscow, 1983.
[12] KH. D. IKRAMOV, *The use of block symmetries to solve algebraic eigenvalue problems*, USSR Comput. Math. Math. Physics, 21 (1990) pp. 41–54.
[13] A. LEE, *On centrohermitian and skew-centrohermitian matrices*, Linear Algebra Appl., 29 (1980), pp. 205–210.
[14] B. N. PARLETT, *A recurrence among the elements of functions of triangular matrices*, Linear Algebra Appl., 14 (1976), pp. 117–121.
[15] G. W. STEWART, *Algorithm* 406 : HQR3 *and* EXCHANG: FORTRAN *subroutines for calculating and ordering eigenvalues of a real upper Hessenberg matrix*, ACM Trans. Math. Software, 2 (1976), pp. 275–280.
[16] G. W. STEWART AND J.-G. SUN, *Matrix Perturbation Theory*, Academic Press, New York, 1990.
[17] M. VUJICIC, F. HERBUT, AND G. VUJICIC, *Canonical forms for matrices under unitary congruence transformations* I*: conjugate-normal matrices*, SIAM J. Appl. Math., 23 (1972), pp. 225–238.
[18] J. R. WEAVER, *Centrosymmetric (cross-symmetric) matrices, their basic properties, eigenvalues, and eigenvectors*, Amer. Math. Monthly, 92 (1985), pp. 711–717.
[19] J.H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, UK, 1965.
[20] D. C. YOULA, *A normal form for a matrix under the unitary congruence group*, Canadian J. Math., 13 (1961), pp. 694–704.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, pp. 1127-1134, October 1995

007

THE GROUP INVERSE ASSOCIATED WITH AN IRREDUCIBLE PERIODIC NONNEGATIVE MATRIX *

STEVE KIRKLAND†

Abstract. Suppose that $\mathbf{M}$ is an irreducible stochastic matrix with period $d \geq 2$. The canonical form for $\mathbf{M}$ that exhibits its periodic structure generates a natural partitioning of $\mathbf{M}$, which in turn generates a partitioning of $(\mathbf{I}-\mathbf{M})^{\#}$, the group generalized inverse of $\mathbf{I}-\mathbf{M}$. We derive a formula for the blocks in the partitioned form of $(\mathbf{I}-\mathbf{M})^{\#}$, discuss possible sign patterns of $(\mathbf{I}-\mathbf{M})^{\#}$, and use the partitioned formula to obtain information about the Markov chain associated with $\mathbf{M}$.

Key words. nonnegative matrix, group inverse, Markov chain

AMS subject classifications. 15A48, 15A09, 15A51

1. Introduction. Square matrices with nonnegative entries have received a good deal of attention, not only because of their utility in applications, but also because of the remarkable properties that they possess. Recall that a square nonnegative matrix $\mathbf{M}$ is *reducible* if there is a permutation matrix $\mathbf{P}$ such that $\mathbf{PMP}^T = \left[\begin{array}{c|c} \mathbf{X} & \mathbf{Y} \\ \hline \mathbf{0} & \mathbf{Z} \end{array}\right]$, where $\mathbf{X}$ and $\mathbf{Z}$ are square (nonvacuous) matrices, and $\mathbf{0}$ is the zero matrix of the appropriate size; if no such $\mathbf{P}$ exists, $\mathbf{M}$ is *irreducible.* In the case that $\mathbf{M}$ is irreducible, there are two possibilities: either some power of $\mathbf{M}$ is positive, in which case it is called *primitive,* or there is a $d \geq 2$ and a permutation matrix $\mathbf{Q}$ such that

$$\mathbf{QMQ}^T = \begin{bmatrix} \mathbf{0} & \mathbf{M}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{M}_2 & \mathbf{0} \cdots & \mathbf{0} \\ \vdots & & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & & & \mathbf{M}_{d-1} \\ \mathbf{M}_d & \mathbf{0} & & \cdots & \mathbf{0} \end{bmatrix}, \tag{1}$$

where the diagonal zero blocks are square, and where each of the products $\mathbf{A}_i = \mathbf{M}_i\mathbf{M}_{i+1}\ldots\mathbf{M}_d\mathbf{M}_1\ldots\mathbf{M}_{i-1}$ is primitive. In the latter case, $\mathbf{M}$ is *periodic* (*with period* d), and we refer to the matrix in (1) as the *periodic normal form* for $\mathbf{M}$. The theorem of Perron and Frobenius states that an irreducible nonnegative matrix $\mathbf{M}$ has an algebraically simple positive eigenvalue r, called the *Perron value,* and that associated with r are left and right eigenvectors (called *left* and *right Perron vectors*) in which every entry is positive.

Thus if $\mathbf{M}$ is an irreducible matrix with Perron value r, $r\mathbf{I}-\mathbf{M}$ is not invertible, but because its null space has dimension 1, $r\mathbf{I}-\mathbf{M}$ has a *group inverse,* i.e., there is a matrix $(r\mathbf{I}-\mathbf{M})^{\#}$ such that (i) $(r\mathbf{I}-\mathbf{M})^{\#}(r\mathbf{I}-\mathbf{M}) = (r\mathbf{I}-\mathbf{M})(r\mathbf{I}-\mathbf{M})^{\#}$, (ii) $(r\mathbf{I}-\mathbf{M})^{\#}(r\mathbf{I}-\mathbf{M})(r\mathbf{I}-\mathbf{M})^{\#} = (r\mathbf{I}-\mathbf{M})^{\#}$, and (iii) $(r\mathbf{I}-\mathbf{M})(r\mathbf{I}-\mathbf{M})^{\#}(r\mathbf{I}-\mathbf{M}) = (r\mathbf{I}-\mathbf{M})$. The group inverse $(r\mathbf{I}-\mathbf{M})^{\#}$ provides information about $\mathbf{M}$ in (at least) two different ways.

(i) Let the left and right Perron vectors of $\mathbf{M}$ be $\mathbf{y}^T$ and $\mathbf{x}$, respectively, normalized so that $\mathbf{y}^T\mathbf{x} = 1$, and let $\mathbf{X}$ and $\mathbf{Y}$ be the diagonal matrices whose diagonal entries are

* Received by the editors December 13, 1993; accepted (in revised form) by C. Meyer October 3, 1994. The research of this author was supported in part by the Natural Sciences and Engineering Research Council of Canada grant OGP0138251.

† Department of Mathematics and Statistics, University of Regina, Regina, Saskatchewan, S4S 0A2 (kirkland@max.cc.uregina.ca).

the corresponding entries in $\mathbf{x}$ and $\mathbf{y}$. Deutsch and Neumann [2] show that the matrix whose (i,j)th entry is the second partial derivative of r with respect to m_{ij} is equal to $2\mathbf{Y}(r\mathbf{I}-\mathbf{M})^{\#T}\mathbf{X}$. Thus the sign pattern of $(r\mathbf{I}-\mathbf{M})^{\#}$ reveals whether r is convex or concave as a function of a particular entry in $\mathbf{M}$. Haviv, Ritov, and Rothblum [3] also use $(r\mathbf{I}-\mathbf{M})^{\#}$ as part of an iterative method for calculating higher order derivatives of r.

(ii) In the case where $\mathbf{M}$ is an irreducible *stochastic* matrix, i.e., each of its row sums is 1, $\mathbf{M}$ can be viewed as the transition matrix of a Markov chain (see [5]). Meyer [6] shows how $(\mathbf{I}-\mathbf{M})^{\#}$ can be used to calculate the matrix consisting of the mean first passage times for the chain. Furthermore, in the case where $\mathbf{M}$ is primitive with left Perron vector $\mathbf{y}^T$ (normalized so that the entries in $\mathbf{y}^T$ sum to 1), Meyer shows that $(\mathbf{I}-\mathbf{M})^{\#} = \lim_{n\to\infty}\mathbf{N}_M^{(n)} - n\mathbf{1}\mathbf{y}^T$, where $\mathbf{1}$ is the all ones vector, and $\mathbf{N}_M^{(n)}$ is the matrix whose (i,j)th entry is the expected number of times that the chain is in state j in the first n steps, given that the chain was initially in state i. Since $\mathbf{y}^T$ is the stationary distribution for the Markov chain, we see that the (i,j)th entry of $(\mathbf{I}-\mathbf{M})^{\#}$ measures the asymptotic difference between the expected number of visits to j given that the chain started in state i, and the expectation of the same quantity given that the chain started in its stationary distribution.

Consider an irreducible periodic matrix $\mathbf{M}$ with Perron value r; without loss of generality, we assume that $\mathbf{M}$ is in periodic normal form. From paragraphs (i) and (ii) above, $(r\mathbf{I}-\mathbf{M})^{\#}$ contains information about $\mathbf{M}$, and since $\mathbf{M}$ is a partitioned matrix, it is natural to wonder whether we can partition $(r\mathbf{I}-\mathbf{M})^{\#}$ conformally with $\mathbf{M}$ and deduce the structure of the blocks in $(r\mathbf{I}-\mathbf{M})^{\#}$. This paper provides such a partitioned formula for $(r\mathbf{I}-\mathbf{M})^{\#}$, discusses the sign patterns of $(r\mathbf{I}-\mathbf{M})^{\#}$, and applies the partitioned formula to obtain information about Markov chains whose transition matrix is irreducible and periodic.

In the sequel we will adopt several conventions. First, we will assume that our matrix $\mathbf{M}$ is irreducible and periodic with period $d \geq 2$. We will also assume that $\mathbf{M}$ is stochastic; we lose no generality with this assumption, since any irreducible nonnegative matrix is similar (via a diagonal matrix with positive diagonal entries) to a positive multiple of a stochastic matrix (see [4, pp. 528–529]). In addition to assuming that $\mathbf{M}$ is stochastic, we will also suppose that $\mathbf{M}$ is in the periodic normal form (1). For each $1 \leq i \leq d$, we let $\mathbf{A}_i = \mathbf{M}_i\mathbf{M}_{i+1}\ldots\mathbf{M}_d\mathbf{M}_1\ldots\mathbf{M}_{i-1}$, which is both primitive and stochastic; in particular, $\mathbf{1}$ is a right Perron vector for $\mathbf{A}_i$. We denote by $\mathbf{u}_i^T$ the left Perron vector of $\mathbf{A}_i$, normalized so that $\mathbf{u}_i^T\mathbf{1} = 1$. Throughout, $\mathbf{1}$ will denote the all ones vector, but its order will be suppressed for notational convenience; the order will always be clear from the context.

2. The main formula. We begin with a useful preliminary result.

PROPOSITION 1. *Suppose that* $\mathbf{A}$ *and* $\mathbf{B}$ *are nonnegative matrices of orders* $n\times m$ *and* $m\times n$, *respectively, and that both have row sums* 1. *Furthermore, suppose that* $\mathbf{AB}$ *and* $\mathbf{BA}$ *are primitive, and that* $\mathbf{u}^T$ *is the left Perron vector of* $\mathbf{AB}$, *normalized so that* $\mathbf{u}^T\mathbf{1} = 1$. *Then* $(\mathbf{I}-\mathbf{AB})^{\#} = \mathbf{I}-\mathbf{1}\mathbf{u}^T + \mathbf{A}(\mathbf{I}-\mathbf{BA})^{\#}\mathbf{B}$.

Proof. Let $\mathbf{G} = \mathbf{I}-\mathbf{1}\mathbf{u}^T + \mathbf{A}(\mathbf{I}-\mathbf{BA})^{\#}\mathbf{B}$. A result in Campbell and Meyer [1, Thm. 8.5.5] implies that if $\mathbf{u}^T\mathbf{G} = \mathbf{0}^T$ and $(\mathbf{I}-\mathbf{AB})\mathbf{G} = \mathbf{I}-\mathbf{1}\mathbf{u}^T$, then $\mathbf{G} = (\mathbf{I}-\mathbf{AB})^{\#}$. Now $\mathbf{u}^T\mathbf{A}$ is a left Perron vector for $\mathbf{BA}$, and since $(\mathbf{I}-\mathbf{BA})^{\#}$ and $(\mathbf{I}-\mathbf{BA})$ have the same null space (see [1]) we find that $\mathbf{u}^T\mathbf{G} = \mathbf{0}^T$. Also, $(\mathbf{I}-\mathbf{AB})\mathbf{G} = \mathbf{I}-\mathbf{1}\mathbf{u}^T+\mathbf{A}(\mathbf{1}\mathbf{u}^T\mathbf{A}-\mathbf{I}+(\mathbf{I}-\mathbf{BA})(\mathbf{I}-\mathbf{BA})^{\#})\mathbf{B}$; since $(\mathbf{I}-\mathbf{BA})(\mathbf{I}-\mathbf{BA})^{\#} = \mathbf{I}-\mathbf{1}\mathbf{u}^T\mathbf{A}$ ([1, Thm. 8.2.3]), the result now follows. □

COROLLARY 1. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be as in Proposition* 1. *Then* $(\mathbf{I}-\mathbf{AB})^{\#}\mathbf{A} = \mathbf{A}(\mathbf{I}-\mathbf{BA})^{\#}$.

Proof. From Proposition 1 we have $(\mathbf{I}-\mathbf{AB})\#\mathbf{A} = \mathbf{A} - \mathbf{1u}^T\mathbf{A} + \mathbf{A}(\mathbf{I}-\mathbf{BA})\#\mathbf{BA} = \mathbf{A}(\mathbf{I}-\mathbf{1u}^T\mathbf{A}+(\mathbf{I}-\mathbf{BA})\#\mathbf{BA})$. Using the fact that $(\mathbf{I}-\mathbf{BA})(\mathbf{I}-\mathbf{BA})\# = \mathbf{I}-\mathbf{1u}^T\mathbf{A}$ now gives the result. □

We now present the main formula.

THEOREM 1. *Suppose that* $\mathbf{M}$ *is an irreducible stochastic matrix with period* $d \geq 2$. *Furthermore, suppose that* $\mathbf{M}$ *is in its periodic normal form, that is,*

$$\mathbf{M} = \begin{bmatrix} \mathbf{0} & \mathbf{M}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{M}_2 & \mathbf{0}\cdots & \mathbf{0} \\ \vdots & & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & & & \mathbf{M}_{d-1} \\ \mathbf{M}_d & \mathbf{0} & & \cdots & \mathbf{0} \end{bmatrix}.$$

For each $1 \leq j \leq d$, *let* $\mathbf{A}_j = \mathbf{M}_j\mathbf{M}_{j+1}\ldots\mathbf{M}_d\mathbf{M}_1\ldots\mathbf{M}_{j-1}$, *and let* $\mathbf{u}_1^T$ *be the left Perron vector for* $\mathbf{A}_1$, *normalized so that* $\mathbf{u}_1^T\mathbf{1} = 1$, *and for* $2 \leq j \leq d$, *let* $\mathbf{u}_j^T = \mathbf{u}_1^T\mathbf{M}_1\ldots\mathbf{M}_{j-1}$, *so that* $\mathbf{u}_j^T$ *is the left Perron vector for* $\mathbf{A}_j$, *normalized so that* $\mathbf{u}_j^T\mathbf{1} = 1$. *Let* $\mathbf{G} = (\mathbf{I}-\mathbf{M})\#$, *and partition* $\mathbf{G}$ *into blocks* $\mathbf{G}_{ij}, 1 \leq i,j \leq d$, *using the same partitioning as that for* $\mathbf{M}$. *Then for each* $1 \leq i,j \leq d$,

$$\mathbf{G}_{ij} = \begin{cases} (\mathbf{I}-\mathbf{A}_i)\#\mathbf{M}_i\ldots\mathbf{M}_{j-1} + \left(\frac{d-1}{2d} - \frac{j-i}{d}\right)\mathbf{1u}_j^T & \text{if } i < j, \\ (\mathbf{I}-\mathbf{A}_i)\# + \left(\frac{d-1}{2d}\right)\mathbf{1u}_i^T & \text{if } i = j, \\ (\mathbf{I}-\mathbf{A}_i)\#\mathbf{M}_i\ldots\mathbf{M}_d\mathbf{M}_1\ldots\mathbf{M}_{j-1} + \left(\frac{d-1}{2d} - \frac{d+j-i}{d}\right)\mathbf{1u}_j^T & \text{if } i > j. \end{cases} \tag{2}$$

Proof. Let $\mathbf{w}^T = (1/d)[\mathbf{u}_1^T\ldots\mathbf{u}_d^T]$, and note that $\mathbf{w}^T$ is the left Perron vector for $\mathbf{M}$, normalized so that $\mathbf{w}^T\mathbf{1} = 1$. Let $\mathbf{G}$ be the matrix whose blocks are given by (2). Again we employ [1, Thm. 8.5.5] to establish the result. From (2) it follows that the jth block of $\mathbf{w}^T\mathbf{G}$ is equal to

$$(1/d) = \sum_{i=1}^{d}\left(\frac{d-1}{2d} - \frac{(j-i)\bmod d}{d}\right)\mathbf{1u}_j^T = \mathbf{0}^T.$$

Consequently, $\mathbf{1w}^T\mathbf{G} = \mathbf{0}$.

Suppose that $1 \leq i < j \leq d$. Then the (i,j)th block of $(\mathbf{I}-\mathbf{M})\mathbf{G}$, $(\mathbf{I}-\mathbf{M})\mathbf{G}_{ij}$, is given by

$$\begin{aligned}\mathbf{G}_{ij} - \mathbf{M}_i\mathbf{G}_{i+1j} &= (\mathbf{I}-\mathbf{A}_i)\#\mathbf{M}_i\ldots\mathbf{M}_{j-1} + \left(\frac{d-1}{2d} - \frac{j-i}{d}\right)\mathbf{1u}_j^T \\ &\quad - \mathbf{M}_i\left((\mathbf{I}-\mathbf{A}_{i+1})\#\mathbf{M}_{i+1}\ldots\mathbf{M}_{j-1} + \left(\frac{d-1}{2d} - \frac{j-i-1}{d}\right)\mathbf{1u}_j^T\right),\end{aligned}$$

where $\mathbf{M}_{i+1}\ldots\mathbf{M}_{j-1}$ is to be interpreted as $\mathbf{I}$ if $i = j-1$. Applying Corollary 1, we see that $(\mathbf{I}-\mathbf{A}_i)\#\mathbf{M}_i\ldots\mathbf{M}_{j-1} = \mathbf{M}_i((\mathbf{I}-\mathbf{A}_{i+1})\#\mathbf{M}_{i+1}\ldots\mathbf{M}_{j-1}$, and hence $(\mathbf{I}-\mathbf{M})\mathbf{G}_{ij} = -(1/d)\mathbf{1u}_j^T$. Similar arguments show that if $1 \leq j < i \leq d$, then $(\mathbf{I}-\mathbf{M})\mathbf{G}_{ij} = -(1/d)\mathbf{1u}_j^T$, while $(\mathbf{I}-\mathbf{M})\mathbf{G}_{ii} = \mathbf{I} - (1/d)\mathbf{1u}_i^T$. Thus $(\mathbf{I}-\mathbf{M})\mathbf{G} = \mathbf{I} - \mathbf{1w}^T$, so that $(\mathbf{I}-\mathbf{M})\# = \mathbf{G}$. □

Theorem 1 may offer some computational advantage in finding $(\mathbf{I}-\mathbf{M})\#$, since it expresses that matrix terms of some group inverses of lower order. Furthermore, Proposition 1 shows that if $j \neq i$, $(\mathbf{I}-\mathbf{A}_j)\#$ can be computed from $(\mathbf{I}-\mathbf{A}_i)\#$.

3. Sign patterns for the group inverse. As was mentioned in (i), for an irreducible stochastic matrix $\mathbf{S}$, the sign pattern of $(\mathbf{I}-\mathbf{S})^{\#}$ reveals the convexity or concavity of the Perron value of $\mathbf{S}$ as a function of its entries. Deutsch and Neumann [2] give a proof of the fact that the diagonal entries of $(\mathbf{I}-\mathbf{S})^{\#}$ are positive, and show that each row and column of $(\mathbf{I}-\mathbf{S})^{\#}$ contains at least one negative entry. They then pose the problem of determining those $\mathbf{S}$ with the property that $(\mathbf{I}-\mathbf{S})^{\#}$ has all nonpositive off-diagonal entries; it is not difficult to see that this is the case if and only if $(\mathbf{I}-\mathbf{S})^{\#}$ is an M-matrix. Our next result characterizes the irreducible stochastic periodic matrices $\mathbf{M}$ such that $(\mathbf{I}-\mathbf{M})^{\#}$ is an M-matrix.

THEOREM 2. *Suppose that* $\mathbf{M}$ *is an irreducible stochastic matrix of the form*

$$\mathbf{M}=\begin{bmatrix} \mathbf{0} & \mathbf{M}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{M}_2 & \mathbf{0}\cdots & \mathbf{0} \\ \vdots & & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & & & \mathbf{M}_{d-1} \\ \mathbf{M}_d & \mathbf{0} & & \cdots & \mathbf{0} \end{bmatrix}$$

for some $d \geq 2$. *If* $d \geq 4$, *then* $(\mathbf{I}-\mathbf{M})^{\#}$ *is not an* M-*matrix. If* $d=3$, *then* $(\mathbf{I}-\mathbf{M})^{\#}$ *is an* M-*matrix if and only if there are positive vectors* $\mathbf{x}^T, \mathbf{y}^T$, *and* $\mathbf{z}^T$, *each of whose entries sum to* 1, *such that* $\mathbf{M}_1=\mathbf{1}\mathbf{x}^T, \mathbf{M}_2=\mathbf{1}\mathbf{y}^T$, *and* $\mathbf{M}_3=\mathbf{1}\mathbf{z}^T$. *If* $d=2$, *then* $(\mathbf{I}-\mathbf{M})^{\#}$ *is an* M-*matrix if and only if* $(\mathbf{I}-\mathbf{A}_1)^{\#}\mathbf{M}_1-(\frac{1}{4})\mathbf{1}\mathbf{u}_2^T \leq \mathbf{0}$ *and* $(\mathbf{I}-\mathbf{A}_2)^{\#}\mathbf{M}_2-(\frac{1}{4})\mathbf{1}\mathbf{u}_1^T \leq \mathbf{0}$, *where* $\mathbf{A}_1=\mathbf{M}_1\mathbf{M}_2, \mathbf{A}_2=\mathbf{M}_2\mathbf{M}_1$, *and for* $i=1,2, \mathbf{u}_i^T$ *is the left Perron vector for* $\mathbf{A}_i$, *normalized so that its entries sum to* 1.

Proof. Let $\mathbf{G}=(\mathbf{I}-\mathbf{M})^{\#}$, and partition $\mathbf{G}$ into blocks $\mathbf{G}_{ij}, 1 \leq i,j \leq d$, using the same partitioning as that for $\mathbf{M}$. From Theorem 1 it follows that $\mathbf{G}_{12}\mathbf{1}=\left(\frac{d-3}{2d}\right)\mathbf{1}$, which is a positive vector if $d \geq 4$. Thus if $d \geq 4, \mathbf{G}$ has some positive off-diagonal entries, and so it cannot be an M-matrix.

If $d=3$, then $\mathbf{G}_{12}\mathbf{1}=\mathbf{0}, \mathbf{G}_{23}\mathbf{1}=\mathbf{0}$, and $\mathbf{G}_{31}\mathbf{1}=\mathbf{0}$. Consequently, if $\mathbf{G}$ is an M-matrix, then each of $\mathbf{G}_{12}, \mathbf{G}_{23}$, and $\mathbf{G}_{31}$ must be an all zero block, for if one of those blocks were to have a nonzero entry, it must necessarily have a positive entry because the row sums are zero. Thus we see that $\mathbf{G}_{12}=(\mathbf{I}-\mathbf{A}_1)^{\#}\mathbf{M}_1=\mathbf{0}$. This implies that each column of $\mathbf{M}_1$ is a null vector for $(\mathbf{I}-\mathbf{A}_1)^{\#}$, which in turn implies that each such column is a scalar multiple of $\mathbf{1}$. Hence there is a nonnegative vector $\mathbf{x}^T$ such that $\mathbf{M}_1=\mathbf{1}\mathbf{x}^T$; from the fact that $\mathbf{M}$ is irreducible and stochastic, it follows that $\mathbf{x}^T$ is positive and that its entries sum to 1. A similar argument establishes the formulae for $\mathbf{M}_2$ and $\mathbf{M}_3$. Conversely, if $\mathbf{M}_1=\mathbf{1}\mathbf{x}^T, \mathbf{M}_2=\mathbf{1}\mathbf{y}^T$, and $\mathbf{M}_3=\mathbf{1}\mathbf{z}^T$ for $\mathbf{x}^T, \mathbf{y}^T$, and $\mathbf{z}^T$ as in the statement of the theorem, an application of Theorem 1 now yields the fact that $(\mathbf{I}-\mathbf{M})^{\#}$ is an M-matrix.

Now suppose that $d=2$. If $\mathbf{G}$ is an M-matrix, then certainly $\mathbf{G}_{12}=(\mathbf{I}-\mathbf{A}_1)^{\#}\mathbf{M}_1-(\frac{1}{4})\mathbf{1}\mathbf{u}_2^T \leq \mathbf{0}$ and $\mathbf{G}_{21}=(\mathbf{I}-\mathbf{A}_2)^{\#}\mathbf{M}_2-(\frac{1}{4})\mathbf{1}\mathbf{u}_1^T \leq \mathbf{0}$. On the other hand, if $(\mathbf{I}-\mathbf{A}_1)^{\#}\mathbf{M}_1-(\frac{1}{4})\mathbf{1}\mathbf{u}_2^T \leq \mathbf{0}$ and $(\mathbf{I}-\mathbf{A}_2)^{\#}\mathbf{M}_2-(\frac{1}{4})\mathbf{1}\mathbf{u}_1^T \leq \mathbf{0}$, then in particular, $(\mathbf{I}-\mathbf{A}_1)^{\#}\mathbf{M}_1\mathbf{M}_2-(\frac{1}{4})\mathbf{1}\mathbf{u}_2^T\mathbf{M}_2 \leq \mathbf{0}$. Since $\mathbf{M}_1\mathbf{M}_2=\mathbf{A}_1$ and $\mathbf{u}_2^T\mathbf{M}_2=\mathbf{u}_1^T$, this last inequality is equivalent to $\mathbf{G}_{11}=(\mathbf{I}-\mathbf{A}_1)^{\#}+(\frac{1}{4})\mathbf{1}\mathbf{u}_1^T \leq \mathbf{I}-(\frac{1}{2})\mathbf{1}\mathbf{u}_1^T$, so that the off-diagonal entries of $\mathbf{G}_{11}$ are negative. A similar argument shows that the off-diagonal entries of $\mathbf{G}_{22}$ are negative, and hence that $\mathbf{G}$ is an M-matrix. □

The next result discusses the conditions under which there is a zero block in the group inverse associated with a periodic matrix.

THEOREM 3. *Suppose that* $\mathbf{M}$ *is an irreducible stochastic matrix of the form*

$$\mathbf{M} = \begin{bmatrix} \mathbf{0} & \mathbf{M}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{M}_2 & \mathbf{0} \cdots & \mathbf{0} \\ \vdots & & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & & & \mathbf{M}_{d-1} \\ \mathbf{M}_d & \mathbf{0} & & \cdots & \mathbf{0} \end{bmatrix}$$

for some $d \geq 2$, *and let* $\mathbf{G} = (\mathbf{I} - \mathbf{M})^{\#}$, *with* $\mathbf{G}$ *partitioned as* $\mathbf{G}_{ij}, 1 \leq i, j \leq d$. *Then* $\mathbf{G}_{ij}$ *is a zero block for some* i *and* j *if and only if* d *is odd,* $(j-i)\mathrm{mod}\, d = (d-1)/2$, *and* $\mathbf{M}_i\mathbf{M}_{i+1}\ldots\mathbf{M}_{j-1} = \mathbf{1}\mathbf{u}_j^T$.

Proof. From Theorem 1 we find that if $\mathbf{G}_{ij} = \mathbf{0}$, then

$$(\mathbf{I} - \mathbf{A}_i)^{\#}\mathbf{M}_i \ldots \mathbf{M}_{j-1} + \left(\frac{d-1}{2d} - \frac{(j-i)\mathrm{mod}\, d}{d}\right)\mathbf{1}\mathbf{u}_j^T = \mathbf{0}.$$

In particular $\mathbf{G}_{ij}\mathbf{1} = \mathbf{0}$, and it follows that

$$\left(\frac{d-1}{2d} - \frac{(j-i)\mathrm{mod}\, d}{d}\right)\mathbf{1} = \mathbf{0}.$$

Hence d must be odd and $(j-i)\mathrm{mod}\, d = (d-1)/2$. Consequently, $\mathbf{0} = \mathbf{G}_{ij} = (\mathbf{I} - \mathbf{A}_i)^{\#}\mathbf{M}_i \ldots \mathbf{M}_{j-1}$, so that each column of $\mathbf{M}_i \ldots \mathbf{M}_{j-1}$ must be a null vector for $(\mathbf{I} - \mathbf{A}_i)^{\#}$; i.e., each column of $\mathbf{M}_i \ldots \mathbf{M}_{j-1}$ is a multiple of 1. Furthermore, by Corollary 1, $(\mathbf{I} - \mathbf{A}_i)^{\#}\mathbf{M}_i \ldots \mathbf{M}_{j-1} = \mathbf{M}_i \ldots \mathbf{M}_{j-1}(\mathbf{I} - \mathbf{A}_j)^{\#}$, so that each row of $\mathbf{M}_i \ldots \mathbf{M}_{j-1}$ is a multiple of $\mathbf{u}_j^T$. It now follows that $\mathbf{M}_i\mathbf{M}_{i+1} \ldots \mathbf{M}_{j-1} = \mathbf{1}\mathbf{u}_j^T$.

The converse is straightforward. □

COROLLARY 2. *If* $\mathbf{G}$ *is as in Theorem* 3, *then* $\mathbf{G}$ *has at most* d *zero blocks.*

4. Applications to Markov chains. Consider a Markov chain with irreducible transition matrix $\mathbf{S}$ and stationary distribution $\mathbf{u}^T$. Let $\mathbf{N}_S^{(n)}$ be the matrix whose (i,j)th entry is the expected number of times that the chain is in state j in the first n steps (the initial step, plus the next $n-1$ steps), given that the chain was initially in state i. Evidently

$$\mathbf{N}_S^{(n)} = \sum_{l=0}^{n-1} \mathbf{S}^l,$$

and as we noted in (ii), Meyer [6] has shown that if $\mathbf{S}$ is primitive, $(\mathbf{I} - \mathbf{S})^{\#} = \lim_{n\to\infty} \mathbf{N}_S^{(n)} - n\mathbf{1}\mathbf{u}^T$, thus yielding an interpretation of the entries in $(\mathbf{I} - \mathbf{S})^{\#}$. Our next result extends this interpretation to *cyclic* Markov chains, that is, those whose transition matrices are irreducible periodic stochastic matrices.

THEOREM 4. *Suppose that* $\mathbf{M}$ *is an irreducible stochastic matrix of the form*

$$\mathbf{M} = \begin{bmatrix} \mathbf{0} & \mathbf{M}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{M}_2 & \mathbf{0} \cdots & \mathbf{0} \\ \vdots & & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & & & \mathbf{M}_{d-1} \\ \mathbf{M}_d & \mathbf{0} & & \cdots & \mathbf{0} \end{bmatrix}$$

for some $d \geq 2$, *and let* $\mathbf{u}^T$ *be the left Perron vector of* $\mathbf{M}$, *normalized so that* $\mathbf{u}^T\mathbf{1} = 1$. *Then*

$$(\mathbf{I} - \mathbf{M})^{\#} = \lim_{n\to\infty} (1/d)\left\{\sum_{k=0}^{d-1} \mathbf{N}_M^{(nd+k)} - (nd+k)\mathbf{1}\mathbf{u}^T\right\}.$$

Proof. Suppose that $1 \leq i < j \leq d$. We find from the special structure of $\mathbf{M}$ that the (i,j) block of $\mathbf{M}^p$, $(\mathbf{M}^p)_{ij}$ say, is equal to $(\mathbf{A}_i)^l \mathbf{M}_i \ldots \mathbf{M}_{j-1}$ if $p = ld + j - i$, while that block is zero if p is not congruent to $j-i$ modulo d. Consequently, if $0 \leq k \leq j-i$, then the (i,j) block of $\mathbf{N}_M^{(nd+k)}$ is given by

$$\sum_{p=0}^{nd+k-1} (\mathbf{M}^p)_{ij} = \sum_{l=0}^{n-1} (\mathbf{A}_i)^l \mathbf{M}_i \ldots \mathbf{M}_{j-1}.$$

Similarly, if $j - i + 1 \leq k \leq d-1$, the (i,j) block of $\mathbf{N}_M^{(nd+k)}$ is given by $\sum_{l=0}^{n} (\mathbf{A}_i)^l$ $\mathbf{M}_i \ldots \mathbf{M}_{j-1}$. It now follows that the (i,j) block of

$$(1/d) \left\{ \sum_{k=0}^{d-1} \mathbf{N}_M^{(nd+k)} - (nd+k)\mathbf{1}\mathbf{u}^T \right\}$$

is equal to $(\sum_{l=0}^{n-1} (\mathbf{A}_i)^l - n\mathbf{1}\mathbf{u}_i^T)\mathbf{M}_i \ldots \mathbf{M}_{j-1} + (1/d)(d-1-j+i)(\mathbf{A}_i)^n \mathbf{M}_i \ldots \mathbf{M}_{j-1} - \frac{d-1}{2d}\mathbf{1}\mathbf{u}_j^T$.

As $n \to \infty, \sum_{l=0}^{n-1} (\mathbf{A}_i)^l - n\mathbf{1}\mathbf{u}_i^T \to (\mathbf{I} - \mathbf{A}_i)^{\#}$, and $(\mathbf{A}_i)^n \to \mathbf{1}\mathbf{u}_i^T$, and hence we see that the (i,j) block of

$$\lim_{n \to \infty} (1/d) \left\{ \sum_{k=0}^{d-1} \mathbf{N}_M^{(nd+k)} - (nd+k)\mathbf{1}\mathbf{u}^T \right\}$$

is equal to

$$(\mathbf{I} - \mathbf{A}_i)^{\#} \mathbf{M}_i \ldots \mathbf{M}_{j-1} + \left(\frac{d-1}{2d} - \frac{j-i}{d} \right) \mathbf{1}\mathbf{u}_j^T,$$

which agrees with the (i,j) block of $(\mathbf{I} - \mathbf{M})^{\#}$. A similar argument goes through when $1 \leq j \leq i \leq d$, yielding that

$$(\mathbf{I} - \mathbf{M})^{\#} = \lim_{n \to \infty} (1/d) \left\{ \sum_{k=0}^{d-1} \mathbf{N}_M^{(nd+k)} - (nd+k)\mathbf{1}\mathbf{u}^T \right\}. \qquad \square$$

For a Markov chain with irreducible stochastic transition matrix $\mathbf{A}$ and stationary distribution vector $\mathbf{u}^T$, Meyer [6] has shown that $\mathbf{E}_A$, the mean first passage matrix (i.e., the matrix whose (i,j) entry is the mean first passage time from state i to state j in the chain) can be expressed in terms of the group inverse associated with $\mathbf{A}$. Specifically, if $\mathbf{J}$ is the all ones matrix and $\mathbf{D}_u$ is the diagonal matrix whose ith diagonal entry is u_i, then

$$\mathbf{E}_A = (\mathbf{I} - (\mathbf{I} - \mathbf{A})^{\#} + \mathbf{J}(\mathbf{I} - \mathbf{A})^{\#}_{\text{diag}})\mathbf{D}_u^{-1}, \tag{3}$$

where $(\mathbf{I} - \mathbf{A})^{\#}_{\text{diag}}$ is the diagonal matrix whose diagonal entries are given by those of $(\mathbf{I} - \mathbf{A})^{\#}$. Our next result gives a partitioned formula for $\mathbf{E}_A$ in the case that the chain is cyclic.

THEOREM 5. *Suppose that* $\mathbf{M}$ *is an irreducible stochastic matrix of the form*

$$\mathbf{M} = \begin{bmatrix} \mathbf{0} & \mathbf{M}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{M}_2 & \mathbf{0} \cdots & \mathbf{0} \\ \vdots & & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & & & \mathbf{M}_{d-1} \\ \mathbf{M}_d & \mathbf{0} & & \cdots & \mathbf{0} \end{bmatrix}$$

for some $d \geq 2$, and let $\mathbf{u}^T$ be the left Perron vector of $\mathbf{M}$, normalized so that $\mathbf{u}^T\mathbf{1} = 1$. Partition $\mathbf{E}_M$ conformally with $\mathbf{M}$, as $(\mathbf{E}_M)_{ij}, 1 \leq i, j \leq d$. Then

$$(\mathbf{E}_M)_{ij} = d[-\mathbf{M}_i \dots \mathbf{M}_{j-1}(\mathbf{I} - \mathbf{A}_j)^{\#} + \mathbf{J}(\mathbf{I} - \mathbf{A}_j)^{\#}_{\text{diag}}]\mathbf{D}_{u_j}^{-1} + (j-i)\mathbf{J}$$
$$= d\mathbf{M}_i \dots \mathbf{M}_{j-1}[\mathbf{E}_{A_j} - (\mathbf{E}_{A_j})_{\text{diag}}] + (j-i)\mathbf{J} \quad \text{if } i < j, \tag{4}$$

$$(\mathbf{E}_M)_{ii} = d[\mathbf{I} - (\mathbf{I} - \mathbf{A}_i)^{\#} + \mathbf{J}(\mathbf{I} - \mathbf{A}_i)^{\#}_{\text{diag}}]\mathbf{D}_{u_i}^{-1} = d\mathbf{E}_{A_i}, \text{ and} \tag{5}$$

(6)
$$(\mathbf{E}_M)_{ij} = d[-\mathbf{M}_i \dots \mathbf{M}_d\mathbf{M}_1 \dots \mathbf{M}_{j-1}(\mathbf{I} - \mathbf{A}_j)^{\#} + \mathbf{J}(\mathbf{I} - \mathbf{A}_j)^{\#}_{\text{diag}}]\mathbf{D}_{u_j}^{-1} + (d+j-i)\mathbf{J}$$
$$= d\mathbf{M}_i \dots \mathbf{M}_d\mathbf{M}_1 \dots \mathbf{M}_{j-1}[\mathbf{E}_{A_j} - (\mathbf{E}_{A_j})_{\text{diag}}] + (d+j-i)\mathbf{J} \quad \text{if } i > j.$$

Proof. Throughout the proof, a subscript ij on a matrix denotes the (i,j)th block of the matrix when it has been partitioned conformally with $\mathbf{M}$. Suppose that $i < j$. From (3), we have $(\mathbf{E}_A)_{ij} = [-(\mathbf{I}-\mathbf{A})^{\#}_{ij} + \mathbf{J}(((\mathbf{I}-\mathbf{A})^{\#})_{jj})_{\text{diag}}](\mathbf{D}_u^{-1})_{jj}$. Applying Theorem 1, Corollary 1, and using the fact that $(\mathbf{D}_u^{-1})_{jj} = d\mathbf{D}_{u_j}^{-1}$, we find that $(\mathbf{E}_A)_{ij} = d[-\mathbf{M}_i \dots \mathbf{M}_{j-1}(\mathbf{I}-\mathbf{A}_j)^{\#} + (j-i)(1/d)\mathbf{1}\mathbf{u}_j + \mathbf{J}(\mathbf{I}-\mathbf{A}_j)^{\#}_{\text{diag}}]\mathbf{D}_{u_j}^{-1} = d[-\mathbf{M}_i \dots \mathbf{M}_{j-1}(\mathbf{I}-\mathbf{A}_j)^{\#} + \mathbf{J}(\mathbf{I}-\mathbf{A}_j)^{\#}_{\text{diag}}]\mathbf{D}_{u_j}^{-1} + (j-i)\mathbf{J}$. By considering Meyer's formula for $\mathbf{E}_{A_j}$, it is readily established that this last expression is equal to $d\mathbf{M}_i \dots \mathbf{M}_{j-1}[\mathbf{E}_{A_j} - (\mathbf{E}_{A_j})_{\text{diag}}] + (j-i)\mathbf{J}$. Analogous arguments yield the desired expressions for the cases $i > j$ and $i = j$. □

We remark that the formulas of Theorem 5 have sensible interpretations in terms of the cyclic Markov chain associated with $\mathbf{M}$. Suppose that i_1 and i_2 are two states corresponding to the ith cyclic set (i.e., the set of indices of the rows and columns associated with the ith diagonal block in the periodic normal form for $\mathbf{M}$). Because of the periodic structure of $\mathbf{M}$, the probability that the chain is in i_2 at the kth step, given that it was initially in i_1, is nonzero only if k is a multiple of d. Furthermore, the d step transition matrix for states in the ith cyclic set is $\mathbf{A}_i$. It now follows that the mean first passage time from i_1 to i_2 is given by the corresponding element of $d\mathbf{E}_{A_i}$, which agrees with formula (5). Similarly, suppose that $1 \leq i < j \leq d$, that i_1 is in the ith cyclic set, and that j_1 is in the jth cyclic set. Then the probability that the chain is in j_1 at the kth step, given that it was initially in i_1, is nonzero only if $k = (j-i) \bmod d$. As above, the d step transition matrix for states in the jth cyclic set is $\mathbf{A}_j$, and so letting S_j denote the jth cyclic set, it follows that we can write the mean first passage time from i_1 to j_1 as

$$(j-i)\{\text{Probability of a transition from } i_1 \text{ to } j_1 \text{ in } j-i \text{ steps}\}$$
$$+ \sum_{l \in S_j \setminus \{j_1\}} \{\text{Probability of a transition from } i_1 \text{ to } l \text{ in } j-i \text{ steps}\}$$
$$\times \{j - i + \text{ mean first passage time from } l \text{ to } j_1\}.$$

It is now readily verified that the above expression is the same as the entry in $d\mathbf{M}_i \dots \mathbf{M}_{j-1}[\mathbf{E}_{A_j} - (\mathbf{E}_{A_j})_{\text{diag}}] + (j-i)\mathbf{J}$ corresponding to states i_1 and j_1, which agrees with formula (4). A similar interpretation can be obtained for the case $i > j$.

It is interesting to note that this type of probabilistic interpretation in fact provides another proof of Theorem 1. Using this type of probabilistic reasoning, we can establish formulas (4)–(6) of Theorem 5 for the blocks of $\mathbf{E}_M$. We can then deduce the formulas for the block of $(\mathbf{I} - \mathbf{M})^{\#}$ by using (3) to express $(\mathbf{I} - \mathbf{M})^{\#}$ in terms of $\mathbf{E}_M$ and $\mathbf{D}_u$.

Acknowledgment. The author is grateful to Professor Michael Neumann for his helpful comments concerning the material in this paper.

REFERENCES

[1] S. L. CAMPBELL AND C. D. MEYER JR., *Generalized Inverses of Linear Transformations*, Dover, New York, 1991.

[2] E. DEUTSCH AND M. NEUMANN, *Derivatives of the Perron root at an essentially nonnegative matrix and the group inverse of an M-matrix*, J. Math. Anal. Appl., 102 (1984), pp. 1–29.

[3] M. HAVIV, Y. RITOV, AND U. G. ROTHBLUM, *Taylor expansions of eigenvalues of perturbed matrices with applications to spectral radii of nonnegative matrices*, Linear Algebra Appl., 168 (1992), pp. 159–188.

[4] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

[5] J. G. KEMENY AND J. L. SNELL, *Finite Markov Chains*, D. Van Nostrand, Princeton, 1960.

[6] C. D. MEYER JR., The role of the group generalized inverse in the theory of finite Markov chains. SIAM Rev., 17 (1975), pp. 443–464.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, pp. 1135–1153, October 1995

008

VARIABLE BLOCK CG ALGORITHMS FOR SOLVING LARGE SPARSE SYMMETRIC POSITIVE DEFINITE LINEAR SYSTEMS ON PARALLEL COMPUTERS, I: GENERAL ITERATIVE SCHEME*

A. A. NIKISHIN† AND A. YU. YEREMIN‡

Abstract. This paper considers a new approach to construction of efficient parallel solution methods of large sparse SPD linear systems. This approach is based on the so-called variable block CG method, a generalization of the standard block CG method, where it is possible to reduce the iteration block size adaptively (at any iteration) by construction of an A-orthogonal projector without restarts and without algebraic convergence of residual vectors. It enables one to find the constructive compromise between the required resource of parallelism, the resulting convergence rate, and the serial arithmetic costs of one block iteration to minimize the total parallel solution time. The orthogonality and minimization properties of the variable CG method are established and the theoretical analysis of the convergence rate is presented. The results of numerical experiments with large FE systems coming from h- and p-approximations of three-dimensional equilibrium equations for linear elastic orthotropic materials show that the convergence rate of the variable block CG method is comparable to that of the standard block CG method even when utilizing a large block size, while the total serial arithmetic costs of the variable block CG method are comparable or even smaller than those of the corresponding point CG method.

Key words. large sparse SPD linear systems, variable block preconditioned CG method, A-orthogonal projector

AMS subject classifications. 65F10, 65F50

1. Introduction.

Iterative solution of large sparse symmetric and unsymmetric linear systems comprises the most time-consuming stage when solving many computationally intensive three-dimensional industrial problems on parallel and especially massively parallel computers. The standard approach to construction of efficient parallel solution methods consists first of all in designing efficient parallel preconditioners. Unfortunately, it is very difficult to achieve parallelism for high-quality preconditioners. For example, incomplete triangular factorizations provide a promising approach to construction with reasonable arithmetic costs of high quality preconditioners but they are very hard to parallelize. High quality factorized sparse approximate inverses based preconditioners [4] possess the good parallelism but their construction is very expensive. Construction of polynomial preconditioners is very inexpensive; they are highly parallelizable but their preconditioning quality is extremely poor especially when solving large industrial problems.

Another way to enhance the parallelism of the iterative solution methods is related to consideration of block iterative schemes like the block CG (BCG) method [5]. Unfortunately, there does not exist any reasonable mathematical approach to choose the optimal block size of such schemes that achieve parallelism, possess the fast convergence rate, and minimize total parallel solution time. In most practical cases an increase of the convergence rate and an improvement of efficiency of the parallel implementation of one block iteration may not compensate its rapid increase with the block size of the arithmetic costs per block iteration. Moreover, the BCG

* Received by the editors April 15, 1993; accepted for publication (in revised form) by A. Greenbaum October 10, 1994.

† Computing Center, Russian Academy of Sciences, Vavilov str. 40, Moscow 117 967, Russia (nikishin@ccas.ru).

‡ Institute of Numerical Mathematics of the Russian Academy of Sciences, Leninsky prosp. 32A, 117 334 Moscow, Russia (badger@ccas.ru).

method may become numerically unstable when increasing the block size.

If an a priori choice of the optimal block size is not feasible one can try to find a constructive compromise between the convergence rate and the arithmetic costs of one block iteration by an adaptive reduction of the block size. There exists an opportunity to reduce adaptively (during the iterative process) the current block size of the standard BCG method. But it is possible only when the current block residual loses full rank, i.e., in the case of the algebraic convergence of at least one residual vector of the block residual (see §2).

In this series of papers we consider another approach to construction of efficient iterative methods for parallel solution of large sparse symmetric positive definite (SPD) linear systems coming from three-dimensional industrial applications. It is based on the so-called variable block CG (VBCG) method, where the current block size can be reduced at each block iteration without any restart and loss of full rank of the block residual. The mathematical idea of the method consists in an adaptive construction of an A-orthogonal projector by reducing the block size of the current block direction vector according to some criterion. This A-orthogonal projector is used on subsequent iterations with the reduced block size to maintain a convergence rate comparable with that of the BCG method with the initial block size. For the resulting VBCG method we prove orthogonality and minimization properties similar to those of the standard CG method. These minimization properties of the variable block preconditioner CG (VBPCG) method enable us to investigate its convergence properties in terms of the convergence properties of the CG method with preconditioning by a projector [2].

The results of numerical experiments with large SPD matrices coming from three-dimensional finite element (FE) applications show that the complexity (measured in terms of the required number of the preconditioned matrix vector multiplications) of the appropriately constructed VBPCG method may coincide with the complexity of the corresponding point preconditioned CG scheme. It means that (1) when constructing efficient iterative solution methods for parallel and especially massively parallel applications, we can essentially exploit the best existing serial preconditioners and (2) the arithmetic costs of the resulting variable block iterative schemes possessing large parallelism may be comparable with the arithmetic costs of the best serial point iterative schemes.

This part of the paper is organized as follows. Section 2 presents a functional approach to derivation of the BCG method which underlies the VBCG method and investigates the reduction of the block size due to the loss of full rank of the block iterates. In §3 we derive the VBCG method, prove its orthogonality and minimization properties, and investigate its convergence properties. In §4 we discuss some realization aspects of VBCG algorithms, present the results of numerical experiments with practical three-dimensional FE systems and concluding remarks.

2. BCG method. The classical approach to derivation of the conjugate gradient method for solving systems of linear equations with an SPD matrix A is essentially based on the notion of "line search" for minimizing a quadratic function of the form

$$\mathcal{F}(x) = (x - x^*)^T A(x - x^*), \tag{2.1}$$

where x^* is the desired solution.

In this section we present a generalization of this approach to the block case and derive the block preconditioned CG (BPCG) method as an iterative process of minimizing a special type quadratic function.

It is well known that the problem of solving the matrix equation

$$AX = B \equiv AX^*,$$

where X, X^*, $B \in \mathcal{R}^{n\times s}$, is equivalent to the unconstrained minimization problem

$$\min F(X) \tag{2.2}$$

for the nonnegative quadratic function

$$F(X) = \mathrm{tr}[(X - X^*)^T A(X - X^*)], \tag{2.3}$$

where the block vector $X \in \mathcal{R}^{n\times s}(s < n)$ consists of $n \times s$ independent variables.

Let $y_{.i}$ denote the ith column of a block vector Y. Then $F(X)$ can be presented as the sum of nonnegative quadratic functions

$$F(X) = \sum_{i=1}^{s} \mathcal{F}_i(x_{.i}),$$

where $\mathcal{F}_i(x_{.i}) = (x_{.i} - x^*_{.i})^T A(x_{.i} - x^*_{.i})$. It can be easily seen that unconstrained minimization problem (2.2) decomposes into s subproblems of finding the minimum of $\mathcal{F}_i(x), x \in \mathcal{R}^n$, which can be solved separately, for instance, by the CG method. It is possible to exploit another step-by-step procedure for solving these subproblems simultaneously, where instead of a line search, minimization over some linear subspace is performed at each step. If X' is an approximation to the minimum point X^* of $F(X)$ and the columns of an $n \times d$ matrix P form a basis of some subspace $\mathcal{P}$, then we construct the next approximation X'' to X^* as follows

$$X'' = X' + P\alpha' \qquad \text{and} \qquad \nabla F'(\alpha') = 0, \tag{2.4}$$

where $\nabla F'(\alpha')$ is the gradient of the function $F'(\alpha) = F(X + P\alpha), \alpha \in \mathcal{R}^{d\times s}$, at the point α'. Procedure (2.4) is equivalent to a plane search on each $\mathcal{F}_i(u)$ independently, i.e., to minimization of $\mathcal{F}_i(x)$ over all x which lie on the hyperplane $\pi_i : x'_{.i} + P\alpha_{.i}$. Therefore, we can write down the optimal value of the step parameter α' in the form

$$\alpha' = -(P^T AP)^{-1}(P^T G'), \tag{2.5}$$

where $G' = AX' - AX^*$ coincides up to a scalar multiplier with the block vector form of the gradient ∇F at the point X'.

Taking into account relations (2.2)–(2.5) together with an appropriate updating strategy for P we can define the BPCG method for solving the matrix equation $AX = B$, where $B, X \in \mathcal{R}^{n\times s}$.

THE BPCG METHOD

Given an initial guess X^0 to the solution matrix X^*.

Initial stage: Set $R^0 = B - AX^0$ and $P^0 = MR^0\gamma_0$, where $M \in \mathcal{R}^{n\times n}$ is an SPD preconditioner, and $\gamma_0 \in \mathcal{R}^{s\times s}$ is a nonsingular matrix.

For $k = 0, 1, \ldots$ *iterate* :

$$X^{k+1} = X^k + P^k\alpha_k, \tag{2.6}$$

$$R^{k+1} = R^k - AP^k\alpha_k, \tag{2.7}$$

$$P^{k+1} = (MR^{k+1} + P^k\beta_k)\gamma_{k+1}, \tag{2.8}$$

where α_k, β_k and $\gamma_k \in \mathcal{R}^{s\times s}$ are determined so that

(a) X^{k+1} minimizes $\mathrm{tr}[(X-X^*)^T A(X-X^*)]$ over all X which lie on the variety $\Pi_k : X^k + \mathcal{P}$ with $\mathcal{P} = \mathrm{span}(P^k)$;

(b) P^{k+1} is A-orthogonal (conjugate) to P^k: $P^{k+1\,T} A P^k = 0$;

(c) α_k and γ_k are nonsingular matrices.

It can be easily shown that conditions (a)–(c) can be satisfied simultaneously for all k independently of the ranks of the matrices R^k and P^k. Moreover, since $\mathrm{span}(P^k) = \mathrm{span}(P^k\delta)$ for any nonsingular matrix $\delta \in \mathcal{R}^{s\times s}$, the entries of the block matrices R^k and X^k are invariant with respect to the choice of γ_k. The following lemma establishes the main properties of the BPCG iterates.

LEMMA 2.1. *For $j < k$ the* BPCG *method iterates satisfy the following conditions:*

$$R^{k\,T} M R^j = 0, \tag{2.9}$$

$$P^{k\,T} A P^j = 0, \tag{2.10}$$

$$R^{k\,T} P^j = 0. \tag{2.11}$$

Proof. We utilize induction to prove the statements of the lemma. Properties of quadratic functions and conditions (a)–(c) for choosing the parameter matrices α_0, β_0 and γ_0 imply the statements of Lemma 2.1 for $k = 1$.

Assume that the statements are valid for all iterations whose numbers are less than or equal to k. Under conditions (a) and (b) we have

$$P^{k+1\,T} A P^k = 0 \qquad \text{and} \qquad R^{k+1\,T} P^k = 0.$$

For $j < k$ by the induction hypothesis we have

$$\begin{aligned} R^{k+1\,T} M R^j &= R^{k\,T} M R^j - \alpha_k^T P^{k\,T} A M R^j \\ &= R^{k\,T} M R^j - \alpha_k^T P^{k\,T} A (P^j \gamma_j^{-1} - P^{j-1}\beta_{j-1}). \end{aligned}$$

Hence, the block residual matrix R^{k+1} is orthogonal to P^j for all $j < k$ since

$$R^{k+1\,T} P^j = R^{k+1\,T} M (R^j \gamma_j + R^{j-1}\gamma_{j-1}\beta_{j-1}\gamma_j + \ldots + R^0\gamma_0\beta_0 \ldots \beta_{j-1}\gamma_j) = 0.$$

The above equalities imply that

$$R^{k+1\,T} M R^k = R^{k+1\,T} (P^k \gamma_k^{-1} - P^{k-1}\beta_{k-1}) = 0.$$

Finally, we conclude that for all $j < k$

$$\begin{aligned} P^{k+1\,T} A P^j &= \gamma_{k+1}^T R^{k+1\,T} M A P^j + \gamma_{k+1}^T \beta_k^T P^{k\,T} A P^j \\ &= \gamma_{k+1}^T R^{k+1\,T} M (R^j - R^{j+1}) \alpha_j^{-1} = 0. \end{aligned}$$

This completes the proof of the lemma. □

Remark. If the matrices R^k and P^k have full rank, then the matrices

$$\alpha_k = (P^{k\,T} A P^k)^{-1} \gamma_k^T (R^{k\,T} M R^k), \tag{2.12}$$

$$\beta_k = \gamma_k^{-1} (R^{k\,T} M R^k)^{-1} (R^{k+1\,T} M R^{k+1}) \tag{2.13}$$

satisfy conditions (a)–(c) and in this case the BPCG method takes the form of the BCG method from [5]. Unfortunately, it is impossible to evaluate the coefficients α_k and β_k according to (2.12) and (2.13) if R^k and P^k lose full rank.

Taking into account the basic properties of quadratic functions we can establish the following minimization property of the BPCG method.

THEOREM 2.2. *R^{k+1} is orthogonal to*

$$\operatorname{span}\{MR^0, MAMR^0, \ldots, (MA)^k MR^0\}$$

and thus X^{k+1} minimizes $\operatorname{tr}[(X - X^*)^T A(X - X^*)]$ *over all X such that*

$$X - X^0 \in \operatorname{span}\{MR^0, MAMR^0, \ldots, (MA)^k MR^0\},$$

where $X^* = A^{-1}B$.

Theorem 2.2 establishes the finite termination property of the BPCG method; however, in practice we exploit this method as an iterative process rather than a finite termination method. O'Leary derived in [5] the following upper estimate of the convergence rate of the original version of the BPCG method. After k iterations of the block CG method the error in the mth component $e^k_{.m} = x^*_{.m} - x^k_{.m}$ satisfies

$$e^{k\,T}_{.m} A e^k_{.m} \le c \cdot \left(\frac{1 - \sqrt{\kappa^{-1}}}{1 + \sqrt{\kappa^{-1}}}\right)^{2k}, \quad 1 \le m \le s, \tag{2.14}$$

where $\kappa = \lambda_n/\lambda_s$ and $\lambda_n \ge \lambda_{n-1} \ge \cdots \ge \lambda_s \ge \cdots \ge \lambda_1$ are the eigenvalues of the preconditioned matrix MA and c is a constant dependent on m but independent of k. This estimate shows that an increase of the block size not only increases parallelism but also may lead to an essentially faster convergence. There does not exist a reasonable approach for choosing a block size that determines the required compromise between the parallelism, the resulting convergence rate, and arithmetic costs of one block iteration rapidly increasing with the block size. Thus when minimizing the total parallel solution time, the increase of the convergence rate and the improvement of efficiency of the parallel implementation of one block iteration may not compensate its large arithmetic costs. Nevertheless, such a compromise can be found by an adaptive reduction of the block size during the iterative process.

The original version of the BPCG method by O'Leary exploits the choice of γ_k computed using the QR or the modified Gram–Schmidt methods to monitor the rank deficiency of P^k. If a rank deficiency is detected, then "we delete the zero or redundant column j of P^k and the corresponding columns of X^k and R^k, and continue the algorithm with $s-1$ vectors" [5]. Let us consider in more detail the case of a rank deficiency of R^k and P^k and describe precisely a process for reducing the block size of the BPCG method.

First of all we note that $\operatorname{rank}(R^k) = \operatorname{rank}(P^k)$ and therefore rank deficiency of R^k leads to the loss of full rank in P^k. We can easily establish the following lemma which allows us to discover how the convergence of the BPCG method forces the loss of full rank of the residual matrix.

LEMMA 2.3. *Let two linear systems of equations*

$$AX = B \quad \text{and} \quad A\hat{X} = \hat{B},$$

where $A \in \mathcal{R}^{n\times n}$ is an SPD *matrix, and $B, \hat{B}, X, \hat{X} \in \mathcal{R}^{n\times s} (s < n)$ be given.*

Let $\hat{B} = B\delta$ and $\hat{X}^0 = X^0\delta$, where $\delta \in \mathcal{R}^{s\times s}$ is nonsingular, and suppose that the BPCG *method applied to these systems with α_i and β_i computed according to* (2.12) *and* (2.13) *does not fail after k steps.*

Then the BPCG *sequences of block approximations and block residuals satisfy the following equalities*

$$\hat{R}^k = R^k\delta \quad \textit{and} \quad \hat{X}^k = X^k\delta.$$

Now we suppose that after k iterations the columns of R^k are linearly dependent, i.e., there exists a nonsingular matrix δ such that

$$\hat{R}^k = R^k\delta = (R^k_* \;\; 0), \tag{2.15}$$

where the column submatrix R^k_* has full rank. If we similarly partition the modified right-hand side $\hat{B}$ and the approximation matrix $\hat{X}^k$, we have

$$\hat{B} = B\delta = (B_* \;\; B^*) \qquad \text{and} \qquad \hat{X}^k = X^k\delta = (X^k_* \;\; X^{*k}), \tag{2.16}$$

and the following matrix equality is valid:

$$AX^{*k} = B^*.$$

In other words, there exists a nonzero matrix $B^* \in \mathcal{R}^{n\times d}$ $(d < s)$ such that $\operatorname{span}(B^*) \subset \operatorname{span}(B)$ and for the corresponding matrix X^{*0}

$$\operatorname{span}(A^{-1}B^* - X^{*0}) \subset \operatorname{span}\{MR^0, MAMR^0, \ldots, (MA)^kMR^0\}.$$

Furthermore, it can be easily seen that

$$P^{k-1\,T}AM\hat{R}^k = (P^{k-1\,T}AMR^k_* \;\; 0),$$

and if γ_k is a block 2×2 diagonal matrix of the corresponding block partitioning

$$\gamma_k = \text{Block Diag}(\gamma_{*k}, \gamma^k_*),$$

then we have

$$P^k = (M\hat{R}^k + P^{k-1}\hat{\beta}_{k-1})\gamma_k = (P^k_* \;\; 0),$$

where

$$P^k_* = (MR^k_* - P^{k-1}(P^{k-1\,T}AP^{k-1})^{-1}(P^{k-1\,T}AMR^k_*))\gamma_{*k}. \tag{2.17}$$

Thus for $i \geq k$ the approximation matrix X^{i+1} can be written in the form

$$\hat{X}^{i+1} = (X^{i+1}_* \;\; X^{*k}), \tag{2.18}$$

where X^{i+1}_* are constructed as follows:

$$X^{i+1}_* = X^i_* + P^i_*\alpha_{*i}, \tag{2.19}$$

$$R^{i+1}_* = R^i_* - AP^i_*\alpha_{*i}, \tag{2.20}$$

$$P^{i+1}_* = (MR^{i+1}_* + P^i_*\beta_{*i})\gamma_{*i+1}, \tag{2.21}$$

$$\alpha_{*i} = (P^{i\,T}_*AP^i_*)^{-1}\gamma^T_{*i}(R^{i\,T}_*MR^i_*), \tag{2.22}$$

$$\beta_{*i} = \gamma^{-1}_{*i}(R^{i\,T}_*MR^i_*)^{-1}(R^{i+1\,T}_*MR^{i+1}_*), \tag{2.23}$$

provided the matrices R^i_* and P^i_* preserve full rank.

In the case of a rank deficiency of R_*^j, $j > k$, we repeat process (2.15)–(2.23) with the residual and the approximation matrices of the form

$$R^j = (R_*^j \ 0)\delta^{-1} \qquad \text{and} \qquad X^j = (X_*^j \ X^{*k})\delta^{-1}.$$

The above analysis shows that the block residual R_k may lose full rank only in the case of the algebraic convergence of a vector component of the modified block residual $\hat{R}_k$. Therefore, one can reduce adaptively the block size only at the final iterations if the convergence rate is to be preserved, but in this case the corresponding reduction of the arithmetic costs will be small. Thus in order to find a constructive compromise between the convergence rate of block iterations and their arithmetic costs we must try to reduce the block size independently of the rank of block iterates.

3. VBPCG algorithm. In this section we describe the block generalization of the CG method for the iterative solution of systems of linear equations $Ax = b$, where a reduction of the block size can be performed at each iteration independently of rank deficiency of R^k. The resulting method will be called the variable block preconditioned conjugate gradient (VBPCG) method.

THE VBPCG METHOD

Given an initial guess x^0 and an initial block size $s(0)$. Construct a right-hand side matrix $B \in \mathcal{R}^{n\times s(0)}$ and an initial guess matrix $X^0 \in \mathcal{R}^{n\times s(0)}$ whose first columns coincide with b and x^0, respectively, while other columns are chosen arbitrarily to produce a full rank matrix $R^0 = B - AX^0$.

Initial stage: Set $R^0 = B - AX^0$ and $P^0 = MR^0$, where M is an SPD preconditioner.

For $k = 0, 1, \ldots$ *iterate:*

1. $\tilde{R}^{k+1} = R^k - AP^k\alpha_k$ *and* $\tilde{X}^{k+1} = X^k + P^k\alpha_k$,

where $\alpha_k \in \mathcal{R}^{s(k)\times s(k)}$ are determined so that

$$\tilde{R}^{k+1\,T}P^k = 0.$$

2. Choose with respect to some criterion a positive integer $s(k+1) \le s(k)$ and a matrix $\epsilon_k \in \mathcal{R}^{s(k)\times s(k+1)}$ so that

$$\operatorname{rank}(\tilde{R}^{k+1}\epsilon_k) = s(k+1).$$

3. Set

$$\begin{aligned} R^{k+1} &= \tilde{R}^{k+1}\epsilon_k, \\ X^{k+1} &= \tilde{X}^{k+1}\epsilon_k, \\ \tilde{P}^k &= P^k\alpha_k\epsilon_k. \end{aligned}$$

4. If $d(k) = s(k) - s(k+1) > 0$ then choose $H^k \in \mathcal{R}^{n\times d(k)}$ so that

$$\begin{aligned} &\text{(i)} \ \operatorname{span}(P^k) = \operatorname{span}(H^k) \oplus \operatorname{span}(\tilde{P}^k), \\ &\text{(ii)} \ \tilde{P}^{k\,T}AH^k = 0 \qquad \text{and} \qquad H^{k\,T}AH^k = I. \end{aligned}$$

5. Update the block direction P^{k+1}

$$P^{k+1} = MR^{k+1} + \tilde{P}^k\beta_k + \sum_{i=0}^{k} H^i\gamma_k^i,$$

where the coefficients β_k and γ_k^i are computed so that

$$P^{k+1\,T} A\tilde{P}^k = 0 \quad \text{and} \quad P^{k+1\,T} AH^i = 0 \quad for\ i \le k.$$

In the following theorem we establish VBPCG properties similar to the orthogonal and the conjugate properties of the BPCG iterates.

THEOREM 3.1. *For $j < k$ the* VBPCG *method iterates satisfy the following conditions:*

$$\begin{aligned} R^{k\,T} MR^j &= 0, \\ P^{k\,T} AP^j &= 0, \\ R^{k\,T} P^j &= 0. \end{aligned}$$

Proof. We prove the statements of the theorem by induction. The statements of Theorem 3.1 are obviously valid for $k = 1$.

Assume that the statements hold true for some $k > 1$. Then by construction for $i \le k$ we have

$$P^{k+1\,T} A\tilde{P}^k = 0, \quad P^{k+1\,T} AH^i = 0, \quad P^{k+1\,T} AP^k = 0.$$

The block residual vector R^{k+1} satisfies the relation

$$R^{k+1\,T} P^k = \epsilon_k^T \tilde{R}^{k+1\,T} P^k = 0.$$

Furthermore, for $j < k$ we have

$$\begin{aligned} R^{k+1\,T} MR^j &= \epsilon_k^T \tilde{R}^{k+1\,T} MR^j = \epsilon_k^T (R^k - AP^k \alpha_k)^T MR^j \\ &= \epsilon_k^T R^{k\,T} MR^j + \epsilon_k^T \alpha_k^T P^{k\,T} A \left(P^j - \tilde{P}^{j-1} \beta_{j-1} - \sum_{i=0}^{j-1} H^i \gamma_{j-1}^i \right). \end{aligned} \tag{3.1}$$

The first term in (3.1) is equal to zero by the induction hypothesis while the second one is equal to zero since

$$\operatorname{span}(H^i) \subset \operatorname{span}(P^i) \quad \text{and} \quad \operatorname{span}(\tilde{P}^{j-1}) \subset \operatorname{span}(P^{j-1}).$$

Note that $P^j \subset \operatorname{span}\{MR^j, MR^{j-1}, \ldots, MR^0\}$ and hence

$$R^{k+1\,T} P^j = 0 \ \ for\ j < k,$$

and

$$R^{k+1\,T} MR^k = R^{k+1\,T} \left(P^k - \tilde{P}^{k-1} \beta_{k-1} - \sum_{i=0}^{k-1} H^i \gamma_{k-1}^i \right) = 0.$$

Finally, we prove that $P^{k+1\,T} AP^j = 0$ for $j < k$. To this end it is sufficient to show that

$$P^{k+1\,T} A\tilde{P}^j = 0.$$

Indeed,

$$P^{k+1\,T}A\tilde{P}^j = (MR^{k+1} + \tilde{P}^k\beta_k + \sum_{i=0}^{k} H^i\gamma_k^i)^T A\tilde{P}^j = R^{k+1\,T}MA\tilde{P}^j$$
$$= R^{k+1\,T}MAP^j\alpha_j\epsilon_j = R^{k+1\,T}M(R^j\epsilon_j - R^{j+1}) = 0.$$

The theorem is thus proved. □

In order to derive formulas for evaluating the matrix coefficients α_k, β_k, and γ_k we note that $\operatorname{rank}(P^k) = s(k)$ since

$$R^{k\,T}P^k = R^{k\,T}\left(MR^k + \tilde{P}^{k-1}\beta_{k-1} + \sum_{i=0}^{k-1} H^i\gamma_{k-1}^i\right) = R^{k\,T}MR^k,$$

and by construction $\operatorname{rank}(R^k) = s(k)$. Therefore, step 5 of the VBPCG method is well defined, i.e., $P^k = 0$ if and only if $R^k = 0$, and the matrix $(P^{k\,T}AP^k)$ is nonsingular. By construction

$$\operatorname{span}(P^k) = \operatorname{span}(H^k) \oplus \operatorname{span}(\tilde{P}^k)$$

and we can rewrite step 5 of the VBPCG method in the form

$$P^{k+1} = MR^{k+1} + P^k\tilde{\beta}_k + \sum_{i=0}^{k-1} H^i\gamma_k^i,$$

where the coefficients $\tilde{\beta}_k$ and γ_k^i are evaluated in such a way that

$$P^{k+1\,T}AP^k = 0 \quad \text{and} \quad P^{k+1\,T}AH^i = 0 \quad \text{for } i < k.$$

Then we can derive the following formulas to compute α_k, $\tilde{\beta}_k$, and γ_k^i.

COROLLARY 3.2. *It holds that*

$$\alpha_k = (P^{k\,T}AP^k)^{-1}(R^{k\,T}MR^k), \tag{3.2}$$

$$\tilde{\beta}_k = (R^{k\,T}MR^k)^{-1}(\tilde{R}^{k+1\,T}MR^{k+1}), \tag{3.3}$$

$$\gamma_k^i = -(H^{i\,T}AMR^{k+1}) \quad \textit{for } i < k. \tag{3.4}$$

Proof. α_k must satisfy the equation $(\tilde{R}^{k+1\,T}P^k) = 0$; hence

$$\alpha_k = (P^{k\,T}AP^k)^{-1}(P^{k\,T}R^k),$$

and relation (3.2) follows from the equality $R^{k\,T}P^k = R^{k\,T}MR^k$.

Let us demonstrate that $\tilde{\beta}_k$ from (3.3) and γ_k^i from (3.4) satisfy the conditions

$$P^{k+1\,T}AP^k = 0 \quad \text{and} \quad P^{k+1\,T}AH^i = 0 \ \text{ for } i < k.$$

Indeed,

$$P^{k+1\,T}AP^k = \left(MR^{k+1} + P^k\tilde{\beta}_k + \sum_{i=0}^{k-1} H^i\gamma_k^i\right)^T AP^k$$
$$= R^{k+1\,T}MAP^k + \tilde{\beta}_k^T P^{k\,T}AP^k = R^{k+1\,T}M(R^k - \tilde{R}^{k+1})\alpha_k^{-1}$$
$$+ \ (\tilde{R}^{k+1\,T}MR^{k+1})^T(R^{k\,T}MR^k)^{-1}(P^{k\,T}AP^k)$$
$$= (R^{k+1\,T}M\tilde{R}^{k+1})(-\alpha_k^{-1} + \alpha_k^{-1}) = 0$$

and

$$P^{k+1\,T} A H^i = R^{k+1\,T} M A H^i - \gamma_k^{i\,T} I = 0.$$

In the last chain of equalities we used the condition

$$(H^{i\,T} A H^i) = I.$$

The corollary is thus proved. □

The whole process can be divided into three stages:

(i) an initial stage, which coincides with the standard BCG scheme with the initial block size;

(ii) a reduction stage, when we consecutively reduce the block size of iterates;

(iii) a final stage, which coincides with the standard BCG scheme with the final block size, except of the additional reprojection of the block directions on the subspace A-orthogonal to $\mathrm{span}\{H^0, H^1, \ldots, H^k\}$.

The arithmetic costs of the reduction stage are dependent on a particular criterion for choosing the matrices ϵ_k and H^k, but, in general, computation of these matrices does not require any additional preconditioned matrix-vector multiplications. Since the number of iterations at this stage is relatively small, their arithmetic costs cannot affect considerably the total arithmetic complexity of the iterative solution. If we denote the final block size by s, the reprojection of the block direction requires $2\,(s(0)-s)\,s\,n$ additions and multiplications. Thus the arithmetic costs of one VBPCG iteration at the final stage have only a small arithmetic overhead as compared with one BPCG iteration of the block size s.

Now we describe relationship between the subspaces generated at k iterations of the VBPCG method.

DEFINITION 3.3. *An $n \times n$ matrix Q is a conjugate projector with respect to A on a subspace $\mathcal{V}$ iff $\mathcal{V}$ is the* span *of columns Q, $Q^2 = Q$, and $Q^T A(I - Q) = 0$.*

LEMMA 3.4. *Let $\bar{Q}_k$ denote the conjugate projector on*

$$\mathcal{H}_k = \mathrm{span}\{H^0, H^1, \ldots, H^k\}.$$

If $\mathcal{P}_k = \mathrm{span}\{P^0, P^1, \ldots, P^k\}$ and $Q_k = I - \bar{Q}_k$, then

$$\mathcal{P}_k = \mathrm{span}\{Q_k M R^0, Q_k M A\, Q_k M R^0, \ldots, (Q_k M A)^k\, Q_k M R^0\} \oplus \mathcal{H}_k.$$

Proof. From step 5 of the VBPCG method we have for all $j < k$

$$Q_k P^{j+1} = Q_k M R^{j+1} + Q_k \tilde{P}^j \beta_j + \sum_{i=0}^{j} Q_k H^i \gamma_k^i.$$

Since $\mathrm{span}(\tilde{P}^i) = \mathrm{span}(Q_k P^i)$ and $Q_k H^i = 0$ for $i \le k$, this equality implies that

$$\mathrm{span}\{\tilde{P}^0, \tilde{P}^1, \ldots, \tilde{P}^k\} = \mathrm{span}\{Q_k M R^0, Q_k M R^1, \ldots, Q_k M R^k\}.$$

Now, from step 1 and step 2 we have for all $j < k$

$$Q_k M R^{j+1} = Q_k M R^j \epsilon_j - Q_k M A \tilde{P}^j.$$

Therefore, we have

$$\mathrm{span}\{\tilde{P}^0, \tilde{P}^1, \ldots, \tilde{P}^k\} = \mathrm{span}\{Q_k M R^0, Q_k M A\, Q_k M R^0, \ldots, (Q_k M A)^k\, Q_k M R^0\}$$

and the statement of Lemma 3.4 follows immediately from the fact that

$$\operatorname{span}(P^k) = \operatorname{span}(H^k) \oplus \operatorname{span}(\tilde{P}^k). \qquad \square$$

It is of primary importance to establish relations between the block vector X^k computed at the kth iteration of the VBPCG algorithm and the solution to the original system $x^* = A^{-1}b$. Let $E \in \mathcal{R}^{s\times d}$ have full rank and let X'' minimize the following quadratic functional

$$F(X) = \operatorname{tr}[(X - X^*)^T A(X - X^*)],$$

over all X which lie on the variety $\Pi : X' + \mathcal{P}$, where $\mathcal{P}$ is a linear subspace and $X' \in \mathcal{R}^{n\times s}$. If $\hat{X}^* = X^*E$, then $\hat{X}'' = X''E$ minimize the quadratic functional

$$\hat{F}(\hat{X}) = \operatorname{tr}[(\hat{X} - \hat{X}^*)^T A(\hat{X} - \hat{X}^*)],$$

over all $\hat{X}$ which lie on the variety $\hat{\Pi} : \hat{X}' + \mathcal{P}$, where $\hat{X}' = X'E$.

Indeed, if P is a basis of the linear subspace $\mathcal{P}$ then

$$\hat{X}'' = \hat{X}' + P\hat{\alpha}',$$

where

$$\hat{\alpha}' = -(P^T AP)^{-1}\,(P^T A(\hat{X}' - \hat{X}^*)).$$

Hence, we have

$$\hat{X}'' = X'E - P\,(P^T AP)^{-1}\,(P^T A(X' - X^*))E = X''E.$$

We have thus proved the following theorem.

THEOREM 3.5. *Let $\bar{Q}_k$ denote the conjugate projector on*

$$\mathcal{H}_k = \operatorname{span}\{H^0, H^1, \ldots, H^k\}$$

$Q_k = I - \bar{Q}_k$ and $E = \epsilon_0\epsilon_1 \ldots \epsilon_k$.

Then X^{k+1} from the VBPCG *method minimizes*

$$\operatorname{tr}[(X - \hat{X}^*)^T A(X - \hat{X}^*)]$$

over all $X \in \mathcal{R}^{n\times s(k+1)}$ such that

$$X - \hat{X}^0 \in \operatorname{span}\{Q_k MR^0, Q_k MA\,Q_k MR^0, \ldots, (Q_k MA)^k\,Q_k MR^0\} \oplus \mathcal{H}_k,$$

where $\hat{X}^* = A^{-1}BE$ and $\hat{X}^0 = X^0E$.

COROLLARY 3.6. *If the first column of $E = \epsilon_0\epsilon_1 \ldots \epsilon_k$ coincides with the first unit vector, then the first column of X^{k+1} provides the optimal approximation to $x^* = A^{-1}b$ with respect to $\mathcal{F}(x)$ and to the constructed subspace $\mathcal{P}_k$.*

In what follows we assume that the condition of Corollary 3.6 is fulfilled.

Now we present estimates of the asymptotic convergence rate of the VBPCG method. Without loss of generality we shall consider the case $M = I$. In Theorem 3.5 we established the close relationship between minimization properties of the VBPCG method and the CG method with preconditioning by the projector (the CGP method) from [2]. The CGP method for solving the system of linear equations $Ax = b$ can be described as follows.

THE CGP METHOD

Define a linear subspace $\mathcal{H}$ and a full rank matrix $H \in \mathcal{R}^{n\times m}$ $(n > m)$ so that $\mathcal{H} = \operatorname{span}(H)$. Set

$$P = H(H^TAH)^{-1}H^TA \quad \textit{and} \quad Q = I - P,$$

and denote by $\mathcal{V}$ the subspace $\operatorname{span}(Q)$.

Initial stage: Set

$$x_0 = PA^{-1}b = H(H^TAH)^{-1}H^Tb \quad \textit{and} \quad p_0 = r_0 = b - Ax_0.$$

For $i = 0, 1, \ldots$ *iterate*:

$$\begin{aligned} q_i &= Qp_i, \\ \alpha_i &= r_i^Tq_i \,/ q_i^TAq_i, \\ x_{i+1} &= x_i + \alpha_i q_i, \\ r_{i+1} &= r_i - \alpha_i Aq_i, \\ \beta_i &= r_{i+1}^TAq_i \,/ q_i^TAq_i, \\ p_{i+1} &= r_{i+1} - \beta_i p_i. \end{aligned}$$

It can be shown [2] that the approximation x_k constructed at the kth iteration of the CGP method with a conjugate projector Q minimizes functional $\mathcal{F}(u)$ (2.1) over all

$$u \in \operatorname{span}\{Qr_0, QAQr_0, \ldots, (QA)^{k-1}Qr_0\} \oplus \mathcal{H},$$

where $\mathcal{H}$ is the conjugate complement to the subspace $\mathcal{V} = \operatorname{span}(Q)$ and $r_0 \in A\mathcal{V}$ is an initial residual. Thus the asymptotic convergence rate of the CGP method can be estimated in terms of the spectral condition number of the positive definite restriction of Q^TAQ to $A\mathcal{V}$ denoted by $Q^TAQ \mid A\mathcal{V}$ [2]. Let us denote by $\mathcal{L}$ the subspace spanned by the eigenvectors corresponding to $\lambda_1, \ldots, \lambda_m$ and by $\mathcal{Q}_{A\mathcal{H}}$ and $\mathcal{Q}_{\mathcal{L}}$ the orthogonal projector on $A\mathcal{H}$ and $\mathcal{L}$, respectively. If

$$\gamma = \|\mathcal{Q}_{A\mathcal{H}} - \mathcal{Q}_{\mathcal{L}}\|_2 \tag{3.5}$$

is the gap between $A\mathcal{H}$ and $\mathcal{L}$ then the spectral condition number of $Q^TAQ \mid A\mathcal{V}$ satisfies the inequality (see [2])

$$\kappa(Q^TAQ \mid A\mathcal{V}) \le \frac{\lambda_n}{\sqrt{(1-\gamma)^2\lambda_{m+1}^2 + \gamma^2\lambda_1^2}}. \tag{3.6}$$

Now taking into account Theorem 3.5 we can exploit estimate (3.6) to analyze the convergence rate of the VBCG method. If after k iterations of the VBCG method the dimension of the constructed subspace $\mathcal{H} = \mathcal{H}_k$ is equal to m, then the asymptotic convergence of the VBCG method exceeds the convergence of the CGP method with projector $Q = Q_k$ since

$$\operatorname{span}\{Q_kr_0, Q_kAQ_kr_0, \ldots, (Q_kA)^{k-1}Q_kr_0\} \subset \mathcal{P}_k.$$

On the other hand, if $\dim(\mathcal{H}_k) = 0$, then the VBCG method coincides with the BCG method and its convergence rate can be estimated in terms of the reduced condition number of A

$$\kappa = \lambda_n/\lambda_{s(0)}.$$

Therefore, with an appropriate choice of H^i (which is equivalent to the choice of ϵ_i) providing a reasonably small value of γ in (3.5) we can make the convergence of the VBPCG method with rapidly decreasing arithmetic costs per iteration as fast as the convergence of the standard BPCG method with constant block size.

The practical approach to an a posteriori qualitative analysis of the preconditioned CG iterations is closely related to computation of the Ritz values by construction of the tridiagonal spectrally equivalent matrix made up with the CG coefficients. This analysis seems to be very helpful when investigating and optimizing the performance of the VBPCG method. To this end we construct in this section a similar block tridiagonal matrix for investigating the spectral characteristics of the VBPCG method (for the sake of simplicity we consider only the unpreconditioned VBCG method).

Let for some $N > 0$ the matrix R^{N+1} be zero, then we define the following matrices:

$$R = (R^0\ R^1\ \dots\ R^N), \quad \tilde{P} = (\tilde{P}^0\ \tilde{P}^1\ \dots\ \tilde{P}^N), \quad \text{and} \quad H = (H^0\ H^1\ \dots\ H^N).$$

In our notation we have the equalities

$$R = \tilde{P}U + H\Gamma \quad \text{and} \quad A\tilde{P} = RL,$$

where $\Gamma = (H^T A R)$ is a upper block triangular matrix by construction, while L and U are lower and upper block bidiagonal matrices, respectively. Hence, we can write

$$AR = RT + (AHH^T A)R$$

or

$$A(I - HH^T A)R = RT,$$

where $T = LU$ is a block tridiagonal matrix. Let $\eta_k^T \eta_k = (R^{k\,T} R^k)$ and

$$D = \text{Block Diag}(\eta_0^{-1}, \eta_1^{-1}, \dots, \eta_N^{-1}),$$

then the columns of the matrix $\bar{R} = RD$ are orthonormal and hence

$$A(I - HH^T A)\bar{R} = \bar{R}D^{-1}TD = \bar{R}\bar{T},$$

where $\bar{T} = D^{-1}TD$.

We have thus proved the following proposition.

PROPOSITION 3.7. *The* VBPCG *method constructs the orthonormal basis in which the matrix of the operator A restricted to the A-orthogonal complement to* span(H) *has block tridiagonal form.*

We consider now the tridiagonal matrix T in more detail. The bidiagonal matrices L and U can be represented in the form

$$L = \begin{bmatrix} \epsilon_0 & & & \\ -I & \epsilon_1 & & \\ & -I & \cdot & \\ & & \cdot & \cdot \\ & & & -I\ \ \epsilon_N \end{bmatrix}, \quad U = \begin{bmatrix} \xi_0 & -\beta_0 & & & \\ & \xi_1 & -\beta_1 & & \\ & & \cdot & & \\ & & & \cdot & -\beta_{N-1} \\ & & & & \xi_N \end{bmatrix},$$

where

$$\xi_i = (\tilde{P}^{i\,T} A \tilde{P}^i)^{-1} (\tilde{P}^{i\,T} A R^i). \tag{3.7}$$

Therefore, the matrix $T = LU$ is of the form

$$T = \begin{bmatrix} \epsilon_0\xi_0 & -\epsilon_0\beta_0 & & & & \\ -\xi_0 & \epsilon_1\xi_1 + \beta_0 & -\epsilon_1\beta_1 & & & \\ & -\xi_1 & \epsilon_2\xi_2 + \beta_1 & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & -\epsilon_{N-1}\beta_{N-1} \\ & & & & -\xi_{N-1} & \epsilon_N\xi_N + \beta_{N-1} \end{bmatrix}.$$

Without loss of generality we can assume that ϵ_N is the identity matrix. If we denote $(\tilde{P}^{i\,T} A \tilde{P}^i)^{-1}$ by ν_i then

$$\xi_i = \nu_i \epsilon_i^T \eta_i^T \eta_i \quad \text{for } i = 0, 1, \ldots, N-1 \quad \text{and} \quad \xi_N = \alpha_N^{-1}. \tag{3.8}$$

Since the matrix $\bar{T} = D^{-1} L U D$ is symmetric, after some simplifications $\bar{T}$ can be written as

$$\bar{T} = \begin{bmatrix} \eta_0\epsilon_0\nu_0\epsilon_0^T\eta_0^T & -\eta_0\epsilon_0\nu_0\eta_1^T & & & \\ -\eta_1\nu_0\epsilon_0^T\eta_0^T & \eta_1\epsilon_1\nu_1\epsilon_1^T\eta_1^T + \eta_1\nu_0\eta_1^T & -\eta_1\epsilon_1\nu_1\eta_2^T & & \\ & -\eta_2^T\nu_1\epsilon_1^T\eta_1 & \eta_2\epsilon_2\nu_2\epsilon_2^T\eta_2^T + \eta_2\nu_1\eta_2^T & \cdot & \\ & & \cdot & \cdot & \cdot \\ & & & \cdot & \bar{T}_{NN} \end{bmatrix},$$

where $\bar{T}_{NN} = \eta_N^{-T}(P^{N\,T} A P^N)\eta_N^{-1} + \eta_N \nu_{N-1} \eta_N^T$.

Note that due to the condition $\operatorname{rank}(R^k) = s(k)$ the matrices ϵ_k have full rank for all $k > 0$. Therefore, equalities (3.7)–(3.8) imply that the matrices U and L also have full rank and the nonzero eigenvalues of the matrix $(A - AHH^TA)$ coincide with the eigenvalues of the matrix T.

4. Model problem and test matrices. In this section we present the results of numerical experiments of the VBPCG method when solving three-dimensional equilibrium equations for linear elastic orthotropic materials approximated both by h- and p-versions of the finite element method and we discuss some realization aspects of the VBPCG algorithm. We are mainly interested in demonstrating the potential capabilities of this approach, so we only consider one VBPCG method in this paper. Comparison of different versions of the VBPCG method and problems related to construction of an optimal VBPCG algorithm will be treated in a forthcoming paper.

4.1. The model problem (see [3] for more detail). Consider a bounded domain Ω in $\mathcal{R}^3$ with the boundary $\partial\Omega = \partial\Omega_U \cup \partial\Omega_T$. As a model problem we use in this paper the three-dimensional equilibrium equations

$$\sigma_{ij,j} + F_i = 0 \tag{4.1}$$

for an orthotropic elastic material, where σ_{ij} is the stress tensor and F_i is the body force.

The constitutive relation for linear elasticity is the generalized Hooke's law

$$\sigma_{ij} = E_{ijkl}\varepsilon_{kl}, \tag{4.2}$$

which is the most general linear relationship between the stress tensor and the vector of small strains with the components $\varepsilon_{kl}(\overline{u}) = 1/2(\partial u_k/\partial x_l + \partial u_l/\partial x_k)$, where $\overline{u} = (u_1, u_2, u_3)$ is the unknown displacement vector. The coefficients E_{ijkl} are the components of the fourth order elasticity tensor. Boundary conditions are given by prescribed surface tractions

$$T_i = \sigma_{ij} n_j \quad \text{on} \quad \partial\Omega_T, \tag{4.3}$$

where n_j is the outward normal to $\partial\Omega_T$ and $\overline{T} = (T_1, T_2, T_3)$, and the homogeneous displacement boundary conditions

$$u_i = 0 \quad \text{on} \quad \partial\Omega_U \ (\text{mes}(\partial\Omega_U) \neq 0). \tag{4.4}$$

A weak solution to problem (4.1)–(4.4) is obtained by solving the following variational problem [1]: find $\overline{u}$, $u_i \in W_2^1$, $i = 1, 2, 3$, where W_2^1 is the Sobolev space, which satisfies kinematic boundary conditions (4.4) such that

$$a(\overline{u}, \overline{v}) = b(\overline{v}), \tag{4.5}$$

where $\overline{v} = (v_1, v_2, v_3), \overline{v} = 0$ on $\partial\Omega_U$, and $v_i \in W_2^1$, $i = 1, 2, 3$, and

$$a(\overline{u}, \overline{v}) = \int_\Omega E_{ijkl}\varepsilon_{ij}(\overline{u})\varepsilon_{kl}(\overline{v})d\Omega, \tag{4.6}$$

$$b(\overline{v}) = \int_\Omega \overline{F} \cdot \overline{v} d\Omega + \int_{\partial\Omega_T} \overline{T} \cdot \overline{v} \, d\partial\Omega. \tag{4.7}$$

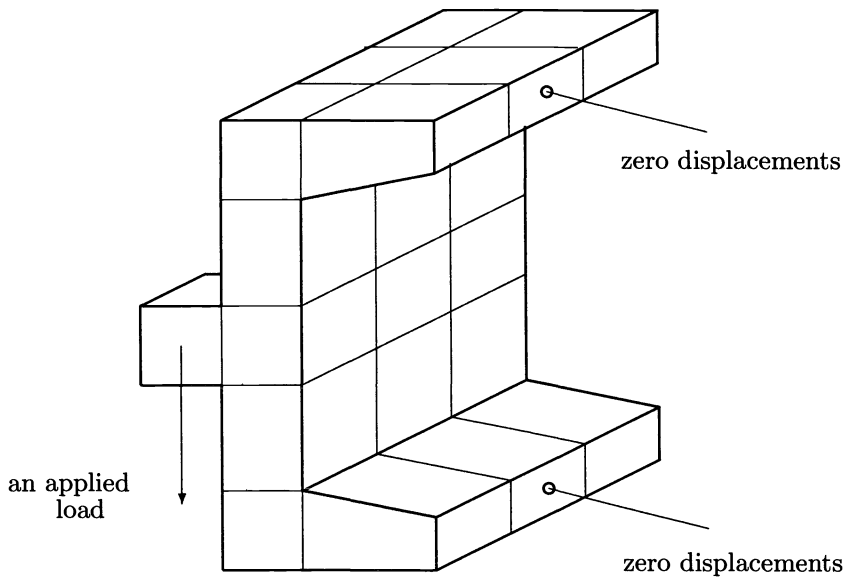


FIG. 4.1. *The domain Ω and the boundary conditions of the channel model problem.*

For our model problem of computing the displacements of a channel, the domain Ω, homogeneous condition (4.4) and the applied load $\overline{F}$ are shown in Fig. 4.1.

The variational problem (4.5)–(4.7) is approximated using the p-version of the conforming FE mesh with the parametric brick-shaped finite elements and the hierarchical shape functions $H_{ijk} = l_i(x)l_j(y)l_k(z)$, where $l_0(t) = 0.5(1-t)$, $l_1(t) = 0.5(1+t)$, $l_m(t) = P_m(t) - P_{m-1}(t)$, $m \geq 2$, and $P_m(t)$ is the Legendre polynomial of degree m.

We thus consider the following FE approximation of variational problem (4.5)–(4.7): find $\overline{u}$ with $u_i \in \mathcal{S}, i = 1, 2, 3$, such that

$$\begin{aligned} &\int_{\Omega} E_{ijkl}\varepsilon_{ij}(\overline{u})\varepsilon_{kl}(\overline{v})d\Omega \\ &= \int_{\partial\Omega_T} \overline{T}\cdot\overline{v}d\partial\Omega + \int_{\Omega} \overline{F}\cdot\overline{v}d\Omega \quad \forall\, \overline{v} \in \mathcal{S}^3, \end{aligned} \tag{4.8}$$

where $\mathcal{S}$ is the space of global basis functions, or, taking into account the symmetry of the elasticity tensor, such that

$$\begin{aligned} &\int_{\Omega} E_{ijkl}\frac{\partial u_i}{\partial x_j}\,\frac{\partial v_k}{\partial x_l}\,d\Omega \\ &= \int_{\partial\Omega_T} \overline{T}\cdot\overline{v}\;d\partial\Omega + \int_{\Omega} \overline{F}\cdot\overline{v}\;d\Omega \quad \forall\, \overline{v} \in \mathcal{S}^3. \end{aligned} \tag{4.9}$$

FE approximation (4.9) gives rise to the linear algebraic system

$$A^{ik}_{mn}u^k_n = f^i_m, \tag{4.10}$$

with the SPD coefficient matrix A, where

$$A^{ik}_{mn} = \int_{\Omega} E_{ijkl}\frac{\partial\phi_m}{\partial x_j}\frac{\partial\phi_n}{\partial x_l}d\Omega \quad \text{and} \quad f^i_m = \int_{\Omega} F_i\phi_m d\Omega + \int_{\partial\Omega_T} T_i\phi_m d\partial\Omega.$$

4.2. Numerical experiments with the VBPCG algorithm. We use the simplest criterion at step 4 of the VBPCG method when constricting the A-orthogonal projector, namely, after a prescribed number of iterations we reduce the number of columns of the block direction one by one.

In Table 4.1 we present the characteristics of four test coefficient matrices originated from $p = 2$ and $p = 3$ FE approximations of the anizotropic channel model problem (see Fig. 4.1) with the Poisson ratios $\nu_{xy} = 0.31$, $\nu_{xz} = 0.32$, and $\nu_{yz} = 0.33$ on two curvilinear meshes, respectively. This table adopts the following notation:

N denotes the size of a test matrix, while NZA denotes its number of nonzero entries.

For all test matrices we exploited the incomplete block symmetric successive over-relaxation (IBSSOR) preconditioning [3]. The test matrices are very ill-conditioned and even the IBSSOR preconditioned test matrices are still ill-conditioned. The spectral characteristics of the IBSSOR preconditioned test matrices are presented in Table 4.2. In this table λ_i denotes the ith eigenvalue of the block tridiagonal matrix made up with the coefficients of the BPCG method applied to solving a linear system with the corresponding test matrix.

Table 4.3 contains the results of numerical experiments with the IBSSOR-VBCG method and adopts the following notation:

TABLE 4.1
Characteristics of test matrices TM1–TM4.

Test Matrix	p	Mesh	N	NZA
TM1	2	M1	3834	473606
TM2	3	M1	6579	1277737
TM3	2	M2	7030	937762
TM4	3	M2	12112	2537096

TABLE 4.2
Spectral characteristics of IBSSOR *preconditioned test matrices* TM1–TM4.

	TM1	TM2	TM3	TM4
λ_1	0.169066e-6	0.160861e-6	0.108627e-6	0.103144e-6
λ_2	0.380e-5	0.368e-5	0.162e-5	0.133e-5
λ_3	0.642e-5	0.606e-5	0.302e-5	0.294e-5
λ_4	0.886e-5	0.731e-5	0.518e-5	0.493e-5
λ_5	0.910e-5	0.758e-5	0.879e-5	0.725e-5
λ_6	0.324e-4	0.258e-4	0.259e-4	0.230e-4
λ_7	0.438e-4	0.388e-4	0.318e-4	0.255e-4
λ_8	0.186e-3	0.172e-3	0.619e-4	0.503e-4
λ_9	0.190e-3	0.181e-3	0.170e-3	0.144e-3
λ_{10}	0.212e-3	0.231e-3	0.240e-3	0.186e-3
λ_{max}	1.055865	1.048247	1.074104	1.097654

BPCG(k) stands for the standard BPCG method with block size k;

VBPCG(k,s) denotes the VBPCG method, where k is the initial block size and s is the final block size;

λ_1 denotes the first Ritz value of the preconditioned matrix while Cond stands for its spectral condition number computed via Ritz values;

Iter denotes the number of preconditioned iterations required to reduce the relative preconditioned residual by a factor of 10^{-11}. A residual of this size is needed in order to obtain the desired level of accuracy in the computed solution. When high levels of accuracy are required, fixed-size BPCG methods typically require about the same number of matrix-vector multiplications as the single vector versions. MVM stands for the number of preconditioned matrix-vector multiplications required to satisfy the above stopping criterion.

Analyzing the data from Table 4.3 the following conclusions can be made.

1. The VBPCG method provides a real opportunity to substantially reduce the number of preconditioned matrix-vector multiplications as compared with the corresponding BPCG method and thus to decrease significantly the serial arithmetic costs of the iterative solution.

2. The VBPCG method can preserve parallelism while significantly reducing the serial arithmetic costs. It should be emphasized that the gain in the serial arithmetic costs is accompanied by only a slight increase in the number of iterations.

3. In some cases the VBPCG method even possesses better serial arithmetic complexity than the corresponding PCG method and hence can be an efficient serial algorithm.

TABLE 4.3
Convergence and spectral characteristics of VBPCG *methods for channel problem.*

Test Matrix	Method	Iter	MVM	λ_1	Cond
TM1	BPCG(1)	1085	1085	0.1690663e–6	6245272.71
	BPCG(10)	93	930	0.1690662e–6	6245278.36
	BPCG(15)	65	975	0.1690636e–6	6245375.80
	VBPCG(10,1)	417	687	0.4119750e–5	256293.52
	VBPCG(10,3)	194	799	0.3265010e–5	323388.10
	VBPCG(15,1)	335	678	0.3873469e–5	272580.47
	VBPCG(15,3)	159	735	0.4637082e–5	227685.02
TM2	BPCG(1)	968	968	0.1608613e–6	6516468.89
	BPCG(10)	105	1050	0.1608613e–6	6516471.09
	BPCG(15)	73	1095	0.1608596e–6	6516541.50
	VBPCG(10,1)	484	844	0.4224703e–5	248123.44
	VBPCG(10,3)	218	941	0.3713804e–5	282257.19
	VBPCG(15,1)	368	753	0.1411960e–4	74239.19
	VBPCG(15,3)	181	861	0.6701381e–5	156419.63
TM3	BPCG(1)	1069	1069	0.1086269e–6	9888013.49
	BPCG(10)	105	1050	0.1086272e–6	9887989.04
	BPCG(15)	73	1095	0.1086251e–6	9888183.01
	VBPCG(10,1)	432	785	0.2872412e–5	373938.30
	VBPCG(10,3)	218	941	0.2221500e–5	483504.16
	VBPCG(15,1)	365	722	0.4134046e-5	259819.04
	VBPCG(15,3)	190	864	0.3934473e–5	272998.22
TM4	BPCG(1)	1347	1347	0.1031435e–6	10642010.56
	BPCG(10)	133	1330	0.1031437e–6	10641990.08
	BPCG(15)	98	1470	0.1031434e–6	10642025.81
	VBPCG(10,1)	664	1141	0,3054796e–5	359321.78
	VBPCG(10,3)	312	1307	0.2845067e–5	385809.79
	VBPCG(15,1)	433	916	0.7444543e–5	147444.20
	VBPCG(15,3)	197	993	0.5628691e–5	195010.65

4. The VBPCG method provides a real opportunity to find a constructive compromise between opposing factors such as convergence rate, parallelism, and arithmetic costs of one block iteration. However, we should note that the performance of the VBPCG method is evidently dependent on spectral characteristics of the preconditioned matrix.

We would like also to note that two further parts of the paper are under preparation. One of them is dedicated to an analysis of efficient implementations of VBPCG algorithms on massively parallel computers while another will consider construction of efficient and reliable mathematical criteria for reducing adaptively the block size. We also hope to present in one of these parts results of numerical experiments with VBPCG algorithms on massively parallel computers.

Acknowledgments. The authors would like to thank Professor D. P. O'Leary for helpful discussions and several improvements and Professor A. Greenbaum for valuable remarks to the final version of this paper. The authors are grateful to Cray Research, Inc. (USA) for providing computer resources to make numerical experiments.

REFERENCES

[1] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1987.

[2] Z. DOSTÁL, *Conjugate gradient method with preconditioning by projector*, Internat J. Comput. Math., 32 (1988), pp. 315–323.

[3] L. YU. KOLOTILINA, I. E. KAPORIN, AND A. YU. YEREMIN, *Block* SSOR *preconditionings for high order* FE *systems* II: *Incomplete* BSSOR *preconditionings*, Linear Algebra Appl., 154–156 (1991), pp. 647–674.

[4] L. YU. KOLOTILINA AND A. YU. YEREMIN, *Factorized sparse approximate inverse preconditioning* II: *Solution of* 3D FE *systems on massively parallel computers*, Research Report EM-RR-3/92, Elegant Mathematics, Inc.(USA), 1992.

[5] D. P. O'LEARY, *The block conjugate gradient algorithm and related methods*, Linear Algebra Appl., 29 (1980), pp. 293–322.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, pp. 1154–1171, October 1995

009

A RESTARTED GMRES METHOD AUGMENTED WITH EIGENVECTORS *

RONALD B. MORGAN†

Abstract. The GMRES method for solving nonsymmetric linear equations is generally used with restarting to reduce storage and orthogonalization costs. Restarting slows down the convergence. However, it is possible to save some important information at the time of the restart. It is proposed that approximate eigenvectors corresponding to a few of the smallest eigenvalues be formed and added to the subspace for GMRES. The convergence can be much faster, and the minimum residual property is retained.

Key words. GMRES, conjugate gradient, Krylov subspaces, iterative methods, nonsymmetric systems

AMS subject classifications. 65F15, 15A18

1. Introduction.

The GMRES method [32] is popular for solving the large nonsymmetric system of linear equations

$$Ax = b. \tag{1}$$

But GMRES is generally used with restarting, and this slows down the convergence. We examine a way to retain some of the information lost at the time of the restart. The convergence can be improved in many situations. This section gives background material on GMRES. Section 2 gives the new method and analyzes its effectiveness for certain cases. Section 3 discusses the implementation and the expenses. Examples and comparisons are given in §4, and §5 looks at the possibility of having a procedure that selects the number of approximate eigenvectors and decides how long they should be used.

For symmetric problems, the conjugate gradient method [13], [17] is often the best iterative method. It extracts an approximate solution from the Krylov subspace $\text{Span}\{b, Ab, A^2b, \ldots, A^{m-1}b\}$. There is an efficient recurrence formula for generating a sequence of orthogonal vectors that span the Krylov subspace. Also the convergence properties are fairly well understood for a Krylov subspace. They depend on the eigenvalue distribution. A simple bound for the minimum residual version [15], [17], [28] of the conjugate gradient method applied to a symmetric positive definite matrix is

$$\begin{aligned} \frac{||r||}{||b||} &\leq 2\Big(\Big(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\Big)^m + \Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^m\Big)^{-1} \\ &\leq 2\Big(1-\frac{2}{\sqrt{\kappa}+1}\Big)^m, \end{aligned} \tag{2}$$

where r is the residual vector $b - A\hat{x}$, and $\hat{x}$ is the approximate solution. Also $\kappa \equiv \frac{\lambda_n}{\lambda_1}$ is the condition number, the ratio of largest to smallest eigenvalues. So convergence

*Received by the editors August 18, 1993; accepted for publication (in revised form) by A. Greenbaum December 3, 1994. This research was supported by National Science Foundation contract CCR-8910665.

†Department of Mathematics, Baylor University, P.O. Box 97328, Waco, Texas 76798-7328 (morgan@baylor.edu).

is generally slow if there is an extremely small eigenvalue. But the placement of the other eigenvalues also influences convergence. Clumping of eigenvalues is favorable. The actual convergence rate often improves as the method proceeds [4], [5], [7], [35]. This is because some of the outlying eigenvalues are effectively eliminated from the spectrum once the Krylov subspace contains a good approximation to the corresponding eigenvector. Another good thing about the conjugate gradient method is that the convergence can usually be improved by preconditioning (multiplying (1) by an approximate inverse to A) [3], [6], [13], [15], [23].

The conjugate gradient method can be generalized to nonsymmetric problems in several ways. The three main approaches are the nonsymmetric Lanczos algorithm [19], [20], [37], the conjugate gradient method applied to the normal equations (CGNE) [8], [16], and GMRES [32]. The nonsymmetric Lanczos method is similar to the conjugate gradient method in that it uses a Krylov subspace and has a recurrence formula. The algorithm is unstable, but improvements have been made [11], [12], [14], [18], [29], [36]. In particular, the QMR version [11], [12] has attracted attention. The CGNE method transforms to another problem (the normal equations), so the convergence properties are different. Often convergence is much slower. Nevertheless there are some problems, particularly indefinite and fairly nonsymmetric ones, for which CGNE is best [26]. GMRES is currently a popular method for large nonsymmetric problems (see, for example, [21], [27]). It uses the Arnoldi algorithm [1], [30], [31], [37] to build an orthonormal basis for the Krylov subspace, so full orthogonalization is needed. The best approximate solution is extracted from the subspace, in that the norm of the residual vector is minimized.

Because full orthogonalization is used, the method becomes more expensive as the subspace grows. Also importantly, the storage requirements increase. Restarting can be used when the subspace reaches a certain size.

RESTARTED GMRES

1. *Start:* Choose x_0 and compute $r_0 = b - Ax_0$ and $v_1 = r_0/||r_0||$.
2. *Iterate:* For $j = 1, 2, \ldots, m$ do:
   $h_{ij} = (Av_j, v_i), i = 1, 2, \ldots, j,$
   $\hat{v}_{j+1} = Av_j - \sum_{i=1}^{j} h_{ij}v_i,$
   $h_{j+1,j} = ||\hat{v}_{j+1}||$, and
   $v_{j+1} = \hat{v}_{j+1}/h_{j+1,j}$.
3. *Form the approximate solution:* $\hat{x} = x_0 + V\hat{d}$, where $\hat{d}$ minimizes $||\beta e_1 - \bar{H}d||$, for all $d \in R^m$. Here $\bar{H}$ is the $(m+1)$ by m matrix with elements h_{ij} defined in step 2, and $\beta = ||r_0||$.
4. *Restart:* Compute $r = b - A\hat{x}$; if satisfied then stop, else let $x_0 = \hat{x}$, $v_1 = r/||r||$, $r_0 = r$, and go to 2.

The convergence of GMRES is similar to that for the conjugate gradient method if the matrix is nearly normal. Again the presence of small eigenvalues slows convergence. Suppose A has spectral decomposition $A \equiv Z\Lambda Z^{-1}$, with all the eigenvalues being real and positive. Assuming that the initial guess x_0 is the zero vector, we have

$$\frac{||r||}{||b||} \leq 2||Z||||Z^{-1}||\Big(1 - \frac{2}{\sqrt{\kappa}+1}\Big)^m, \tag{3}$$

(see [32] for similar but more general results). Again $\kappa = \frac{\lambda_n}{\lambda_1}$, but here it is not necessarily the same as the standard condition number. For more highly nonnormal matrices, convergence properties are more complicated. Some analysis has been done,

especially if all of the eigenvalues are in an ellipse not containing the origin (see [9], [22], [30–32]).

The disadvantage with restarting is that some information is lost at the time of the restart. The subspace is discarded, and this slows down the convergence. Other methods such as nonsymmetric Lanczos and CGNE avoid restarting. But they have their own disadvantages as mentioned earlier. Another attempt at avoiding restarting is incomplete orthogonalization [30], [31], but convergence properties are not well understood.

2. Adding approximate eigenvectors to the subspace. We attempt to improve GMRES by reducing the ill effects of restarting. Some information can be retained at the time of the restart. This is done by saving vectors from the old subspace and adding them to the new subspace that is generated. For instance, one could save the last few Arnoldi vectors (the v_i's). However, there are other vectors that are more helpful to the convergence.

We note that information about the eigenvalues and eigenvectors of A is available during GMRES. They can be calculated with the Arnoldi method for eigenvalues [1], [30], [37]. Eigenvalue calculations have been used before in conjunction with GMRES to implement hybrid methods (see [10], for example).

We investigate saving approximate eigenvectors of A corresponding to the smallest eigenvalues in magnitude. These vectors are added to the new subspace. The motivation for this is that if a converged eigenvector is added to the subspace, the corresponding eigenvalue is effectively eliminated from the spectrum or deflated. Convergence proceeds according to the modified spectrum. This is demonstrated in the following theorem for the case of real and positive eigenvalues. We let $\kappa_e \equiv \frac{\lambda_n}{\lambda_{k+1}}$, the "effective condition number," and assume that the initial guess x_0 is zero.

THEOREM 1. *Suppose A has spectral decomposition $A \equiv Z\Lambda Z^{-1}$, with all the eigenvalues being real and positive. Assume that the minimum residual solution $\hat{x}$ is extracted from the subspace* $\mathrm{Span}\{b, Ab, \ldots, A^{m-1}b, z_1, z_2, \ldots, z_k\}$, *where the z_i's are columns of Z. Then*

$$\frac{||r||}{||b||} \leq 2||Z||||Z^{-1}||\Big(1 - \frac{2}{\sqrt{\kappa_e}+1}\Big)^m, \tag{4}$$

where $r \equiv b - A\hat{x}$ is the residual vector.

Proof. Any vector $\hat{x}$ from the subspace $\mathrm{Span}\{b, Ab, \ldots, A^{m-1}b, z_1, z_2, \ldots, z_k\}$ can be written in the form

$$\hat{x} = \sum_{i=1}^{k} \alpha_i z_i + p(A)b,$$

where p is a polynomial of degree $m-1$ or less. Expand b in terms of the eigenvectors:

$$b = \sum_{i=1}^{n} \beta_i z_i \tag{5}$$

and define the polynomial q as $q(x) = 1 - xp(x)$. Then we can calculate that

$$\begin{aligned} r = b - A\hat{x} &= -\sum_{i=1}^{k} \alpha_i \lambda_i z_i + q(A)b \\ &= \sum_{i=1}^{k} \gamma_i z_i + \sum_{i=k+1}^{n} \beta_i q(\lambda_i) z_i, \end{aligned} \tag{6}$$

where $\gamma_i = \beta_i q(\lambda_i) - \alpha_i \lambda_i$.

Since the solution minimizes the residual norm, it will be at least as good as any choice we make. Pick q to be the shifted-and-scaled Chebyshev polynomial that is small over the interval $[\lambda_{k+1}, \lambda_n]$. Then pick $\alpha_i = \frac{\beta_i q(\lambda_i)}{\lambda_i}$, so that each γ_i is zero. Now

$$r = \sum_{i=k+1}^{n} \beta_i q(\lambda_i) z_i,$$

and the desired result follows from the standard bound in (3).

Next, the effect of saving eigenvectors is examined for a couple of specific distributions of eigenvalues. First suppose the eigenvalues are distributed $1, 2, 3, \ldots, n$. Some linear equations problems do have a spectrum roughly similar to this (for example, see the model problem [15] from finite difference descretization of Poisson's equation). Suppose the eigenvectors corresponding to the k smallest eigenvalues are added to the subspace. Then the convergence bound improves from

$$\frac{||r||}{||b||} \leq 2||Z||||Z^{-1}||\Big(1 - \frac{2}{\sqrt{n}+1}\Big)^m$$

to

$$\frac{||r||}{||b||} \leq 2||Z||||Z^{-1}||\Big(1 - \frac{2\sqrt{k+1}}{\sqrt{n}+\sqrt{k+1}}\Big)^m.$$

We can roughly compare convergence by comparing $\sqrt{\kappa}$ to $\sqrt{\kappa_e}$. The ratio is

$$\frac{\sqrt{\kappa}}{\sqrt{\kappa_e}} = \frac{\sqrt{n}}{\sqrt{\frac{n}{k+1}}} = \sqrt{k+1}. \tag{7}$$

So convergence is roughly $\sqrt{k+1}$ times as fast with the eigenvectors added to the subspace. For example, with $k = 3$, the rate of convergence is about twice as fast. However, to get quadruple the convergence requires $k = 15$. The returns are diminishing as more eigenvectors are added.

Next, consider the eigenvalue distribution $\frac{-n}{2}, \frac{-n}{2}+1, \ldots, -2, -1, 1, 2, \ldots, \frac{n}{2}$. This is a much tougher problem than the previous one, because it is indefinite. The residual norm is reduced by roughly a factor of $1 - \frac{2}{n}$ in each iteration [2], [24]. If k is even and the k eigenvectors with smallest eigenvalues in magnitude are added to the subspace, then the factor improves to $1 - \frac{k+2}{n}$. This means convergence is approximately $\frac{k+2}{2}$ times better with the eigenvectors. Adding eigenvectors to the subspace is even more important in this indefinite case than it was in the previous positive definite example. With $k = 8$, convergence is about five times better. This compares to three times better in the previous example.

The results in the preceding two paragraphs may not always apply. We will discuss three problems with them. First, the distribution of eigenvalues may not be so favorable. For example, there may not be small eigenvalues. Then convergence could still be slow for indefinite and highly nonsymmetric problems, yet saving approximate eigenvectors would not be beneficial. Also, if the eigenvalues are not so evenly spaced, the results may not be as good.

Second, we have only analyzed bounds on convergence and estimates of rates, not the actual rates of convergence. As mentioned earlier, the convergence of the conjugate gradient method does not depend strictly on the condition number. An outlying

eigenvalue does not have a perpetual effect on the covergence rate. It will at most add a number of iterations, then the outlying eigenvalue is taken care of and the convergence proceeds according to the other eigenvalues. This is because the underlying polynomial can have a zero at that eigenvalue. However, Cline [5] observed in some experiments that adding an extremely small eigenvalue to a particular distribution of eigenvalues requires from 5 to 19 extra iterations. With restarted GMRES, reducing the number of iterations by a few is worthwhile, because this reduction occurs during every restart cycle.

A third problem with the earlier analysis is that it may be awhile before the approximate eigenvectors become very accurate. However, an approximate eigenvector can have beneficial effects long before it has attained full accuracy. This is shown in the following theorem for the case of one approximate eigenvector.

THEOREM 2. *Suppose A has spectral decomposition $A \equiv Z\Lambda Z^{-1}$, with Λ diagonal. Suppose the* GMRES *with eigenvectors method is used with one approximate eigenvector y_1. Let $\psi \equiv \angle(y_1, z_1)$, and let β_1 be the coefficient of z_1 in the expansion of b; see* (5). *Then*

$$||r|| \leq ||Z||||Z^{-1}|| \max_{i \neq 1} |q(\lambda_i)| \, ||b|| + \frac{||A||}{\lambda_1} \tan\psi \, |q(\lambda_1)| \, |\beta_1|, \tag{8}$$

where q is a polynomial of degree m or less such that $q(0) = 1$.

Proof. Similar to (6), we can derive

$$r = q(A)b - \alpha_1 A y_1,$$

where q is a polynomial of degree m or less, such that $q(0) = 1$. Decompose y_1 as

$$y_1 = \cos\psi \, z_1 + \sin\psi \, u,$$

where y_1, z_1, and u are all unit vectors and $u \perp z_1$. Then

$$\begin{aligned} r &= q(A)b - \alpha_1 \lambda_1 \cos\psi z_1 - \alpha_1 \sin\psi Au \\ &= \sum_{i=2}^{n} \beta_i q(\lambda_i) z_i + \big(\beta_1 q(\lambda_1) - \alpha_1 \lambda_1 \cos\psi\big) z_1 - \alpha_1 \sin\psi Au. \end{aligned}$$

Pick $\alpha_1 = \frac{\beta_1 q(\lambda_1)}{\lambda_1 \cos\psi}$, and use the minimum residual property. Then

$$\begin{aligned} ||r|| &\leq \left\| \sum_{i=2}^{n} \beta_i q(\lambda_i) z_i - \frac{\beta_1 q(\lambda_1) \sin\psi Au}{\lambda_1 \cos\psi} \right\| \\ &\leq \left\| \sum_{i=2}^{n} \beta_i q(\lambda_i) z_i \right\| + \left\| \frac{\beta_1 q(\lambda_1) \tan\psi Au}{\lambda_1} \right\| \\ &\leq ||Z||||Z^{-1}|| \max_{i \neq 1} |q(\lambda_i)| \, ||b|| + \frac{||A||}{\lambda_1} \tan\psi \, |q(\lambda_1)| \, |\beta_1|. \end{aligned}$$

The second term in the right-hand side of (8) occurs because of the inaccuracy of the approximate eigenvector. Roughly, it appears that this term will not be significant as long as the accuracy of the approximate eigenvector is greater than the amount of

improvement brought by the polynomial q (as long as $\tan\psi$ is somewhat less than $\max_{i\neq 1} |q(\lambda_i)|$).

If the eigenvalues are all real and positive, (8) can be made more specific by choosing the polynomial q to be a shifted and scaled Chebyshev polynomial that is small over the interval $[\lambda_2, \lambda_n]$. Then

$$||r|| \leq ||Z||||Z^{-1}|| \left(1 - \frac{2}{\sqrt{\kappa_e}+1}\right)^m ||b|| + \frac{||A||}{\lambda_1} \tan\psi \, |\beta_1|, \tag{9}$$

where $\kappa_e \equiv \frac{\lambda_n}{\lambda_2}$. And this can be put in a form more similar to (4):

$$\frac{||r||}{||b||} \leq ||Z||||Z^{-1}|| \left(\left(1 - \frac{2}{\sqrt{\kappa_e}+1}\right)^m + \frac{||A||}{\lambda_1} \tan\psi \right). \tag{10}$$

3. Implementation. The implementation presented here first generates the Krylov subspace, then adds the approximate eigenvectors. There is still an upper-Hessenberg matrix for the linear equations problem, but the eigenvalue problem is more complicated.

Let m be the dimension of the Krylov subspace, and suppose k approximate eigenvectors are used. Let $l = m + k$. Let W be the n by l matrix whose first m columns are the orthonormalized Arnoldi vectors (the v_i vectors in step 2 of GMRES) and whose last k vectors are the approximate eigenvectors y_i, for $i = 1, \ldots, k$. Let Q be the n by $l+1$ matrix whose first $m+1$ columns are Arnoldi vectors and whose last k columns are formed by orthogonalizing the vectors Ay_i, for $i = 1, \ldots, k$, against the previous columns of Q. Then

$$AW = Q\bar{H}, \tag{11}$$

where $\bar{H}$ is an $(l+1)$ by l upper-Hessenberg matrix (this is similar to (3) in [32], for the standard Arnoldi iteration on which GMRES is based).

The restarted linear equations problem is

$$A(x - x_0) = r_0.$$

The approximate solution $\hat{x} - x_0$ is a combination of the columns of W, so

$$\hat{x} - x_0 = W\hat{d}.$$

The minimum residual solution can be calculated in the same way as for standard GMRES. Let

$$P\bar{H} = R, \tag{12}$$

where P is orthogonal and R is upper triangular. Then

$$\begin{aligned} ||r|| &= ||b - A\hat{x}|| \\ &= ||r_0 - A(\hat{x} - x_0)|| \\ &= ||r_0 - AW\hat{d}|| \\ &= ||r_0 - Q\bar{H}\hat{d}|| \\ &= ||Q^* r_0 - \bar{H}\hat{d}|| \\ &= ||PQ^* r_0 - R\hat{d}||. \end{aligned} \tag{13}$$

The minimal solution is then found by solving for $\hat{d}$ that makes the first l entries be zero. Note Q^*r_0 is a multiple of the first coordinate vector. As in standard GMRES, the residual norm is a byproduct. It is the magnitude of the last entry of PQ^*r_0.

We wish to find approximate eigenvectors from the subspace spanned by the columns of W. Since W is not orthonormal, the generalized Rayleigh–Ritz procedure with reduced eigenvalue problem

$$W^*AWg_i = \theta_i W^*Wg_i$$

could be used. However, we choose a version of Rayleigh-Ritz that finds good approximations to the eigenvalues nearest to zero [24], [25]. This version uses the reduced problem

$$W^*A^*Wg_i = \frac{1}{\theta_i} W^*A^*AWg_i. \tag{14}$$

Let $F \equiv W^*A^*W$ and $G \equiv W^*A^*AW$. Then the reduced eigenvalue problem is the l by l generalized eigenvalue problem

$$Fg_i = \frac{1}{\theta_i} Gg_i. \tag{15}$$

The g_i's associated with the k largest $\frac{1}{\theta_i}$'s (or the k smallest θ_i's) are needed. An approximate eigenvector is $y_i = Wg_i$. And $Ay_i = AWg_i = Q\bar{H}g_i$. If y_i is complex, the real and imaginary parts are used separately.

Little calculation is required for G, because

$$\begin{aligned} G &= W^*A^*AW \\ &= \bar{H}^*Q^*Q\bar{H} \\ &= \bar{H}^*\bar{H} \\ &= R^*R. \end{aligned} \tag{16}$$

The first m columns of F are the same as the first m columns of $\bar{H}^*$. Entries in the intersection of the last k rows and the last k columns can be cheaply computed using the previous F, since $f_{ij} = y_i^*A^*y_j = g_i^*W_{\text{old}}^*A^*W_{\text{old}}g_j = g_i^*F_{\text{old}}g_j$. The remaining entries are calculated as $f_{ij} = y_i^*A^*y_j = (Ay_i)^*y_j$, so they are more expensive.

The small generalized eigenvalue problem (15) is solved with EISPACK [33] in the examples in the next section. However an iterative method, such as subspace iteration, could also be used. Only the eigenvectors associated with the largest values of $\frac{1}{\theta_i}$ are needed, G is already in a factored form, good starting vectors are the last k coordinate vectors, and full convergence is not necessary.

The implementation is a little different for the first run, before any restart. Standard GMRES is used, except eigenvector calculations are added on at the end. F is the same as $\bar{H}^*$ except that the last column is removed, and G can be found with (16). For simplicity, the listing of the algorithm is given just for the second and subsequent runs.

One restarted run of GMRES with eigenvectors

1. *Initial definitions and calculations:* The Krylov subspace has dimension m, k is the number of approximate eigenvectors, and $l = m + k$. Let $q_1 = r_0/||r_0||$ and

$w_1 = q_1$. Let $y_1, y_2, \ldots, y_k$ be the approximate eigenvectors. Let $w_{m+i} = y_i$, for $i = 1, \ldots, k$. For $j = m+1, \ldots, l$ do: $f_{ij} = g_i^* F_{\text{old}} g_j, i = m+1, \ldots, l$.

2. *Generation of Arnoldi vectors:* For $j = 1, 2, \ldots, m$ do:
   $h_{ij} = (Aq_j, q_i), i = 1, 2, \ldots, j,$
   $f_{ji} = h_{ij}, i = 1, 2, \ldots, j,$
   $f_{j,m+i} = (Aq_j, y_i), i = 1, 2, \ldots, k,$
   $\hat{q}_{j+1} = Aq_j - \sum_{i=1}^{j} h_{ij} q_i,$
   $h_{j+1,j} = ||\hat{q}_{j+1}||$, and
   $q_{j+1} = \hat{q}_{j+1}/h_{j+1,j}$.
   If $j < m$, let $w_{j+1} = q_{j+1}$ and $f_{j,j+1} = h_{j+1,j}$.
3. *Addition of approximate eigenvectors:* For $j = m+1, \ldots, l$ do:
   $h_{ij} = (Aw_j, q_i), i = 1, 2, \ldots, j,$
   $f_{ji} = h_{ij}, i = 1, 2, \ldots, m,$
   $\hat{q}_{j+1} = Aw_j - \sum_{i=1}^{j} h_{ij} q_i,$
   $h_{j+1,j} = ||\hat{q}_{j+1}||$, and
   $q_{j+1} = \hat{q}_{j+1}/h_{j+1,j}$.
4. *Form the approximate solution:* Let $\beta = ||r_0||$. Find $\hat{d}$ that minimizes $||\beta e_1 - \bar{H} d||$ for all $d \in R^l$. The orthogonal factorization $P\bar{H} = R$, for R upper triangular, is used. Then $\hat{x} = x_0 + W\hat{d}$.
5. *Form the new approximate eigenvectors:* Calculate $G = R^* R$. Solve $Fg_i = \frac{1}{\theta_i} Gg_i$, for the appropriate g_i (separate g_i into real and complex parts if it is complex and treat as two distinct vectors). Form $y_i = Qg_i$ and $Ay_i = Q\bar{H}g_i$. Let $F_{\text{old}} = F$.
6. *Restart:* Compute $r = b - A\hat{x}$; if satisfied with the residual norm then stop, else let $x_0 = \hat{x}$ and go to 2.

We now examine the expenses and storage requirements for the GMRES with eigenvectors method as compared to standard GMRES. We consider only the major expenses. Suppose the subspace is currently a Krylov subspace of dimension j. If we choose the next vector for the subspace to be one more Arnoldi vector, then there is one matrix-vector product needed. The orthogonalization requires about $2jn$ multiplications. If instead we let the next vector be an approximate eigenvector, no matrix-vector product is required. The other costs are approximately $5jn$ multiplications. This includes $2jn$ for the orthogonalization of Ay_i, jn for computing a portion of F, and $2jn$ for forming y_i and Ay_i. This can be reduced to $4jn$ if a matrix-vector product is used for Ay_i instead of forming it from the columns of Q. It is also possible to reduce costs by another jn if Ay_i is not explicitly orthogonalized (the entries of $\bar{H}$ can still be calculated). This last option has not been tested.

We compare the storage for a Krylov subspace of dimension $m + k$ in standard GMRES to storage for the GMRES with eigenvectors method using a Krylov subspace of dimension m and k approximate eigenvectors. The major storage requirement for GMRES($m + k$) is $m + k + 2$ vectors of length n. For GMRES with eigenvectors, the major storage requirement is for $m+2k+2$ vectors of length n. So using an approximate eigenvector requires about twice the storage of using an additional Arnoldi vector. This is because both y_i and Ay_i are stored.

The relative efficiency of the two methods depends on how expensive the matrix-vector product is compared to the orthogonalization costs. We consider two extreme cases, although many problems will fall somewhere in between. The first case is where the matrix-vector product is the main expense, and the second is where the matrix-vector product is fairly inexpensive and orthogonalization costs dominate. Saving approximate eigenvectors is particularly worthwhile for the first case, since no matrix-vector product is required for the approximate eigenvectors. The benefits of the

approximate eigenvectors are essentially free from expense. However, since storage is often limited, using one approximate eigenvector means two less Arnoldi vectors can be used.

For the second case of expensive orthogonalization, a matrix-vector product would be used to get Ay_i. So the expense is about $4jn$ for an approximate eigenvector. Therefore using an approximate eigenvector instead of an Arnoldi vector costs about twice as much. To be useful, an approximate eigenvector must be as effective as two Arnoldi vectors.

4. Examples. In the following examples, the right-hand sides have all entries 1.0. The first four examples are bidiagonal matrices with 0.1 in each superdiagonal position. The initial guesses x_0 are zero vectors. The calculations are done in double precision on either an IBM 3090-170J or a Vax 6510. We call the iteration between restarts a "run."

Example 1. We let the matrix have $1, 2, 3, \ldots, 999, 1000$ on the main diagonal and as mentioned above, the super diagonal elements have 0.1's. For this matrix, the quantity $||Z||||Z^{-1}||$ in (3) and (4) is small (about 1.2). The new GMRES with eigenvectors method using $m = 21$ and $k = 4$ (21 Krylov vectors and four approximate eigenvectors) is compared to GMRES(25). Thus the same size subspaces are used. After 12 runs, the eigenvector method has a residual norm of 0.42e-9 compared to 0.15e-4 for standard GMRES. See Fig. 1 for a graph of the convergence. After iteration 100, the eigenvector method converges more than twice as fast. This is roughly as predicted by (7), even though the Krylov portion of the subspace is smaller for the eigenvector method than for the regular GMRES. At iteration 100, after four runs, the eigenvector approximations are not very accurate. The approximate eigenvalues are 1.01, 2.20, 3.86, and 6.10, and the corresponding residual norms range from 0.13 to 1.9. But already the eigenvectors are accurate enough to assist convergence. After eight runs, the approximate eigenvalues are more accurate with from 8 to 2 significant digits and residual norms from 0.13e-3 to 0.17.

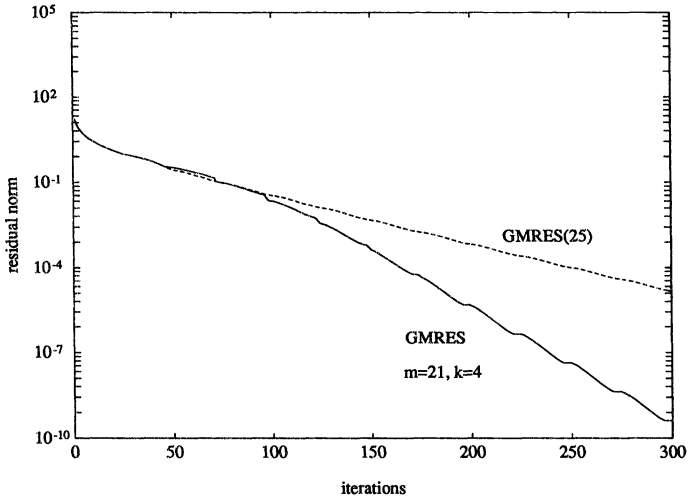


FIG. 1. GMRES *vs.* GMRES *with eigenvectors.*

Next, methods requiring about the same storage are compared. The eigenvector method with $m = 17$ and $k = 4$ reaches residual norm of 0.22e-6 after 12 runs (see Table 1). This is still better than GMRES(25), even though smaller subspaces are used and far less matrix-vector products are required. If an equal number of matrix-vector

TABLE 1.
Eigenvalues $1, 2, 3, \ldots, 1000$.

			Residual norms		
m	k	l	Initial	After 12 runs	After 300 matrix-vector products
25	0	25	0.31e+2	0.15e-4	0.15e-4
21	4	25	0.31e+2	0.42e-9	0.35e-11
17	4	21	0.31e+2	0.22e-6	0.18e-10
21	2	23	0.31e+2	0.67e-7	0.20e-8
19	3	22	0.31e+2	0.76e-7	0.14e-9
13	6	19	0.31e+2	0.19e-4	0.23e-11
9	8	17	0.31e+2	0.25e-2	0.40e-11

products are taken, the eigenvector method with $m = 17$ and $k = 4$ is much further ahead. After 300 matrix-vector products, it attains 0.18e-10 versus 0.15e-4. Table 1 also gives results with different choices of k but with the same storage (the same $m+2k$). Using just two approximate eigenvectors gives the lowest residual norm after 12 runs. However, if one is most interested in the number of matrix-vector products, using six eigenvectors is better, even though the Krylov subspace has dimension of only $m = 13$.

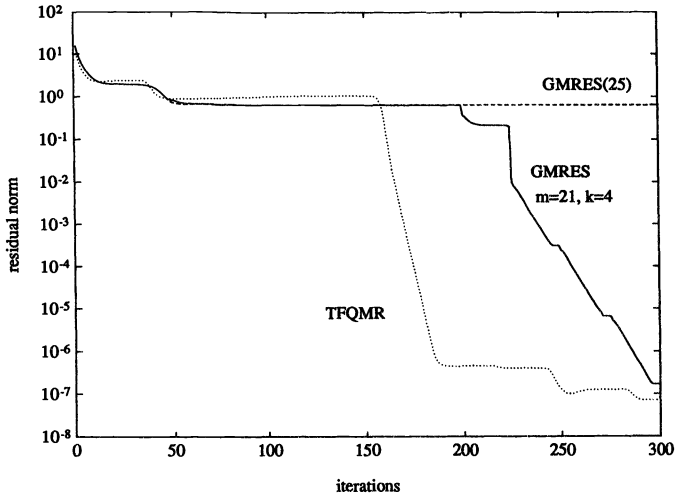


FIG. 2. *Comparison for Example* 2 (TFQMR *uses two* mvp *per iteration*).

Example 2. The next matrix has some very small eigenvalues that make the problem difficult. The entries on the main diagonal are $0.01, 0.02, 0.03, 0.04, 10, 11, 12, 13, \ldots, 1005$ (taking on all integer values from 10–1005). See Table 2 and Fig. 2 for the computational results. Figure 2 also includes the TFQMR method; this is discussed in Example 7. It appears that the regular GMRES method will not converge. Going past 12 runs, there is no further improvement in the residual norm. The four small eigenvalues make this problem too difficult for Krylov subspaces of dimension 25. The GMRES with eigenvectors method also stalls out for a while. Considering the case $m = 21$ and $k = 4$, the residual norm only improves from 0.65–0.63 during runs four through seven. But finally after run seven, there are rough approximations to *all four* of the small eigenvalues. With the corresponding approximate eigenvectors in the subspace, the convergence rate is soon fairly rapid.

Example 3. Here an indefinite matrix is used. The diagonal entries are $-2, -1, 1, 2, 3, 4, \ldots, 997, 998$. In this situation, the eigenvector method is much better than regular

TABLE 2.
Eigenvalues $0.01, 0.02, 0.03, 0.04, 10, 11, 12, \ldots, 1005$.

Residual norms

m	k	l	Initial	After 12 runs	After 300 matrix-vector products
25	0	25	0.32e+2	0.64	0.64
21	4	25	0.32e+2	0.17e-6	0.91e-10
17	4	21	0.32e+2	0.18e-2	0.47e-9

TABLE 3.
Eigenvalues $-2, -1, 1, 2, \ldots, 998$.

Residual norms

m	k	l	Initial	After 5 runs	After 10 runs	After 15 runs	After 20runs	After 500 mvp's
25	0	25	0.32e+2	0.90	0.58	0.38	0.24	0.24
21	4	25	0.32e+2	0.22	0.83e-4	0.21e-7	0.54e-11	0.14e-13
17	4	21	0.32e+2	0.53	0.24e-2	0.50e-5	0.11e-7	0.64e-13

GMRES; see Table 3. With $k = 4$, approximations are developed to the eigenvalues -2, -1, 1, and 2. This eliminates the two negative eigenvalues and effectively turns the indefinite problem into a definite one.

Example 4. This example is designed to be difficult for the GMRES with eigenvectors method. Let the matrix have diagonal elements $1, 1.01, 1.02, 1.03, 1.04, 2, 3, 4, 5, \ldots, 995, 996$ (see Table 4). Removing the four smallest eigenvalues does not have much effect on the spread of eigenvalues. The two methods are roughly equivalent when using the same size subspace. With equal storage, standard GMRES is one order of magnitude better after 15 runs and the eigenvector method is a little better after each has taken 375 matrix-vector products. Even in this difficult situation, using eigenvectors does not substantially decrease efficiency. In another test with five approximate eigenvectors ($m = 20$, $k = 5$), the method does not improve. The reason is that approximations to the five smallest eigenvalues do not develop in time to substantially help. The eigenvalues are so close together that they are difficult to compute (after 15 runs, the five approximate eigenpairs have residual norms no smaller than 0.05).

TABLE 4.
Eigenvalues $1, 1.01, 1.02, 1.03, 1.04, 2, 3, \ldots, 996$.

Residual norms

m	k	l	Initial	After 5 runs	After 10 runs	After 15 runs	After 375 mvp's
25	0	25	0.32e+2	0.10e-1	0.58e-4	0.48e-6	0.48e-6
21	4	25	0.32e+2	0.67e-2	0.17e-4	0.12e-6	0.16e-7
17	4	21	0.32e+2	0.22e-1	0.24e-3	0.50e-5	0.77e-7

Example 5. Let the matrix be block diagonal with eigenvalues equally spaced around a circle in the complex plane with center at (1,0) and radius 0.99, starting at 0.01. The matrix is normal with blocks of size 2 or 1. With $n = 100$, the approximate eigenvectors are very helpful. Theoretical convergence results for GMRES often consider an ellipse or circle [32] containing the eigenvalues. The smallest circle containing all of the eigenvalues does not change when a few of the eigenvalues nearest the origin are eliminated. Nevertheless, the convergence is improved. With a change to $n = 200$,

TABLE 5.
Eigenvalues in a circle.

Residual norms for $n = 100$

m	k	l	Initial	After 5 runs	After 10 runs	After 15 runs	After 20 runs	After 500 mvp's
25	0	25	0.10e+2	0.47	0.13	0.38e-1	0.11e-1	0.11e-1
21	4	25	0.10e+2	0.52	0.45e-2	0.85e-5	0.16e-7	0.26e-9
17	4	21	0.10e+2	0.65	0.12	0.98e-3	0.77e-5	0.13e-8

Residual norms for $n = 200$

m	k	l	Initial	After 10 runs	After 20 runs	After 30 runs	After 40 runs	After 1000 mvp's
25	0	25	0.14e+2	0.19	0.15e-1	0.12e-2	0.10e-3	0.10e-3
21	4	25	0.14e+2	0.25	0.30e-1	0.36e-2	0.44e-3	0.92e-4
17	4	21	0.14e+2	0.38	0.68e-1	0.12e-1	0.22e-2	0.94e-4

the problem of finding the eigenvalues is tougher, because they are closer together. Here the eigenvalue problem is apparently more difficult than the linear equations problem and the eigenvector approximations never become accurate enough to really help (see Table 5). After run 10, the approximation to the smallest eigenvalue has residual norm 0.26e-1 and this does not improve during the next 30 runs. For comparison, during the test with $n = 100$, the approximation to the smallest eigenvalue has residual norm 0.13e-3 after 10 runs, and it slowly improves.

Example 6. This example has a standard test matrix (see Table 6). The problem is from the finite difference discretization of the partial differential equation $u_{xx}+u_{yy}+Du_x = -(41)^2$ on the unit square with $u = 0$ on the boundary. Central differences are used. The mesh spacing is $h = \frac{1}{41}$, so $n = 1600$. Tests are done with increasing degrees of nonsymmetry: $D = 1$, $D = 41$ and $D = (41)^2$. For the first two, the results are similar to Example 1. Using eigenvectors is definitely worthwhile. For the last test, the approximate eigenvectors are not particularly useful. There are no eigenvalues near to the origin, and the algorithm has trouble computing the ones that are closest to the origin. Perhaps this is because there are several about the same distance away.

The rest of this section has comparisons with the quasiminimal residual, or QMR, method of Freund and Nachtigal [12] and the QMR transpose-free variant, TFQMR [11]. Because these methods do not restart, they have an advantage for difficult problems that require large subspaces. The quasiminimization in QMR controls much of the instability, but it does not insure that the subspace is used as effectively as in GMRES.

In the tests here, TFQMR uses standard weights [11] and QMR has unit weights. The right-hand side of all 1.0's is used for both of the initial vectors. The look-ahead feature is not used [12], [22]. Convergence of QMR and TFQMR is monitored with the approximate residual norms given in [11], [12]. The matrices from Example 5 are left out because of their small size: while QMR reaches convergence faster than the GMRES methods, it is only after the dimension of the subspace is larger than the size of the problem (actually QMR is unstable, but it does converge when an initial vector is changed).

Example 7. Table 7 has comparisons between GMRES(25), the modified version of GMRES with 21 Krylov vectors and four approximate eigenvectors, and the two versions of QMR. Both the number of iterations and the number of matrix-vector products are given. The QMR methods require two matrix-vector products per iteration (and one for the residual norm at the end), while GMRES requires one

TABLE 6.
Finite difference matrix.

Residual norms for D=1

m	k	l	Initial	After 8 runs	After 200 matrix-vector products
25	0	25	0.40e+2	0.13e-3	0.13e-3
21	4	25	0.40e+2	0.52e-10	0.28e-12
17	4	21	0.40e+2	0.33e-7	0.12e-11

Residual norms for D=41

m	k	l	Initial	After 8 runs	After 200 matrix-vector products
25	0	25	0.40e+2	0.70e-4	0.70e-4
21	4	25	0.40e+2	0.33e-9	0.64e-11
17	4	21	0.40e+2	0.95e-7	0.90e-11

Residual norms for D=$(41)^2$

m	k	l	Initial	After 20 runs	After 500 matrix-vector products
25	0	25	0.40e+2	0.98e-7	0.98e-7
21	4	25	0.40e+2	0.71e-8	0.16e-9
17	4	21	0.40e+2	0.57e-6	0.32e-9

matrix-vector product per iteration, and the GMRES with eigenvectors method requires one for each Krylov vector but no matrix-vector products for the approximate eigenvectors. For these examples, TFQMR always converges in the least number of iterations, while GMRES with eigenvectors always uses the fewest matrix-vector products. Figure 2 also shows TFQMR converging in less iterations for the matrix from Example 2, but GMRES with eigenvectors uses less than half the number of matrix-vector products (see Table 7). For one more comparison not included in the table, the matrix in Example 2 is modified to have ten small eigenvalues from 0.01–0.10 and the rest from 10–999. For this problem, GMRES with eigenvectors using $m = 15$ and $k = 10$ was compared to TFQMR. The modified GMRES approach required fewer iterations, 707 compared to 1109, and it used less than one-fifth the number of matrix-vector products, 437–2219. These tests do not indicate that one method is better than the other, but they do show that GMRES with eigenvectors is worth considering, especially in situations where the matrix-vector product is expensive.

Example 8. The GMRES with eigenvectors method may also be particularly useful when there are several similar systems of linear equations or several right-hand sides. One such case occurs in solving time-dependent differential equations. In the following tests, a simple time-dependent problem is considered. Let the differential equation be $u_t = u_{xx} + u_{yy} + u_x$, on the unit square with t going from 0.0–1.0. The initial condition is u(x,y,0) = 1.0, the boundary condition is $u = 0$ on the boundary. The backward difference method is used with time steps of 0.1, and discretization of the spacial variables is as in Example 6. The termination criterion while solving the systems of linear equations is $||r|| < 10^{-4}$. Table 8 gives the number of iterations at each time step and the total number of iterations and matrix-vector products for the QMR methods, GMRES(20), and GMRES with eigenvectors with m=17 and k=3. The QMR methods have a tendency to start slowly, then converge rapidly. This can be a disadvantage when several systems are solved to low accuracy. Meanwhile the GMRES with eigenvectors method has an advantage, because it can use the approximate eigenvectors from the previous time step to help at the current one. GMRES with eigenvectors performs better than the QMR methods for this problem.

TABLE 7.
Comparison to QMR.

Iterations and matrix-vector products to reach $||r|| < 10^{-6}$

			GMRES m=25, k=0	GMRES w/e.vectors m=21, k=4	QMR	TFQMR
Ex. 1		it's	370	214	180	119
		mvp's	370	186	361	239
Ex. 2		it's	-	286	339	250
		mvp's	-	246	679	501
Ex. 3		it's	-	339	215	160
		mvp's	-	291	431	321
Ex. 4		it's	355	325	252	162
		mvp's	355	277	505	325
Ex. 6	D=1	it's	278	132	125	95
		mvp's	278	116	251	191
	D=41	it's	300	149	100	(83)*
		mvp's	300	134	201	(167)
	$D=41^2$	it's	441	382	366	218
		mvp's	441	326	733	437

* Because of instability, a different left initial vector was used.

TABLE 8.
Time-dependent problem.

Iterations for each time step

	GMRES m=20, k=0	GMRES w/e.vectors m=17, k=3	QMR	TFQMR
t = 0.1	130	91	92	62
t = 0.2	113	34	71	70
t = 0.3	48	27	60	67
t = 0.4	28	20	38	49
t = 0.5	14	14	29	44
t = 0.6	15	12	22	27
t = 0.7	13	14	11	24
t = 0.8	11	11	13	16
t = 0.9	6	6	11	18
t = 1.0	1	2	2	1
Total iter.'s	379	231	349	378
Total mvp's	379	213	708	766

5. Attempt at an automatic procedure. Here we deal with two questions. How many approximate eigenvectors should be used, and should the approximate

eigenvectors be discarded at some point? However, it is difficult to give answers that apply to all matrices.

For determining the proper number of approximate eigenvectors to use, we consider the model eigenvalue distribution $1, 2, \ldots, n$. If we assume that the storage is fixed, then methods with the same value of $m + 2k$ should be compared. We compute k that gives the lowest value of

$$\left(\left(\frac{\sqrt{\kappa_e}+1}{\sqrt{\kappa_e}-1}\right)^m + \left(\frac{\sqrt{\kappa_e}-1}{\sqrt{\kappa_e}+1}\right)^m\right)^{-1},$$

where $\kappa_e \equiv \frac{\lambda_n}{\lambda_{k+1}}$. This formula comes from Theorem 1, but with the more accurate bound given in the first part of (2). After doing some comparisons, we find that if $m + 2k$ is given, the best value is approximately

$$k = \frac{m+2k}{7}.$$

For values of $m + 2k$ greater than 50, slightly more should be used, and for values less than 20, the number should be rounded down.

It would be desirable to have a code that adaptively increases or decreases the number of approximate eigenvectors being used. However, it is difficult to determine if adding another eigenvector will help when no accurate approximation is available for the next eigenvalue. For now we just consider the possibility of releasing the approximate eigenvectors and going back to standard GMRES. This switch should be done if the eigenvectors are not helping. Even beneficial eigenvectors may lose their effectiveness once components of the residual vector in the directions of those eigenvectors have been purged.

One possibility is to check how effective the addition of the approximate eigenvectors is in lowering the residual norm. This information is readily available. The amount the eigenvectors lower the residual norm can be compared with the amount the residual norm decreases in the previous k Krylov iterations. This suggests the test: switch when

$$||r_m|| - ||r_{m+k}|| < ||r_{m-k}|| - ||r_m||.$$

However, it turns out that the beneficial effect of the eigenvectors is not fully reflected in how they lower the residual norm. They also enable the Krylov vectors to be more effective. This makes it difficult to determine whether the eigenvectors are useful or not. A factor can be added in

$$||r_m|| - ||r_{m+k}|| < 0.2\big(||r_{m-k}|| - ||r_m||\big). \tag{17}$$

This is effective for Examples 1 and 2, but it releases much too early for Example 3.

We consider the addition of some more complicated tests that involve the accuracy of the eigenvectors. For the approximate eigenvector y_i, denote the eigenvector residual norm by rne_i. Then

$$rne_i = \frac{||Ay_i - \rho_i y_i||}{||y_i||},$$

where

$$\rho_i = \frac{y_i^* A y_i}{y_i^* y_i} = \frac{g_i^* F^* g_i}{y_i^* y_i}.$$

The eigenvector residual norm can be computed explicitly since Ay_i and y_i have been formed, and there is also a formula involving F^*, G, ρ_i, and g_i. To determine if the

eigenvectors are not helpful, we check that the improvement of the best approximate eigenvector is at least one-tenth as much as the improvement of the linear equations approximate solution during a restarted run of the method. So the criterion is

$$(18)\quad -\log_{10}(rne_1)_{\text{new}} + \log_{10}(rne_1)_{\text{old}} < 0.1\big(-\log_{10}||r_{m+k}||_{\text{new}} + \log_{10}||r_{m+k}||_{\text{old}}\big).$$

The switch is done if both (17) and (18) are satisfied.

In addition, we check to see if the eigenvectors are no longer useful. We use the test

$$(19)\quad -\log_{10}\Big(\frac{rne_{(k-1)}}{\text{largest element in matrix}}\Big) - \log_{10}||r_{m+k}|| > -\log_{10}(r\text{tol}),$$

where $r$tol is the desired residual norm for the linear equations problem and $rne_{(k-1)}$ is the residual norm of the eigenvector that is second to last in accuracy. This test roughly follows from Theorem 2. The idea is that the magnitudes of the components of the solution to the restarted problem are approximately $||r_{m+k}||$. If an approximate eigenvector is fully used, then the component in that direction will be reduced from this size by approximately the accuracy of the eigenvector. So the resulting size of the component is reflected in the left-hand side of inequality (19). Once these components have been reduced to the desired magnitude specified by $r$tol, the approximimate eigenvectors are no longer needed. The switch is done if both (17) and (19) are satisfied.

In the tests that follow, the method begins with $m = 21$ and $k = 4$, then switches to GMRES(25). However, we note that if storage is the limiting factor, the switch could have been to GMRES(29). For the problem in Example 1 with $r\text{tol} = 1.\text{e} - 9$, the switch is made when (17) and (19) are satisfied after eight runs. Then after 12 runs the residual norm is 0.26e-9. This is just as good as if eigenvectors are kept for all of the runs. See Table 9. For Example 2 with $r\text{tol} = 1.\text{e} - 6$, the switch is made after 11 of 12 runs, and the method does better on the last run without the eigenvalues. For Example 3 with $r\text{tol} = 1.\text{e} - 10$, the switch is after 11 runs. The residual norm after 20 runs is 0.47e-8, not as good as the residual norm of 0.54e-11 without switching. In this case the eigevectors are very important and the switch test is triggered too soon. For Example 5 with $n = 100$ and for Example 6 with $D = 1$ and $D = 41$, the switch is not particularly significant.

Next for Example 4, Example 5 with $n = 200$, and Example 6 with $D = (41)^2$, the eigenvectors are not particularly useful and the switch is done when (17) and (18) are both satisfied. For Example 5, this happens after just 11 of 40 runs, because the approximate eigenvectors are not improving. The switch also works well for Example 6 with $D = (41)^2$.

More complicated adaptive procedures can be implemented. One possibility is to adaptively choose the number of eigenvectors to be used. Also the eigenvectors could be released individually as they converge. However, even the simpler procedures described in this section may not work for all problems.

6. Conclusion. Forming and using approximate eigenvectors can improve the convergence of restarted GMRES. Even just a few eigenvectors can make a big difference if the matrix has both small and large eigenvalues. Once the eigenvectors converge, the corresponding eigenvalues are essentially removed or deflated from the spectrum. And the approximate eigenvectors can improve convergence even before they are accurate.

TABLE 9.
Discarding eigenvectors.

	rtol	total runs	switch after	res. norm	res. norm w/o switch
Ex. 1	1.e-9	12	8	0.26e-9	0.50e-9
Ex. 2	1.e-6	12	11	0.36e-8	0.17e-6
Ex. 3	1.e-10	20	11	0.47e-8	0.54e-11
Ex. 5, n=100	1.e-8	20	17	0.81e-8	0.16e-7
Ex. 6, D=1	1.e-10	8	6	0.30e-9	0.52e-10
Ex. 6, D=41	1.e-10	8	8	0.33e-9	0.33e-9
Ex. 4		15	9	0.84e-7	0.12e-6
Ex. 5, n=200		40	11	0.13e-3	0.44e-3
Ex. 6, D=$(41)^2$		20	5	0.94e-8	0.71e-8

This method is useful for any problem that is difficult because of small eigenvalues. However, there are several situations where it is particularly beneficial. If the matrix-vector product is expensive, approximate eigenvectors can be used with relatively little extra expense. The method is also particularly effective when the spectrum of the matrix is well-behaved except for a few eigenvalues, such as in the case of having only a few negative eigenvalues or only a few eigenvalues with negative real parts. Also, if GMRES is used with a problem that has more than one right-hand side, then the eigenvectors can be computed once and used efficiently for all of the right-hand sides.

The method is not really needed for easy problems where few restarts are used. It also may not help if the problem is hard because of eigenvalues scattered around the complex plane. Another possibly related situation is when the small eigenvalues are less separated from rest of the spectrum than the spectrum is separated from zero. Then the eigenvalue problem is tougher than the linear equations problem. If the eigenvectors are not converging, then they probably should be discarded.

Acknowledgments. The author would like to thank Wayne Joubert for helpful discussions and thank the referees for their many constructive comments.

REFERENCES

[1] W. E. ARNOLDI, *The principle of minimized iterations in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

[2] S. F. ASHBY, *Polynomial Preconditioning for Conjugate Gradient Methods*, Ph.D. thesis, Dept. of Computer Science, University of Illinois, Urbana, 1987.

[3] O. AXELSSON, *A survey of preconditioned iterative methods for linear systems of equations*, BIT, 25 (1985), pp. 166–187.

[4] O. AXELSSON AND G. LINDSKOG, *On the rate of convergence of the preconditioned conjugate gradient method*, Numer. Math., 48 (1986), pp. 499–523.

[5] A. K. CLINE, *Several Observations on the Use of Conjugate Gradient Methods*, ICASE Report 76-22, NASA Langley Research Center, Hampton, VA, 1976.

[6] P. CONCUS, G. H. GOLUB, AND G. MEURANT, *Block preconditioners for the conjugate gradient method*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 220–252.

[7] P. CONCUS, G. H. GOLUB, AND D. P. O'LEARY, *A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations*, Sparse Matrix Computations, J. R. Bunch and D. J. Rose, eds., Academic Press, New York, 1976, pp. 309–332.

[8] E. CRAIG, *The N-step iteration procedure*, J. Math. Phys., 34 (1955), pp. 64–73.

[9] H. C. ELMAN, *Iterative Methods for Large Sparse Nonsymmetric Systems of Linear Equations*, Ph.D. thesis, Computer Science Dept., Yale University, New Haven, CT, 1982.

[10] H. C. ELMAN, Y. SAAD, AND P. E. SAYLOR, *A hybrid Chebyshev Krylov subspace algorithm*

for solving nonsymmetric systems of linear equations, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 840–855.

[11] R. W. FREUND, *A transpose-free quasi-minimal residual algorithm for non-Hermitian linear systems*, SIAM J. Sci. Statist. Comput., 14 (1993), pp. 470–482.

[12] R. W. FREUND AND N. NACHTIGAL, QMR: *a quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.

[13] G. H. GOLUB AND D. P. O'LEARY, *Some history of the conjugate gradient and Lanczos methods*, SIAM Rev., 31 (1989), pp. 50–100.

[14] M. GUTKNECHT, *A completed theory of the unsymmetric Lanczos process and related algorithms, Part* I, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 594–639.

[15] L. A. HAGEMAN AND D. M. YOUNG, *Applied Iterative Methods*, Academic Press, New York, 1981.

[16] M. R. HESTENES, *The conjugate gradient method for solving linear systems*, in Proc. Symp. Appl. Math., Vol. 6, Numerical Analysis, American Mathematical Society, New York, 1956.

[17] M. R. HESTENES AND E. STIEFEL, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 409–436.

[18] W. JOUBERT, *Lanczos methods for the solution of nonsymmetric systems of linear equations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 926–943.

[19] C. LANCZOS, *An iterative method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. Nat. Bur. Standards, 45 (1950), pp. 255–282.

[20] ———, *Solution of systems of linear equations by minimized iterations*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 33–53.

[21] J. G. LEWIS AND D. J. PIERCE, *Recent research in iterative methods at Boeing*, Comput. Phys. Comm., 53 (1989), pp. 213–221.

[22] T. A. MANTEUFFEL, *The Tchebychev iteration for nonsymmetric linear systems*, Numer. Math., 28 (1977), pp. 307–327.

[23] J. A. MEIJERINK AND H. A. VAN DER VORST, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. Comp., 31 (1977), pp. 148–162.

[24] R. B. MORGAN, *Computing interior eigenvalues of large matrices*, Linear Algebra Appl., 154-156 (1991), pp. 289–309.

[25] R. B. MORGAN AND M. ZENG, *Estimates for interior eigenvalues of large nonsymmetric matrices*, submitted.

[26] N. M. NACHTIGAL, S. C. REDDY, AND L. N. TREFETHEN, *How fast are nonsymmetric matrix iterations?*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 778–795.

[27] A. NAVARRA, *An application of* GMRES *to indefinite linear problems in meteorology*, Comput. Phys. Comm., 53 (1989), pp. 321–327.

[28] C. C. PAIGE AND M. A. SAUNDERS, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.

[29] B. N. PARLETT, D. R. TAYLOR, AND Z. A. LIU, *A look-ahead Lanczos algorithm for unsymmetric matrices*, Math. Comp., 44 (1985), pp. 105–124.

[30] Y. SAAD, *Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices*, Linear Algebra Appl., 34 (1980), pp. 269–295.

[31] Y. SAAD, *Krylov subspace methods for solving large unsymmetric linear systems*, Math. Comp., 37 (1981), pp. 105–126.

[32] Y. SAAD AND M. H. SCHULTZ, GMRES: *a generalized minimum residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[33] B. T. SMITH, J. M. BOYLE, Y. IKEBE, B. C. KLEMA, AND C. B. MOLER, *Matrix Eigensystems Routines: EISPACK Guide*, 2nd ed., Springer-Verlag, New York, NY, 1970.

[34] P. SONNEVELD, CGS, *a fast Lanczos-type solver for nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 36–52.

[35] A. VAN DER SLUIS AND H. A. VAN DER VORST, *The rate of convergence of conjugate gradients*, Numer. Math., 48 (1986), pp. 543–560.

[36] H. A. VAN DER VORST, Bi-CGSTAB: *A fast and smoothly converging variant of* Bi-CG *for the solution of nonsymmetric linear equations*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 631–644.

[37] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, p. 1172, October 1995

010

COMMENTS ON LARGE LEAST SQUARES PROBLEMS INVOLVING KRONECKER PRODUCTS*

HONGYUAN ZHA†

Abstract. In this note we point out that the least squares solution of $(A \otimes B)x = t$ can be obtained numerically without referring to Kronecker products.

Key words. linear least squares problem, Kronecker product

AMS subject classification. 65F15

In [1], least squares solution of $(A \otimes B)x = t$ is considered and a QR decomposition based method is developed using the machinery of Kronecker products. The paper also contains many interesting properties of matrices in Kronecker product form. However, we want to point out that the method in [1] can be derived in a more direct and concise way without even mentioning Kronecker products.

It is easy to see that the least squares problem is equivalent to

$$\min_X \ \|T - BXA^T\|_F, \tag{1}$$

where X and T is given in (1.6) and (1.7) in [1]. The minimal F-norm solution of (1) are given by $X = B^+T(A^+)^T$ as is pointed out in [1], where $^+$ denotes Moore–Penrose inverse. If the QR decompositions of A and B are available (assuming A and B are of full column rank), i.e.,

$$AP_2 = Q_A \begin{pmatrix} R_A \\ 0 \end{pmatrix}, \quad BP_1 = Q_B \begin{pmatrix} R_B \\ 0 \end{pmatrix},$$

where P_1 and P_2 are permutation matrices. Let

$$Q_B^T T Q_A = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix},$$

where T_{11} has the same row dimension as R_B and same column dimension as R_A. Then

$$X = P_1 R_B^{-1} T_{11} R_A^{-T} P_2.$$

If the SVD of A and B are available, similar procedure can be used.

Remark. A more complicated least squares problem that might benefit from a Kronecker product formulation is the following, $\min_X \|T - B_1 X A_1^T - \cdots - B_k X A_k^T\|_F$, which is easily seen to be equivalent to $\min_x \|t - (A_1 \otimes B_1 + \cdots + A_k \otimes B_k)x\|_F$. It will be interesting to see how the Kronecker product structures can be explored in the above least squares problem.

REFERENCES

[1] D.W. FAUSETT AND C.T. FULTON, *Large least squares problems involving Kronecker products*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 219–227.

* Received by the editors March 9, 1994; accepted for publication (in revised form) by C. Van Loan October 11, 1994.

† Department of Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802 (zha@cse.psu.edu). The work of this author was supported in part by National Science Foundation grant CCR-9308399.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, pp. 1173–1183, October 1995

011

TRACE AND EIGENVALUE INEQUALITIES FOR ORDINARY AND HADAMARD PRODUCTS OF POSITIVE SEMIDEFINITE HERMITIAN MATRICES*

BO-YING WANG† AND FUZHEN ZHANG‡

Dr. Zhang dedicates this paper to his mother who passed away on February 5, 1994.

Abstract. Let A and B be $n \times n$ positive semidefinite Hermitian matrices, let α and β be real numbers, let $\circ$ denote the Hadamard product of matrices, and let A_k denote any $k \times k$ principal submatrix of A. The following trace and eigenvalue inequalities are shown:

$$\operatorname{tr}(A \circ B)^\alpha \le \operatorname{tr}(A^\alpha \circ B^\alpha), \quad \alpha \le 0 \text{ or } \alpha \ge 1,$$

$$\operatorname{tr}(A \circ B)^\alpha \ge \operatorname{tr}(A^\alpha \circ B^\alpha), \quad 0 \le \alpha \le 1,$$

$$\lambda^{1/\alpha}(A^\alpha \circ B^\alpha) \le \lambda^{1/\beta}(A^\beta \circ B^\beta), \quad \alpha \le \beta, \alpha\beta \ne 0,$$

$$\lambda^{1/\alpha}[(A^\alpha)_k] \le \lambda^{1/\beta}[(A^\beta)_k], \quad \alpha \le \beta, \alpha\beta \ne 0.$$

The equalities corresponding to the inequalities above and the known inequalities

$$\operatorname{tr}(AB)^\alpha \le \operatorname{tr}(A^\alpha B^\alpha), \quad |\alpha| \ge 1,$$

and

$$\operatorname{tr}(AB)^\alpha \ge \operatorname{tr}(A^\alpha B^\alpha), \quad |\alpha| \le 1$$

are thoroughly discussed. Some applications are given.

Key words. trace inequality, eigenvalue inequality, Hadamard product, Kronecker product, Schur-convex function, majorization

AMS subject classifications. 15A18, 15A39, 15A42, 15A45

1. Introduction.

Let A be an $n \times n$ complex matrix. We denote $\lambda(A) = (\lambda_1(A), \ldots, \lambda_n(A))$, where the $\lambda_i(A)$'s are the eigenvalues of A; furthermore, we arrange $\lambda_1(A) \ge \cdots \ge \lambda_n(A)$ if they are all real. As usual, $A \circ B = (a_{ij}b_{ij})$ is the Hadamard (entrywise or Schur) product of A and B when A and B are of the same size. For real vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ with components in decreasing order, we write $x \le y$ if $x_i \le y_i$, $i = 1, \ldots, n$; $x \prec_w y$ if x is weakly majorized by y, i.e., $\sum_{i=1}^k x_i \le \sum_{i=1}^k y_i$, $k = 1, \ldots, n$; and $x \prec y$ if $x \prec_w y$ and $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$.

For any scalar α and any $n \times n$ diagonalizable matrix A with spectral decomposition $A = UDU^*$, where $D = \operatorname{diag}\{\lambda_1(A), \ldots, \lambda_n(A)\}$ and U is unitary, we define (for more general definition, see [HJ, p. 411])

$$A^\alpha = UD^\alpha U^* = U\operatorname{diag}\{(\lambda_1(A))^\alpha, \ldots, (\lambda_n(A))^\alpha\}U^*$$

whenever all the $(\lambda_i(A))^\alpha$'s make sense, and denote

$$\lambda^\alpha(A) = (\lambda(A))^\alpha = \lambda(A^\alpha).$$

* Received by the editors August 12, 1993; accepted for publication (in revised form) by R. Horn October 28, 1994.

† Department of Mathematics, Beijing Normal University, Beijing 100875, China. The work of this author was supported by a National Science Foundation grant of China.

‡ Department of Mathematics, Nova Southeastern University, Fort Lauderdale, Florida 33314 (zhang@alpha.nova.edu).

We write $A \geq 0$ if A is a positive semidefinite Hermitian matrix, and $A \geq B$ if A and B are Hermitian and $A - B \geq 0$. Throughout this paper we assume that $A \geq 0$, $B \geq 0$, α and β are positive numbers unless A and B are both positive definite, in which case α and β can be any real numbers, and m is a positive integer. It is well known [HH, Corollary 2.3] that the product of two positive semidefinite Hermitian matrices is diagonalizable and has nonnegative eigenvalues.

While studying the moments of the eigenvalues of Schrödinger Hamiltonians in quantum mechanics, Lieb and Thirring [LT] first showed (in the setting of operators on a separable Hilbert space) that

$$\operatorname{tr}(AB)^\alpha \leq \operatorname{tr}(A^\alpha B^\alpha) \tag{1}$$

for any real number $\alpha \geq 1$.

The inequalities in (1) were extended to unbounded operators by Araki [Ar]. Upper and lower bounds for $\operatorname{tr}(AB)^m$ and $\operatorname{tr}(A^m B^m)$ when m is a positive integer were obtained by Marcus [M], Le Couteur [C], and proved again by Bushell and Trustrum [BT]:

$$\sum_{i=1}^{n} \lambda_i^m(A)\lambda_{n-i+1}^m(B) \leq \operatorname{tr}(AB)^m \leq \operatorname{tr}(A^m B^m) \leq \sum_{i=1}^{n} \lambda_i^m(A)\lambda_i^m(B). \tag{2}$$

In a recent paper, Wang and Gong [WG] generalized the above results in terms of majorization, and proved

$$\log \lambda^{1/\alpha}(A^\alpha B^\alpha) \prec \log \lambda^{1/\beta}(A^\beta B^\beta), \quad 0 < \alpha < \beta,$$

as consequences

$$\lambda^{1/\alpha}(A^\alpha B^\alpha) \prec_w \lambda^{1/\beta}(A^\beta B^\beta), \quad 0 < \alpha \leq \beta, \tag{3}$$

$$\lambda^{1/\beta}(A^\beta B^\beta) \prec_w \lambda^{1/\alpha}(A^\alpha B^\alpha), \quad \alpha \leq \beta < 0, \tag{4}$$

$$\lambda^\alpha(AB) \prec_w \lambda(A^\alpha B^\alpha), \quad |\alpha| \geq 1, \tag{5}$$

and

$$\lambda(A^\alpha B^\alpha) \prec_w \lambda^\alpha(AB), \quad |\alpha| \leq 1. \tag{6}$$

We are concerned with analogues of these inequalities for the entrywise product. A simple example shows that an analogue of (2)

$$\sum_{i=1}^{n} \lambda_i^m(A)\lambda_{n-i+1}^m(B) \leq \operatorname{tr}(A \circ B)^m$$

does not hold in general: take $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, and $m = 2$. However, the inequality

$$\operatorname{tr}(A^m \circ B^m) \leq \sum_{i=1}^{n} \lambda_i^m(A)\lambda_i^m(B)$$

is valid, due to the majorization (see, e.g., [H, p. 146], [BS], or [Z])

$$\lambda(A \circ B) \prec_w \lambda(A) \circ \lambda(B), \quad \text{whenever } A, B \geq 0.$$

Consequently, we always have

$$\lambda(A^m \circ B^m) \prec_w \lambda(A^m) \circ \lambda(B^m).$$

It will be seen shortly that

$$\lambda^{1/\alpha}(A^\alpha \circ B^\alpha) \leq \lambda^{1/\beta}(A^\beta \circ B^\beta) \tag{7}$$

for any nonzero real numbers α and β such that $\alpha \leq \beta$. In particular

$$\lambda^m(A \circ B) \leq \lambda(A^m \circ B^m), \quad m = 1, 2, \ldots.$$

In this paper we first give necessary and sufficient conditions for equalities in (1), (5), and (6) to hold, then show some eigenvalue inequalities for principal submatrices and matrix powers. Finally we discuss an analogue of the Lieb–Thirring inequality (1) for the Hadamard product and present some applications.

2. Trace inequalities for ordinary product. This section is devoted to the discussion of the Lieb–Thirring inequality (1) and majorizations (5) and (6). Necessary and sufficient conditions for trace equalities to hold, i.e., for $\prec_w$ in (5) and (6) to become $\prec$, are given.

In the following (and thereafter), A and B are automatically understood to be positive definite when α (or β) is negative or equal to 0.

THEOREM 2.1. *Let A and B be positive semidefinite Hermitian matrices. Then*

$$\operatorname{tr}(AB)^\alpha \leq \operatorname{tr}(A^\alpha B^\alpha), \quad \textit{whenever } |\alpha| \geq 1, \tag{8}$$

and

$$\operatorname{tr}(AB)^\alpha \geq \operatorname{tr}(A^\alpha B^\alpha), \quad \textit{whenever } |\alpha| \leq 1. \tag{9}$$

Equality holds for some value of α if and only if $\alpha = -1, 0, 1$, or $AB = BA$.

Proof. The inequalities follow from (5) and (6) which have appeared in [WG]. We need consider only the equality case. Sufficiency is obvious if one recalls that A and B are simultaneously unitarily diagonalizable when A and B are normal and commute. To prove necessity, noticing that $\operatorname{tr}(AB)^\alpha = \operatorname{tr}(A^{-1}B^{-1})^{-\alpha}$ when $\alpha < 0$, we may assume that

$$\operatorname{tr}(AB)^\alpha = \operatorname{tr}(A^\alpha B^\alpha), \quad \text{for some } \alpha > 0, \alpha \neq 1,$$

and break down the proof into cases (a) $\alpha \geq 2$, (b) $1 < \alpha < 2$, and (c) $0 < \alpha < 1$. Equality holds trivially when $\alpha = 0$, A and B are nonsingular.

(a) $\alpha \geq 2$. In this case we claim that $\operatorname{tr}(AB)^\alpha = \operatorname{tr}(A^\alpha B^\alpha)$ implies that $AB = BA$.

If $\alpha = 2$, i.e., $\operatorname{tr}(AB)^2 = \operatorname{tr}(A^2B^2)$, we assume, without loss of generality, that A is a diagonal matrix with diagonal entries $a_1, \ldots, a_n$. Then

$$\operatorname{tr}(A^2B^2) - \operatorname{tr}(AB)^2 = \sum_{i,j} a_i^2 |b_{ij}|^2 - \sum_{i,j} a_i a_j |b_{ij}|^2 = \sum_{i<j} (a_i - a_j)^2 |b_{ij}|^2 = 0.$$

Thus $a_i b_{ij} = a_j b_{ij}$ for every pair of i and j, i.e., $AB = BA$.

For $\alpha > 2$, we show that $\operatorname{tr}(AB)^\alpha = \operatorname{tr}(A^\alpha B^\alpha)$ implies $\operatorname{tr}(AB)^2 = \operatorname{tr}(A^2B^2)$, which leads to $AB = BA$, as we have just seen.

If $\operatorname{tr}(AB)^2 \neq \operatorname{tr}(A^2B^2)$, we apply the strictly increasing and strictly Schur-convex function (see [MO, p. 60, A.8.a]) $\sum_{i=1}^n t_i^{\alpha/2}$ to the weak majorization $\lambda^2(AB) \prec_w \lambda(A^2B^2)$, and get

$$\operatorname{tr}(AB)^\alpha < \sum_{i=1}^n \lambda_i^{\alpha/2}(A^2B^2) \leq \sum_{i=1}^n \lambda_i(A^\alpha B^\alpha) = \operatorname{tr}(A^\alpha B^\alpha),$$

where the last inequality follows from (5), a contradiction.

(b) $1 < \alpha < 2$. In this case we claim that

$$\operatorname{tr}(AB)^x = \operatorname{tr}(A^xB^x) \quad \text{for all } 1 < x < \alpha.$$

In fact, if $\operatorname{tr}(AB)^{x_0} \neq \operatorname{tr}(A^{x_0}B^{x_0})$ for some x_0 and $1 < x_0 < \alpha$, applying the strictly increasing and strictly Schur-convex function $\sum_{i=1}^n t_i^{\alpha/x_0}$ to $\lambda^{x_0}(AB) \prec_w \lambda(A^{x_0}B^{x_0})$, where $\prec_w$ is strict, we have

$$\begin{aligned}
\operatorname{tr}(AB)^\alpha &= \sum_{i=1}^n \lambda_i^\alpha(AB) \\
&< \sum_{i=1}^n \lambda_i^{\alpha/x_0}(A^{x_0}B^{x_0}) \\
&\leq \sum_{i=1}^n \lambda_i(A^\alpha B^\alpha), \quad \text{(use (5))}
\end{aligned}$$

a contradiction. Thus $\operatorname{tr}(AB)^x - \operatorname{tr}(A^xB^x)$ is identically zero for $1 < x < \alpha$.

Now expanding $\operatorname{tr}(AB)^x - \operatorname{tr}(A^xB^x)$ as a series of x and using the fact [Co, pp. 31, 78] that if a series converges to zero on an open interval, then it converges to zero on the whole real number line, we have $\operatorname{tr}(AB)^x - \operatorname{tr}(A^xB^x) = 0$, that is, $\operatorname{tr}(AB)^x = \operatorname{tr}(A^xB^x)$ for all real $x > 0$, particularly for 2, thus $AB = BA$.

(c) $0 < \alpha < 1$. We show that

$$\operatorname{tr}(AB)^x = \operatorname{tr}(A^xB^x), \quad \text{for all } \alpha < x < 1.$$

Otherwise, $\operatorname{tr}(AB)^{x_0} > \operatorname{tr}(A^{x_0}B^{x_0})$ for some x_0 and $\alpha < x_0 < 1$. Applying the strictly increasing and strictly Schur-convex function $\sum_{i=1}^n e^{\frac{\alpha}{x_0}t_i}$ to $\log\lambda(A^{x_0}B^{x_0}) \prec \log\lambda^{x_0}(AB)$ (see [WG, Theorem 6]) when both of A and B are nonsingular, we have

$$\begin{aligned}
\operatorname{tr}(A^\alpha B^\alpha) &= \sum_{i=1}^n \lambda_i(A^\alpha B^\alpha) \\
&\leq \sum_{i=1}^n \lambda_i^{\alpha/x_0}(A^{x_0}B^{x_0}) \quad \text{(use (6))} \\
&< \sum_{i=1}^n [\lambda_i(AB)]^\alpha \\
&= \operatorname{tr}(AB)^\alpha,
\end{aligned}$$

a contradiction. Due to the same reason as in (b), $\operatorname{tr}(AB)^x = \operatorname{tr}(A^xB^x)$ for all real $x > 0$, thus $AB = BA$ when A and B are nonsingular. The singular case can be accomplished by the usual technique of continuity. □

Notice that when two normal matrices commute they are simultaneously unitarily diagonalizable. The corollary below is immediate.

COROLLARY 2.2. *Let A and B be positive semidefinite Hermitian matrices. If* $\operatorname{tr}(AB)^\alpha = \operatorname{tr}(A^\alpha B^\alpha)$, $\alpha \neq -1, 0, 1$, *then* $(AB)^\alpha = A^\alpha B^\alpha$.

3. Eigenvalue inequalities for principal submatrices. For an $n \times n$ matrix A, we use A_k to designate any $k \times k$ principal submatrix of A, $1 \leq k \leq n$. A result of Ando [A, Corollary 4.2] yields the following lemma when one notices that the map $A \to A_k$ is normalized positive linear (see [A] for the definition).

LEMMA 3.1. *Let A be an $n \times n$ positive semidefinite Hermitian matrix. Then*

$$A_k \leq [(A^\alpha)_k]^{1/\alpha}, \qquad 1 \leq \alpha < \infty \tag{10}$$

and

$$A_k \geq [(A^{-\alpha})_k]^{-1/\alpha}, \qquad 1 \leq \alpha < \infty. \tag{11}$$

The following theorem says that $\lambda^{1/x}[(A^x)_k]$ is a monotone vector-valued function of x.

THEOREM 3.2. *Let A be an $n \times n$ positive semidefinite Hermitian matrix. Then*

$$\lambda^{1/\alpha}[(A^\alpha)_k] \leq \lambda^{1/\beta}[(A^\beta)_k], \quad \textit{whenever } \alpha \leq \beta,\ \alpha\beta \neq 0, \tag{12}$$

with equality if and only if $\alpha = \beta$ or $A = P(A_k \oplus H)P^T$ for some $H \geq 0$ and some permutation matrix P.

Proof. For $0 < \alpha \leq 1$, using (10), we have

$$(A^\alpha)_k \leq [(A^{\alpha\frac{1}{\alpha}})_k]^\alpha = (A_k)^\alpha.$$

For $-1 \leq \alpha < 0$, using (11), we have

$$(A^\alpha)_k \geq [(A^{\alpha\frac{1}{\alpha}})_k]^\alpha = (A_k)^\alpha.$$

Thus

$$(A_k)^\alpha \geq (A^\alpha)_k, \qquad 0 < \alpha \leq 1, \tag{13}$$

and

$$(A_k)^\alpha \leq (A^\alpha)_k, \qquad -1 \leq \alpha < 0. \tag{14}$$

For $\alpha \leq \beta$ with the same sign, using (13), we get

$$(A^\beta)_k^{\alpha/\beta} \geq (A^\alpha)_k, \quad \text{when } 0 < \alpha/\beta \leq 1,$$

and

$$(A^\alpha)_k^{\beta/\alpha} \geq (A^\beta)_k, \quad \text{when } 0 < \beta/\alpha \leq 1,$$

in either case

$$\lambda^{1/\alpha}[(A^\alpha)_k] \leq \lambda^{1/\beta}[(A^\beta)_k].$$

If $\alpha \leq \beta$ with different signs, using (14),

$$(A^\alpha)_k^{\beta/\alpha} \leq (A^\beta)_k, \qquad \text{when } -1 \leq \beta/\alpha < 0,$$

and

$$(A^\beta)_k^{\alpha/\beta} \leq (A^\alpha)_k, \qquad \text{when } -1 \leq \alpha/\beta < 0,$$

in either case we have

$$\lambda^{1/\alpha}[(A^\alpha)_k] \leq \lambda^{1/\beta}[(A^\beta)_k].$$

Thus inequality (12) follows immediately.

Now we discuss the equality case in (12). Without loss of generality, we may assume that A_k lies in the upper-left corner of A, i.e., we partition A as $A = \begin{pmatrix} A_k & C \\ C^* & H \end{pmatrix}$, where H is some positive semidefinite Hermitian matrix. We first consider the case where $\alpha = 1$ or $\beta = 1$ and $\alpha \neq \beta$. Suppose

$$\lambda(A_k) = \lambda^{1/s}[(A^s)_k]$$

for some $s \neq 0, 1$. Then

$$\lambda(A_k) = \lambda^{1/x}[(A^x)_k]$$

for all $x \neq 0$ between s and 1, because of (12). Thus we can always find an interval I between s and 1 on the positive real number line, such that

$$\lambda^x(A_k) = \lambda(A^x)_k, \quad x \in I,$$

that is,

$$\operatorname{tr}(A_k)^x - \operatorname{tr}(A^x)_k = 0, \quad x \in I,$$

which is the same as

$$\operatorname{tr}(BA)^x - \operatorname{tr}(B^x A^x) = 0, \quad x \in I,$$

where $B = \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix}$. Hence $AB = BA$ by Theorem 1, which leads to $A = A_k \oplus H$ as required.

For general α and β with $\alpha < \beta$ and $\alpha\beta \neq 0$, if

$$\lambda^{1/\alpha}[(A^\alpha)_k] = \lambda^{1/\beta}[(A^\beta)_k],$$

we rewrite it as

$$\lambda[(A^\alpha)_k] = \lambda^{\alpha/\beta}[(A^\beta)_k] = \lambda^{\alpha/\beta}\{[(A^\alpha)^{\beta/\alpha}]_k\}.$$

The earlier argument yields $A^\alpha = (A^\alpha)_k \oplus \tilde{H}$ for some $\tilde{H} \geq 0$. Thus

$$A = (A^\alpha)^{1/\alpha} = [(A^\alpha)_k]^{1/\alpha} \oplus (\tilde{H})^{1/\alpha} = A_k \oplus H,$$

where $H = (\tilde{H})^{1/\alpha}$. □

COROLLARY 3.3. *If A is an $n \times n$ positive semidefinite Hermitian matrix, then*

$$\lambda^{\alpha}(A_k) \leq \lambda[(A^{\alpha})_k], \quad \alpha \leq 0 \quad \text{or} \quad 1 \leq \alpha,$$

and

$$\lambda^{\alpha}(A_k) \geq \lambda[(A^{\alpha})_k], \quad 0 < \alpha < 1,$$

with equality if and only if $\alpha = 0, 1$, or $A = P(A_k \oplus H)P^T$.

Lemma 3.1 and Theorem 3.2 yield the following corollary.

COROLLARY 3.4. *Let $A = \begin{pmatrix} A_1 & A_2 \\ A_2^* & A_3 \end{pmatrix} \geq 0$ be an $n \times n$ matrix, and write $A^{\alpha} = \begin{pmatrix} B_1 & B_2 \\ B_2^* & B_3 \end{pmatrix}$, where A_1 and B_1 are corresponding $k \times k$ principal submatrices of A and A^{α}, respectively. Then*

$$A_1 \leq B_1^{1/\alpha}, \quad \alpha \geq 1,$$

$$A_1^{\alpha} \geq B_1, \quad 0 < \alpha < 1,$$

$$A_1^{\alpha} \leq B_1, \quad -1 < \alpha < 0,$$

$$A_1 \geq B_1^{1/\alpha}, \quad \alpha \leq -1.$$

Equality in each case holds if and only if one of the following conditions is satisfied:

1. $\alpha = 1$;
2. $\operatorname{tr} A_1^{\alpha} = \operatorname{tr} B_1$;
3. $A_2 = B_2 = 0$, *i.e.*, $A = A_1 \oplus A_3$.

Moreover (2) *and* (3) *are equivalent when $\alpha \neq 0, 1$. Thus* (2) *is the same as $A_1^{\alpha} = B_1$ when $\alpha \neq 1$.*

A direct computation gives the inequality $(A_k)^2 \leq (A^2)_k$. However $(A_k)^3 \leq (A^3)_k$ does not hold in general, as the following example shows.

Take A to be the 4-by-4 matrix with (1,1)-entry 2 and 1 elsewhere, and $k = 2$. Then $(A^3)_2 - (A_2)^3 = \begin{pmatrix} 16 & 14 \\ 14 & 12 \end{pmatrix}$, which is not positive semidefinite.

It is well known that $A \circ B$ is the principal submatrix of the Kronecker product $A \otimes B$ lying in the intersections of rows and columns $1, n+2, \ldots, n^2$ of $A \otimes B$. Considering $A \otimes B$ in Theorem 3.2 in place of A and noticing that $(A \otimes B)^t = A^t \otimes B^t$ for any real number t, we have the following theorem.

THEOREM 3.5. *Let A and B be positive semidefinite Hermitian matrices. Then*

$$\lambda^{1/\alpha}(A^{\alpha} \circ B^{\alpha}) \leq \lambda^{1/\beta}(A^{\beta} \circ B^{\beta}), \quad \text{whenever } \alpha \leq \beta, \ \alpha\beta \neq 0. \tag{15}$$

It is immediate that for $A, B, \ldots, C$ positive semidefinite Hermitian matrices

$$\lambda^{1/\alpha}(A^{\alpha} \circ B^{\alpha} \circ \cdots \circ C^{\alpha}) \leq \lambda^{1/\beta}(A^{\beta} \circ B^{\beta} \circ \cdots \circ C^{\beta}), \quad \alpha \leq \beta, \ \alpha\beta \neq 0.$$

Taking $\beta = 1, \alpha = 1$, and $\beta = 1$ in Theorem 3.5, respectively, we get the following corollary.

COROLLARY 3.6. *Let $A, B \geq 0$. Then*

$$\lambda^{\alpha}(A \circ B) \leq \lambda(A^{\alpha} \circ B^{\alpha}), \quad \alpha \leq 0 \text{ or } \alpha \geq 1,$$

and

$$\lambda^{\alpha}(A \circ B) \geq \lambda(A^{\alpha} \circ B^{\alpha}), \quad 0 < \alpha < 1.$$

It is noted in §4 that equality holds in (15) or Corollary 3.6 if and only if A and B have the structures described in Theorem 4.1 or $\alpha = \beta \neq 0$ in (15) , or if $\alpha = 0, 1$ in Corollary 3.6.

4. Trace inequalities for Hadamard product. The following is an analogue of Theorem 2.1 for the Hadamard product.

THEOREM 4.1. *Let $A, B \geq 0$. Then for any real number α*

$$\operatorname{tr}(A \circ B)^{\alpha} \leq \operatorname{tr}(A^{\alpha} \circ B^{\alpha}), \quad \text{if } \alpha \leq 0 \quad \text{or} \quad 1 \leq \alpha, \tag{16}$$

and

$$\operatorname{tr}(A \circ B)^{\alpha} \geq \operatorname{tr}(A^{\alpha} \circ B^{\alpha}), \quad \text{if } 0 < \alpha < 1. \tag{17}$$

Equality occurs if and only if one of the following conditions is satisfied:

(i) $\alpha = 0$ *or* 1;

(ii) $(A \circ B)^{\alpha} = A^{\alpha} \circ B^{\alpha}$;

(iii) *there exists a permutation matrix P such that*

$$A \otimes B = P[(A \circ B) \oplus H]P^{T}$$

for some $H \geq 0$;

(iv) *there exists a permutation matrix P such that $PAP^{T} = D_A \oplus 0 \oplus \tilde{A}$ and $PBP^{T} = D_B \oplus \tilde{B} \oplus 0$,where D_A and D_B are invertible diagonal matrices of the same size, $\tilde{A}$ and $\tilde{B}$ are positive semidefinite Hermitian matrices each with the same size as 0 in the other direct sum*;

(v) $(A \circ B)(X \circ Y) = (AX) \circ (BY)$ *for all $n \times m$ matrices X and Y, where m is an integer.*

Moreover, (ii), (iii), (iv) *and* (v) *are equivalent when $\alpha \neq 0, 1$.*

Proof. The trace inequalities (16) and (17) follow from Corollary 3.6. We need discuss only the equality case. We assume $\alpha \neq 0, 1$, and show that "equality"$\Leftrightarrow$(ii)$\Leftrightarrow$(iii)$\Rightarrow$(iv)$\Rightarrow$(v), (iv)$\Rightarrow$(ii), and (v)$\Rightarrow$(iv).

Consider the Kronecker product $(A \otimes B)^{\alpha} = A^{\alpha} \otimes B^{\alpha}$ and note that $A^{\alpha} \circ B^{\alpha}$ is a principal submatrix of $A^{\alpha} \otimes B^{\alpha}$, consequently of $(A \otimes B)^{\alpha}$, lying in the same position as $A \circ B$ does in $A \otimes B$. If $\operatorname{tr}(A \circ B)^{\alpha} = \operatorname{tr}(A^{\alpha} \circ B^{\alpha})$, then (ii), equivalently (iii), results from Corollary 3.4. To obtain (iv), we notice that for any permutation matrix Q

$$\operatorname{tr}Q(A \circ B)^{\alpha}Q^{T} = \operatorname{tr}Q(A^{\alpha} \circ B^{\alpha})Q^{T}$$

and

$$\operatorname{tr}(QAQ^{T} \circ QBQ^{T})^{\alpha} = \operatorname{tr}(QAQ^{T})^{\alpha} \circ (QBQ^{T})^{\alpha}.$$

Thus we may assume $b_{11} \neq 0$ if $B \neq 0$ and consider the first row of $A \otimes B = (a_{ij}B)$. $a_{11}b_{11}$ appears in $A \circ B$; for $j > 1$, $a_{1j}b_{11}$ lies on none of columns $1, n+2, \ldots, n^2$. In other words, if $R(A \otimes B)R^{T} = \begin{pmatrix} A \circ B & A_2 \\ A_2^{*} & A_3 \end{pmatrix}$ for some permutation matrix R, then

$a_{1j}b_{11}$ is contained in A_2. Applying Corollary 3.4 or by (iii), we have $A_2 = 0$. Hence $a_{1j}b_{11} = 0$, and $a_{1j} = 0 = \overline{a_{j1}}$ for $j > 1$. Interchanging the roles of A and B, we obtain $b_{1j} = b_{j1} = 0$ for $j > 1$ if $a_{11} \neq 0$. Repeating the argument for all $b_{ii} \neq 0$, we see that for some permutation matrix S

$$SBS^T = \begin{pmatrix} b'_1 & * & \cdots & * & & \\ * & \ddots & \ddots & \vdots & & \\ \vdots & \ddots & \ddots & * & & \\ * & \cdots & * & b'_k & & \\ & & & & 0_{n-k} & \end{pmatrix}$$

and

$$SAS^T = \begin{pmatrix} a'_1 & 0 & \cdots & 0 & & \\ 0 & \ddots & \ddots & \vdots & & \\ \vdots & \ddots & \ddots & 0 & & \\ 0 & \cdots & 0 & a'_k & & \\ & & & & \tilde{A}_{n-k} & \end{pmatrix},$$

where $b'_1, \ldots, b'_k$ are the nonzero b_{ii}'s and $\tilde{A}_{n-k}$ is an $(n-k)$-square positive semidefinite Hermitian matrix. Let $a_1, \ldots, a_s$ be those of $a'_1, \ldots, a'_k$ which are nonzero, then we have a permutation matrix P such that

$$PAP^T = \begin{pmatrix} a_1 & & 0 & & & \\ & \ddots & & & & \\ 0 & & a_s & & & \\ & & & 0_t & & \\ & & & & \tilde{A}_{n-s-t} & \end{pmatrix}$$

and

$$PBP^T = \begin{pmatrix} b_1 & & 0 & & & \\ & \ddots & & & & \\ 0 & & b_s & & & \\ & & & \tilde{B}_t & & \\ & & & & 0_{n-s-t} & \end{pmatrix},$$

where $b_1, \ldots, b_s$ are not equal to zero. (iv) follows. Thus we have proved the implications "equality"$\Leftrightarrow$(ii)$\Leftrightarrow$(iii)$\Rightarrow$(iv).

Direct computations give (iv)$\Rightarrow$(ii) and (iv)$\Rightarrow$(v). To see (v)$\Rightarrow$(iv), we take $X = A$ and $Y = B$ in (v). Then $(A \circ B)^2 = A^2 \circ B^2$ which results in (iv), as seen. □

Going back to Theorem 3.5, we see equality in (15) holds, that is,

$$\lambda(A^\alpha \circ B^\alpha) = \lambda^{\alpha/\beta}[(A^\alpha)^{\beta/\alpha} \circ (B^\alpha)^{\beta/\alpha}],$$

if and only if either $\alpha = \beta \neq 0$ or A and B have the structures described in the previous theorem, by applying Theorem 4.1 to A^α and B^α.

5. Applications. The Lieb–Thirring inequality (1) may be investigated for a variety of real-valued matrix functions in the place of the trace function. We consider, as an example, the matrix function—sum of principal minors. Let $E_k(X)$ denote the sum of all the $\binom{n}{k}$ k-square principal minors of the $n \times n$ matrix X, let $E_k(x)$ denote the kth elementary symmetric function of the row vector x, and let $C_k(X)$ denote the kth compound matrix of X. Then (see [MM, pp.18, 24])

$$E_k(X) = \mathrm{tr}C_k(X) = E_k(\lambda(X)). \tag{18}$$

THEOREM 5.1. *Let A and B be positive semidefinite Hermitian matrices. Then*

$$E_k(AB)^\alpha \le E_k(A^\alpha B^\alpha), \quad |\alpha| \ge 1, \tag{19}$$

$$E_k(AB)^\alpha \ge E_k(A^\alpha B^\alpha), \quad |\alpha| \le 1, \tag{20}$$

$$E_k(A \circ B)^\alpha \le E_k(A^\alpha \circ B^\alpha), \quad \alpha \le 0 \quad \textit{or} \quad 1 \le \alpha, \tag{21}$$

and

$$E_k(A \circ B)^\alpha \ge E_k(A^\alpha \circ B^\alpha), \quad 0 \le \alpha \le 1. \tag{22}$$

Equality holds in (19) *or* (20) *if and only if $\alpha = -1, 0, 1$ or the kth compound matrices of A and B commute, and equality holds in* (21) *or* (22) *if and only if $\alpha = 0, 1$, the rank of $A^\alpha \circ B^\alpha$ is less than k, or A and B have the structures described in Theorem* 4.1.

Proof. Noting that $C_k(XY) = C_k(X)C_k(Y))$ and applying (18), we have for $|\alpha| > 1$,

$$\begin{aligned} E_k(AB)^\alpha &= \mathrm{tr}C_k(AB)^\alpha \\ &= \mathrm{tr}(C_k(A)C_k(B))^\alpha \\ &\le \mathrm{tr}(C_k(A))^\alpha (C_k(B))^\alpha \quad \text{(by Theorem 1)} \\ &= E_k(A^\alpha B^\alpha). \end{aligned}$$

Equality holds if and only if $C_k(A)C_k(B) = C_k(B)C_k(A)$. The inequality is reversed when $|\alpha| \le 1$.

For the case of the entrywise product and $\alpha \le 0$ or $1 \le \alpha$, we have

$$\begin{aligned} E_k(A \circ B)^\alpha &= E_k(\lambda^\alpha(A \circ B)) \\ &\le E_k(\lambda(A^\alpha \circ B^\alpha)) \quad \text{(by Corollary 3.3)} \\ &= E_k(A^\alpha \circ B^\alpha). \end{aligned}$$

Equality occurs if and only if either $\lambda^\alpha(A \circ B) = \lambda(A^\alpha \circ B^\alpha)$ or each term of $E_k(\lambda(A^\alpha \circ B^\alpha))$ vanishes. The former results in the structures of A and B given in Theorem 4.1 when $\alpha \ne 0, 1$, and the latter is equivalent to $\lambda(A^\alpha \circ B^\alpha)$ containing at least $n - k + 1$ zeros, that is, to $\mathrm{rank}(A^\alpha \circ B^\alpha) < k$. The case $0 \le \alpha \le 1$ is similarly discussed. □

Remark 1. Theorems 2.1 and 4.1 are obtained if one takes $k = 1$ in the previous theorem. If $k = n$, then (19) is the identity $\det(AB)^\alpha = \det(A^\alpha B^\alpha)$, and (21)

becomes $\det(A \circ B)^\alpha \leq \det(A^\alpha \circ B^\alpha)$, both sides of which vanish when one of A and B is singular, since $\operatorname{rank}(A^\alpha \circ B^\alpha) \leq \operatorname{rank}(A^\alpha \otimes B^\alpha) = \operatorname{rank}(A)\operatorname{rank}(B)$.

Remark 2. Regarding Theorem 3.5, we can also prove, by using a result of Ando [A, Theorems 10 and 11], that for $A, B \geq 0$,

$$(A^\alpha \circ B^\alpha)^{1/\alpha} \leq (A^\beta \circ B^\beta)^{1/\beta}, \quad \alpha \leq \beta \leq -1 \quad \text{or} \quad 1 \leq \alpha \leq \beta.$$

The inequality above does not hold for all $\alpha \leq \beta, \alpha\beta \neq 0$, as the following example shows:

Take $\alpha = 1/3, \beta = 1, A = B = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}^3$. Then $(A^{1/3} \circ B^{1/3})^3 \not\leq A \circ B$, since

$$\det[A \circ B - (A^{1/3} \circ B^{1/3})^3] = \det[\begin{pmatrix} 169 & 64 \\ 64 & 25 \end{pmatrix} - \begin{pmatrix} 73 & 22 \\ 22 & 7 \end{pmatrix}] = -36 < 0.$$

In general, $(A \circ B)^3 \not\leq A^3 \circ B^3$. However, the inequality $(A \circ B)^2 \leq A^2 \circ B^2$ holds, as seen in [A], [H], or [Z].

Acknowledgments. Dr. Zhang wishes to thank Professor E. H. Lieb for drawing his attention to [Ar] and Professor R. A. Horn and the referee for helpful suggestions and valuable comments.

REFERENCES

[A] T. Ando, *Concavity of certain maps on positive definite matrices and applications to Hadamard products,* Linear Algebra Appl., 26(1979), pp. 203–241.

[Ar] H. Araki, *On an inequality of Lieb and Thirring,* Lett. in Math. Physics, 19(1990), pp. 167–170.

[BS] R. B. Bapat and V. S. Sunder, *On majorization and Schur products,* Linear Algebra Appl., 72(1985), pp. 107–117.

[BT] P. J. Bushell and G. B. Trustrum, *Trace inequalities for positive definite matrix power products,* Linear Algebra Appl., 132(1990), pp. 173–178.

[C] K. J. Le Couteur, *Representation of the function $Tr(\exp(A - \lambda B))$ as a Lapalace transform with positive weight and some matrix inequalities,* J. Phys. A, 13(1980), pp. 3147–3159.

[Co] J. B. Conway, *Functions of One Complex Variable,* Second Edition, Springer-Verlag, New York, 1978.

[H] R. A. Horn, *The Hadamard product,* Proc. of Symposia in Applied Mathematics, Vol. 40, American Mathematical Society, Providence, 1990.

[HH] Y. Hong and R. A. Horn, *The Jordan canonical form of a product of a Hermitian and a positive semidefinite matrix,* Linear Algebra Appl., 147(1991), pp. 373–386.

[HJ] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis,* Cambridge University Press, New York, 1991.

[LT] E. H. Lieb and W. Thirring, *Studies in Mathematical Physics,* Essays in Honor of Valentine Bartmann, Princeton University Press, Princeton, NJ, 1976.

[M] M. Marcus, *An eigenvalue inequality for the product of normal matrices,* Amer. Math. Monthly, 63(1956), pp. 173–174.

[MM] M. Marcus and H. Minc, *A Survey of Matrix Theory and Matrix Inequalities,* Dover Books, New York, 1992.

[MO] A. W. Marshall and I. Olkin, *Inequalities: Theory of majorization and its applications,* Academic Press, San Diego, CA, 1979.

[WG] B.-Y. Wang and M.-P. Gong, *Some eigenvalue inequalities for positive semidefinite matrix power products,* Linear Algebra Appl., 184(1993), pp. 249–260.

[Z] F. Zhang, *Another proof of a singular value inequality concerning Hadamard products,* Linear Multilinear Algebra, 22(1988), pp. 307–311.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, pp. 1184–1196, October 1995

012

A BASIS-KERNEL REPRESENTATION OF ORTHOGONAL MATRICES*

XIAOBAI SUN† AND CHRISTIAN BISCHOF†

Abstract. In this paper we introduce a new representation of orthogonal matrices. We show that any orthogonal matrix can be represented in the form $Q = I - YSY^T$, which we call the basis-kernel representation of Q. In particular, we point out that the kernel S can be chosen to be triangular and that a familiar representation of an orthogonal matrix as a product of Householder matrices can be readily deduced from a basis-kernel representation with triangular kernel. We also show that there exists, in some sense, a minimal orthogonal transformation between two subspaces of same dimension, an important application of which is on block elimination problems. We explore how the basis Y determines the subspaces that Q acts on in a nontrivial fashion, and how S determines the way Q acts on this subspace. Especially, there is a canonical representation that explicitly shows that Q partitions $\mathbf{R}^n$ into three invariant subspaces in which it acts as the identity, a reflector, and a rotator, respectively. We also present a generalized Cayley representation for arbitrary orthogonal matrices, which illuminates the degrees of freedom we have in choosing orthogonal matrices acting on a predetermined subspace.

Key words. orthogonal matrices, block elimination, orthogonality condition, basis-kernel representation, Cayley transform, householder matrices

AMS subject classifications. 15A04, 15A21, 15A23, 65F25

1. Introduction.

Orthogonal transformations are a well-known tool in numerical linear algebra and are used extensively in decompositions such as the QR factorization, tridiagonalization, bidiagonalization, Hessenberg reduction, or the eigenvalue or singular value decomposition of a matrix (see, for example, [7], [11]). The orthogonal transformations employed are usually compositions of the following elementary transformations.

Givens rotator.

$$G = G(\theta) = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}. \tag{1}$$

In the two-dimensional plane, application of G to a vector x amounts to a clockwise rotation of x by an angle of θ.

Jacobi reflector.

$$J = J(\theta) = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{pmatrix}. \tag{2}$$

In the two-dimensional plane, application of J to a vector x amounts to reflecting x with respect to the line spanned by the vector

$$(\cos(\theta/2), \sin(\theta/2))^{\mathrm{T}}.$$

Householder reflector.

$$H = H(v) = I - \beta v v^{\mathrm{T}}, \qquad \beta v^{\mathrm{T}} v \beta = 2\beta. \tag{3}$$

* Received by the editors March 16, 1992; accepted for publication (in revised form) by C. Van Loan December 7, 1994. This work was supported by the Applied and Computational Mathematics Program, Advanced Research Projects Agency contract DM28E04120, and by the Office of Scientific Computing, U.S. Department of Energy Contract W-31-109-Eng-38. This paper is PRISM Working Note 19, available via anonymous ftp to `ftp.super.org` in the directory pub/prism.

† Mathematics and Computer Science Division, Argonne National Laboratory, 9700 South Case Avenue, Argonne, Illinois 60439 (`xiaobai, bischof@mcs.anl.gov`).

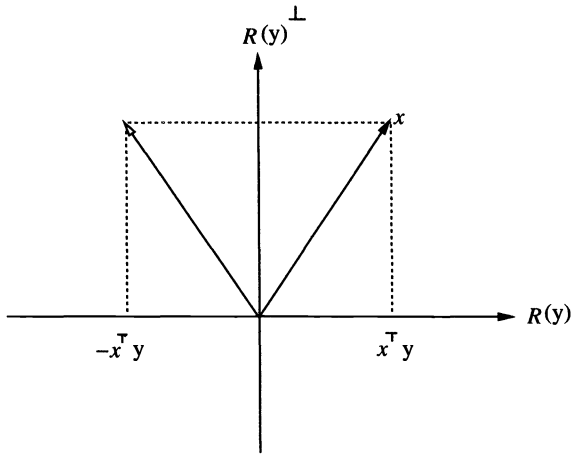


FIG. 1. *Reflectors.*

This representation of Householder matrices is used in the LINPACK [5] and LAPACK [1] libraries. The condition on v and β in (3) covers all choices for v and β that result in an orthogonal matrix H. In particular, it includes the degenerate case $\beta = 0$ where H is the identity matrix I. Note that the application of H to a vector x amounts to a reflection of x with respect to the hyperplane $\mathcal{R}(v)^{\perp}$, the orthogonal complement of the range $\mathcal{R}(v)$.

Each of the three well-known elementary transformations, when applied to a matrix, implies a low-rank (rank 1 or 2) update of the matrix.

Givens rotators form a group under matrix multiplication with the identity matrix as the unit element of the group; in particular, the product of any two Givens rotators is again a Givens rotator. Note that unless $\theta = 0 \bmod 2\pi$, $G(\theta)$ has no eigenvalue at 1. That is, except for the identity, a Givens reflector rotates every nonzero vector in the entire two-dimensional space.

In contrast, Jacobi reflectors are not closed under matrix multiplication. As a matter of fact, the product of any two reflectors is a rotator. A Jacobi reflector can be represented as a rank-1 modification to the identity matrix, namely,

$$J(\theta) = I - (I - J) = I - 2yy^{\mathrm{T}}, \quad \text{where } y = \begin{pmatrix} \sin(\theta/2) \\ -\cos(\theta/2) \end{pmatrix}. \tag{4}$$

Unlike Givens rotation, a Jacobi reflector divides $\mathbf{R}^2$ into two complementary subspaces, acting as the identity on one of them and reflecting on the other:

$$Jx = \begin{cases} x & x \in \mathcal{R}(y)^{\perp}, \\ -x & x \in \mathcal{R}(y). \end{cases}$$

For an arbitrary vector $x \in \mathcal{R}^2$, $J(\theta)x$ is therefore a reflection of x with respect to the line $\mathcal{R}(y)^{\perp} = \mathcal{R}([\cos(\theta/2), \sin(\theta/2)]^{\mathrm{T}})$. We may also say Jx is the reflection of x along $\mathcal{R}(y)$, or simply along y. For the special Jacobi reflector $J(0)$, $J(0) = J(2\pi) = I - 2e_2e_2^{\mathrm{T}}$. This is illustrated in Fig. 1.

A Givens rotator $G(\theta)$ can always be represented as a product of two Jacobi reflectors,

$$G(\theta) = J(\alpha)J(\beta) \quad \text{with} \quad \beta - \alpha = \theta \bmod 2\pi.$$

In particular, $G(\theta) = J(0)J(\theta)$. That is, $G(\theta)$ can be decomposed as a reflection with respect to $(\cos(\theta/2), \sin(\theta/2))^{\mathrm{T}}$ followed by another reflection with respect to $(1,0)^{\mathrm{T}}$. Thus $G(\theta)$ can be represented as a rank-2 modification to the identity matrix,

$$G(\theta) = I - YSY^{\mathrm{T}}, \tag{5}$$

with, for instance,

$$Y = \begin{pmatrix} 0 & \sin(\theta/2) \\ 1 & -\cos(\theta/2) \end{pmatrix} \quad \text{and} \quad S = \begin{pmatrix} 2 & 4\cos(\theta/2) \\ & 2 \end{pmatrix}.$$

An orthogonal matrix Q is a reflector if $Q^2 = I$, and Householder reflectors are a direct generalization of Jacobi reflectors. For each vector x, $H(v)x$ is the reflection of x with respect to the hyperplane $\mathcal{R}(v)^{\perp}$. The concept of reflectors was further developed by Schreiber and Parlett [9] to *block reflectors*, for example,

$$Q = I - 2YY^{\mathrm{T}}, \qquad Y^{\mathrm{T}}Y = I, \quad Y \in \mathcal{R}^{m\times k}. \tag{6}$$

Note that the reflectors we have mentioned so far are all symmetric.

The representations (3), (4), and (6) for reflectors and (5) for rotators are all special cases of the representation

$$Q = I - YSY^{\mathrm{T}}, \qquad Y \in \mathcal{R}^{m\times k}, \quad S \in \mathcal{R}^{k\times k} \tag{7}$$

for an $m \times m$ orthogonal matrix. With a triangular matrix S, this form of representation appears first as the *compact WY representation* by Schreiber and Van Loan [10], as a way of expressing the product of k Householder matrices in a computationally more advantageous form.

If S is nonsingular and Y is of rank k, then Q acts on the space $\mathcal{R}(Y)^{\perp}$ as the identity and changes every nonzero vector in $\mathcal{R}(Y)$, which we call the active space of Q. From the preceding discussion we see that Jacobi and Householder reflectors have one-dimensional active subspaces, whereas, except for the identity, Givens rotations have two-dimensional active subspaces.

We show in this paper that the representation (7), which we call *the basis-kernel representation*, is a universal representation for *any* orthogonal matrix. This is proved in the next section, and there we also introduce the so-called orthogonality conditions on Y and S, which must be satisfied for the matrix Q of (7) to be orthogonal. We prove further that any orthogonal matrix can be expressed in basis-kernel form with a triangular kernel, and we show how the familiar representation of orthogonal matrices as products of Householder matrices can be readily deduced from this representation. Our theory is also used to show that, for orthogonal transformations mapping a matrix A to a matrix B, there is a "minimal" transformation Q in that its associated basis Y has a minimal number of columns. In §3 we describe in detail how the basis Y and the kernel S characterize Q. We derive a canonical form that makes explicit how Q partitions $\mathbf{R}^n$ into a couple of subspaces in which it acts as the identity, a reflector and/or a rotator, respectively. In §4 we present a generalized form, applicable to arbitrary orthogonal matrices, of the Cayley representation [6]. The generalized Cayley form reveals that, given a subspace $\mathcal{L}$ of dimension k, there are $k(k-1)/2$ degrees of freedom in choosing a nonsymmetric matrix active upon the subspace $\mathcal{L}$ while there is one and only one symmetric matrix. Finally, we comment on our results and outline directions of future research.

2. The basis-kernel representation of orthogonal matrices.

THEOREM 2.1. *For any $m \times m$ orthogonal matrix Q there exist a full-rank $m \times k$ matrix Y and a nonsingular $k \times k$ matrix S, $k \leq m$, such that*

$$Q := Q(Y,S) = I - YSY^{\mathrm{T}}. \tag{8}$$

Proof. If $I-Q$ is nonsingular, we may choose $Y = I$ and $S = I-Q$. Otherwise, let X and Y be orthonormal bases of $\mathcal{N}(I-Q)$ and $\mathcal{R}(I-Q)$, the null space and range of $I-Q$, respectively. Then,

$$Q = (X,Y)\begin{pmatrix} I & 0 \\ 0 & I-S \end{pmatrix}\begin{pmatrix} X^{\mathrm{T}} \\ Y^{\mathrm{T}} \end{pmatrix},$$

for some orthogonal matrix $I-S$ that has no eigenvalue at 1. Therefore, S is nonsingular and $Q = I - YSY^{\mathrm{T}}$. □

As already mentioned in the preceding section, we call $\mathcal{R}(Y)$ of (8) the *active subspace* of Q (which is uniquely defined by Q as to be seen later) and denote it with $\mathcal{A}(Q)$. We define the *degree* of Q as the dimension of $\mathcal{A}(Q)$. We call S the *kernel* of Q, Y the *basis*, and (8) the *basis-kernel representation* of Q. So, for example, a Householder matrix (3) is an orthogonal matrix of degree 1.

Now let X_y and X_s be two j-by-k matrices, $j \geq k$, such that $X_y^{\mathrm{T}}X_s = I$. Then, $YSY^{\mathrm{T}} = (YX_y^{\mathrm{T}})(X_sSX_s^{\mathrm{T}})(X_yY^{\mathrm{T}})$. It is important to realize that a particular orthogonal matrix Q has many representations in the form of (8), and Y and S need not necessarily be of full rank. For convenience, we call them all basis-kernel representations of Q.

2.1. The orthogonality conditions. Like the condition on v and β in (3) for a Householder reflector, there is a condition on Y and S that guarantees the orthogonality of $Q(Y,S)$, even when Y and S are not of full rank.

LEMMA 2.2.

1. *The orthogonality condition*

$$SY^{\mathrm{T}}YS^{\mathrm{T}} = S + S^{\mathrm{T}} \tag{9}$$

or

$$S^TY^{\mathrm{T}}YS = S + S^{\mathrm{T}} \tag{10}$$

is a sufficient condition for the orthogonality of $Q(Y,S)$.

2. *The condition* (9) *and the condition* (10) *are equivalent.*

3. *When S is nonsingular, the orthogonality conditions can be expressed in the unified form*

$$Y^TY = S^{-1} + S^{-T}. \tag{11}$$

Proof. Parts 1 and 3. If we write $Q = I - YSY^T$, then the condition (9) implies $QQ^{\mathrm{T}} = I$ and the condition (10) implies $Q^{\mathrm{T}}Q = I$. The expression of (11) follows immediately from the conditions in Part 1 when S is nonsingular.

Part 2. Now assume S is of rank $r < k$. Let $S = U\begin{pmatrix} \Sigma & \\ & 0 \end{pmatrix}V^{\mathrm{T}}$ be a singular value decomposition of S with $\Sigma \in \mathbf{R}^{r\times r}$ nonsingular. Then,

$$U^{\mathrm{T}}SU = \begin{pmatrix} \Sigma & \\ & 0 \end{pmatrix}\begin{pmatrix} V_1^{\mathrm{T}} \\ V_2^{\mathrm{T}} \end{pmatrix}(U_1\ U_2) = \begin{pmatrix} \tilde{S}_{11} & \tilde{S}_{12} \\ 0 & 0 \end{pmatrix},$$

where $\tilde{S}_{11} = \Sigma V_1^{\mathrm{T}} U_1$ is a square matrix, and $\tilde{S}_{12} = \Sigma V_1^{\mathrm{T}} U_2$. The orthogonality condition (9) can then be expressed as

$$(12)\quad \begin{pmatrix} \tilde{S}_{11} & \tilde{S}_{12} \\ 0 & 0 \end{pmatrix} (YU)^{\mathrm{T}}(YU) \begin{pmatrix} \tilde{S}_{11}^{\mathrm{T}} & 0 \\ \tilde{S}_{12}^{\mathrm{T}} & 0 \end{pmatrix} = \begin{pmatrix} \tilde{S}_{11} & \tilde{S}_{12} \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \tilde{S}_{11}^{\mathrm{T}} & 0 \\ \tilde{S}_{12}^{\mathrm{T}} & 0 \end{pmatrix}.$$

The last equation implies that $\tilde{S}_{12} = 0$ and that $\tilde{S}_{11}$ must be nonsingular. Thus, $S = U_1 \tilde{S}_{11} U_1^{\mathrm{T}}$. Multiplying (12) by $\tilde{S}_{11}^{-1}$ and $\tilde{S}_{11}^{-T}$ from the left and right, respectively, we obtain

$$(YU_1)^{\mathrm{T}}(YU_1) = \tilde{S}_{11}^{-1} + \tilde{S}_{11}^{-\mathrm{T}}.$$

Therefore,

$$\tilde{S}_{11}^{\mathrm{T}}(YU_1)^{\mathrm{T}}(YU_1)\tilde{S}_{11} = \tilde{S}_{11} + \tilde{S}_{11}^{\mathrm{T}},$$

and hence the condition (10). In the same fashion, (10) implies $\tilde{S}_{12} = 0$. □

Given Y, we now show some examples of choices for S such that the orthogonality condition is satisfied.

EXAMPLE 2.3. *$Q(Y, S)$ is orthogonal if*

$$S = 2(Y^{\mathrm{T}}Y)^{\dagger},$$

where $B^{\dagger}$ denotes a pseudoinverse of the matrix B [13]. Such a singular and symmetric kernel was first introduced in [9].

EXAMPLE 2.4. *$Q(Y, S)$ is orthogonal if Y has no zero column and*

$$S = [\mathrm{tril}(Y^{\mathrm{T}}Y) + \mathrm{diag}(Y^{\mathrm{T}}Y)/2]^{-1},$$

or

$$S = [\mathrm{triu}(Y^{\mathrm{T}}Y) + \mathrm{diag}(Y^{\mathrm{T}}Y)/2]^{-1},$$

where $\mathrm{tril}(A)$ ($\mathrm{triu}(A)$) *is the strictly lower (upper) triangular part of matrix A, and* $\mathrm{diag}(A)$ *is the diagonal of matrix A.* Note that the triangularity of S and the orthogonality condition (11) together imply that S is unique. One can see that, given Y, the triangular kernel is easy to compute. As a matter of fact, it is the procedure for computing the compact WY representation proposed in [14], [8].

2.2. Regularity assumption. The discussion following Theorem 2.1 and the examples above have shown that Y and S need not necessarily be of full rank. On the other hand, we know from Theorem 2.1 that for an orthogonal matrix, there is always a basis-kernel representation with full rank Y and nonsingular S. Such a representation we call a *regular* basis-kernel representation. Under the regularity assumption, the active space of Q is $\mathcal{R}(Y)$ and the degree of Q is the number of Y's columns.

THEOREM 2.5. *A nonregular basis-kernel transformation can be transformed into a regular one.*

Proof. Suppose Y is rank deficient. Let $YP = \tilde{Y}R$, with

$$R = \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix},$$

be a rank-revealing QR decomposition of Y (see, for example, [3], [4]), that is, R_{11} is nonsingular and $\operatorname{rank}(R_{11}) = \operatorname{rank}(Y)$. Then, $Q(Y,S) = Q(\tilde{Y},\tilde{S})$ with $\tilde{S} = RP^{\mathrm{T}}SPR^{\mathrm{T}}$. Thus, we can assume without loss of generality that Y is of full rank.

Now suppose S is singular. We know from the proof of Lemma 2.2 that $S = U\bar{S}U^{\mathrm{T}}$ for some U and $\bar{S}$ of full rank. Thus, $Q(Y,S) = Q(\bar{Y},\bar{S})$ with $\bar{Y} = YU$. □

We therefore assume in the rest of the paper that a basis-kernel representation of an orthogonal matrix is regular unless explicitly stated otherwise.

2.3. Triangular kernels. For a given Y, the triangular kernel of Example 2.4 presents another way of computing the compact WY form of a product of Householder reflectors. In fact, *any* orthogonal matrix can be expressed in basis-kernel form with an upper or lower triangular kernel.

THEOREM 2.6. *Any orthogonal matrix Q can be expressed as $Q = I - YSY^T$ with a triangular kernel S.*

Proof. Let $Q = Q(Y,S)$ be an orthogonal matrix of degree k. It is sufficient to prove the claim that there is a (unit) lower matrix L such that $S = L^{\mathrm{T}}RL$ for some upper triangular matrix R, since $Q(YL^{\mathrm{T}},R)$ will be a basis-kernel representation of Q with triangular kernel. The claim holds for orthogonal matrices of degree $k = 1$. Let $Q(Y,S)$ be an orthogonal matrix of degree $k > 1$. Suppose the claim holds for all matrices of degree less than k. Partition S^{-1},

$$T = S^{-1} = \begin{pmatrix} \tau & a^{\mathrm{T}} \\ b & \tilde{T}_{-1} \end{pmatrix}.$$

The orthogonality condition (11) implies $2\tau = e_1^{\mathrm{T}}(Y^{\mathrm{T}}Y)e_1 \neq 0$. Thus,

$$L_1 T L_1^{\mathrm{T}} = \begin{pmatrix} \tau & (a-b)^{\mathrm{T}} \\ 0 & T_{-1} \end{pmatrix}, \tag{13}$$

with

$$L_1 = \begin{pmatrix} 1 & 0 \\ -b/\tau & I \end{pmatrix}, \text{ and } T_{-1} = \tilde{T}_{-1} - ba^{\mathrm{T}}/\tau.$$

Substituting (13) into (11) results in

$$L_1(Y^{\mathrm{T}}Y)L_1^{\mathrm{T}} = \begin{pmatrix} \tau & (a-b)^{\mathrm{T}} \\ 0 & T_{-1} \end{pmatrix} + \begin{pmatrix} \tau & 0 \\ (a-b) & T_{-1}^{\mathrm{T}} \end{pmatrix}. \tag{14}$$

Now let

$$Y_{-1} = Y \begin{pmatrix} -b^{\mathrm{T}}/\tau \\ I \end{pmatrix}, \quad \text{and} \quad S_{-1} = T_{-1}^{-1}.$$

We know from (14) that $I - Y_{-1}S_{-1}Y_{-1}$ is an orthogonal matrix of degree $k-1$. With the induction hypothesis, there is a unit lower triangular matrix L_{-1} and an upper triangular matrix R_{-1} such that $S_{-1} = L_{-1}^{\mathrm{T}}R_{-1}L_{-1}$. With

$$L = \begin{pmatrix} 1 & \\ & L_{-1} \end{pmatrix} L_1, \quad \text{and} \quad R = \begin{pmatrix} \tau^{-1} & (b-a)^{\mathrm{T}}L_{-1}R_{-1}^{-1} \\ & R_{-1}^{-1} \end{pmatrix},$$

we then have $S = L^{\mathrm{T}}RL$.

Similarly, we can find nonsingular upper triangular matrices R and lower triangular matrices L such that $S = R^{\mathrm{T}}LR$, $S = L^{\mathrm{T}}R^{\mathrm{T}}L$, or $S = R^{\mathrm{T}}L^{\mathrm{T}}R$. The last two decompositions follow from the fact that S^{T} is the kernel of Q^{T}. □

Example 2.4 shows that, for a fixed Y, the upper (lower) triangular kernel is unique. An orthogonal matrix, however, has more than one representation with an upper (lower) triangular kernel. Let $Q(Y, S)$ be a representation with upper triangular kernel S. There is an orthogonal matrix U such that $U^{\mathrm{T}}SU$ is also upper triangular [7, p. 385], and hence $Q(YU, U^{\mathrm{T}}SU)$ is another representation of Q with triangular kernel.

From the compact WY representation we know that the product of k Householder matrices can be expressed in basis-kernel form. The converse holds true as well.

COROLLARY 2.7. *Any orthogonal matrix of degree k can be expressed as a product of exactly k nontrivial Householder reflectors.*

Proof. We prove the corollary by induction on the degree k of the orthogonal matrices. The corollary holds for the case of $k = 1$ since an orthogonal matrix of degree 1 is by itself a Householder matrix. Let $k > 1$, and assume that the theorem is true for all orthogonal matrices of degree $\leq k - 1$. Let Q be an orthogonal matrix of degree k and $Q = I - YSY^{\mathrm{T}}$ with an upper triangular kernel S. The orthogonality condition implies

$$S = (\mathrm{triu}(Y^{\mathrm{T}}Y, 1) + \mathrm{diag}(Y^{\mathrm{T}}Y)/2)^{-1}.$$

If we partition Y as $Y = (y, Y_{-1})$, then

$$S = \begin{pmatrix} s & -sy^{\mathrm{T}}Y_{-1}S_{-1} \\ & S_{-1} \end{pmatrix},$$

and hence

$$Q = I - ysy^{\mathrm{T}} + ysy^{\mathrm{T}}Y_{-1}S_{-1}Y_{-1}^{\mathrm{T}} - Y_{-1}S_{-1}Y_{-1}^{\mathrm{T}} = (I - ysy^{\mathrm{T}})(I - Y_{-1}S_{-1}Y_{-1}^{\mathrm{T}}),$$

where $(I - ysy^{\mathrm{T}})$ is a nontrivial Householder matrix and $(I - Y_{-1}S_{-1}Y_{-1}^{\mathrm{T}})$ is an orthogonal matrix of degree $k - 1$ and can be expressed, by the induction hypothesis, as a product of exactly $k - 1$ Householder matrices. □

Notice how easy it is to determine the representation of Q in terms of Householder matrices from a basis-kernel representation with triangular kernel. The Householder vectors are simply the columns of the basis Y, and the scaling factors are the corresponding diagonal elements of the kernel S. Since the basis-kernel representation with triangular kernel is not unique, the representation of an orthogonal matrix as product of Householder matrices is not unique, either.

Generalizing the proof of Corollary 2.7, we note the following result for factorization and composition of arbitrary orthogonal matrices in basis-kernel representation with (block) triangular kernel.

COROLLARY 2.8.

$$Q_1(Y_1, S_1)Q_2(Y_2, S_2) = I - (Y_1, Y_2) \begin{pmatrix} S_1 & -S_1(Y_1^{\mathrm{T}}Y_2)S_2 \\ & S_2 \end{pmatrix} (Y_1, Y_2)^{\mathrm{T}}.$$

Using this formula, one can, for example, quickly assemble random orthogonal matrices in a "binary tree"-like fashion from lower-degree random orthogonal matrices, deriving, in effect, a parallel block version of the Householder-oriented approach by Stewart [12].

2.4. Block orthogonal transformations. The following theorem shows that, if there is an orthogonal transformation that transforms an $m \times k$ matrix A into a matrix B, $k < m$, the degree of Q concerned need not be larger than k.

THEOREM 2.9. *Let A and B be two m-by-k matrices, $k < m$. If $B = QA$ for some orthogonal matrix Q, then Q is either of degree no greater than k or can be replaced by an orthogonal factor of its own with degree no greater than k.*

Proof. Let $Q = Q(Y, S)$ be a basis-kernel representation of Q. Suppose the degree of Q is n, $n > k$. Let $Y^T A = U \begin{pmatrix} M \\ 0 \end{pmatrix}$ be a QR-factorization of $Y^T A$, with $M \in \mathbf{R}^{r \times k}$, where $r \leq k$ is the rank of $Y^T A$ and $U \in \mathbf{R}^{n \times n}$. Then $Q(Y, S) = Q(\bar{Y}, \bar{S})$, where $\bar{Y} = YU$ and $\bar{S} = U^T S U$. Partitioning $\bar{Y} = [\bar{Y}_1, \bar{Y}_2]$, where $\bar{Y}_1$ is $m \times r$, we then have $\bar{Y}_2^T A = 0$. From the proof of Theorem 2.6, $\bar{S} = LRL^{\mathrm{T}}$ for some lower triangular matrix L and upper triangular matrix R. Thus, $Q(\bar{Y}, \bar{S}) = Q(\tilde{Y}, R)$ with $\tilde{Y} = \bar{Y}L$. If we partition $\tilde{Y} = (\tilde{Y}_1, \tilde{Y}_2)$ in the same fashion as $\bar{Y}$, then $\tilde{Y}_2^{\mathrm{T}} A = 0$. Partition

$$R = \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix}$$

conformingly, we have from Corollary 2.8

$$Q = Q(\tilde{Y}_1, R_{11}) Q(\tilde{Y}_2, R_{22}) = Q_1 Q_2,$$

and

$$B = QA = Q(\tilde{Y}_1, R_{11}) A,$$

since $Q_2 A = A - \tilde{Y}_2 R_{22} \tilde{Y}_2^{\mathrm{T}} A = A$. □

Not surprising, for any two vectors a and b with $\|a\|_2 = \|b\|_2$, there is always an orthogonal matrix Q of degree 1(i.e., a Householder matrix) such that $b = Qa$.

In matrix computations, the following block elimination problem is fundamental. Given an $m \times k$ matrix A, determine an orthogonal matrix Q such that

$$QA = \begin{pmatrix} C \\ 0 \end{pmatrix}, \tag{15}$$

where C is a k-by-k matrix. The usual Householder-based approach constructs an orthogonal matrix Q and an upper triangular matrix C in a column-by-column fashion as a product of k Householder matrices. Using the WY representation, one then can deduce a basis-kernel representation with Householder vectors and a triangular kernel.

Theorem 2.9 and its proof lead to the following conclusions.

(i) The elimination problem (15) can be solved with an orthogonal matrix of degree at most k.

(ii) Finding ways to determine orthogonal matrices directly in terms of their basis and kernel (as compared to products of Householder matrices or Givens rotations) seems preferable to arrive at computationally more advantageous procedures. This issue is explored further in [2].

(iii) The minimal degree of a solution Q to a transformation problem between subspaces of dimension k could be even lower than k(such as the case of $r < k$ indicated in the above proof), which would result in a lower degree, and hence computationally less expensive transformation. See also [2].

3. Geometric properties. In the introduction, we reviewed the geometric properties of reflectors "active" in one-dimensional or multidimensional subspaces and of rotators in two-dimensional subspaces. In §2, we showed that the basis-kernel representation is a natural approach for representing, composing, and decomposing orthogonal matrices. This section shows that the basis-kernel representation also makes it easy to understand geometric properties of orthogonal matrices.

3.1. The basis and active subspace. The following theorem shows how Y defines the active space and S specifies the transformation in the active subspace.

THEOREM 3.1.

(i) $Qx = x \Leftrightarrow x \in \mathcal{R}(Y)^{\perp}$.

(ii) *For any $u \in \mathcal{R}(Y)$, there exists one and only one vector b such that $u = YS^{\mathrm{T}}b$, and $Qu = -v$, where $v = YSb$.*

Proof. Part 1. For any x such that $Qx = x$, we have $YSY^{\mathrm{T}}x = 0$. Since Y has full rank, $YSY^{\mathrm{T}}x = 0$ if and only if if $SY^{\mathrm{T}}x = 0$. Thus, $x \in \mathcal{R}(Y)^{\perp}$ if and only if S is nonsingular.

Part 2. Since Y is a basis for its own column space, for any vector u in $\mathcal{R}(Y)$ there exists a unique vector c such that $u = Yc$. By the orthogonality condition we have

$$Qu = (I - YSY^{\mathrm{T}})Yc = Yc - Y(I + SS^{T})c = -YSS^{T}c = -YSb,$$

where $b = S^{-T}c$. Hence $u = YS^{T}b$. □

Thus, when $k < m$, the matrix Q has eigenvalues at 1, and the orthogonal complement of $\mathcal{R}(Y)$ is the invariant subspace of Q corresponding to its eigenvalues at 1. Furthermore, on $\mathcal{R}(Y)$, vectors $u = YS^{\mathrm{T}}b$ and $v = YSb$ in $\mathcal{R}(Y)$ are images of each other under the mappings Q and Q^{-1}, respectively.

With respect to the composition of orthogonal matrices, Corollary 2.8 shows that, if $\mathcal{R}(Y_1) \cap \mathcal{R}(Y_2) = \{0\}$, then $\mathcal{A}(Q_1Q_2) = \mathcal{R}(Q_1) \oplus \mathcal{R}(Q_1)$, or degree$(Q_1Q_2)$ = degree(Q_1)+degree(Q_2). On the other hand, if $Y_2 = Y_1$ and $S_2 = S_1^{\mathrm{T}}$, then the degree of $Q_1Q_2 = I$ is zero. In general, we have the following.

COROLLARY 3.2. *Let Q_1 and Q_2 be two orthogonal matrices. Then,*

$$\mathcal{A}(Q_1Q_2) \subseteq \mathcal{A}(Q_1) \oplus \mathcal{A}(Q_2).$$

3.2. The kernel. While the basis Y determines the space acted upon by Q, the kernel S specifies the action taken in this subspace.

THEOREM 3.3.

1. $\lambda(Q) = \lambda(-SS^{-\mathrm{T}}) \cup \{1\}$.
2. $\det(Q) = \begin{cases} 1, & \text{if } k \text{ is even,} \\ -1, & \text{otherwise,} \end{cases}$
3. $Qx = -x \Rightarrow x \in \mathcal{R}(Y)$ *if and only if S is symmetric.*

Proof. Part 1. When S is nonsingular, the orthogonality condition can be expressed as

$$S(Y^{\mathrm{T}}Y) = SS^{-\mathrm{T}} + I.$$

For any vector $y \in \mathcal{R}(Y)$, there exists a unique vector b such that $y = Yb$, and

$$Qy = (I - YSY^{\mathrm{T}})Yb = Yb - Y(SS^{-\mathrm{T}} + I)b = -YSS^{-\mathrm{T}}b. \tag{16}$$

In particular,

$$QY = -Y(SS^{-\mathrm{T}}).$$

By Theorem 3.1, $\mathcal{R}(Y)$ is the invariant subspace of Q corresponding to all of its eigenvalues not equal to 1. Therefore $\lambda(Q) = \lambda(-SS^{-\mathrm{T}}) \cup \{1\}$.

Part 2. We know from Part 1 that $\det(Q) = \det(-SS^{-\mathrm{T}})$. We then have

$$\det(Q) = (-1)^k \det(S)\det(S^{-1}) = (-1)^k.$$

Part 3. From Part 2 of Theorem 3.1 and Part 1 of Theorem 3.3, it remains to show that $Qx = -x$ for any $x \in \mathcal{R}(Y)$ implies that S is symmetric. We see from (16) that

$$\begin{aligned} & Qx = -x \quad \forall x \in \mathcal{R}(Y), \\ \Leftrightarrow\ & YSS^{-\mathrm{T}}b = Yb \quad \forall b \in \mathbf{R}^k, \\ \Leftrightarrow\ & SS^{-\mathrm{T}} = I \end{aligned}$$

and the symmetry of S follows. □

Note that the determinant of H does not depend on the symmetry of H and that S cannot be skew-symmetric.

Theorem 3.3 implies that reflectors and symmetric orthogonal matrices are really one and the same.

COROLLARY 3.4. *An orthogonal matrix is a reflector if and only if it is symmetric and not equal to the identity.*

Theorem 3.3 also illustrates how Q acts upon the subspace $\mathcal{R}(Y)$. The matrix $(-SS^{-T})$ is the representation of Q in $\mathcal{R}(Y)$ with respect to the basis Y, and it has eigenvalues on the unit circle in the complex plane, but not at 1. Let g_j be an eigenvector of $-SS^{-\mathrm{T}}$ corresponding to its eigenvalue $\cos(\theta_j) + i\sin(\theta_j)$. Then,

$$Q(Yg_j) = Y(-SS^{-\mathrm{T}})g_j = (Yg_j)(\cos(\theta_j) + i\sin(\theta_j)).$$

That is, for an arbitrary vector in $\mathcal{R}(Y)$, its components along Yg_j are "rotated" by θ_j, respectively. When Q is a block reflector, the components are rotated uniformly by the same angle π; that is, the sign of vectors in $\mathcal{R}(Y)$ is simply flipped.

If Q should act as other than a reflection on $\mathcal{R}(Y)$, S must be nonsymmetric and $-SS^{-\mathrm{T}}$ must have truly complex eigenvalues, which exist in conjugate pairs. Taking into account Lemma 3.6, we then have the following corollary.

COROLLARY 3.5. *If Q is nonsymmetric, then its kernel S can be expressed with respect to properly chosen Y via*

$$-SS^{-T} = \operatorname{diag}\left(\begin{bmatrix} \cos(\Theta) & \sin(\Theta) \\ -\sin(\Theta) & \cos(\Theta) \end{bmatrix}, B\right), \tag{17}$$

where $B = -I$ or the empty matrix, and $\Theta = \operatorname{diag}(\theta_j)$, $\sin(\theta_j) \neq 0$. The first diagonal block of (17) can be viewed as a block Givens rotator. Corollary 3.5 shows that an orthogonal matrix divides its active subspace into two subspaces: it acts as a reflector in one of them and a rotator in the other. An orthogonal matrix of odd degree always has a nontrivial subspace that it acts on as a reflector.

As it turns out, there is a close relationship between SS^{-T} and Y when Y is orthonormal.

LEMMA 3.6. *Let Q be an orthogonal matrix and $Q = I - YSY^{\mathrm{T}}$ be a regular basis-kernel representation of Q. The following statements are equivalent:*

Y is orthonormal;

$I - S$ is orthogonal;

SS^{-T} is orthogonal.

Proof. We have seen from Theorem 2.1 that if Y is orthonormal, then $I - S$ is orthogonal. Now suppose that $I - S$ is orthogonal. Then $S = I + SS^{-T}$. At the same time, the orthogonality condition (9) implies that

$$S(Y^{\mathrm{T}}Y) = I + SS^{-\mathrm{T}}.$$

Together, they imply that $Y^{\mathrm{T}}Y = I$. □

Corollary 3.5 and Lemma 3.6 allow us to derive a particularly simple canonical form for S^{-1}.

THEOREM 3.7. *For any orthogonal matrix of degree k there exist an orthonormal basis Y and a kernel S such that*

$$S^{-1} = \frac{1}{2}\begin{pmatrix} I & & \\ & I & D \\ & -D & I \end{pmatrix},$$

where D is either zero or a nonsingular diagonal matrix.

Proof. Let $Q = Q(Y, S)$, and, invoking Corollary 3.5, assume that Y is orthonormal and (17) holds. From the proof of Lemma 3.6, we have

$$S^{-1} = (I + SS^{\mathrm{T}})^{-1}.$$

The theorem is true for the special case that $SS^{\mathrm{T}} = I$ with $D = 0$. As another special case consider SS^{T} to be a 2-by-2 Givens rotation $G(\theta)$ with $\sin(\theta) \neq 0$. We then have

$$I + G(\theta) = \begin{pmatrix} 1+\cos(\theta) & \sin(\theta) \\ -\sin(\theta) & 1+\cos(\theta) \end{pmatrix} = 2\cos(\theta/2)\begin{pmatrix} \cos(\theta/2) & \sin(\theta/2) \\ -\sin(\theta/2) & \cos(\theta/2) \end{pmatrix},$$

and since $\sin(\theta) = 2\sin(\theta/2)\cos(\theta/2) \neq 0$,

$$(I + G(\theta))^{-1} = \begin{pmatrix} 1 & -\cot(\theta/2) \\ \cot(\theta/2) & 1 \end{pmatrix}.$$

The claim of the theorem in general easily follows from (17). □

4. The generalized Cayley representation. For any skew-symmetric matrix B, the matrices

$$(I + B)(I - B)^{-1} \text{ and } (I + B)(B - I)^{-1} \tag{18}$$

are orthogonal. The former does not have eigenvalue at -1, and the latter does not have eigenvalue at 1. Conversely, an orthogonal matrix Q can be represented in one of the above forms with some skew-symmetric matrix B as long as Q does not have eigenvalues at both 1 and -1. Representation (18) is known as the Cayley representation [6] or the Cayley transform of B.

Note that the Cayley representation does not include symmetric orthogonal matrices except I and $-I$, nor does it include the nonsymmetric matrices that have both a nontrivial "inactive" subspace and a nontrivial "active" reflection subspace. We can, however, generalize this representation to cover all orthogonal matrices, by combining the traditional Cayley representation and our basis-kernel representation.

THEOREM 4.1. *Let Y be an orthonormal matrix with k columns. Then Q is an orthogonal matrix with active subspace $\mathcal{R}(Y)$ if and only if*

$$Q = I - Y(I - (B + I)(B - I)^{-1})Y^{\mathrm{T}} \tag{19}$$

for some skew symmetric matrix B. Moreover, Q is symmetric if and only if $B = 0$.

Proof. It can be checked directly that, for a skew-symmetric matrix B, Q of (19) is orthogonal. On the other hand, if Q is an orthogonal matrix with active subspace $\mathcal{R}(Y)$, then Q can be represented as $Q = I - YSY^{\mathrm{T}}$ for some S that satisfies the equation $I = Y^TY = S^{-1} + S^{-\mathrm{T}}$. Thus, $B = I - 2S^{-1}$ is skew-symmetric and $S = I + SS^{-\mathrm{T}} = I - (B+I)(B-I)^{-1}$. The orthogonal matrix Q is symmetric if and only if $S = 2I$ and if and only if $B = 0$. □

Note that, for the special case that Q has full degree (i.e., no eigenvalue at 1), the generalized Cayley representation (19) becomes the traditional one when one chooses $Y = I$.

Theorem 4.1 implies that, given a subspace $\mathcal{Y}$ of dimension k, we have $k(k-1)/2$ degrees of freedom in choosing a nonsymmetric orthogonal matrix so that $\mathcal{A}(Q) = \mathcal{Y}$, but there is only one symmetric orthogonal matrix whose active subspace is $\mathcal{Y}$.

5. Conclusions. This paper introduced the basis-kernel representation $Q = I - YSY^T$ of an orthogonal matrix. We showed that any orthogonal matrix can be represented in this form, in particular with a triangular kernel, and showed the relation to the familiar representation of orthogonal matrices as products of Householder matrices.

We also showed how the basis Y determines the subspace that Q acts on in a nontrivial fashion, and how the kernel S determines the action taken on this subspace. This led to a particularly simple representation of $-SS^T$ and S^{-1} which explicitly shows how Q acts on its active subspace as a composition of rotators and reflectors. We also showed that reflectors are exactly the symmetric orthgonal matrices.

We generalized the Cayley representation to cover all orthogonal matrices and pointed out that, given a subspace, there is great freedom in choosing nonsymmetric orthogonal matrices acting upon it, but that the symmetric orthogonal matrix is uniquely determined by an active subspace.

We would like to point out that the basis-kernel representation, and the theory we have developed for it, deals directly with Y and S, whereas the usual approaches to orthogonal matrix computations deal principally with elementary operations such as Givens rotators, Jacobi reflectors, or Householder reflectors. Thus, we believe that this representation opens the door to different approaches for deriving orthogonal matrices with desired properties. For example, the proof of Theorem 2.9 hinted at the possibility for finding lower-rank orthogonal matrices for block elimination problems than the orthogonal matrices provided by the usual approaches. These issues are explored further in [2].

Acknowledgment. We thank Beresford Parlett for some stimulating discussions.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. DuCroz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, *LAPACK User's Guide*, Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[2] C. Bischof and X. Sun, *On orthogonal block elimination*, Tech. Report MCS-P450-0794, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 1994.

[3] C. H. Bischof and P. C. Hansen, *Structure-preserving and rank-revealing QR factorizations*, SIAM J. Sci. Comput., 12 (1991), pp. 1332–1350.

[4] T. F. CHAN, *An improved algorithm for computing the singular value decomposition*, ACM Trans. Math. Software, 8 (1982), pp. 72–83.
[5] J. J. DONGARRA, J. R. BUNCH, C. B. MOLER, AND G. W. STEWART, *LINPACK Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, 1979.
[6] F. R. GANTMACHER, *Applications of the Theory of Matrices*, Interscience Publications, Inc., New York, 1959.
[7] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd edition, The Johns Hopkins University Press, Baltimore, 1989.
[8] C. PUGLISI, *Modification of the Householder method based on the compact WY representation* SIAM J. Sci. Comput., 3 (1992), pp. 723–726.
[9] R. SCHREIBER AND B. PARLETT, *Block reflectors: Theory and computation*, SIAM J. Numer. Anal., 25 (1988), pp. 189–205.
[10] R. SCHREIBER AND C. VAN LOAN, *A storage efficient WY representation for products of Householder transformations*, SIAM J. Sci. Comput., 10 (1989), pp. 53–57.
[11] G. W. STEWART, *Introduction to Matrix Computation*, Academic Press, New York, 1973.
[12] ———, *The efficient generation of random orthogonal matrices with an application to condition estimators*, SIAM J. Numer. Anal., 17 (1980), pp. 403–409.
[13] G. W. STEWART AND J. SUN, *Matrix Perturbation Theory*, Academic Press, Inc., New York, 1991.
[14] H. F. WALKER, *Implementation of the GMRES method using Householder transformations*, SIAM J. Sci. Comput., 9 (1988), pp. 152–163.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, pp. 1197–1209, October 1995

013

ON THE CONVERGENCE OF THE JACOBI METHOD FOR ARBITRARY ORDERINGS*

WALTER F. MASCARENHAS†

Abstract. This paper presents new results concerning the effect of the ordering on the rate of convergence of the Jacobi iteration for computing eigenvalues of symmetric matrices. We start by showing that the diagonal elements converge for any ordering. Next we emphasize that different parts of the matrix converge at different speeds. Taking advantage of this phenomenon, we propose a strategy that leads to a convergence exponent of $3^{4/5} \approx 2.41$. Then we show that choosing the rotations to sort the diagonal can improve the convergence by a constant factor and we present experimental results on the performance of this new strategy.

Key words. eigenvalues, Jacobi method, convergence, convergence exponent

AMS subject classifications. 65F15, 65H15

1. Introduction.

The theory of the convergence of the Jacobi method for computing eigenvalues of symmetric matrices was a lively field of study in the 1960s. This theory described some cyclic orderings for which the method converges, and the rate of convergence was proven to be quadratic. In the next decade these questions became less interesting because the QR method was found to be more efficient for the computers available at that time. However, the introduction of parallel computers has renewed interest in the Jacobi method, since it can easily be made to work efficiently on such machines. Another point in favor of the Jacobi method is its accuracy, as shown recently in [DV].

Instead of looking at the parallelism and accuracy of the Jacobi method, in this work we are mainly concerned with its convergence. We do not know if the ideas in §3 are parallelizable, although progress in this direction was made in [EM].

In §2 we give the first new result about convergence: The diagonal of the iterates is convergent.[1] The proof of convergence of the diagonal is very general; it works for any ordering, cyclic or not, and even for angles outside of the traditional interval $(-\pi/4, \pi/4)$. However, it does not prove that the diagonal elements converge to the eigenvalues for we must leave open the possibility that the off-diagonal elements never decrease to zero. In the case where there are no repetitions among the entries of the limit of the diagonal, the proof in [P, p.181] can be easily adapted to show quadratic convergence towards the eigenvalues, regardless of the ordering (see [M]). The case where we have repeated components in the limit of the diagonal is much more complex [BP], [CVD]. In the author's Ph.D. thesis [M], he presented contrived orderings and matrices for which repetitions in the diagonal can lead to divergence. The proof of divergence for the examples above is quite tedious and the present paper is oriented to the generic case instead, where there are no repetitions in the limit diagonal. Thus in §§3 and 4, we make the nonrepetition hypothesis.

Nonrepetition hypothesis. The entries of the limit of the diagonal are distinct.

We would like to emphasize that the nonrepetition hypothesis is *not* equivalent to the absence of multiple eigenvalues. We strongly believe that, for contrived orderings

* Received by the editors March 26, 1990; accepted for publication (in revised form) by F. T. Luk November 5, 1994.

† Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, Minnesota 55455 (walter@ime.unicamp.br).

[1] We have been informed by R. Schreiber that he and G. Schroff have proved the same result, but have never published it.

and matrices, it may well be that the iterates get trapped close to some nondiagonal matrix with repeated diagonal entries, even if our original matrix does not have multiple eigenvalues. If this is indeed the case, then the nonrepetition hypothesis points out what can go wrong more precisely than does the usual assumption about the multiplicity of eigenvalues.

Section 3 is developed around the basic fact that different parts of the matrix converge at different speeds. Using this fact we present an ordering with superquadratic convergence, with exponent $3^{4/5} \approx 2.41$. We also argue in §3 that the row ordering has convergence rate better than quadratic, but the convergence exponent depends on the size of the matrix, decaying quickly to 2 as the dimension of the matrix increases. For the ordering by diagonals, on the other hand, the decay of the convergence exponent with the dimension is much slower and we have an exponent of at least ≈ 2.2 even for matrices of dimension 256. In §4 we present experiments comparing the performance of the strategy proposed in §3 with that of the standard row ordering and the ordering by diagonals. The conclusion of this section is that the asymptotic rate of convergence is of little relevance in practice. In practice, the transient regime is more important than the asymptotic regime and the performance of the Jacobi method depends strongly on the initial matrix. In §4 we also present an idea that has a practical impact: choosing the angles in each rotation in order to sort the diagonal. We explain why this idea works for the strategy of §3, making it competitive with the more traditional orderings.

We refer the reader to [P] for basic information on the Jacobi method. One remark about notation: the tilde ($\tilde{}$) will be used to denote the value of a quantity after one rotation is performed. For example, when pivoting in (i,j), $i<j$, we have (see [P])

$$\tilde{a}_{ir} = \cos\theta\, a_{ir} - \sin\theta\, a_{jr}, \tag{1.1}$$

$$\tilde{a}_{jr} = \cos\theta\, a_{jr} + \sin\theta\, a_{ir}, \tag{1.2}$$

if $r \notin \{i,j\}$. The rotation angle θ in the expressions above is such that

$$\theta = \pm\frac{\pi}{4} \quad \text{if} \quad a_{ii} = a_{jj}, \tag{1.3}$$

$$\tan(2\theta) = \frac{2a_{ij}}{a_{jj} - a_{ii}} \quad \text{if} \quad a_{ij} \neq a_{jj}. \tag{1.4}$$

There are two angles θ that satisfy this last equation. We will call the θ in the interval $[-\pi/4, \pi/4)$ *the inner angle.* The other angle is called *the outer angle.* It is also possible to choose θ satisfying (1.3) and (1.4) in such a way that $\tilde{a}_{ii} > \tilde{a}_{jj}$, what tends to sort the diagonal in nondecreasing order. We call this last angle *the sorting angle.*

If we take θ to be the inner angle then the diagonal entries are updated by

$$\tilde{a}_{ii} = a_{ii} - \tan\theta\, a_{ij}, \tag{1.5}$$

$$\tilde{a}_{jj} = a_{jj} + \tan\theta\, a_{ij}. \tag{1.6}$$

If θ is the outer angle then

$$\tilde{a}_{ii} = a_{jj} + \tan\theta_u\, a_{ij}, \tag{1.7}$$

$$\tilde{a}_{jj} = a_{ii} - \tan\theta_u\, a_{ij}, \tag{1.8}$$

where θ_u is the inner angle.

2. The convergence of the diagonal. In this section we consider the evolution of the diagonal elements under the Jacobi iteration. Understanding this evolution is important because, first, the diagonal elements should converge to the eigenvalues and, second, most analysis of convergence makes some kind of assumption about the behavior of the diagonal. Our fundamental result is that the (sorted) diagonal always converges, regardless of the ordering. If v is a vector, let sort(v) denote the vector of elements of v sorted in nondecreasing order. The theorem follows.

THEOREM 1. *Let O be any ordering for the Jacobi method, applied to an arbitrary matrix A, with any rule for choosing among the two possible angles in each rotation. The vectors sort(d) of sorted diagonal elements converge to some limit sort$^\infty$.*

It should be pointed out immediately that this theorem says nothing about the relation between sort$^\infty$ and spec(A), the spectrum of A. Indeed, as shown in [M], there exist orderings for which the Jacobi iteration does not always converge to a diagonal matrix. However, Theorem 1 provides a framework to proceed in the study of the convergence of the Jacobi method for arbitrary orderings, because it allows us to state the nonrepetition hypothesis. A natural continuation of this work would be to show that, given an ordering, the nonrepetition hypothesis holds for "almost all" matrices. We could use a slight modification of the proof [P, p.181] to show that an arbitrary ordering of the Jacobi method works "almost always."

We present now the proof of Theorem 1.

Proof. Let us analyze the effect that a Jacobi rotation with pivot (i, j) has on the diagonal. The first thing to notice is that if $r \neq i, j$, then d_r does not change. Define $d_+ = \max\{d_i, d_j\}$ and $d_- = \min\{d_i, d_j\}$. Then we have the situation shown in Fig. 1.

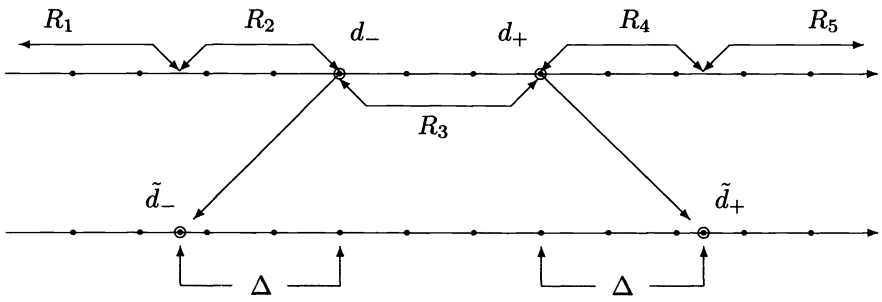


FIG. 1

The R_i are various subsets of the real line whose significance will become clear in the arguments below. In words, Fig. 1 shows that the bigger of the two diagonal elements involved in the rotation gets even bigger, and the smaller gets smaller by the same amount. Since the trace of A is preserved by any similarity transformation, it is clear that the diagonal elements move by the same amount and in opposite directions. However, the fact that the biggest moves to the right and the smallest to the left, or equivalently $\Delta > 0$, is an interesting property of Jacobi rotations. This simple observation is the building block of this section.

The diagram can be formalized by the equations

$$\tilde{d}_+ = d_+ + \Delta, \tag{2.1}$$

$$\tilde{d}_- = d_- - \Delta, \tag{2.2}$$

$$\Delta \geq 0. \tag{2.3}$$

This is a consequence of (1.1)–(1.8). In fact, in the case $a_{ii} = a_{jj}$ then (2.1)–(2.3) hold for $\Delta = |\tan\theta\ a_{ij}|$. In the case when $a_{ii} \neq a_{jj}$ and we take the inner angle we have

$$\operatorname{sign}(\tan\theta\ a_{ij}) = \operatorname{sign}(\tan(2\theta)\ a_{ij}) = \operatorname{sign}\left(\frac{a_{ij}^2}{a_{jj} - a_{ii}}\right) = \operatorname{sign}(a_{jj} - a_{ii}).$$

Therefore, if $a_{ii} > a_{jj}$ then $\tan\theta\ a_{ij} < 0$ and (1.5) leads to

$$\tilde{d}_+ = \tilde{a}_{ii} = \tilde{a}_{ii} - \tan\theta\ a_{ij} = \tilde{d}_+ + \Delta,$$

for $\Delta = -\tan\theta\ a_{ij} \geq 0$. Analogously, in the case $a_{jj} > a_{ii}$, $\Delta = +\tan\theta\ a_{ij} > 0$. A similar analysis using (1.7) and (1.8) shows that (2.1)–(2.3) are also true if we take θ to be the outer angle.

Let us now look at the sum

$$\Sigma = \sum_{r,s=1,n} |d_r - d_s|.$$

The nice feature of Σ is that if $|d_i - d_m|$ decreases, then this is compensated by a corresponding increase in $|d_j - d_m|$, as can be seen in Fig. 1. Adding these terms gives the inequality

$$\tilde{\Sigma} \geq \Sigma + 2\Delta. \tag{2.4}$$

To prove (2.4) we notice that each of the diagonal elements belongs to one of the five regions indicated in Fig. 1:

$$\begin{aligned} R_1 &= (-\infty, d_- - \Delta), \\ R_2 &= [d_- - \Delta, d_-], \\ R_3 &= (d_-, d_+), \\ R_4 &= [d_+, d_+ + \Delta], \\ R_5 &= (d_+ + \Delta, +\infty). \end{aligned}$$

Equation (2.4) follows from the statements below, which the reader can verify.

1. If $d_m \in R_1 \cup R_5$ then $|\tilde{d}_m - \tilde{d}_+| + |\tilde{d}_m - \tilde{d}_-| = |d_m - d_+| + |d_m - d_-|$.
2. If $d_m \in R_3$ then $|\tilde{d}_m - \tilde{d}_+| + |\tilde{d}_m - \tilde{d}_-| = |d_m - d_+| + |d_m - d_-| + 2\Delta$.
3. If $d_m \in R_2$ then $|\tilde{d}_m - \tilde{d}_+| = |d_m - d_+| + \Delta$ and $|\tilde{d}_m - \tilde{d}_-| \geq |d_m - d_-| - \Delta$. Therefore $|\tilde{d}_m - \tilde{d}_+| + |\tilde{d}_m - \tilde{d}_-| \geq |d_m - d_+| + |d_m - d_-|$.
4. The same result above holds for $d_m \in R_4$.
5. Finally, $|\tilde{d}_+ - \tilde{d}_-| = |d_+ - d_-| + 2\Delta$.

Defining $\|v\|_\infty = \max|v_i|$, an analysis similar to the one above shows

$$\|\operatorname{sort}(d) - \operatorname{sort}(\tilde{d})\|_\infty \leq \Delta, \tag{2.5}$$

Equation (2.5) follows by looking at the maximum change in $\operatorname{sort}(d)$ in regions R_2 and R_4. Applying (2.4) to the sequence A^r we have

$$2\Delta_r \leq \Sigma_{r+1} - \Sigma_r.$$

Summing this last equation from $r = 1$ to $r = k$ gives

$$2\sum_{r=1}^{k} \Delta_r \leq \Sigma_{k+1} - \Sigma_1.$$

Since $\|A^k\|_F = \|A\|_F$, the Σ_k are bounded. Therefore

$$\sum_{k=1}^{\infty} \Delta_k \leq C < \infty.$$

As a consequence of (2.5),

$$\sum_{k=1}^{k} \|\text{sort}(d^k) - \text{sort}(\tilde{d^k})\|_\infty \leq \sum_{k=1}^{k} \Delta_k.$$

Thus the series $\sum_{k=1}^{\infty} \|\text{sort}(d^k) - \text{sort}(\tilde{d^k})\|_\infty$ is convergent. This implies that $\text{sort}(d^k)$ is (absolutely) convergent, which finishes the proof of the theorem. □

The theorem doesn't state that each diagonal element will converge. The diagonal elements might, for example, be permuted in each rotation. The following corollary shows that this is not the case if we always choose the inner angle.

COROLLARY 1. If at each rotation we choose the inner angle, then the diagonal converges.

Proof. In this case (1.5) implies that $\max |d_i^k - d_i^{k+1}| \leq \Delta_k$. Since $\sum \Delta_k$ converges, d^k also converges. □

The diagonal will also converge if in each rotation we choose either the inner or the sorting angle, but the proof is more involved (see [M]).

3. Orderings with higher order of convergence. The goal of this section is to show that, for matrices satisfying the nonrepetition hypothesis, the usual characterization of the convergence of the Jacobi method as "quadratic" [P] is not a complete picture of the actual dynamics of the off-diagonal elements. In actuality, each entry will decay at its own speed, some faster than others. The usual theorems about quadratic convergence show that the decay is at least quadratic, but they do not say anything about the fast decaying entries. This observation suggests trying to identify which elements are decaying more slowly and pivoting them more often than the others. That is exactly what is done when we look for the maximum pivot or use thresholds, and that is why these tricks work. Of course this identification of the slowly decaying elements should be cheap if it is supposed to be practical.

In this section we are concerned only with the asymptotic rate of convergence. In the next section we will discuss practical aspects of the convergence. We will present examples where the slower parts are easily localized that will lead to a strategy for obtaining a higher order of convergence. The idea is that by pivoting twice in the slower part and once in the faster part we can reduce the off-diagonal to its cube. The slow part will amount to one quarter of the matrix. Therefore the total work involved to get this cubic reduction is 1.25 sweeps. If we think in terms of sweeps this corresponds to a convergence exponent of $3^{4/5} \approx 2.41$. The results are formalized in Theorem 2 below. A simple example of the difference in the sizes of the off-diagonal entries is as follows. Suppose we have a symmetric matrix of the form

$$M = \begin{pmatrix} O(\epsilon) & O(\epsilon) \\ & O(\epsilon) \end{pmatrix},$$

where the $O(\epsilon)$ refers to the off-diagonal part. Assume that the diagonal elements are far apart. Take the ordering to be any block ordering for which we perform the rotations in the diagonal blocks first and only then in the upper-right corner block.

As usual, after we pivot in the diagonal blocks we have

$$\tilde{M} = \begin{pmatrix} O(\epsilon^2) & O(\epsilon) \\ & O(\epsilon^2) \end{pmatrix}.$$

Now when pivoting in the upper-right corner block we will have, for a_{uv} in one of the diagonal blocks,

$$\tilde{a}_{uv} = \cos\theta a_{uv} \pm \sin\theta a_{rs}$$

for some a_{rs} in the upper-right corner block (see (1.1)). But from equation (9-5-2) in [P] we have

$$|\theta| \leq \frac{|a_{ij}|}{|a_{ii} - a_{jj}|},$$

which implies that $\theta = O(a_{ij}) = O(\epsilon)$ if the diagonal entries are far apart. Therefore, since a_{rs} is $O(\epsilon)$, $\tilde{a}_{uv} = O(\epsilon^2)$ and, as a conclusion, the diagonal blocks will remain $O(\epsilon^2)$ until the end of the sweep.

Now comes the surprising part: At the end of the sweep the upper-right corner block will be $O(\epsilon^3)$. To see why this is true, note that if a_{uv} is the upper-right corner block then

$$\tilde{a}_{uv} = \cos\theta a_{uv} \pm \sin\theta a_{rs}$$

for some a_{rs} in one of the diagonal blocks. Thus

$$\tilde{a}_{uv} = \cos\theta a_{uv} + O(\epsilon^3), \tag{3.1}$$

since $\theta = O(\epsilon)$ and $a_{rs} = O(\epsilon^2)$. Observe, again, that at some point a_{uv} will be the pivot. Thus just after this rotation $\tilde{a}_{uv} = 0$. An easy inductive argument using (3.1) shows that at the end of the sweep we will have $\tilde{a}_{uv} = O(\epsilon^3)$.

Therefore, after the sweep is finished the matrix will look like

$$\bar{M} = \begin{pmatrix} O(\epsilon^2) & O(\epsilon^3) \\ & O(\epsilon^2) \end{pmatrix}. \tag{3.2}$$

This shows that the upper-right corner block will decay faster than the rest of the matrix.

We now show how to exploit this observation so as to obtain a strategy for high order of convergence. The idea is to repeat the argument above by breaking the $O(\epsilon^2)$ blocks in smaller sublocks, as in the following matrix.

$$M_0 = \begin{bmatrix} O(\epsilon^2) & O(\epsilon^2) & O(\epsilon^3) & \\ & O(\epsilon^2) & & \\ & & O(\epsilon^2) & O(\epsilon^2) \\ & & & O(\epsilon^2) \end{bmatrix}$$

Suppose that we apply the ordering above to the submatrix formed by the three blocks in the upper-left corner and also to the submatrix corresponding to the three blocks in the lower-right corner. By the same argument we get a matrix of the form

$$M_1 = \begin{bmatrix} O(\epsilon^4) & O(\epsilon^6) & O(\epsilon^3) & \\ & O(\epsilon^4) & & \\ & & O(\epsilon^4) & O(\epsilon^6) \\ & & & O(\epsilon^4) \end{bmatrix}$$

At this point it would be natural to pivot in the $O(\epsilon^3)$ block. If we do this we get

$$M_2 = \begin{bmatrix} O(\epsilon^4) & O(\epsilon^6) & O(\epsilon^7) & \\ & O(\epsilon^4) & & \\ & & O(\epsilon^4) & O(\epsilon^6) \\ & & & O(\epsilon^4) \end{bmatrix}$$

The ϵ^3 terms did not get cubed because when they interact with the ϵ^4 terms they become ϵ^7. However, suppose that instead we pivot first in the $O(\epsilon^4)$ blocks and only then in the $O(\epsilon^3)$ block. This will cost an additional $\approx .25$ of a sweep since the $O(\epsilon^4)$ blocks contain roughly .25 of the off-diagonal entries of the matrix. The same kind of analysis shows that this will lead to

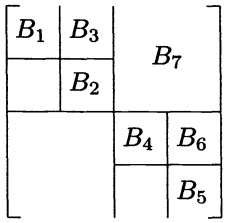


FIG. 2. *Partition of a matrix in three levels: Level* 1, B_7; *Level* 2, B_3 *and* B_6; *Level* 3, B_1, B_2, B_4, *and* B_5.

$$M_3 = \begin{bmatrix} O(\epsilon^8) & O(\epsilon^6) & O(\epsilon^3) & \\ & O(\epsilon^8) & & \\ & & O(\epsilon^8) & O(\epsilon^6) \\ & & & O(\epsilon^8) \end{bmatrix}$$

and then to

$$M_4 = \begin{bmatrix} O(\epsilon^8) & O(\epsilon^6) & O(\epsilon^9) & \\ & O(\epsilon^8) & & \\ & & O(\epsilon^8) & O(\epsilon^6) \\ & & & O(\epsilon^8) \end{bmatrix}$$

The $O(\epsilon^3)$ terms in M_2 are responsible for the $O(\epsilon^6)$ entries at the diagonal blocks in M_3.

To summarize, by applying the equivalent of 1.25 sweeps to M_0 we get to M_3, and we have reduced the off-diagonal part to its cube.

Let us call a 2.41 *ordering* any strategy in which we divide the off-diagonal of the matrix into blocks as in Fig. 2 above and pivot according to the block ordering

$$\{1, 2, 3, 4, 5, 6, 1, 2, 4, 5, 7\}.$$

Inside of each block the pivots are chosen according to any of the usual cyclic orderings, pivoting each entry only once.

In order to make a more formal statement about what was said above, suppose that A^k is the sequence of iterates obtained applying a given 2.41 ordering to a matrix A, and $B_i^k, i = 1, \ldots, 7$ are the off-diagonal blocks defined when we decompose A^k as in the matrix above. Let us also call N the number of rotations needed to perform

one sweep of O. We are interested in knowing what happens to A at the end of s sweeps, or, in other words, in the properties of A^{sN}. In order to measure how big A_{off}^{sN} is in a scale-invariant fashion, we define

$$\Gamma_s = \max\left\{\left(\frac{\|B_7^{sN}\|_F}{\Delta d_A}\right)^{\frac{1}{3}}, \left(\frac{\|B_i^{sN}\|_F}{\Delta d_A}\right)^{\frac{1}{2}}, \ i = 1, \ldots, 6\right\}.$$

We can then state the main theorem of this section.

THEOREM 2. *Let $A = A^0$ be an $n \times n$ symmetric matrix, $n \geq 8$, and assume we apply the Jacobi method according to a 2.41 ordering to it. If*

$$\Gamma_0 \leq \frac{1}{16\alpha_n\sqrt{\alpha_n}}, \tag{3.3}$$

where $\alpha_n = \frac{n}{4} + 1 \geq 3$, then

$$\Gamma_{s+1} \leq c_n \Gamma_s^3$$

for $c_n = 5.6\alpha_n^3$, which implies

$$\Gamma_s \leq (C_n\Gamma_0)^{3^s} \tag{3.4}$$

with $C_n = \sqrt{c_n}$.

This theorem states that if the ratio of the off-diagonal entries to the difference of the diagonal entries is small enough (3.3) then it will be cubed after one sweep of a 2.41 ordering. The first proof of Theorem 2 was presented in [M]. After that [HR] gave a proof that provides sharper bounds and also shows global convergence for the 2.41 ordering that performs the orderings by rows inside of each block. These proofs are quite tedious and we refer the reader to [HR].

The 2.41 orderings are just one example of orderings with superquadratic convergence. Their nicest property is simplicity, since it is clear why and when we should pivot again in the slowest part. We also believe that they are optimal in the sense that any strategy that reduces the off-diagonal to its cube requires at least as many rotations as one sweep of a 2.41 ordering. We do not have a proof of this optimality, but in [M] we present some heuristics in this direction. Another question is to determine which orderings are optimal in the sense of having the biggest convergence exponent. We have not made any progress in solving this problem, which we expect to be very hard. A natural idea to get higher order of convergence would be to repeat (3.2) and partition the matrix in more levels (see Fig. 2). Unfortunately, this idea did not work and our efforts to find strategies with convergence exponent higher than 2.41 have failed.

We present now a method for analyzing the terminal behavior of any particular ordering. Assume that each entry is of order $\epsilon^{e_{ij}}$, for $\epsilon \ll \delta$. If the nonrepetition hypothesis holds then (1.3) implies that when pivoting at a_{ij} we will also have $\theta = O(\epsilon^{e_{ij}})$. Supposing then that a_{rs} and a_{uv} interact during this rotation, we will have

$$\tilde{a}_{rs} = \cos\theta a_{rs} \pm \sin\theta a_{uv} = O(\epsilon^l)$$

for $l = \min\{e_{rs}, e_{uv} + e_{ij}\}$. Therefore the exponents will evolve according to the formulae,

$$e_{ir}^{k+1} = \min\{e_{ir}^k, e_{ij}^k + e_{jr}^k\}, \tag{3.5}$$

$$e_{jr}^{k+1} = \min\{e_{jr}^k, e_{ij}^k + e_{ir}^k\}, \tag{3.6}$$

$$e_{ri}^{k+1} = \min\{e_{ri}^k, e_{ij}^k + e_{rj}^k\}, \tag{3.7}$$

$$e_{rj}^{k+1} = \min\{e_{rj}^k, e_{ij}^k + e_{ri}^k\}, \tag{3.8}$$

$$e_{ij}^{k+1} = +\infty, \tag{3.9}$$

and

$$e_{uv}^{k+1} = e_{uv}^k \tag{3.10}$$

for $\{u, v\} \cap \{i, j\} = \emptyset$.

These considerations lead to the following lemma, whose easy inductive proof is left to the reader.

LEMMA 1. Let O be an ordering for which there are e_{ij}'s and λ such that if we let the e_{ij}'s evolve according to (3.5)–(3.10), we get

$$\bar{e}_{ij} \geq \lambda e_{ij}$$

after N rotations. Suppose that A^0 is such that $\Delta d_{A^k} > \delta > 0$ for all $k > k_0$. Then $k > k_0$ implies

$$\Gamma_{k+N} \leq C\Gamma_k^{\lambda}$$

for

$$C = \frac{1}{2}\left(2 + \frac{1}{\delta}\right)^{2^N}$$

and

$$\Gamma_s = \max\{\ |a_{ij}^s|^{e_{ij}}\ \}.$$

The size of C makes this result a lemma instead of a theorem. If the ordering O in Lemma 1 has period N, then a lower bound for its convergence exponent is

$$(\lambda_{n,O})^{\frac{n(n-1)}{2N}},$$

where

$$\lambda_{n,O} = \max\left\{\ \min\left\{\frac{\bar{e}_{ij}}{e_{ij}},\ i < j = 2, \ldots, n\right\},\ e_{ij} \geq 1\right\}.$$

Another detail in Lemma 1 is that before we can apply it we need to find the appropriate e_{ij} and λ. Determining $\lambda_{n,O}$ is in itself an interesting problem. We provide below a naive algorithm for finding $\lambda_{n,O}$. Observe that the equation for the e_{ij} and $\lambda_{n,O}$ is similar to an eigenvalue problem: $F(e) \geq \lambda e$. Our idea is to use the power method for solving this problem. This leads to the following algorithm.

ALGORITHM 1

1. Set $e_{ij}^0 = 1.0$ for all i, j.
2. Apply one sweep according to O and (3.5)–(3.10) in order to obtain $\bar{e}_{ij}^k$.
3. Estimate $\lambda_{n,O}$ by $\lambda_k = \min\{\frac{\bar{e}_{ij}^k}{e_{ij}^k}\}$.

4. Set $m_k = \min_{i,j}\{\bar{e}_{ij}^k\}$ and normalize the e_{ij} by

$$e_{ij}^{k+1} = \frac{\bar{e}_{ij}^k}{m_k}.$$

5. Repeat 2, 3, and 4 until convergence.

We have not been able to prove the convergence of Algorithm 1, but some partial results in this direction are presented in [M].

Algorithm 1 can be used with Lemma 1 in order to give a rigorous proof of higher-order convergence, under the nonrepetition hypothesis, for specific orderings and dimensions. By reducing λ a little bit, so that we can take into account the rounding errors, we can even use a λ computed numerically. Our experiments with Algorithm 1 suggest that for the row ordering the convergence exponent approaches 2 rather quickly as the dimension n increases. On the other hand, Algorithm 1 shows that the *ordering by diagonals*, with pivots given by

(3.11)

$$\{(1,n),(1,n-1),(2,n),\ldots,(i,j),(i+1,j+1),\ldots,(i,n),(1,n-i),\ldots,(n-1,n)\},$$

has convergence exponent bigger than 2 even for n moderately large. In fact, we got $\lambda = 2.25$ for $n = 128$ and $\lambda = 2.2$ for $n = 256$. However, we point out that Algorithm 1 gives only a lower bound for the convergence exponent.

An interesting observation is that experiments with a variation of Algorithm 1, where the rotations were performed in order to eliminate the smaller exponents, also seem to lead to quadratic convergence when the dimension n is large. This has the surprising implication that for large n, distinct eigenvalues, and infinite precision, the classical "greedy" strategy in which we pivot the biggest element in the off-diagonal is not optimal! Any 2.41 ordering is better. Of course infinite precision is not what we have in practice, and looking to the biggest possible pivot, or variations on the theme, seems to lead to optimal convergence in a practical sense [GK], with the caveat that this idea is hard to implement on massively parallel machines.

4. Comparing orderings in practice. This section describes how the ideas discussed in the preceding sections perform in practice, where by "practice" we mean the first few sweeps. We use two classes of matrices: matrices with random entries coming from a uniform distribution in $[-1,1]$ and matrices with eigenvalues $1,\alpha,\ldots,\alpha^{n-1}$, for some α, which we call *graded matrices*. The experiments show that the performance depends both on the ordering and on the matrices. They also show, unfortunately, that the superquadratic convergence of 2.41 orderings is not relevant in practice. This happens mainly because the dynamics of the first few sweeps is dictated more by the constant factors in front of the "$O(\epsilon)$" estimates from §3 than by the asymptotic rate of convergence. As one way to improve these constant factors, especially for 2.41 orderings, we propose choosing the rotations in such a way to sort the diagonal. We will explain why this idea works so well for 2.41 orderings, making them competitive with the row ordering for matrices with random entries.

The experiments use the ordering by rows and the ordering by diagonals (see (3.11)), and a 2.41 ordering, with the ordering by diagonals inside of each block. The results of the first experiment are listed in Table 1. Each entry on this table corresponds to averages over 100 matrices, generated using a standard random number generator for the uniform distribution in $[-1,1]$. The entries on and above the diagonal were generated independently.

TABLE 1
$n \times n$ random matrices.

n	Work until convergence								
	Row	Row sort at first	Row sort always	Diagonal	Diagonal sort at first	Diagonal sort always	2.41	2.41 sort at first	2.41 sort always
8	4.89	4.90	4.93	4.66	4.74	4.82	5.19	5.04	5.08
32	6.67	6.63	6.46	6.40	6.05	6.09	6.13	6.02	5.69
128	7.95	7.94	7.55	6.97	6.91	6.90	9.14	8.66	7.34

Comparing columns 2, 5, and 6 in Table 1 we see that that the ordering by diagonals is the winner among matrices with random entries and the 2.41 orderings do poorly. The other columns of Table 1 show the effect of permuting the rows and columns of the matrix in order to sort the diagonal. For $n = 128$, sorting improved the performance of all orderings, but the effect on the 2.41 orderings is much more pronounced.

The main reason why 2.41 orderings do not perform well is that pivoting in the $O(\epsilon^3)$ terms at the first level (see Fig. 2) destroys the $O(\epsilon^8)$ terms that we have already obtained at the third level. Sorting the diagonal attenuates this increase of the third level since it reduces the angles of rotations on the first level by increasing the difference between the corresponding diagonal entries.

In the same way that each ordering has its slower and faster entries, the effect of sorting also depends on the ordering. For the 2.41 orderings it is clear that sorting can lead to improvement. For the other orderings we have empirically verified that sorting also improves convergence and we believe that a similar phenomenon of coupling between angles and slow-fast parts happens, but in a weaker form. Again, 2.41 orderings have the nice property of being simple and allowing us to get an understanding of sorting in its simplest effects.

The next experiments show that the effect of sorting also depends on the matrix. Our test matrices are simplified versions of the matrices in [DV]. Each test matrix was obtained by applying $5n(n-1)$ rotations, at randomly chosen entries and by randomly chosen angles, to a diagonal matrix with diagonal entries in a geometric progression from 1 to κ^{-1}, where κ is the condition number. Each number in Tables 2 and 3 is the average over 500 experiments.

TABLE 2
$n \times n$ graded matrices, condition number $= 10^{20}$.

n	Total work performed until convergence								
	Row	Row sort at first	Row sort always	Diagonal	Diagonal sort at first	Diagonal sort always	2.41	2.41 sort at first	2.41 sort always
8	4.38	3.90	3.32	5.40	5.28	5.35	4.76	4.48	4.17
32	7.27	6.98	5.23	12.03	12.10	10.24	10.00	9.77	7.40
64	9.15	8.85	6.77	15.55	15.57	11.91	14.11	13.49	9.20

TABLE 3
$n \times n$ graded matrices, condition number $= 10^{60}$.

n	Total work performed until convergence								
	Row	Row sort at first	Row sort always	Diagonal	Diagonal sort at first	Diagonal sort always	2.41	2.41 sort at first	2.41 sort always
8	2.79	2.31	2.19	3.44	3.34	3.59	2.89	2.65	2.66
32	4.82	4.62	3.55	8.60	8.60	8.26	6.90	6.70	4.61
64	6.15	5.91	4.60	11.47	11.79	10.80	10.18	10.05	6.58

Notice that the effect of sorting is much more pronounced for graded matrices. Another interesting point in Tables 2 and 3 is the awful performance of the ordering by diagonals, which performs best in the first experiment. The disparity between the convergence for the two different classes of matrices makes it clear that the "best" strategy in practice is hard to characterize, since the performance of the different orderings depend strongly on the matrix. Sorting, on the other hand, leads to improvement in both classes.

REFERENCES

[BP] K. W. BRODLIE AND M. J. D. POWELL, *On the convergence of cyclic Jacobi methods*, J. Instit. Math. Appl., 15 (1975), pp. 279–287.

[CVD] J. P. CHARLIER AND P. VAN DOOREN, *On Kogbetlianz's algorithm in the presence of clusters*, Linear Algebra Appl., 95 (1987), pp. 135–160.

[DV] J. DEMMEL AND K. VESELIĆ, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.

[EM] P.J. EBERLEIN AND M. MANTHARAM, *Block Recursive Algorithm to Generate Jacobi-sets*, Tech. Report, Dept. of Comp. Sci., State University of New York at Buffalo.

[HR] V. HARI AND N.H.RHEE, *On the Global and Cubic Convergence of a Quasi-Cyclic Jacobi Method*, Tech. Report 91-27, AHPCRC, University of Minnesota; Numer. Math., submitted.

[M] W. MASCARENHAS, *On the Convergence of the Jacobi Method for Arbitrary Orderings*, Ph.D. thesis, Numerical Analysis Report 91-2, Department of Mathematics, Massachusetts Institute of Technology, Cambridge.

[P] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1980.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, pp. 1210–1222, October 1995

014

MULTISPLITTING PRECONDITIONERS BASED ON INCOMPLETE CHOLESKI FACTORIZATIONS*

R. BRU†, C. CORRAL†, A. MARTÍNEZ†, AND J. MAS†

Abstract. Let $Ax = b$ be a linear system where A is a symmetric positive definite matrix. Preconditioners for the conjugate gradient method based on multisplittings obtained by incomplete Choleski factorizations of A are studied. The validity of these preconditioners when A is an M-matrix is proved and a parallel implementation is presented.

Key words. multisplitting, preconditioner, incomplete Choleski factorization, parallel algorithm

AMS subject classifications. 15A12, 65F10, 65F35, 65Y05

1. Introduction.

Let A be an $n \times n$ nonsingular symmetric M-matrix (see [4] for definition and properties of M-matrix). We consider the solution of the linear system

$$Ax = b \tag{1}$$

by the preconditioned conjugate gradient (PCG) method. This minimization method of conjugate directions consists of solving the new linear system

$$\hat{A}\hat{x} = \hat{b},$$

where $\hat{A} = SAS^T$, $\hat{x} = S^{-T}x$, and $\hat{b} = Sb$, provided that $\text{cond}(\hat{A}) < \text{cond}(A)$ or that a better clustering of the matrix eigenvalues is obtained. The matrix

$$K = (S^T S)^{-1} \tag{2}$$

is said to be the *preconditioning matrix* or *preconditioner*.

The PCG method can be written in terms of the matrix K and the main difference from the conjugate gradient method is the need to solve an auxiliary system $Ks = r$ in each iteration, for obtaining the next conjugate direction.

There exist several techniques for constructing the preconditioner (see [14], [9], and [11]). A usual technique consists of considering a splitting of the matrix A

$$A = P - Q, \tag{3}$$

where P is a nonsingular matrix and the spectral radius of $P^{-1}Q$ is less than 1 (i.e., $\rho(P^{-1}Q) < 1$). From this splitting one constructs the preconditioner matrix K by using a partial sum of the power series of A^{-1}. That is

$$K = P(I + T + \cdots + T^{m-1})^{-1}, \tag{4}$$

where $T = P^{-1}Q$. This preconditioner is called *m-step polynomial preconditioner*. In addition, for this type of preconditioners one can solve the auxiliary system $Ks = r$ by doing m steps

$$Ps^{(i)} = Qs^{(i-1)} + r, \quad i = 1, 2, \ldots, m$$

* Received by the editors May 10, 1993; accepted for publication (in revised form) by T. Manteuffel November 7, 1994.

† Departament de Matemàtica Aplicada, Universitat Politècnica de València, 46071 València, Spain (rbru@mat.upv.es). Supported in part by Spanish CICYT grant TIC91-1157-CO3-01 and ESPRIT III Basic Research Programme of the EC contract 9072 (project GEPPCOM) .

of the iteration given by the splitting (3) for the system $As = r$, choosing $s^{(0)} = 0$.

Adams and Ong [2] give an additive polynomial preconditioner, by considering two different splittings of the matrix A based on the symmetric successive overrelaxation (SSOR) method, and then averaging the updates of each splitting. Huang and O'Leary in [8] study a Krylov multisplitting algorithm based on multisplitting of a symmetric positive definite matrix. Other related work is found in [6].

A usual way for choosing the splitting (3) is based on the incomplete Choleski factorization. Our goal in this paper is to construct an additive polynomial preconditioner based on a multisplitting defined by O'Leary and White [12], obtaining that multisplitting by means of the incomplete Choleski factorizations. Thus, this method can be clasified as generalized preconditioned conjugate gradient (GPCG) following the taxonomy of conjugate gradient methods given in [3]. In §2 we deal with this preconditioner and we give sufficient conditions for K to be symmetric and positive definite. In §3 we work with a particular multisplitting and study the parallel implementation of the PCG method in this case. Finally, we give some numerical results on a distributed memory multicomputer (Parsys SN1040).

2. Multisplitting preconditioner. Let $\mathcal{G}$ be the set of all pairs of indices (i, j),

$$\mathcal{G} = \{(i, j) : 1 \leq i, j \leq n\}$$

and let G be a subset of $\mathcal{G}$ such that

$$\begin{array}{ll} \text{(a)} & (i, i) \in G, \quad 1 \leq i \leq n. \\ \text{(b)} & \text{If } (i, j) \in G, \text{ then } (j, i) \in G. \end{array} \tag{5}$$

The set G is said to be the *nonzero set* of the following factorization. According to Meijerink and Van der Vorst [11], when A is a symmetric M-matrix then there exists a unique lower triangular matrix L and a nonnegative matrix N with

$$\begin{aligned} l_{ij} &= 0 \text{ if } (i, j) \notin G, \\ n_{ij} &= 0 \text{ if } (i, j) \in G, \end{aligned}$$

such that

$$A = M - N = LL^T - N$$

holds. This is the so-called *incomplete Choleski factorization* (ICF) of A. In addition, this defines a regular splitting. Moreover, the nonzero entries of M match the corresponding entries of A.

Consider p subsets, $G_1, G_2, \ldots, G_p$, of $\mathcal{G}$ satisfying (5). Then from the p incomplete Choleski factorizations, obtained from these subsets as nonzero sets of the factorizations, we obtain the p regular splittings

$$A = M_k - N_k, \quad k = 1, 2, \ldots, p. \tag{6}$$

Let us recall the definition of a multisplitting as introduced by O'Leary and White [12].

DEFINITION 1. *Let A, M_k, N_k, and D_k, $k = 1, 2, \ldots, p$, be matrices of size $n \times n$. Then (M_k, N_k, D_k) is said to be a multisplitting of A if*

(i) $A = M_k - N_k$, *where M_k is nonsingular,* $k = 1, 2, \ldots, p$,

(ii) $\sum_{k=1}^{p} D_k = I$, *where* D_k *are nonnegative diagonal matrices.*

In the case that the p splittings (i) *are regular, the multisplitting is called regular.*

Now we are ready to construct the m-step additive polynomial preconditioner based on a multisplitting. As mentioned in the Introduction, we define the preconditioner $K = \mathcal{M}_m$ in such a way that solving $Ks = r$ amounts to doing m steps of the iteration scheme

$$s^{(i)} = \sum_{k=1}^{p} D_k M_k^{-1} N_k s^{(i-1)} + \sum_{k=1}^{p} D_k M_k^{-1} r, \qquad i = 1, 2, \ldots, m \tag{7}$$

with $s^{(0)} = 0$. Then the updated vector from m steps is given by

$$s^{(m)} = (I + H + \cdots + H^{m-1})Wr,$$

where

$$W = \sum_{k=1}^{p} D_k M_k^{-1} \quad \text{and} \quad H = \sum_{k=1}^{p} D_k M_k^{-1} N_k.$$

Therefore the m-step additive polynomial preconditioner related to the above multisplitting defined in (6) is given by

$$K^{-1} = \mathcal{M}_m^{-1} = (I + H + \cdots + H^{m-1})W. \tag{8}$$

Let us check whether this preconditioner is valid. The next results give sufficient conditions on the multisplitting obtained by incomplete Choleski factorizations so that $\mathcal{M}_m^{-1}$ is symmetric positive definite.

THEOREM 1. *Let A be a symmetric positive definite matrix. Consider the subsets of indices* $G_1, G_2, \ldots, G_p$ *satisfying* (5) *such that the corresponding incomplete Choleski factorizations yield the multisplitting* (M_k, N_k, D_k). *Suppose that* $D_k M_k = M_k D_k$ *for* $k = 1, 2, \ldots, p$. *Then the preconditioning matrix* $\mathcal{M}_m^{-1}$ *defined by* (8) *is symmetric.*

Proof. Since the matrices $M_k = L_k L_k^T$ are symmetric and since they commute with D_k, $k = 1, 2, \ldots, p$, the matrix W is symmetric also.

The matrix $\mathcal{M}_m^{-1}$ can be written as

$$\mathcal{M}_m^{-1} = \sum_{i=0}^{m-1} H^i W.$$

Since

$$H = \sum_{k=1}^{p} D_k M_k^{-1} N_k = I - WA,$$

we have

$$\mathcal{M}_m^{-1} = \sum_{i=0}^{m-1} [(I - WA)^i W],$$

that is, $\mathcal{M}_m^{-1}$ is a linear combination of

$$(WA)^j W, \quad j = 0, 1, \ldots, m-1,$$

which are symmetric since W and A are symmetric and the result follows. □

The condition $D_k M_k = M_k D_k$ holds if the matrices D_k are scalar. Furthermore, the same condition holds for the class of multisplittings defined below.

DEFINITION 2. *We say that the multisplitting* (M_k, N_k, D_k) *is* block diagonal conformable *if, for every* k*:*

(i) M_k *is block diagonal,* $M_k = \begin{bmatrix} M_{11}^{(k)} & & & \\ & M_{22}^{(k)} & & \\ & & \ddots & \\ & & & M_{qq}^{(k)} \end{bmatrix}$.

(ii) D_k *is block scalar, that is* $D_k = \begin{bmatrix} D_{11}^{(k)} & & & \\ & D_{22}^{(k)} & & \\ & & \ddots & \\ & & & D_{qq}^{(k)} \end{bmatrix}$,

where $D_{ii}^{(k)} = d_i^{(k)} I$, *and*

(iii) *the size of* $M_{ii}^{(k)}$ *coincides with the size of* $D_{ii}^{(k)}$, $i = 1, 2, \ldots, q$.

We quote that the size of one block $M_{ii}^{(k)}$ (and $D_{ii}^{(k)}$), $i = 1, 2, \ldots, q$, can be different from the size of another block $M_{jj}^{(k)}$ or $M_{ii}^{(l)}$.

The following result establishes a necessary and sufficient condition for the preconditioning matrix to be positive definite.

THEOREM 2. *Let* A *be a symmetric positive definite matrix. Consider the subsets of indices* $G_1, G_2, \ldots, G_p$ *satisfying* (5) *such that yield the block diagonal conformable multisplitting* (M_k, N_k, D_k). *Then the matrix* $\mathcal{M}_m^{-1}$ *defined by* (8) *is positive definite if, and only if, the matrix* $I + H + \cdots + H^{m-1}$ *has only positive eigenvalues.*

Proof. We saw in the proof of Theorem 1 that the matrix W is symmetric. Since D_k and M_k^{-1} commute, and since M_k^{-1} is positive definite and D_k is positive semidefinite, then $D_k M_k^{-1}$ is positive semidefinite for all k. And so, W is a positive semidefinite matrix. Let us see, indeed, it is positive definite. Suppose that

$$x^T W x = 0.$$

Then, for all k

$$x^T D_k M_k^{-1} x = 0.$$

Pick an arbitrary component x_l of the vector x. By the properties of the multisplitting, there exists at least an index k such that the l diagonal entry of D_k is positive. Let $D_{ii}^{(k)}$ be the diagonal block of D_k containing this entry and let $\tilde{x}$ be the corresponding subvector of x. Then

$$\tilde{x}^T D_{ii}^{(k)} \left(M_{ii}^{(k)} \right)^{-1} \tilde{x} = 0.$$

Since $D_{ii}^{(k)} \left(M_{ii}^{(k)} \right)^{-1}$ is positive definite, $\tilde{x} = 0$. Hence, $x_l = 0$ and the matrix W is positive definite.

Then, from (8) we can write

$$\mathcal{M}_m^{-1} W^{-1} = I + H + \cdots + H^{m-1},$$

and the result follows by Theorem A.2.7 of [13]. □

THEOREM 3. *Let A be a symmetric positive definite matrix. Consider the subsets of indices $G_1, G_2, \ldots, G_p$ satisfying* (5) *that yield the block diagonal conformable multisplitting (M_k, N_k, D_k). Then*

(i) *if m is odd, the matrix $I + H + \cdots + H^{m-1}$ has only positive eigenvalues.*

(ii) *If m is even, the matrix $I + H + \cdots + H^{m-1}$ has only positive eigenvalues if, and only if, $\rho(H) < 1$.*

Proof. Since A and W are positive definite and $H = I - WA$, the matrix H has real eigenvalues $1 - \alpha$, with $\alpha > 0$. Any eigenvalue of $I + H + \cdots + H^{m-1}$ can be written as

$$1 + \lambda + \cdots + \lambda^{m-1} = \frac{1 - \lambda^m}{1 - \lambda},$$

where λ is an eigenvalue of H.

If m is odd then $\frac{1-\lambda^m}{1-\lambda} > 0$ and hence $I + H + \cdots + H^{m-1}$ has positive eigenvalues.

Consider the case that m is even. If $\rho(H) < 1$ then $\frac{1-\lambda^m}{1-\lambda} > 0$ for each eigenvalue of H.

Conversely, suppose that $\rho(H) \geq 1$. There exists an eigenvalue λ of H such that $\lambda \leq -1$. Hence the matrix $I + H + \cdots + H^{m-1}$ has a nonpositive eigenvalue. □

Therefore, to use the above preconditioner, we must make the matrices $\mathcal{M}_m^{-1}$ positive definite. It is sufficient that A is a symmetric M-matrix.

THEOREM 4. *Let A be a symmetric M-matrix. Let $G_1, G_2, \ldots, G_p$ be subsets of indices satisfying* (5) *such that yield the block diagonal conformable multisplitting (M_k, N_k, D_k). Then*

(i) *the multisplitting (M_k, N_k, D_k) is regular.*

(ii) *The matrix $\mathcal{M}_m^{-1}$ is positive definite.*

Proof. (i) Since G_k satisfies (5), by Theorem 2.4 of [11], the corresponding splitting is regular, and the multisplitting is as well.

(ii) Since the splittings $A = M_k - N_k$ are regular, $H = \sum_{k=0}^{p} D_k M_k^{-1} N_k$ is nonnegative, and so $\rho(H)$ is an eigenvalue of H.

Since the eigenvalues of H are $1 - \alpha$, with $\alpha > 0$, we deduce that $\rho(H) < 1$. Then the proof follows by Theorems 2 and 3. □

Note that $\rho(H) < 1$ obtained above is a particular case of Theorem 1(a) of O'Leary and White [12].

Remark. We want to point out that the conditions on Theorem 1(a) of [12] also guarantee the convergence of the parallel chaotic methods given by Bru, Elsner, and Neumann [5] and hence the iterative scheme (7) can be replaced by the synchronous and asynchronous schemes given in [5] provided that, for the last one, the sequence of integers of Theorem 2.2 of [5] is regulated.

3. Parallel implementation on a distributed memory multiprocessor.

In this section we implement one of the possible parallel algorithms of the PCG method with the multisplitting preconditioner given in the last section.

3.1. Implementation description.

We define a particular multisplitting and we evaluate the performance of the resulting parallel algorithm on the distributed

memory multiprocessor Parsys SN1040 with 16 transputers connected by a bidirectional ring, for the solution of the Laplace equation satisfying boundary conditions in the unit square. A five-point discretization of this equation yields the linear system $Ax = b$, where A has the following structure:

$$\mathrm{A} = \begin{bmatrix} B & -I & & & & \\ -I & B & -I & & & \\ & -I & B & -I & & \\ & & \ddots & \ddots & \ddots & \\ & & & -I & B & -I \\ & & & & -I & B \end{bmatrix},$$

I is the $n \times n$ identity matrix and B is the $n \times n$ tridiagonal matrix

$$B = \begin{bmatrix} 4 & -1 & & & & \\ -1 & 4 & -1 & & & \\ & -1 & 4 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 4 & -1 \\ & & & & -1 & 4 \end{bmatrix}.$$

Therefore, A is a pentadiagonal matrix of size $n^2 \times n^2$. It is clear that A is an M-matrix. For simplicity we consider that n is a multiple of the number of processors p.

Different iterative methods can be applied to solve this problem. In particular, it is well known that the multigrid method is very efficient. However, the later method does not seem to be as parallel as the method presented. As we explain below, we took the algorithm given in [11] as sequential algorithm as speedup reference. Therefore, we selected our splittings similarly to that in [11].

We considered nonzero subsets G contained in the set of pairs of indices illustrated in Fig. 1.

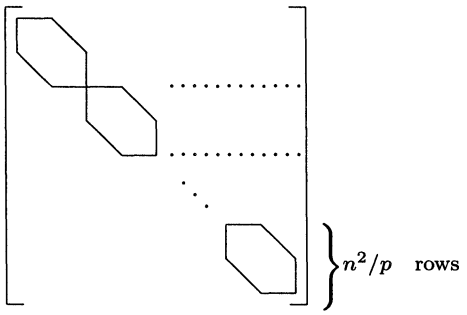


FIG. 1.

Let

$$A = M - N$$

be the corresponding splitting to one of that subset G. Then we construct the

multisplitting (M_k, N_k, D_k), where

$$D_1 = \begin{bmatrix} I & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}, \quad D_2 = \begin{bmatrix} 0 & & & \\ & I & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}, \ldots, D_p = \begin{bmatrix} 0 & & & \\ & 0 & & \\ & & \ddots & \\ & & & I \end{bmatrix},$$

with I and 0 the $\frac{n^2}{p} \times \frac{n^2}{p}$ identity and null matrices, respectively, $M_k = M$ and $N_k = N$ for $k = 1, 2, \ldots, p$.

It is worth noting that there are different multisplittings that yield the same iterative scheme, with distinct matrices M_k. Our choice has been suggested by the splitting used by Meijerink and van der Vorst in [11], but note that the nonzero set G described above (and so the splitting) is not the same as the nonzero set in [11]. In fact, in [11] one considers an incomplete Choleski factorization to be "more" complete than our ICF. In fact, the nonzero set of our factorizations are strict subsets of the nonzero set of [11], which includes properly the nonzero set of A, as can be seen by comparing Figs. 1 and 2, that contains the nonzero set of the ICF of [11]. Finally, we quote that the above multisplitting is block diagonal conformable.

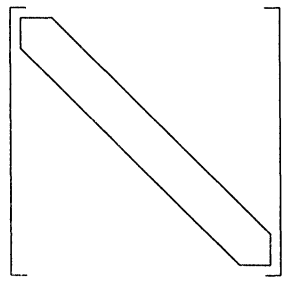


FIG. 2.

This multisplitting allows us to distribute the work and the matrix A among the processors in such a way that each processor stores n^2/p consecutive rows of A. Thus, from the symmetry of the matrix, each processor has access to the corresponding n^2/p columns of A.

According to the above comments, the incomplete Choleski factorization of A can be computed independently by blocks in each processor.

As we mentioned in §2, once we compute the incomplete Choleski factorization, the algorithm requires computing the iterative scheme (7). We discuss below its implementation.

The iteration steps in any processor are defined by

$$Ms^{(i)} = Ns^{(i-1)} + r, \qquad i = 1, 2, \ldots$$

with $s^{(0)} = 0$. Since $M = LL^T$ we must solve the triangular systems

$$Ly = Ns^{(i-1)} + r \quad \text{and} \quad L^T s^{(i)} = y.$$

Because each processor has a block of the matrix L and since the structure of the matrices D_k of the multisplitting, then the solution of those systems can be implemented by solving p independent subsystems of n^2/p unknowns corresponding to the p (decoupled) diagonal blocks of L.

In this way, the computation of the incomplete Choleski factorization and the solution of the triangular systems do not require communication among the different processors. Therefore, each processor computes a block of the vector s and broadcasts the result in each step.

The algorithm also requires some inner products. In this implementation only the inner product (p, Ap) has been parallelized, where p is the conjugate direction. The residual vector is also updated in parallel.

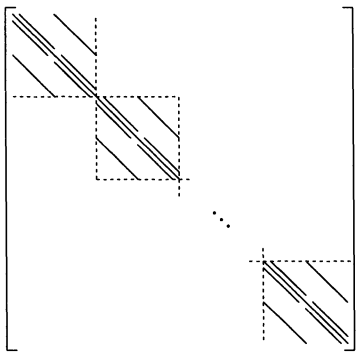


FIG. 3. $N1 = 1$, $N2 = 0$.

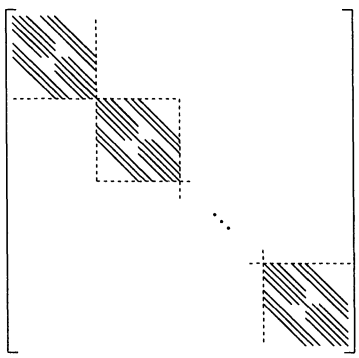


FIG. 4. $N1 = N2 = 2$.

3.2. Results. In our numerical experiments we always considered the same number of blocks $M_{ii}^{(k)}$ as the number of splittings. To simplify, that number equals the number of processors. We chose three types of nonzero sets, varying the number of nonzero diagonals of the matrix L. All nonzero sets, as we mentioned early in this section, are contained in the subsets sketched in Fig. 1.

Let $N1$ denote the number of diagonals that we will add from the main diagonal and let $N2$ denote the number of diagonals that we will add from the last nonzero subdiagonal of A towards the main diagonal. The first type of nonzero set corresponds to the choice $N1 = 1$ and $N2 = 0$, which is equivalent to consider G as a strict subset of the nonzero set of A. This is shown in Fig. 3. The second type of nonzero sets considered corresponds to $N1 = N2 = 2$, which is illustrated in Fig. 4. The last type, $N1 = n-1, N2 = 0$, corresponds to the complete Choleski factorization of each block of M. Note that some nonzero elements of A will be zero in the incomplete Choleski factorization.

We experimented with different matrix sizes, $n^2 = 256, 576, 1024$. The behaviour of the number of iterations as a function of the number of steps is similar. However, the efficiency increases with the size of the matrix. Next we discuss the results for the size $n^2 = 1024$, the biggest possible size that we can perform on one processor.

Table 1 shows the number of iterations in function of the number of steps m for each type of nonzero sets and the number of blocks $M_{ii}^{(k)}$ and Table 2 displays the estimation in norm 1 of the condition numbers of the matrices $\hat{A}$ in function of the number of blocks, of the nonzero sets and of the number of steps (the estimation of the condition number of the matrix A is 422.6537). We used MATLAB to obtain the condition numbers.

TABLE 1
Number of iterations/number of steps, matrix 1024 × 1024.

Number of Steps	Number of iterations											
	1 Block			4 Blocks			8 Blocks			16 Blocks		
	1,0	2,2	31,0	1,0	2,2	31,0	1,0	2,2	31,0	1,0	2,2	31,0
0	49	49	49	49	49	49	49	49	49	49	49	49
1	22	11	1	25	19	17	29	23	23	32	32	31
2	13	7	1	15	11	11	17	13	14	18	17	19
3	11	6	1	13	10	10	14	13	13	16	18	18
4	9	5	1	10	8	8	11	9	10	13	12	13
5	8	4	1	9	7	8	10	10	10	12	14	14
6	8	4	1	9	6	6	9	7	8	11	10	11
7	7	4	1	8	6	6	9	8	8	10	12	11
8	7	3	1	7	5	6	8	6	7	9	8	9
9	6	3	1	7	5	5	8	6	7	9	10	10
10	6	3	1	7	5	5	7	6	6	8	7	9

TABLE 2
Condition numbers of $\hat{A}$, *matrix* 1024 × 1024.

	N. Steps	(1,0)	(2,2)	(31,0)
1 Block	1	61.5564	12.1822	1.0000
	2	30.5612	6.3692	1.0000
	3	21.0536	4.3942	1.0000
	4	16.0138	3.3699	1.0000
	5	12.9838	2.7481	1.0000
4 Blocks	1	114.4251	172.8624	263.2012
	2	54.0487	80.7103	109.4664
	3	40.8713	57.9777	70.7226
	4	30.6652	39.6803	46.1034
	5	25.1980	31.4235	34.5530
8 Blocks	1	133.8638	151.4745	160.9327
	2	60.2741	65.5247	69.8067
	3	44.7025	48.6135	50.8713
	4	33.0003	34.3462	35.6532
	5	26.9623	28.1091	29.0656
16 Blocks	1	134.6827	109.9920	108.2051
	2	59.4193	47.7501	46.9271
	3	44.7839	37.0339	36.4563
	4	33.0271	26.5663	26.0868
	5	27.1491	22.4489	22.1085

Figure 5 displays the condition number of the matrices $\hat{A}$ in function of the number of steps and the number of blocks for the nonzero set corresponding to $N1 = N2 = 2$ and Fig. 6 shows the condition number for the three types of nonzero sets with four blocks $M_{ii}^{(k)}$.

One can observe that the condition number decreases when the number of steps increases, very fast when the number of steps is small. This behaviour is similar to the behaviour of the number of iterations.

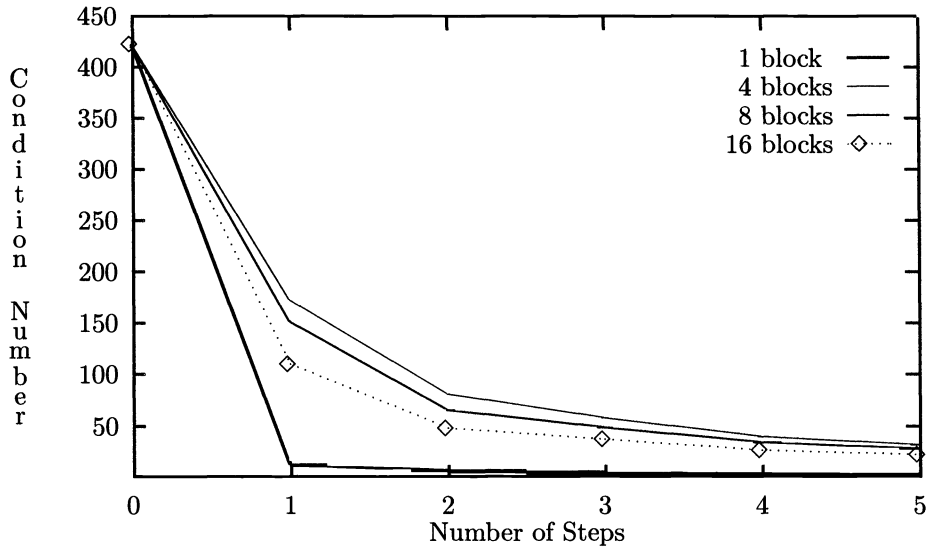


FIG. 5. *Number of steps and condition number of* $\hat{A}$, $N1 = N2 = 2$.

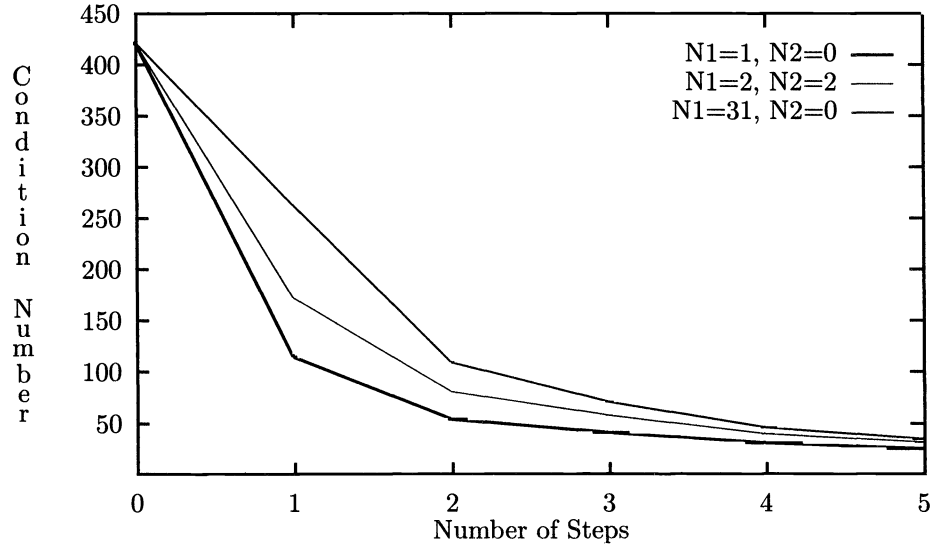


FIG. 6. *Number of steps and condition number of* $\hat{A}$, *four blocks.*

Figure 7 displays the above results for the nonzero set $N1 = N2 = 2$. In all graphics the zero value in the X-axis corresponds to the CG method without preconditioning.

Note that when the number of splittings (and so the number of blocks) increases, the number of iterations (see Table 1) and the condition number of the matrices (see Table 2) rise. This is due to the fact that when the number of blocks increases the nonzero sets in the factorizations becames smaller, and so, the factorizations are "more" incomplete. Then one can expect a bigger number of iterations as Fig. 7 shows.

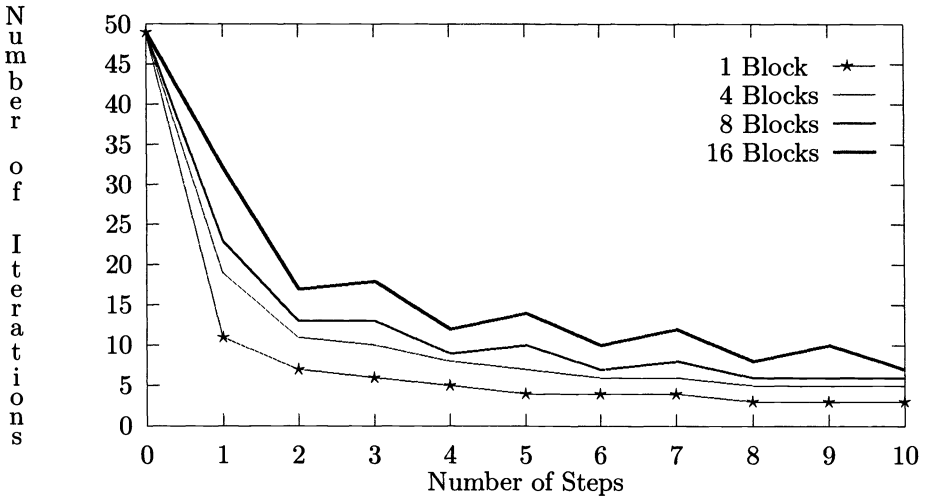


FIG. 7. *Number of steps and iterations, matrix* 1024 × 1024, $N1 = 2$, $N2 = 2$.

Figure 8 shows the number of iterations versus the number of steps for a 1024 × 1024 matrix and 16 blocks (processors) for the three types of nonzero subsets. Finally Fig. 9 shows the parallel execution time versus the number of steps for the same case. The unit time is 10^{-6} seconds.

It seems that the optimum number of steps is one or two, for all types of nonzero sets (this happens for any number of blocks $M_{ii}^{(k)}$). From Table 1 and Fig. 8 one notes that for this number of steps the optimum number of iterations corresponds to

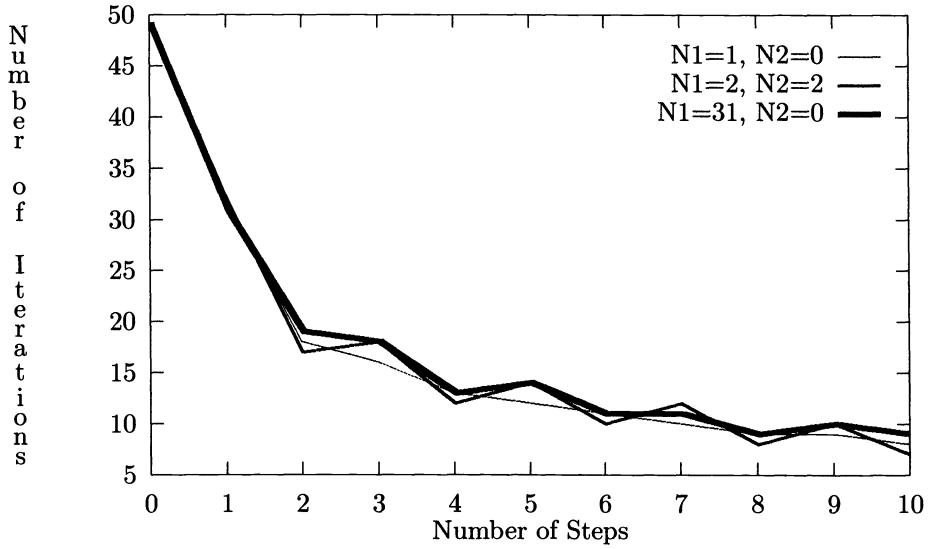


FIG. 8. *Number of steps and iterations, matrix* 1024 × 1024, 16 *blocks.*

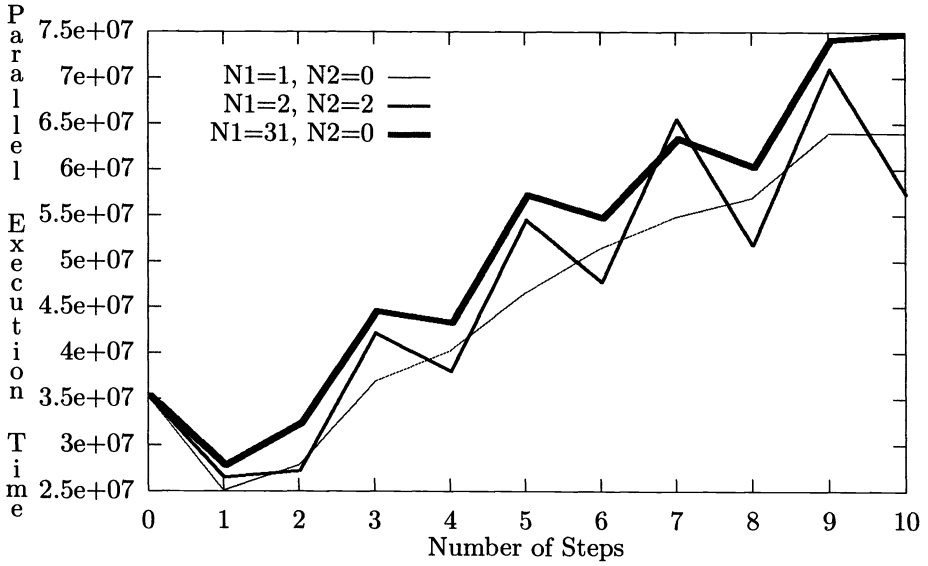


FIG. 9. *Number of steps and parallel execution time, matrix* 1024×1024, 16 *blocks.*

$N1 = n - 1, N2 = 0$. But, for $N1 = N2 = 2$ we get similar results. Then the last nonzero set seems preferable.

From Figs. 7–9 we observe that when the number of steps is odd, then the number of iterations and the parallel execution time increase with respect to the previous even number of steps. By definition of $\hat{A}$, the matrices $\hat{A}$ and $\mathcal{M}_m^{-1}A$ are similar and since

$$\mathcal{M}_m^{-1}A = \left(I + H + H^2 + \cdots + H^{m-1}\right)(I - H) = I - H^m,$$

the condition number of $\hat{A}$ is

$$\text{cond}(\hat{A}) = \max_{i,j} \frac{\lambda_i\left(\mathcal{M}_m^{-1}A\right)}{\lambda_j\left(\mathcal{M}_m^{-1}A\right)} = \frac{1 - \min_i \lambda_i\left(H^m\right)}{1 - \max_j \lambda_j\left(H^m\right)}. \tag{9}$$

Then, if H has negative eigenvalues and m is odd, the numerator of (9) is greater than one, however, if m is even the numerator of (9) is always less than one, and then we can expect a better relative decreasing for even values of m. This fact can explain the numerical behaviour displayed in the corresponding graphics. In fact, when the fill-in of the incomplete Choleski factorization is total, it is easy to check that H is indefinite, in fact, all diagonal entries are zero. Furthermore, we made several numerical experiments with small Laplace matrices with MATLAB and always H was indefinite.

We used the algorithm given in [11] as sequential algorithm as speedup reference, that is $S_p = T_1/T_p$, where T_1 is the execution time of the algorithm given in [11] and T_p is the parallel execution time of our algorithm in p processors. In both algorithms the same number of steps are considered.

Instead of speedups, the efficiencies $E_p = S_p/p$ are given in Table 3.

TABLE 3
Efficiency, matrix 1024×1024.

	No. Steps	(1,0)	(2,2)	(31,0)
4 Processors	1	.4336	.4343	.5770
	2	.7727	.6586	.5604
	3	.8157	.6527	.5066
	4	.8908	.6934	.5218
8 Processors	1	.2252	.2495	.4629
	2	.5638	.4888	.4785
	3	.6658	.4662	.4230
	4	.7358	.5780	.4507
16 Processors	1	.1120	.1093	.3325
	2	.3870	.2935	.3648
	3	.4626	.2830	.3178
	4	.5186	.3811	.3634

One observes that the efficiency increases with the number of steps. This fact is due to our preconditioner "more" incomplete than the preconditioner given in [11], as we discussed in the above subsection. Then the increasing of number of steps improves the accuracy of the solution more than in the other algorithm.

Finally note that the efficiency decreases notoriously when the number of processors increases. This is because of the inadequate use of the capabilities of the processors when the number of processors increases for a fix matrix. One must consider also that the topology used, the bidirectional ring, is not the most efficient.

Acknowledgment. The authors thank Professor L. Elsner for very helpful comments.

REFERENCES

[1] L. ADAMS, *M-step preconditioned conjugate gradient methods*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 452–462.
[2] L. ADAMS AND E. ONG, *Additive polynomial preconditioners for parallel computers*, Parallel Comput., 9, 1988/89, pp. 333–345.
[3] S.F. ASHBY, T.A. MANTEUFFEL, AND P.E. SAYLOR, *A taxonomy for conjugate gradient methods*, SIAM J. Numer. Anal., 27 (1990), pp. 1542–1568.
[4] A. BERMAN AND R.J. PLEMMONS, *Nonnegative Matrices in the Mathematical Science*, Academic Press, New York, 1979.
[5] R. BRU, L. ELSNER, AND M. NEUMANN, *Models of parallel chaotic iteration methods*, Linear Algebra Appl., 103 (1988), pp 175–192.
[6] R. BRU, C. CORRAL, AND J. MAS, *A preconditioned conjugate gradient method on a distributed memory multiprocessor*, Appl. Math. Lett., 8 (1995), pp. 49–53.
[7] M. HESTENES AND E. STIEFEL, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Stand., 49 (1952), pp. 409–436.
[8] C. HUANG AND D. O'LEARY, *A Krylov multisplitting algorithm for solving linear systems of equations*, Linear Algebra Appl., 194 (1993), pp. 9–29.
[9] O. JOHNSON, C. MICCHELLI, AND G. PAUL, *Polynomial preconditioners for conjugate gradient calculations*, SIAM, J. Numer. Anal., 20 (1983), pp. 362–376.
[10] T.A. MANTEUFFEL, *An incomplete factorization technique for positive definite linear systems*, Math. Comput., 34 (1980), pp. 473–497.
[11] J.A. MEIJERINK AND H.A. VAN DER VORST, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. Comput., 31 (1977), pp. 148–162.
[12] D. O'LEARY AND R. WHITE, *Multi-splittings of matrices and parallel solution of linear equations*, SIAM J. Alg. Discrete Methods, 6 (1985), pp. 630–640.
[13] J. ORTEGA, *Introduction to Parallel and Vector Solution of Linear Systems*, Plenum Press, New York, 1988.
[14] Y. SAAD, *Krylov subspace methods on supercomputers*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 1200–1232.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, pp. 1223–1240, October 1995

015

ON THE SYMMETRIC AND UNSYMMETRIC SOLUTION SET OF INTERVAL SYSTEMS*

GÖTZ ALEFELD† AND GÜNTER MAYER‡

Abstract. We consider the solution set S of real linear systems $Ax = b$ with the $n \times n$ coefficient matrix A varying between a lower bound $\underline{A}$ and an upper bound $\overline{A}$, and with b similarly varying between $\underline{b}$, $\overline{b}$. First we list some properties on the shape of S if all matrices A are nonsingular. Then we restrict A to be nonsingular and symmetric deriving a complete description for the boundary of the corresponding symmetric solution set S_{sym} in the 2×2 case. Finally we derive a new criterion for the feasibility of the Cholesky method with which bounds for S_{sym} can be found.

Key words. linear interval equations, unsymmetric solution set, enclosures for the solution set of linear interval systems, symmetric linear systems, symmetric solution set, interval Cholesky method, criteria of feasibility for the interval Cholesky method

AMS subject classifications. 65F05, 65G10

1. Introduction.

In [2] we introduced the interval Cholesky method in order to find an interval enclosure $[x]^C$ of the *symmetric solution set*

$$S_{\text{sym}} := \{x \in \boldsymbol{R}^n \mid Ax = b,\ A = A^T \in [A],\ b \in [b]\}, \tag{1.1}$$

where $[A] = [A]^T$ is a given $n \times n$ matrix with real compact intervals as entries, and where $[b]$ is a given vector with n real compact intervals as components. We showed that $[x]^C$ need not enclose the *solution set*

$$S := \{x \in \boldsymbol{R}^n \mid Ax = b,\ A \in [A],\ b \in [b]\} \supseteq S_{\text{sym}}, \tag{1.2}$$

where in this definition the symmetry of A is dropped. This phenomenon is not astonishing, since, in general, S_{sym} differs from S as was shown in [2] by a simple example.

In this paper (§4) we want to intensify our study on the symmetric solution set S_{sym}. To this end, in §3 we repeat some characteristic properties of S. Parts of them are stated and proved in [4]. We will prove them again in a much shorter way than in [4] following the lines in [8]. We then turn over to properties of S_{sym}. For 2×2 matrices S_{sym} can be represented in each orthant O as the intersection of S, O, and two sets of which the boundary is formed by conic sections. Thus, one deduces at once that in the general $n \times n$ case, the boundary ∂S_{sym} can be curvilinear in contrast to ∂S, which is shown in [4] to be the surface of a polytope.

In the second part of our paper (§5) we prove new criteria for the feasibility of the interval Cholesky method. Assuming the midpoint matrix $\check{A}$ of $[A]$ to be symmetric and positive definite we will show, for example, that the method results in an enclosing interval $[x]^C$ if the spectral radius of $\frac{1}{2}|\check{A}^C|d([A])$ is less than 1, where $d([A]) \in \boldsymbol{R}^{n \times n}$ denotes the diameter of $[A]$ and where $|\check{A}^C|$ is a matrix which is defined later.

We mention that symmetric interval systems have also been considered by Jansson [5]. In his paper the symmetric solution set is enclosed by an iterative process.

* Received by the editors May 16, 1994; accepted for publication (in revised form) by N. Higham November 8, 1994.

† Institut für Angewandte Mathematik, Universität Karlsruhe, D-76128 Karlsruhe, Germany (goetz.alefeld@mathematik.uni-karlsruhe.de).

‡ Fachbereich Mathematik, Universität Rostock, D-18051 Rostock, Germany (guenter.mayer@mathematik.uni-rostock.de).

2. Preliminaries. We start this section with some notations that we use throughout the paper.

By $\boldsymbol{R}^n$, $\boldsymbol{R}^{m\times n}$, $\boldsymbol{IR}$, $\boldsymbol{IR}^n$, $\boldsymbol{IR}^{m\times n}$, we denote the set of real vectors with n components, the set of real $m \times n$ matrices, the set of intervals, the set of interval vectors with n components, and the set of $m \times n$ interval matrices, respectively. By "interval" we always mean a real compact interval. Interval vectors and interval matrices are vectors and matrices, respectively, with interval entries. We write intervals in brackets with the exception of degenerate intervals (so-called *point intervals*) which we identify with the element being contained, and we proceed similarly with interval vectors and interval matrices. Examples are the ith column $e^{(i)}$ of the $n \times n$ identity matrix I and the null matrix O. As usual, we identify $\boldsymbol{R}^{n\times 1}$ and $\boldsymbol{IR}^{n\times 1}$ with $\boldsymbol{R}^n$ and $\boldsymbol{IR}^n$, respectively. We use the notation $[a] = [\underline{a}, \overline{a}] \in \boldsymbol{IR}$ simultaneously without further reference and, in an analogous way, we write $[x] = [\underline{x}, \overline{x}] = ([x]_i) \in \boldsymbol{IR}^n$ and $[A] = [\underline{A}, \overline{A}] = ([a]_{ij}) \in \boldsymbol{IR}^{n\times n}$. For $[a]$, $[b] \in \boldsymbol{IR}$ we define

$$
\begin{aligned}
\check{a} &:= (\underline{a} + \overline{a})/2 && \text{midpoint,}\\
|[a]| &:= \max\{|\underline{a}|, |\overline{a}|\} && \text{absolute value,}\\
d([a]) &:= \overline{a} - \underline{a} && \text{diameter,}\\
q([a],[b]) &:= \max\{|\underline{a} - \underline{b}|\,,\ |\overline{a} - \overline{b}|\} && \text{distance,}\\
\beta([a],[b]) &:= |[a]| + q([a],[b]). && \qquad (2.1)
\end{aligned}
$$

For interval vectors and interval matrices, these quantities are defined entrywise, i.e., they are real vectors and matrices, respectively. In particular, $|x| = (|x_i|) \in \boldsymbol{R}^n$ for point vectors x. We equip $\boldsymbol{R}^n$ and also $\boldsymbol{R}^{n\times n}$ with the natural partial ordering $\leq$. In addition we write $x < y$ or, equivalently, $y > x$ for vectors $x = (x_i)$, $y = (y_i) \in \boldsymbol{R}^n$ if $x_i < y_i$ for $i = 1, \ldots, n$. With the definition

$$
\langle [a] \rangle := \begin{cases} 0 & \text{if } 0 \in [a] \in \boldsymbol{IR},\\ \min\{|\underline{a}|, |\overline{a}|\} & \text{otherwise,} \end{cases}
$$

we construct the *comparison matrix* $\langle [A] \rangle := (c_{ij}) \in \boldsymbol{R}^{n\times n}$ of $[A]$ by setting

$$
c_{ij} := \begin{cases} \langle [a]_{ij} \rangle & \text{if } i = j,\\ -|[a]_{ij}| & \text{if } i \neq j. \end{cases}
$$

We call $[A] \in \boldsymbol{IR}^{n\times n}$ *regular* if no matrix $\tilde{A} \in [A]$ is singular, and we write $\rho(A)$ for the spectral radius of $A \in \boldsymbol{R}^{n\times n}$. Intervals $[a]$ are named *zero symmetric* if $\underline{a} = -\overline{a}$. For interval vectors and interval matrices zero-symmetry is defined entrywise.

We close this section by noting equivalent formulations of nonempty intersections of intervals and by recalling two properties of the function β above, which are proved in [6, Lemma 1.7.5, p. 28].

LEMMA 2.1. *Let* $[a]$, $[b] \in \boldsymbol{IR}$. *Then the following properties are equivalent.*

(a) $[a] \cap [b] \neq \emptyset$.

(b) $\underline{a} \leq \overline{b}$ *and* $\overline{a} \geq \underline{b}$.

(c) $|\check{a} - \check{b}| \leq \frac{1}{2} d([a]) + \frac{1}{2} d([b])$.

LEMMA 2.2. *With* β *from* (2.1) *the following properties hold.*

(a) *If* $[a]_i$, $[b]_i \in \boldsymbol{IR}$, $[a]_i \subseteq [b]_i$ *for* $i = 1, \ldots, n$, *then*

$$
\beta([a]_1 \cdot \ldots \cdot [a]_n\ ,\ [b]_1 \cdot \ldots \cdot [b]_n) \leq \beta([a]_1, [b]_1) \cdot \ldots \cdot \beta([a]_n, [b]_n)\,.
$$

(b) *If* $[a],\ [b] \in \boldsymbol{IR},\ [a] \subseteq [b]$ *and* $\langle [a] \rangle > q([a],[b])$, *then*

$$\beta\left([a]^{-1},[b]^{-1}\right) \leq \left(\langle [a] \rangle - q([a],[b])\right)^{-1},$$

where $[c]^{-1} := \{\, c^{-1} \mid c \in [c] \,\}$ *for* $[c] \in \boldsymbol{IR},\ 0 \notin [c]$.

3. The solution set S. In this section we recall some properties of the solution set S defined in (1.2). To this end, we always assume that a fixed regular interval matrix $[A] \in \boldsymbol{IR}^{n\times n}$ and a fixed interval vector $[b] \in \boldsymbol{IR}^n$ are given. Then the elements of S can be characterized in two equivalent ways.

THEOREM 3.1. *The following three properties are equivalent.*

(a) $x \in S$;

(b) $|\check{A}x - \check{b}| \leq \frac{1}{2}d([A])|x| + \frac{1}{2}d([b])$;

(c) $[A]x \cap [b] \neq \emptyset$.

The equivalence (a) $\Leftrightarrow$ (b) is known as Oettli–Prager criterion [7], the equivalence (a) $\Leftrightarrow$ (c) is due to Beeck [3]. We will omit the proof.

To derive some more properties on S we decompose $\boldsymbol{R}^n$ into its closed orthants $O_k,\ k = 1,\ldots,2^n$, which are uniquely determined by the signs $s_{k_j} \in \{-1,+1\},\ j = 1,\ldots n$, of the components of their interior points. Hence, if O denotes some orthant, fixed by the signs $s_1,\ldots,s_n$, then $x = (x_i) \in O$ fulfills

$$x_j \begin{cases} \geq 0 & \text{if } s_j = 1, \\ \leq 0 & \text{if } s_j = -1. \end{cases} \tag{3.1}$$

For $[A]$, $[b]$ as above, and for $i,j = 1,\ldots,n$, let

$$c_{ij} := \begin{cases} \underline{a}_{ij} & \text{if } s_j = 1, \\ \overline{a}_{ij} & \text{if } s_j = -1, \end{cases} \tag{3.2}$$

and

$$d_{ij} := \begin{cases} \overline{a}_{ij} & \text{if } s_j = 1, \\ \underline{a}_{ij} & \text{if } s_j = -1. \end{cases} \tag{3.3}$$

Denote by $\underline{H}_i$, $\overline{H}_i$, the half spaces

$$\left.\begin{aligned} \underline{H}_i &:= \left\{ y \in \boldsymbol{R}^n \mid \sum_{j=1}^{n} c_{ij}y_j \leq \overline{b}_i \right\} \\ \overline{H}_i &:= \left\{ y \in \boldsymbol{R}^n \mid \sum_{j=1}^{n} d_{ij}y_j \geq \underline{b}_i \right\} \end{aligned}\right\} \quad i = 1,\ldots,n. \tag{3.4}$$

Note that $\underline{H}_i$, $\overline{H}_i$ depend on the choice of the orthant O. By means of these half spaces we can represent $S \cap O$ in the following way (cf. also [8, Cor. 1.2]).

THEOREM 3.2. *Let* $[A] \in \boldsymbol{IR}^{n\times n}$ *be regular and let* O *denote any orthant of* $\boldsymbol{R}^n$. *Then*

$$S \cap O = \bigcap_{i=1}^{n} (\underline{H}_i \cap \overline{H}_i) \cap O. \tag{3.5}$$

In particular, if $S \cap O$ *is nonempty, it is convex, compact, connected, and a polytope.*

S is compact, connected, but not necessarily convex. It is the union of finitely many convex polytopes.

Proof. Let $[a] \in \boldsymbol{IR}$, $\xi \in \boldsymbol{R}$. Then

$$\xi \cdot [a] = \begin{cases} [\xi\underline{a}, \xi\overline{a}] & \text{if } \xi \geq 0, \\ [\xi\overline{a}, \xi\underline{a}] & \text{if } \xi < 0. \end{cases}$$

Hence (3.5) follows from Lemma 2.1(a), (b), from Theorem 3.1(a), (c), and from the definition of $\underline{H}_i$, $\overline{H}_i$.

Since O, $\underline{H}_i$, $\overline{H}_i$ are convex, the same holds for $S \cap O$ because of (3.5). This in turn shows that $S \cap O$ is connected. The compactness and the connectivity of S follows from the same property of $[A] \times [b]$ and from the continuity of the function

$$g : \begin{cases} [A] \times [b] & \to & \boldsymbol{R}^n, \\ (A, b) & \mapsto & A^{-1}b, \end{cases}$$

the range of which is S. Now S being compact the same holds for $S \cap O$ since O is closed. The remaining property of S follows trivially from

$$S = \bigcup_{j=1}^{2^n} (S \cap O_j)$$

and from (3.5), where O_j, $j = 1, \ldots, 2^n$, denote the orthants of $\boldsymbol{R}^n$ numbered arbitrarily. □

That S can be nonconvex is seen by the following example.

Example 3.3. Let $[A] = \begin{pmatrix} 1 & 0 \\ [-1,1] & 1 \end{pmatrix}$, $[b] = \begin{pmatrix} [-1,1] \\ 0 \end{pmatrix}$. Then S is given by $S = \{(x,y) \mid |y| \leq |x| \leq 1\}$ as illustrated in Fig. 1.

THEOREM 3.4. *Let $[A]$ be a point matrix. Then S is a parallelepiped; in particular, S is convex.*

Proof. Let $[A] = [A, A]$, and denote the columns of A^{-1} by $c^1, \ldots, c^n$. Then

$$S = \left\{ A^{-1}\underline{b} + \sum_{j=1}^{n} t_j c^j \mid 0 \leq t_j \leq d([b]_j),\ j = 1, \ldots, n \right\}.$$

This proves the theorem. □

We remark that a necessary and sufficient criterion for the convexity of S can be found in [9].

4. On the symmetric solution set S_{sym}. We now turn over to the symmetric solution set S_{sym} defined in (1.1). We again assume $[A] \in \boldsymbol{IR}^{n \times n}$ to be regular, and, in addition, to fulfill

$$[A] = [A]^T,$$

which is equivalent to $\underline{A} = \underline{A}^T$ and $\overline{A} = \overline{A}^T$.

We first prove two simple properties of S_{sym}.

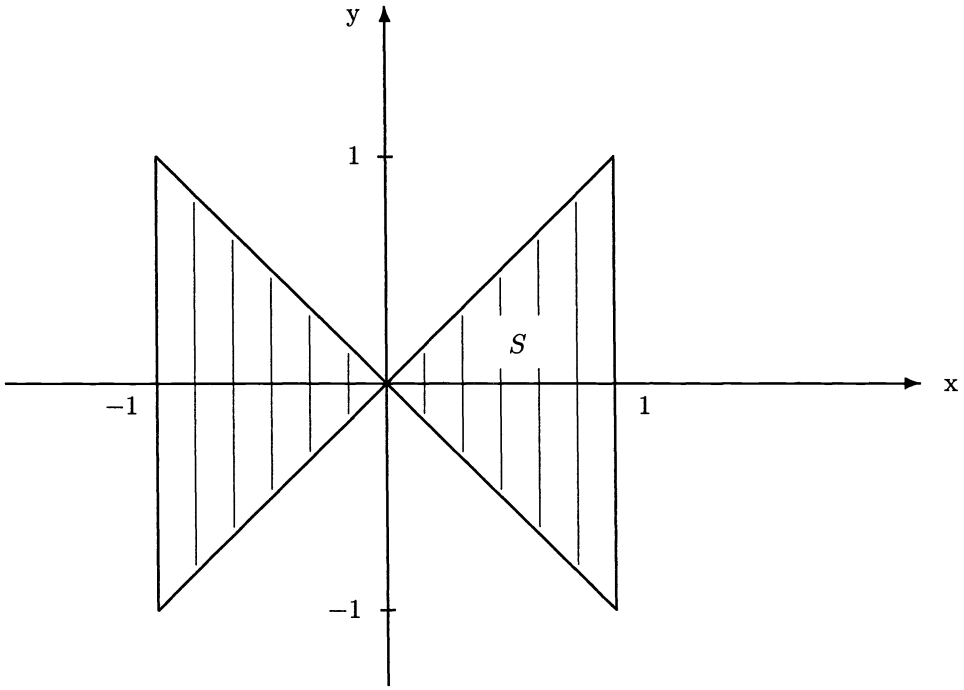


FIG. 1. *The shape of the solution set S in Example* 3.3.

THEOREM 4.1. *Let* $[A] = [A]^T \in \boldsymbol{IR}^{n\times n}$ *be regular. Then* S_{sym} *is compact and connected.*

Proof. Define $[A]_{\text{sym}} := \{A \in [A] |\ A = A^T\}$. Then

$$f : \begin{cases} [A]_{\text{sym}} \times [b] & \to & \boldsymbol{R}^n, \\ (A, b) & \mapsto & A^{-1}b \end{cases} \tag{4.1}$$

is continuous. Let $\{A_k\}$ be an infinite sequence from $[A]_{\text{sym}}$. Since the $(1,1)$-entries of A_k are all contained in the compact set $[a]_{11}$, there is a subsequence $\{A_k^{(1)}\}$ of $\{A_k\}$ such that its $(1,1)$-entries are convergent. By the same reason one can choose a subsequence $\{A_k^{(2)}\}$ of $\{A_k^{(1)}\}$ such that the $(1,2)$-entries are convergent. It is obvious that the $(1,1)$-entries of $\{A_k^{(2)}\}$ keep this property. Repeating the arguments by running through the indices (i,j), $1 \le i \le j \le n$ and taking into account the symmetry of A_k shows that there is a convergent subsequence of $\{A_k\}$, which proves $[A]_{\text{sym}}$ to be compact. Therefore, $[A]_{\text{sym}} \times [b]$ is compact, and the same holds for the range S_{sym} of f.

If $A_1,\ A_2 \in [A]_{\text{sym}}$ then the line segment $A_1 + t(A_2 - A_1) \in [A]_{\text{sym}},\ 0 \le t \le 1$. Hence $[A]_{\text{sym}}$ is connected and also $[A]_{sym} \times [b]$. Using the continuous function f from (4.1) once more shows S_{sym} to be connected. □

We next investigate S_{sym} in the 2×2 case more carefully. To this end, as in §3, we fix an orthant O given by the signs $s_1, \ldots, s_n$ of the components of its interior

points. We define $\underline{H}_i$, $\overline{H}_i$ as in (3.2)–(3.4) and e, $f \in \boldsymbol{R}^n$ by

$$e_i := \begin{cases} \underline{b}_i & \text{if } s_i = 1, \\ \overline{b}_i & \text{if } s_i = -1, \end{cases}$$

$$f_i := \begin{cases} \overline{b}_i & \text{if } s_i = 1, \\ \underline{b}_i & \text{if } s_i = -1. \end{cases}$$

For $n = 2$ we use the sets

$$C^- := \left\{ y \in \boldsymbol{R}^2 \mid \underline{a}_{11} y_1^2 - \overline{a}_{22} y_2^2 - f_1 y_1 + e_2 y_2 \leq 0 \right\}, \tag{4.2}$$

$$C^+ := \left\{ y \in \boldsymbol{R}^2 \mid \overline{a}_{11} y_1^2 - \underline{a}_{22} y_2^2 - e_1 y_1 + f_2 y_2 \geq 0 \right\}. \tag{4.3}$$

Obviously, each of these two sets has a conic section as boundary provided that $\underline{a}_{11}^2 + \overline{a}_{22}^2 \neq 0$ for C^- and, similarly, $\overline{a}_{11}^2 + \underline{a}_{22}^2 \neq 0$ for C^+. As for the hyperplanes $\underline{H}_i$, $\overline{H}_i$ in §3 we point out that C^-, C^+ depend on the choice of the orthant O. However, the type of the conic section is independent of O if one does not distinguish between hyperbolas and pairs of intersecting straight lines, and if one considers a single point as an ellipse. If each symmetric matrix from $[A]$ is positive definite then $\underline{a}_{ii} > 0$, $i = 1, 2$, hence the boundary of C^- and C^+ is formed by hyperbolas in the above-mentioned generalized sense.

We now describe S_{sym} in the 2×2 case by means of S, C^-, and C^+ .

THEOREM 4.2. *Let $[A] = [A]^T \in \boldsymbol{IR}^{2\times 2}$ be regular and let O denote any orthant of $\boldsymbol{R}^2$. Then*

$$S_{\text{sym}} \cap O = S \cap O \cap C^- \cap C^+ . \tag{4.4}$$

In particular, if $S_{\text{sym}} \cap O$ is nonempty, it is compact, but not necessarily convex.

Proof. The compactness follows from Theorem 4.1. The nonconvexity is shown by Example 4.4. It remains to prove (4.4).

$\subseteq$: Let $x \in S_{\text{sym}} \cap O$. Then $x \in S \cap O$, and there exists a *symmetric* matrix $A \in [A]$ and a vector $b \in [b]$ such that $Ax = b$. With $[t] := [a]_{12} = [a]_{21}$ and $t := a_{12} = a_{21}$ we get

$$a_{11} x_1 + t x_2 = b_1 , \tag{4.5}$$

$$t x_1 + a_{22} x_2 = b_2 . \tag{4.6}$$

Multiplying (4.5) by x_1 and (4.6) by x_2 and substituting $t x_1 x_2$ we obtain

$$a_{11} x_1^2 - a_{22} x_2^2 = b_1 x_1 - b_2 x_2 .$$

Thus

$$x^T \begin{pmatrix} [a]_{11} & 0 \\ 0 & -[a]_{22} \end{pmatrix} x \cap x^T \begin{pmatrix} [b]_1 \\ -[b]_2 \end{pmatrix} \neq \emptyset , \tag{4.7}$$

whence, by Lemma 2.1, we get equivalently

$$\underline{a}_{11} x_1^2 - \overline{a}_{22} x_2^2 \leq f_1 x_1 - e_2 x_2 ,$$
$$\overline{a}_{11} x_1^2 - \underline{a}_{22} x_2^2 \geq e_1 x_1 - f_2 x_2 .$$

This means $x \in C^-$ and $x \in C^+$, respectively. Therefore, $S_{\text{sym}} \cap O \subseteq S \cap O \cap C^- \cap C^+$.

$\supseteq$: Let

$$x \in S \cap O \cap C^- \cap C^+ \,. \tag{4.8}$$

Since $x \in S$, there are $A \in [A]$, $b \in [b]$ such that

$$Ax = b \tag{4.9}$$

holds. We are going to show that $A \in [A]$ in (4.9) can be chosen to be symmetric when changing $b \in [b]$ appropriately. To simplify the notation we use

$$t_1 := a_{12} \in [a]_{12} \quad \text{and} \quad t_2 := a_{21} \in [a]_{21} = [a]_{12} =: [t]$$

for the two off-diagonal entries of A in (4.9).

If $t_1 = t_2$ then $x \in S_{\text{sym}} \cap O$. Therefore, assume $t_1 \neq t_2$, say

$$t_1 < t_2 \,. \tag{4.10}$$

If $x_1 = 0$ then A can be replaced in (4.9) by the symmetric matrix

$$A_{\text{sym}} := \begin{pmatrix} a_{11} & t_1 \\ t_1 & a_{22} \end{pmatrix}$$

thus showing $x \in S_{\text{sym}} \cap O$. Analogously one proceeds for $x_2 = 0$.

Let now $x_1 \neq 0$ and $x_2 \neq 0$. We first consider the case $x_1 > 0$, $x_2 > 0$, which, by (4.8), means that O is the first quadrant of $\boldsymbol{R}^2$. Our proof is based on the equivalence of (4.9) with

$$t_1 = \frac{b_1 - a_{11}x_1}{x_2} \in [t], \qquad t_2 = \frac{b_2 - a_{22}x_2}{x_1} \in [t] \,. \tag{4.11}$$

Assume $x \notin S_{\text{sym}} \cap O$. This means that $b \in [b]$ and $A \in [A]$ from (4.9) *cannot* be replaced such that (4.9) is satisfied for some symmetric matrix $A_{\text{sym}} \in [A]$ and some suitably modified vector $b \in [b]$. Taking into account (4.10) we consequently obtain

$$\underline{t} \leq t_1 \leq t_{\max} := \frac{\overline{b}_1 - \underline{a}_{11}x_1}{x_2} < t_{\min} := \frac{\underline{b}_2 - \overline{a}_{22}x_2}{x_1} \leq t_2 \leq \overline{t}, \tag{4.12}$$

whence

$$\overline{b}_1 x_1 - \underline{a}_{11}x_1^2 < \underline{b}_2 x_2 - \overline{a}_{22}x_2^2 \,.$$

Since we supposed O to be the first quadrant this implies $x \notin C^-$, which contradicts (4.8).

Replacing (4.10) by $t_1 > t_2$ and assuming $x \notin S_{\text{sym}} \cap O$ yields

$$\overline{t} \geq t_1 \geq t_{\min} := \frac{\underline{b}_1 - \overline{a}_{11}x_1}{x_2} > t_{\max} := \frac{\overline{b}_2 - \underline{a}_{22}x_2}{x_1} \geq t_2 \geq \underline{t}$$

from which we get the contradiction $x \notin C^+$. Therefore,

$$S \cap O \cap C^- \cap C^+ \subseteq S_{\text{sym}} \cap O \tag{4.13}$$

holds if O is the first quadrant O_1 .

Let now $x \in O \neq O_1$, $x_1 \neq 0$, $x_2 \neq 0$, $s_1 := \operatorname{sign}(x_1)$, $s_2 := \operatorname{sign}(x_2)$, $D_x := \operatorname{diag}(s_1, s_2) \in \boldsymbol{R}^{2\times 2}$. Then (4.9) is equivalent to

$$\hat{A}\hat{x} = \hat{b} \tag{4.14}$$

with $\hat{A} := D_x A D_x \in D_x[A]D_x =: [\hat{A}]$, $\hat{x} := D_x x \in O_1$, $\hat{b} := D_x b \in D_x[b] =: [\hat{b}]$. Let S, S_{sym}, C^-, C^+, e_i, f_i be associated with the given quantities $[A]$, $[b]$, and O, and let $\hat{S}$, $\hat{S}_{\text{sym}}$, $\hat{C}^-$, $\hat{C}^+$, $\hat{e}_i$, $\hat{f}_i$ be the corresponding quantities associated with $[\hat{A}]$, $[\hat{b}]$, and O_1. Since

$$s_1 f_1 = \left\{ \begin{array}{lll} \overline{b}_1 & \text{if} & s_1 = \quad 1 \\ -\underline{b}_1 & \text{if} & s_1 = -1 \end{array} \right\} = \max\{s_1[b]_1\} = \hat{f}_1$$

and

$$s_2 e_2 = \left\{ \begin{array}{lll} \underline{b}_2 & \text{if} & s_2 = \quad 1 \\ -\overline{b}_2 & \text{if} & s_2 = -1 \end{array} \right\} = \min\{s_2[b]_2\} = \hat{e}_2,$$

we get from $y \in C^-$ the inequality

$$\begin{aligned} 0 &\geq (s_1 \underline{a}_{11} s_1)(s_1 y_1)^2 - (s_2 \overline{a}_{22} s_2)(s_2 y_2)^2 - (s_1 f_1)(s_1 y_1) + (s_2 e_2)(s_2 y_2) \\ &= \hat{\underline{a}}_{11}\hat{y}_1^2 - \hat{\overline{a}}_{22}\hat{y}_2^2 - \hat{f}_1\hat{y}_1 + \hat{e}_2\hat{y}_2, \end{aligned}$$

where $\hat{y} := D_x y$. Hence $y \in C^-$ implies $\hat{y} \in \hat{C}^-$, and analogously $y \in C^+$ yields $\hat{y} \in \hat{C}^+$. Therefore, $x \in S \cap O \cap C^- \cap C^+$ results in $\hat{x} \in \hat{S} \cap O_1 \cap \hat{C}^- \cap \hat{C}^+$ whence

$$\hat{x} \in \hat{S}_{\text{sym}} \cap O_1 \tag{4.15}$$

as we have proved above. Since (4.15) implies $\hat{A}_{\text{sym}}\hat{x} = \hat{b}$ for some symmetric matrix $\hat{A}_{\text{sym}} \in [\hat{A}]$ and some right-hand side $\hat{b} \in [\hat{b}]$, it yields $x \in S_{\text{sym}} \cap O$ via (4.14). □

The generalization of Theorem 4.2 for the case $n > 2$ is not straightforward since the elimination process performed in the proof does not seem to work in this case.

Since $x \in C^- \cap C^+$ is equivalent to (4.7), we obtain immediately the subsequent corollary from Theorem 3.1(a), (c) and from Theorem 4.2.

COROLLARY 4.3. *For regular matrices* $[A] = [A]^T \in \boldsymbol{IR}^{2\times 2}$ *and* $[b] \in \boldsymbol{IR}^2$ *the following properties are equivalent.*

(a) $x \in S_{\text{sym}}$.

(b) $[A]x \cap [b] \neq \emptyset$ *(i. e., $x \in S$) and*

$$x^T \begin{pmatrix} [a]_{11} & 0 \\ 0 & -[a]_{22} \end{pmatrix} x \cap x^T \begin{pmatrix} [b]_1 \\ -[b]_2 \end{pmatrix} \neq \emptyset .$$

Note that in contrast to Theorem 4.2 no orthant enters explicitly in Corollary 4.3. Therefore, it can be viewed as an analogue of Theorem 3.1.

We now illustrate Theorem 4.2 by two examples. In particular we show that S_{sym} can be nonconvex in the orthants and that its boundary can be curvilinear.

Example 4.4. Let

$$[A] := \begin{pmatrix} 5 & [-4,0] \\ [-4,0] & 5 \end{pmatrix} \quad \text{and} \quad [b] := \begin{pmatrix} 9 \\ 0 \end{pmatrix}.$$

With

$$A_{t_1,t_2} := \begin{pmatrix} 5 & t_1 \\ t_2 & 5 \end{pmatrix}, \; t_1, t_2 \in [-4,0],$$

we get

$$A_{t_1,t_2}^{-1} = \frac{1}{25 - t_1 t_2} \begin{pmatrix} 5 & -t_1 \\ -t_2 & 5 \end{pmatrix} \geq 0$$

and

$$A_{t_1,t_2}^{-1} \cdot \begin{pmatrix} 9 \\ 0 \end{pmatrix} = \frac{1}{25 - t_1 t_2} \begin{pmatrix} 45 \\ -9t_2 \end{pmatrix}, \; t_1, t_2 \in [-4, 0] \; . \tag{4.16}$$

Hence S and S_{sym} are completely contained in the first quadrant O_1. With the notations of §§3 and 4 we obtain

$$\begin{aligned} \underline{H}_1 &= \{y \in \boldsymbol{R}^2 \mid 5y_1 - 4y_2 \leq 9\} \; , & \overline{H}_1 &= \{y \in \boldsymbol{R}^2 \mid 5y_1 \geq 9\} \; , \\ \underline{H}_2 &= \{y \in \boldsymbol{R}^2 \mid -4y_1 + 5y_2 \leq 0\} \; , & \overline{H}_2 &= \{y \in \boldsymbol{R}^2 \mid 5y_2 \geq 0\} \; , \end{aligned}$$

hence $S = \underline{H}_1 \cap \overline{H}_1 \cap \underline{H}_2 \cap \overline{H}_2 \cap O_1$ is the triangle with the vertices $(1.8, 0)$, $(1.8, 1.44)$, and $(5, 4)$. To describe S_{sym} we list the sets

$$\begin{aligned} C^- &= \{y \in \boldsymbol{R}^2 \mid 5y_1^2 - 5y_2^2 - 9y_1 \leq 0\} \; , \\ C^+ &= \{y \in \boldsymbol{R}^2 \mid 5y_1^2 - 5y_2^2 - 9y_1 \geq 0\} \; . \end{aligned}$$

Then $K := C^- \cap C^+$ is the hyperbola

$$K: \quad \left(y_1 - \frac{9}{10}\right)^2 - y_2^2 = \frac{81}{100} \; .$$

By (4.16) or by Theorem 4.2 one can see that S_{sym} is that part of the right branch of K which lies between the points $(1.8, 0)$ and $(5, 4)$. The sets S and S_{sym} are illustrated in Fig. 2.

Our next example shows that parts of a parabola, of a circle, and straight lines can also form the boundary of S_{sym}.

Example 4.5. Let

$$[A] := \begin{pmatrix} 1 & [1,2] \\ [1,2] & [-1,0] \end{pmatrix}, \quad [b] := \begin{pmatrix} 4 \\ [1,2] \end{pmatrix}, \quad A_{\alpha,\beta,\gamma} := \begin{pmatrix} 1 & \alpha \\ \beta & \gamma \end{pmatrix} \in [A]$$

with $\alpha, \beta \in [1,2]$, $\gamma \in [-1,0]$. Since $\det A_{\alpha,\beta,\gamma} = \gamma - \alpha\beta \leq -1$, the interval matrix $[A]$ is regular with

$$A_{\alpha,\beta,\gamma}^{-1} = \frac{1}{\det A_{\alpha,\beta,\gamma}} \begin{pmatrix} \gamma & -\alpha \\ -\beta & 1 \end{pmatrix} \; .$$

With $b_1 \geq 2b_2 \geq 2$ we get $A_{\alpha,\beta,\gamma}^{-1} \cdot b \geq 0$ for any choice $A_{\alpha,\beta,\gamma} \in [A]$, $b \in [b]$. Hence S and S_{sym} are completely contained in the first quadrant O_1. Using the notation above we obtain for O_1 the following sets:

$$\begin{aligned} \underline{H}_1 &= \{y \in \boldsymbol{R}^2 \mid y_1 + y_2 \leq 4\}, \overline{H}_1 = \{y \in \boldsymbol{R}^2 \mid y_1 + 2y_2 \geq 4\}, \\ \underline{H}_2 &= \{y \in \boldsymbol{R}^2 \mid y_1 - y_2 \leq 2\}, \overline{H}_2 = \{y \in \boldsymbol{R}^2 \mid 2y_1 \geq 1\} \; , \end{aligned}$$

$$\begin{aligned} C^- &= \{y \in \boldsymbol{R}^2 \mid y_1^2 - 4y_1 + y_2 \leq 0\} = \{y \in \boldsymbol{R}^2 \mid y_2 \leq 4 - (y_1 - 2)^2\} \; , \\ C^+ &= \{y \in \boldsymbol{R}^2 \mid y_1^2 + y_2^2 - 4y_1 + 2y_2 \geq 0\} = \{y \in \boldsymbol{R}^2 \mid (y_1 - 2)^2 + (y_2 + 1)^2 \geq 5\} \; . \end{aligned}$$

The set $S = \underline{H}_1 \cap \overline{H}_1 \cap \underline{H}_2 \cap \overline{H}_2 \cap O_1$ is the convex hull of the points $(\frac{1}{2}, \frac{7}{4})$, $(\frac{1}{2}, \frac{7}{2})$, $(\frac{8}{3}, \frac{2}{3})$ and $(3, 1)$. The boundary of $S_{\text{sym}} = S \cap O_1 \cap C^- \cap C^+$ is formed by the following four curves.

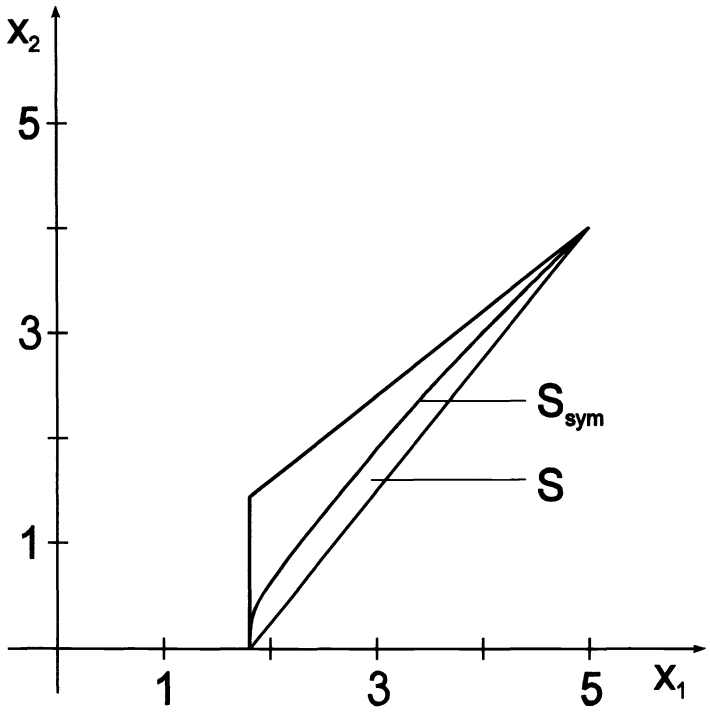


FIG. 2. *The shape of the solution sets S and S_{sym} in Example* 4.4.

(i) The straight line between $(\frac{1}{2}, \frac{7}{4})$ and $(\frac{8}{5}, \frac{6}{5})$.
(ii) The straight line between $(1, 3)$ and $(3, 1)$.
(iii) The part of the parabola $y_2 = 4 - (y_1 - 2)^2$ between $(\frac{1}{2}, \frac{7}{4})$ and $(1, 3)$.
(iv) The part of the circle $(y_1 - 2)^2 + (y_2 + 1)^2 = 5$ between $(\frac{8}{5}, \frac{6}{5})$ and $(3, 1)$.

The situation is illustrated in Fig. 3.

5. Computing enclosures for S_{sym}. As was shown in [2], S_{sym} can be enclosed by the vector $[x]^C$, which results from the following interval version of the well-known Cholesky method, for which we assume $[A] = [A]^T \in \boldsymbol{IR}^{n\times n}$, and $[b] \in \boldsymbol{IR}^n$.

Step 1. "LL^T decomposition"

for $j := 1$ **to** n **do**

$$[l]_{jj} := \left([a]_{jj} - \sum_{k=1}^{j-1} [l]_{jk}^2 \right)^{\frac{1}{2}} ;$$

for $i := j + 1$ **to** n **do**

$$[l]_{ij} := \left([a]_{ij} - \sum_{k=1}^{j-1} [l]_{ik}[l]_{jk} \right) / [l]_{jj} ;$$

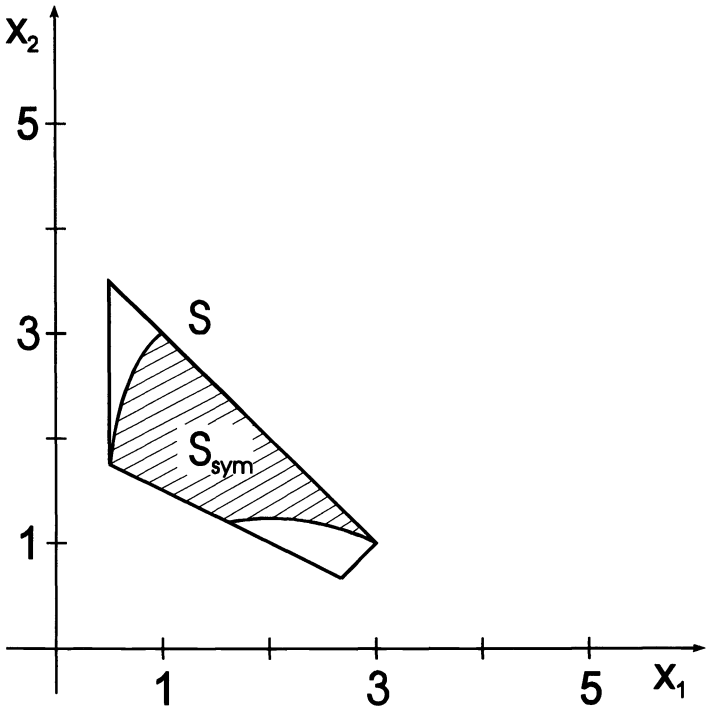


FIG. 3. *The shape of the solution sets S and S_{sym} in Example* 4.5.

Step 2. Forward substitution

for $i := 1$ **to** n **do**

$$[y]_i := \left([b]_i - \sum_{j=1}^{i-1} [l]_{ij}[y]_j \right) / [l]_{ii} \ ;$$

Step 3. Backward substitution

for $i := n$ **downto** 1 **do**

$$[x]_i^C := \left([y]_i - \sum_{j=i+1}^{n} [l]_{ji}[x]_j^C \right) / [l]_{ii} \ ;$$

$\mathrm{ICh}([A],[b]) := [x]^C$.

Here,

$$[a]^2 := \{a^2 \mid a \in [a]\} \tag{5.1}$$

and

$$[a]^{1/2} := \sqrt{[a]} := \{\sqrt{a} \mid a \in [a]\}$$

for intervals $[a]$.

In contrast to the classical, i.e., noninterval Cholesky method, it is an open question when the interval Cholesky method is feasible. In [2] several criteria are given that guarantee the existence of $[x]^C$. We add here two new ones as well as a nonexistence criterion, which we formulate first.

THEOREM 5.1. *If* $[A] = [A]^T \in \boldsymbol{IR}^{n\times n}$ *contains at least one symmetric matrix* A *which is not positive definite, then* $[x]^C$ *does not exist.*

Proof. We first recall that a real symmetric matrix has an LL^T-decomposition with positive diagonal entries l_{ii} if and only if this matrix is positive definite (see [11]). L can be computed by the Cholesky method. Assume now that $A = A^T \in [A] = [A]^T$ is not positive definite. Then the Cholesky method will break down. This is the case if and only if for some index j either l_{jj} cannot be computed because of

$$a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 < 0$$

(see Step 1) or y_i cannot be computed because of $l_{ii} = 0$ (see Step 2). By the inclusion monotonicity of the interval arithmetic, either $[l]_{jj}$ does not exist, or $0 \in [l]_{ii}$ and the interval Cholesky method will break down. □

Example 4.5 illustrates Theorem 5.1: Since

$$A = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \in [A] = \begin{pmatrix} 1 & [1,2] \\ [1,2] & [-1,0] \end{pmatrix}$$

is not positive definitive, $[x]^C$ does not exist for $[A]$. Note, however, that the interval Gaussian algorithm is feasible for this interval matrix.

Before formulating our new feasibility criterion we need some preparations.

By Theorem 3.4 in [2] we have for $[y]$ from Step 2 in the interval version of the Cholesky method

$$[y] = [D^n]\left([L^{n-1}]\left([D^{n-1}]\left(\ldots\left([L^2]\left([D^2]\left([L^1]\left([D^1][b]\right)\right)\right)\right)\ldots\right)\right)\right)$$

and

$$[x]^C = [D^1]\left([L^1]^T\left([D^2]\left(\ldots\left([L^{n-2}]^T\left([D^{n-1}]\left([L^{n-1}]^T\left([D^n][y]\right)\right)\right)\right)\ldots\right)\right)\right), \tag{5.2}$$

where the diagonal matrices $[D^s]$ and the lower triangular matrices $[L^s]$ are defined for $s = 1, \ldots, n-1$ by

$$[d^s]_{ij} := \begin{cases} 1 & \text{if } i = j \neq s, \\ 1/[l]_{ss} & \text{if } i = j = s, \\ 0 & \text{otherwise,} \end{cases}$$

$$[l^s]_{ij} := \begin{cases} 1 & \text{if } i = j, \\ -[l]_{is} & \text{if } i > j = s, \\ 0 & \text{otherwise,} \end{cases}$$

with $[l]_{ij}$ from the Cholesky method. (Note that $[l]_{ij}$ is computed in the jth step of the "LL^T-decomposition"). By (5.2) it is easy to see that the mapping

$$f : \begin{cases} \boldsymbol{IR}^n & \to & \boldsymbol{IR}^n, \\ [b] & \mapsto & ICh([A],[b]) \end{cases} \tag{5.3}$$

is a *sublinear* one in the sense of [6, p. 98], i.e.,

(i) $[b] \subseteq [c] \Rightarrow f([b]) \subseteq f([c])$,

(ii) $\alpha \in \boldsymbol{R} \Rightarrow f(\alpha[b]) = \alpha f([b])$,

(iii) $f([b] + [c]) \subseteq f([b]) + f([c])$ for $[b]$, $[c] \in \boldsymbol{IR}^n$.

An easy computation yields

$$| [D^n] | \cdot | [L^{n-1}] | \cdot | [D^{n-1}] | \cdot \ldots \cdot | [L^2] | \cdot | [D^2] | \cdot | [L^1] | \cdot | [D^1] | = \langle [L] \rangle^{-1}$$

again with $[L] = ([l]_{ij})$ from the Cholesky method. Hence, for the particular "right-hand side" $[\hat{b}] := [-1, 1]e$, where $e = (1, \ldots, 1)^T$, one gets

$$\mathrm{ICh}([A], [\hat{b}]) = [x]^C = \langle [L]^T \rangle^{-1} \left(\langle [L] \rangle^{-1} [\hat{b}] \right) = \left(\langle [L]^T \rangle^{-1} \langle [L] \rangle^{-1} \right) [\hat{b}] \, .$$

With the abbreviation

$$| [A]^C | := \langle [L]^T \rangle^{-1} \langle [L] \rangle^{-1}, \tag{5.4}$$

one therefore obtains for any $[b] \subseteq [\hat{b}]$ the inclusion

$$\mathrm{ICh}([A], [b]) \subseteq | [A]^C | \, [\hat{b}] \, .$$

Thus, $| [A]^C |$ can be thought of as a *measure* for the width of the enclosure $\mathrm{ICh}([A], [b])$ of S_{sym} that does not depend on the right-hand side $[b]$ as long as $[b]$ is contained in $[\hat{b}]$. The condition $[b] \subseteq [\hat{b}]$ can be considered as a sort of normalization. If it no longer holds, replace $[\hat{b}]$ by $t[\hat{b}]$ with $t > 0$ as small as possible such that $[b] \subseteq t[\hat{b}]$ is valid. Then

$$\mathrm{ICh}([A], [b]) \subseteq t | [A]^C | \, [\hat{b}],$$

hence $t| [A]^C |$ is a corresponding measure.

By (5.2) we also get

$$\left| \left(\mathrm{ICh} \left([A], [-e^{(1)}, e^{(1)}] \right), \ldots, \mathrm{ICh} \left([A], [-e^{(n)}, e^{(n)}] \right) \right) \right| = | [A]^C | \, ,$$

hence $| [A]^C |$ is the *absolute value* of the sublinear mapping f in the sense of [6, p. 100]. By an elementary rule of the diameter d (cf. [1]) one proves at once the property

$$d(f([b])) \geq | [A]^C | \, d([b])$$

of f which is then called *normal* in [6, p. 102].

We next recall an equivalent definition of Step 1 in the interval Cholesky method.

DEFINITION 5.2. ([2]) *Let either* $[A] = ([a]_{11}) \in \boldsymbol{IR}^{1 \times 1}$ *or*

$$[A] = \begin{pmatrix} [a]_{11} & [c]^T \\ [c] & [A'] \end{pmatrix} = [A]^T \in \boldsymbol{IR}^{n \times n}, \; n > 1, \; [c] \in \boldsymbol{IR}^{n-1},$$
$$[A'] \in \boldsymbol{IR}^{(n-1) \times (n-1)}.$$

(a) $\Sigma_{[A]} := [A'] - (1/[a]_{11}) \, [c][c]^T \in \boldsymbol{IR}^{(n-1) \times (n-1)}$ *is termed the* Schur complement (*of the* $(1,1)$ *entry* $[a]_{11}$) *provided* $n > 1$ *and* $0 \notin [a]_{11}$. *In the product* $[c][c]^T$ *we assume that* $[c]_i [c]_i$ *is evaluated as* $[c]_i^2$ (see (5.1)). $\Sigma_{[A]}$ *is not defined if* $n = 1$ *or if* $0 \in [a]_{11}$.

(b) *We call the pair* $([L], [L]^T)$ *the* Cholesky decomposition *of* $[A]$ *if* $0 < \underline{a}_{11}$ *and if either* $n = 1$ *and* $[L] = (\sqrt{[a]_{11}})$ *or*

$$[L] = \begin{pmatrix} \sqrt{[a]_{11}} & 0 \\ \dfrac{[c]}{\sqrt{[a]_{11}}} & [L'] \end{pmatrix}, \tag{5.5}$$

where $([L'], ([L'])^T)$ *is the Cholesky decomposition of* $\Sigma_{[A]}$ *provided that it exists.*

As was shown in [2] the matrix $[L]$ of the Cholesky method and that of the Definition 5.2(b) are identical.

The proof of our main result, Theorem 5.4, is heavily based on the following lemma.

LEMMA 5.3. *Let the Cholesky decomposition* $([L], [L]^T)$ *of* $[A] = [A]^T \in \boldsymbol{IR}^{n\times n}$ *exist, and let* $[B] = [B]^T \supseteq [A]$ *be such that for a suitable* $u > 0$ *we have*

$$q([A], [B])u < \langle [L]\rangle\langle [L]^T\rangle u \ . \tag{5.6}$$

Then the Cholesky method is feasible for $[B]$.

Proof by induction. The proof proceeds similarly as for Lemma 4.5.14 in [6].

Let $n = 1$. Then (5.6) implies $u > 0$. Again (5.6) together with $0 < \underline{a}_{11}$ yields

$$(\underline{a}_{11} - \underline{b}_{11})u \le q([A], [B])u < \langle [a]_{11}\rangle u = \underline{a}_{11}u \ ,$$

hence

$$0 < \underline{b}_{11}u$$

follows. This shows $0 < \underline{b}_{11} = \langle [b]_{11}\rangle$ which proves the existence of $\mathrm{ICh}([B], [b])$ for $n = 1$.

Assume now that the statement is true for some dimension $n \ge 1$, and let (5.6) hold for

$$[A] = \begin{pmatrix} [a]_{11} & [c]^T \\ [c] & [A'] \end{pmatrix} \subseteq [B] = \begin{pmatrix} [b]_{11} & [d]^T \\ [d] & [B'] \end{pmatrix} \in \boldsymbol{IR}^{(n+1)\times(n+1)} \ . \tag{5.7}$$

We first show $\underline{b}_{11} > 0$. With

$$q([A], [B]) = (q_{ij}) = \begin{pmatrix} q_{11} & r^T \\ r & Q' \end{pmatrix} \tag{5.8}$$

we get from (5.6)

$$\sum_{j=1}^{n+1} q_{1j}u_j < \langle [a]_{11}\rangle\, u_1 - \sum_{j=2}^{n+1} |\, [a]_{1j}\,|\, u_j \ ,$$

hence

$$\underline{a}_{11} - q_{11} = \langle [a]_{11}\rangle - q_{11} > \left\{ \sum_{j=2}^{n+1} (q_{1j} + |\, [a]_{1j}\,|)\, u_j \right\} / u_1 \ge 0 \ .$$

Together with (5.7) this implies $0 < \underline{b}_{11} = \langle [b]_{11}\rangle$, whence the Schur complement $\Sigma_{[B]} \supseteq \Sigma_{[A]}$ exists.

By our assumptions, the Schur complement $\Sigma_{[A]}$ has a Cholesky decomposition $([L'], [L']^T)$. If we can show that

$$q(\Sigma_{[A]}\,,\,\Sigma_{[B]})u' < \langle [L']\rangle\langle [L']^T\rangle u' \tag{5.9}$$

holds for some vector $u' > 0$ then $\Sigma_{[B]}$ has a Cholesky decomposition, say $([\hat{L}'], [\hat{L}']^T)$, by the hypothesis of our induction, and with

$$[\hat{L}] := \begin{pmatrix} \sqrt{[b]_{11}} & O \\ \dfrac{[d]}{\sqrt{[b]_{11}}} & [\hat{L}'] \end{pmatrix},$$

we obtain the Cholesky decomposition $([\hat{L}], [\hat{L}]^T)$ of $[B]$.

To prove (5.9) we apply β from (2.1) componentwise, and use the notation from (5.8) as well as that of Lemma 2.2. We then get

(5.10)

$$\begin{aligned} q(\Sigma_{[A]}\,,\,\Sigma_{[B]}) &= q\left([A'] - [c][c]^T[a]_{11}^{-1}, [B'] - [d][d]^T[b]_{11}^{-1}\right) \\ &\le Q' + q([c][c]^T[a]_{11}^{-1}, [d][d]^T[b]_{11}^{-1}) \\ &= Q' - |\,[c][c]^T[a]_{11}^{-1}\,| + \beta([c][c]^T[a]_{11}^{-1}, [d][d]^T[b]_{11}^{-1}) \\ &= Q' - |\,[c]\,|\,|\,[c]^T\,|\,\langle [a]_{11}\rangle^{-1} + \beta([c][c]^T[a]_{11}^{-1}, [d][d]^T[b]_{11}^{-1}) \\ &\le Q' - |\,[c]\,|\,|\,[c]^T\,|\,\langle [a]_{11}\rangle^{-1} + \beta([c],[d])\cdot\beta([c]^T,[d]^T)\cdot\beta([a]_{11}^{-1},[b]_{11}^{-1}) \\ &= Q' - |\,[c]\,|\,|\,[c]^T\,|\,\langle [a]_{11}\rangle^{-1} + (|\,[c]\,|+r)(|\,[c]\,|+r)^T\beta([a]_{11}^{-1},[b]_{11}^{-1})\ . \end{aligned}$$

We now want to apply Lemma 2.2 (b) on the last factor in (5.10). To this end we must show

$$\langle [a]_{11}\rangle \ > \ q([a]_{11},[b]_{11}) = q_{11}\ . \tag{5.11}$$

Therefore, we set $u = \begin{pmatrix} u_1 \\ u' \end{pmatrix}$ in (5.6). With (5.5) and with the notation (5.8), we then obtain

$$\begin{pmatrix} q_{11} & r^T \\ r & Q' \end{pmatrix}\begin{pmatrix} u_1 \\ u' \end{pmatrix} < \begin{pmatrix} \sqrt{\langle [a]_{11}\rangle} & O \\ -\dfrac{|\,[c]\,|}{\sqrt{\langle [a]_{11}\rangle}} & \langle [L']\rangle \end{pmatrix}\begin{pmatrix} \sqrt{\langle [a]_{11}\rangle} & -\dfrac{|\,[c]^T\,|}{\sqrt{\langle [a]_{11}\rangle}} \\ O & \langle [L']^T\rangle \end{pmatrix}\begin{pmatrix} u_1 \\ u' \end{pmatrix},$$

whence

$$q_{11}u_1 + r^Tu' < \langle [a]_{11}\rangle u_1 - |\,[c]^T\,|u' \tag{5.12}$$

and

$$ru_1 + Q'u' < -|\,[c]\,|u_1 + |\,[c]\,|\,|\,[c]^T\,|\,\langle [a]_{11}\rangle^{-1}u' + \langle [L']\rangle\langle [L']^T\rangle u'\ . \tag{5.13}$$

Since $u_1 > 0$, the inequality (5.12) implies (5.11), and Lemma 2.2(b) and (5.10) yield

$$q(\Sigma_{[A]}\,,\,\Sigma_{[B]}) \le Q' - |\,[c]\,|\,|\,[c]^T\,|\,\langle [a]_{11}\rangle^{-1} + (|\,[c]\,|+r)(|\,[c]\,|+r)^T(\langle [a]_{11}\rangle - q_{11})^{-1}\ .$$

Together with (5.11), (5.12), (5.13), this implies

$$\begin{aligned} q(\Sigma_{[A]}\,,\,\Sigma_{[B]})u' &\le Q'u' - |\,[c]\,|\,|\,[c]^T\,|\langle [a]_{11}\rangle^{-1}u' \\ &\quad +(|\,[c]\,|+r)(\langle [a]_{11}\rangle - q_{11})^{-1}(|\,[c]\,|+r)^Tu' \\ &< -ru_1 - |\,[c]\,|u_1 + \langle [L']\rangle\ \langle [L']^T\rangle u' \\ &\quad +(|\,[c]\,|+r)(\langle [a]_{11}\rangle - q_{11})^{-1}(\langle [a]_{11}\rangle u_1 - q_{11}u_1) \\ &= \langle [L']\rangle\langle [L']^T\rangle u'. \end{aligned}$$

This proves (5.9) and terminates the induction. □

We are now ready to prove our main result.

THEOREM 5.4. *Let* $[A]$, $[B] \in \boldsymbol{IR}^{n\times n}$, $[A] = [A]^T$, $[B] = [B]^T$, *and suppose that* $\mathrm{ICh}([A],[b])$ *exists. If*

$$\rho(|\,[A]^C|\,q([A],[B])) < 1, \tag{5.14}$$

then the Cholesky method is feasible for $[B]$.

Proof. Let $Q := q([A],[B])$, $[C] := [A] + [-Q,Q]$. Then $[B] \subseteq [C]$, and $\mathrm{ICh}([B],[b])$ exists if $\mathrm{ICh}([C],[b])$ does. By (5.14) the inverse of $I - |\,[A]^C\,|\,Q$ exists and can be represented as Neumann series

$$(I - |\,[A]^C|Q)^{-1} = \sum_{k=0}^{\infty} \left(|\,[A]^C\,|Q\right)^k \geq 0\,.$$

With any $v \in \boldsymbol{R}^n$ satisfying $v > 0$ define

$$u := \left(I - |\,[A]^C\,|Q\right)^{-1}\ |\,[A]^C\,|v\,. \tag{5.15}$$

Since $|\,[A]^C\,| \geq 0$ and $\left(I - |\,[A]^C\,|Q\right)^{-1} \geq 0$ are regular each of their rows contains at least one positive entry. Therefore $|\,[A]^C|v > 0$ and $u > 0$. Now (5.15) yields

$$|\,[A]^C|Qu = u - |\,[A]^C\,|v,$$

whence

$$\begin{aligned} Qu &= \langle[L]\rangle\langle[L]^T\rangle u - v \\ &< \langle L\rangle\langle L\rangle^T u, \end{aligned}$$

with $([L],[L]^T)$ being the Cholesky decomposition of $[A]$. Hence, Lemma 5.3 guarantees the feasibility of the Cholesky method for $[C]$ and therefore also for $[B]$. □

We illustrate Theorem 5.4 by a simple example.

Example 5.5. Let

$$[B] := \begin{pmatrix} 4 & 2 & 2 \\ 2 & 4 & [0,2] \\ 2 & [0,2] & 4 \end{pmatrix}.$$

Then $\langle[B]\rangle \cdot (1,1,1)^T = 0$, hence $\langle[B]\rangle$ is singular. In particular, $\langle[B]\rangle$ is not an M–matrix (which requires $\langle[B]\rangle^{-1} \geq 0$; cf. [2]), whence, by definition, $[B]$ is not an H–matrix. Therefore, Theorem 4.2 in [2] does not apply. Consider now

$$[A] := \begin{pmatrix} 4 & 2 & 2 \\ 2 & 4 & 1 \\ 2 & 1 & 4 \end{pmatrix} \subseteq [B]\,.$$

Since $\langle[A]\rangle$ is irreducibly diagonally dominant, the interval Cholesky method is feasible for $[A]$ by Corollary 4.3 (ii) in [2], for example. A simple computation yields

$$[L] = \begin{pmatrix} 2 & 0 & 0 \\ 1 & \sqrt{3} & 0 \\ 1 & 0 & \sqrt{3} \end{pmatrix}, \qquad \langle[L]\rangle^{-1} = \frac{\sqrt{3}}{6}\begin{pmatrix} \sqrt{3} & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 2 \end{pmatrix}$$

and

$$|\,[A]^C| = \langle [L]^T\rangle^{-1}\langle [L]\rangle^{-1} = \frac{1}{12}\begin{pmatrix} 5 & 2 & 2 \\ 2 & 4 & 0 \\ 2 & 0 & 4 \end{pmatrix}.$$

From

$$q([A],[B]) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

we get the matrix

$$|\,[A]^C| q([A],[B]) = \begin{pmatrix} 0 & \frac{1}{6} & \frac{1}{6} \\ 0 & 0 & \frac{1}{3} \\ 0 & \frac{1}{3} & 0 \end{pmatrix},$$

which has the eigenvalues $-\frac{1}{3}$, 0, $\frac{1}{3}$. Therefore, Theorem 5.4 applies. The elements $[\hat{l}]_{ij}$ that result from the interval Cholesky method for $[B]$ are given by

$$[\hat{L}] = \begin{pmatrix} 2 & 0 & 0 \\ 1 & \sqrt{3} & 0 \\ 1 & [-1,1]/\sqrt{3} & [\sqrt{8},3]/\sqrt{3} \end{pmatrix}.$$

Our example also illustrates the following corollary.

COROLLARY 5.6. *Let the midpoint matrix* $\check{A}$ *of* $[A] = [A]^T \in \boldsymbol{IR}^{n\times n}$ *be positive definite, and assume that*

$$\rho\left(\frac{1}{2}|\check{A}^C|\; d([A])\right) < 1\ .$$

Then the interval Cholesky method is feasible for $[A]$.

Proof. Because of $[A] = [A]^T$, the matrix $\check{A}$ is symmetric. Since it is positive definite by assumption, the interval Cholesky method is feasible for $\check{A}$ when viewed as a point matrix. Taking into account $q(\check{A},[A]) = \frac{1}{2}d([A])$, the assertion is a direct consequence of Theorem 5.4. □

6. Concluding remarks. We stress the fact that the main purpose of this paper is to give criteria for the feasibility of the interval Cholesky method. If this feasibility is guaranteed—for example, this is the case if one of the criteria presented in this paper or in [2] holds—the question arises immediately how close the symmetric solution set S_{sym} is included. Especially, what is the relation between the results of applying the Gaussian algorithm (or some other method) and the interval Cholesky method, respectively? In [2] it was shown by simple examples that generally no comparison is possible. The examples from [2] can be generalized to arbitrary large dimensions $n > 2$ without any difficulties. Hence up to now it is not clear under which conditions on the given interval matrix the interval Cholesky method is superior to the interval Gaussian algorithm or vice versa. The investigation of this question and/or some

statistics about the width of the bounds for systems of larger dimension will be part of further research.

We also mention that for a given real system a very careful analysis of the floating-point Cholesky decomposition was performed in [10]. If the matrix as well as the right-hand side are afflicted with tolerances then bounds are computed for the set of all solutions for data within tolerances.

Acknowledgments. The authors are grateful to two anonymous referees for a series of comments and remarks that improved the paper considerably.

REFERENCES

[1] G. ALEFELD AND J. HERZBERGER, *Introduction to Interval Computations*, Academic Press, New York, 1983.

[2] G. ALEFELD AND G. MAYER, *The Cholesky method for interval data*, Linear Algebra Appl., 194 (1993), pp. 161–182.

[3] H. BEECK, *Über Struktur und Abschätzungen der Lösungsmenge von linearen Gleichungssystemen mit Intervallkoeffizienten*, Computing, 10 (1972), pp. 231–244.

[4] D. J. HARTFIEL, *Concerning the solution set of $Ax = b$ where $P \leq A \leq Q$ and $p \leq b \leq q$*, Numer. Math., 35 (1980), pp. 355–359.

[5] C. JANSSON, *Interval linear systems with symmetric matrices, skew-symmetric matrices and dependencies in the right-hand side*, Computing, 46 (1991), pp. 265–274.

[6] A. NEUMAIER, *Interval Methods for Systems of Equations*, Cambridge University Press, Cambridge, 1990.

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

[8] J. ROHN, *Interval linear systems*, Freiburger Intervall-Berichte, 84/7 (1984), pp. 33–58.

[9] J. ROHN, *On nonconvexity of the solution set of a system of linear interval equations*, BIT, 30 (1989), pp. 161–165.

[10] S. M. RUMP, *Inclusion of the solution of large linear systems with positive definite symmetric M-matrix*, in Computer Arithmetic and Enclosure Methods, L. Atanassova and J. Herzberger, eds., Elsevier, Amsterdam, 1992, pp. 339–347.

[11] G. STRANG, *Linear Algebra and Its Applications*, Third ed., Harcourt Brace Jovanovich, San Diego, 1988.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, pp. 1241–1267, October 1995

016

A DOMAIN DECOMPOSITION METHOD FOR FIRST-ORDER PDES*

LINA HEMMINGSSON†

Abstract. In this report a nonoverlapping domain decomposition method to solve first-order, time-dependent partial differential equations is developed. The time discretization used is implicit, which gives a large system of equations to solve for each time step. Preconditioners with a fast inversion based on a fast modified sine transform are defined. Theoretical analysis of the method is presented, indicating that the ratio in the grid might be crucial for the convergence. Finally numerical results from a parallel implementation on an Intel Paragon are presented, showing very nice properties. Especially a nonuniform decomposition of the domain leads to very good results.

Key words. first-order PDEs, domain decomposition, fast modified sine transform

AMS subject classifications. 65F10, 65N22, 65M55

1. Introduction.

We consider systems of time-dependent, first-order partial differential equations in two space dimensions

$$\frac{\partial u}{\partial t} + B_1(x_1, x_2)\frac{\partial u}{\partial x_1} + B_2(x_1, x_2)\frac{\partial u}{\partial x_2} = g, \qquad x \in \Omega,\ t \geq 0$$

(1) +boundary conditions

+initial conditions.

Here B_1 and B_2 are $n_c \times n_c$-matrices, and u and g are n_c-vectors.

Systems like (1) are often discretized explicitly in time. In problems with different time scales, where we are only interested in the slowly varying part of the solution, the restriction on the time step due to the Courant–Friedrichs–Lewy (CFL)-criterion for the fast oscillations becomes unrealistic in some applications. If we are only interested in the slowly varying part of the solution, we can discretize implicitly in time with a large time step and still obtain an accurate solution [8], [9], [20]. Then a semi-implicit or an implicit discretization in time is preferable. This gives a large system of equations

$$Au = b \tag{2}$$

to solve for each time step.

The time step used may be large compared to the smallest space step. Hence A may be strongly nondiagonally dominant. Moreover it is nonsymmetric. Holmgren and Otto have investigated a number of iterative methods and block or semicirculant preconditioning matrices for (2), showing very good results, [14], [15]. In [11], [12], preconditioners based on a Toeplitz structure are examined, showing similar results.

Large scale problems will still be very time expensive, and the necessity to use a more efficient method becomes obvious. Over the past several years much effort has been focused on investigating domain decomposition methods on multiprocessor

* Received by the editors June 24, 1994; accepted for publication (in revised form) by H. Weinberger December 1, 1994. This work was supported by the Swedish National Board for Industrial and Technical Development, NUTEK, contract 9303806.

† Department of Scientific Computing, Box 120, S-751 04 Uppsala, Sweden (lina@tdb.uu.se).

computers, [3]–[5], [16]. In these techniques, the domain is divided into several subdomains. Each processor is given one or more subdomains and they can work in parallel with communication only on the interfaces.

In this report a nonoverlapping domain decomposition method to solve (2) is developed and examined. For semi-implicit/implicit time discretizations, the implicit part in the semidiscrete equations often have constant coefficients. The idea presented in this paper is to divide the original domain into subdomains where the implicit part of the semidiscrete PDE has well-known fast solvers defined by a fast modified sine transform [10]–[12]. To avoid the difficulties introduced by the numerical boundary conditions, the boundaries are treated separately by a nonuniform decomposition of the domain. The fast solver allows for variable coefficients in one space dimension, which means that we can have a nonuniform space grid in one space dimension. This is suitable for instance when we are computing a two-dimensional viscous channel flow. In the other space dimension we can treat piecewise constant coefficients and still employ the fast solver. For the case with variable coefficients in both space dimensions, we suggest a solution method based on an iterative solver in the subdomains.

2. The domain decomposition setting.

2.1. Introduction. The first known domain decomposition algorithm was introduced by Schwarz in 1869 [19]. The idea was to use known solvers for elliptic equations on simple geometries such as circles and rectangles to solve the equation on a more complex domain. This is known as the Schwartz alternating procedure.

In this report we study the case with nonoverlapping subdomains. In Fig. 1 the original domain is divided into four rectangular subdomains. The idea is to reduce the original system to a smaller system for the unknowns on the interior boundaries (dashed lines). This smaller system is called the Schur complement system. The solution of this system includes the solution of the different subdomain systems. These subdomain systems can be solved using known solution methods for rectangles. Once the Schur complement system is solved, each subdomain system can be solved separately with no communication. This can be exploited using a parallel computer. The inherent parallelism in this solution method leads to the suggestion that one might even use this method on more regular domains. Implementing the method on a parallel computer could lead to a considerable gain in time. It also gives the possibility of solving larger problems as each processor only stores the unknowns of its awarded subdomains. We shall study a domain whose closure is decomposed into p rectangular subdomains. For the sake of simplicity we will only consider a decomposition in which the interior boundaries are parallel to the x_1-axis. Denote the subdomains by Ω_q and the interior boundaries by Γ_q. Also denote $\bigcup_{q=1}^{p} \Omega_q = \Omega$ and $\bigcup_{q=1}^{(p-1)} \Gamma_q = \Gamma$. The unknowns in Ω_q are denoted by $u_1^{(q)}$ and the unknowns on Γ_q by $u_0^{(q)}$.

If we arrange the system so that we write the unknowns on Γ first and those of Ω last, we obtain a system of equations for the solution vector $(u_0, u_1)^T$, which we express in block form as

$$\mathcal{A} \begin{pmatrix} u_0 \\ u_1 \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}, \tag{3}$$

where

$$\mathcal{A} = \begin{pmatrix} A_{00} & A_{01} \\ A_{10} & A_{11} \end{pmatrix},$$

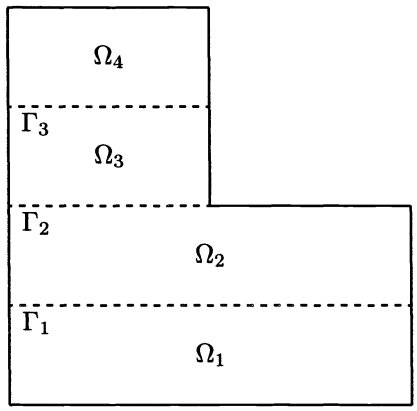


FIG. 1. *The domain decomposed in four subdomains.*

or more explicitly expressed in terms of the different subdomains

$$\mathcal{A} = \left(\begin{array}{ccc|ccc} A_{00}^{(1)} & & & A_{01}^{(11)} & A_{01}^{(12)} & \\ & \ddots & & & \ddots & \ddots \\ & & A_{00}^{(p-1)} & & & A_{01}^{((p-1)(p-1))} \quad A_{01}^{((p-1)p)} \\ \hline A_{10}^{(11)} & & & A_{11}^{(1)} & & \\ A_{10}^{(21)} & \ddots & & & A_{11}^{(2)} & \\ & \ddots \quad A_{10}^{((p-1)(p-1))} & & & & \ddots \\ & & A_{10}^{(p(p-1))} & & & A_{11}^{(p)} \end{array}\right).$$

Those blocks which are not written explicitly are understood to be zero.

2.2. The Schur complement method (SCM). In a Schur complement method, [3]–[5], [16], (3) is reduced to a system for the unknowns on Γ

$$Cu_0 = g, \tag{4}$$

where

$$C = A_{00} - A_{01}A_{11}^{-1}A_{10}, \tag{5}$$

and

$$g = b_0 - A_{01}A_{11}^{-1}b_1.$$

Equation (4) is solved for u_0 and this u_0 is inserted into (3). The remaining unknowns $u_1^{(q)}$ can then be computed locally in Ω_q. This computation is the same as solving the differential equation on Ω with Dirichlet boundary conditions u_0 on Γ. The inherent parallelism in the algorithm can be exploited by using a computer with a parallel architecture.

We now state the algorithm to solve our problem.

ALGORITHM SCM

1. Solve $A_{11}w = b_1$.
2. Solve $Cu_0 = b_0 - A_{01}w$.
3. Solve $A_{11}v = A_{10}u_0$.
4. $u_1 = w - v$.

Steps 1 and 3,

$$\text{Solve } A_{11}x_1 = y_1 \tag{6}$$

are local subdomain solves,

$$\text{Solve } A_{11}^{(q)}x_1^{(q)} = y_1^{(q)}, \qquad q = 1, \ldots, p, \tag{7}$$

that can be effected totally in parallel with no communication. Similarly, Step 4 is a local operation, while Step 2 demands the solution of a system that is spread over the processors.

In this paper we use an iterative method to solve (4) and (6). We consider PGCR(ℓ), *preconditioned generalized conjugate residuals*, with restarting length ℓ, [6], which is a method that works well for nonsymmetric, nondiagonally dominant systems of equations.

In some cases it might be somewhat inconvenient to use an iterative method to solve (4), and (6). First we must decide how accurately to carry out Steps 1 and 2 in Algorithm SCM. Then we must carry out Step 3, such that the convergence criterion used is fulfilled. Thus the accuracy in w and u_0 affects the total number of iterations needed to obtain the accuracy called for and, consequently, the performance of the method. Furthermore, we must deal with the subdomain solves in the matrix-vector-multiplications Cx that we perform at each iteration step. Studying C we see that each such multiplication requires local subdomain solves in each subdomain. Hence a stop criterion is needed for these inner iterations, which is a parameter that also affects the performance. In §6.2 we describe how to partly circumvent these problems, when we have constant coefficients in one space dimension in the PDE.

3. Iterative methods. Consider a general system of equations

$$Bx = y. \tag{8}$$

PGCR is a minimal residual iteration [7], which in step k fulfills

$$\|r_k\|_2 = \min_{p_k \in \mathcal{P}_k,\, p_k(0)=1} \|p_k(\hat{B}^{-1}B)r_0\|_2,$$

where $\mathcal{P}_k$ is the set of all polynomials of degree k, $\hat{B}$ is the left preconditioner, $r_k \equiv \hat{B}^{-1}(Bx_k - y)$, and x_k is the approximation of x obtained in iteration k. If $\hat{B}^{-1}B$ is diagonalizable we obtain

$$\frac{\|r_k\|_2}{\|r_0\|_2} \le \text{cond}_2(W_{\hat{B}^{-1}B}) \min_{(p_k \in \mathcal{P}_k,\, p_k(0)=1)} \max_{1 \le \ell \le \eta} |p_k(\lambda_\ell)| \equiv \text{cond}_2(W_{\hat{B}^{-1}B}) \cdot \varepsilon_k, \tag{9}$$

where $W_{\hat{B}^{-1}B}$ is the eigenvector matrix and λ_ℓ are the eigenvalues of $\hat{B}^{-1}B$. We also define the asymptotic convergence factor ρ as

$$\rho \equiv \lim_{k \to \infty} \varepsilon_k^{1/k}. \tag{10}$$

From (9) we conclude that we will have finite termination in η iterations, where η is the number of distinct eigenvalues to the preconditioned system. Moreover, if we can precondition our system such that the eigenvalues of $\hat{B}^{-1}B$ are contained in μ dense clusters, we have a good approximation to the solution in μ iterations. Clustering of the eigenvalues may be even more important than a condition number improvement, [1], [2], and [21]. In §5, preconditioners to (4), and (6) which yield highly clustered spectra are presented. We also show in §6.1, that it might not be necessary to precondition (4).

4. The model problem.

4.1. The differential equation. The model problem studied is a scalar two-dimensional equation

$$\frac{\partial u}{\partial t} + \sigma_1(x_1)\frac{\partial u}{\partial x_1} + \sigma_2(x_2)\frac{\partial u}{\partial x_2} = g, \tag{11}$$
$$0 < x_1 \leq 1\,,\ 0 < x_2 \leq 1\,,\ t > 0,$$

where $\sigma_1(x_1) > 0$ and $\sigma_2(x_2) > 0$. Equation (11) is well-posed if we prescribe $u(x_1, 0, t)$, $u(0, x_2, t)$, and $u(x_1, x_2, 0)$. We could also consider a periodic boundary condition in either space direction. Note that since we are aiming at solving systems of PDEs, we cannot use any particular features of the scalar equation.

4.2. Discretization. Introduce a uniform grid as

$$x_{1,k} = kh_1, \qquad k = 1, \ldots, n,$$
$$x_{2,j} = jh_2, \qquad j = 1, \ldots, m.$$

Let $u_{k,j}$ denote the approximate solution at the point $(x_{1,k}, x_{2,j})$. Equation (11) is discretized in time using the trapezoidal rule with time step Δt. For the space derivatives we use centered differences in the interior of the domain and one-sided differences at the outflow boundaries ($k = n$ or $j = m$).

$$\begin{aligned}
&\frac{\partial u}{\partial x_1} \approx D_{0,x_1} u_{k,j}, \\
&\qquad\qquad\qquad\qquad k = 1, \ldots, n-1 \text{ and } j = 1, \ldots, m-1, \\
&\frac{\partial u}{\partial x_2} \approx D_{0,x_2} u_{k,j}, \\
&\frac{\partial u}{\partial x_1} \approx D_{-,x_1} u_{k,j}, \qquad j = 1, \ldots, m \text{ and } k = n, \\
&\frac{\partial u}{\partial x_2} \approx D_{-,x_2} u_{k,j}, \qquad k = 1, \ldots, n \text{ and } j = m.
\end{aligned} \tag{12}$$

The relations obtained from the last two definitions will be referred to as the numerical outflow boundary conditions.

To give the equations a simpler appearance we define the following quantities:

$$\kappa_1 = \frac{\Delta t}{h_1},$$

$$\kappa_2 = \frac{\Delta t}{h_2},$$

$$\tilde{\kappa}_{1,k} = \sigma_1(x_{1,k})\kappa_1, \qquad k = 1, \ldots, n,$$

$$\tilde{\kappa}_{2,j} = \sigma_2(x_{2,j})\kappa_2, \qquad j = 1, \ldots, m,$$

$$u^\nu = (\, u^\nu_{1,1} \quad u^\nu_{2,1} \quad \ldots \quad u^\nu_{n,1} \quad u^\nu_{1,2} \quad \ldots \quad u^\nu_{n,m}\,)^T .$$

Rearranging the equations leads to a system of equations to solve for each time step

$$Au^{\nu+1} = b. \tag{13}$$

Here b contains known quantities and A has the block tridiagonal form

$$A = \begin{pmatrix} D_1 & G_1 & & & \\ -G_2 & D_2 & G_2 & & \\ & \ddots & \ddots & \ddots & \\ & & -G_{m-1} & D_{m-1} & G_{m-1} \\ & & & -2G_m & D_m + 2G_m \end{pmatrix},$$

where

$$D_j = \begin{pmatrix} 4 & \tilde{\kappa}_{1,1} & & & \\ -\tilde{\kappa}_{1,2} & 4 & \tilde{\kappa}_{1,2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\tilde{\kappa}_{1,n-1} & 4 & \tilde{\kappa}_{1,n-1} \\ & & & -2\tilde{\kappa}_{1,n} & 4 + 2\tilde{\kappa}_{1,n} \end{pmatrix}, \qquad j = 1, \ldots, m, \tag{14}$$

$$G_j = \tilde{\kappa}_{2,j} I_n, \qquad j = 1, \ldots, m, \tag{15}$$

where I_n denotes the identity matrix of order n. In the following we consider the domain decomposition defined in (3) applied to the system of equations (13).

Next we will show that for constant coefficients, the coefficient matrix C defined in (5) is nonsingular. For constant coefficients we define

$$\tilde{\kappa}_d \equiv \sigma_d \kappa_d, \qquad d = 1, 2$$

which yields

$$A = 4I_{n\times m} + \tilde{\kappa}_1 I_m \otimes A_1 + \tilde{\kappa}_2 A_2 \otimes I_n, \tag{16}$$

where A_1 is a matrix of order n and A_2 a matrix of order m defined by

$$A_d = \begin{pmatrix} 0 & 1 & & & \\ -1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 0 & 1 \\ & & & -2 & 2 \end{pmatrix}.$$

If A is nonsingular we know that a unique solution exists, and hence the Schur complement C in (4) will also be nonsingular. In order to prove nonsingularity, we restate a lemma from [17]. Let $\mathcal{E}(a,b)$ denote the closed ellipse centered at the origin with semimajor axis b oriented along the imaginary axis and semiminor axis a. Also let $\mathcal{E}^+(a,b)$ denote the region $\{z|z \in \mathcal{E}(a,b) \text{ and } \Re e(z) \geq 0\}$.

LEMMA 4.1. *The eigenvalues* $\lambda_{1,k}$, $k = 1,\ldots,n$ *of* A_1 *and the eigenvalues* $\lambda_{2,j}$, $j = 1,\ldots,m$ *of* A_2 *satisfy*

(i) $\lambda_{d,k} \neq \lambda_{d,j}$, $k \neq j$,
(ii) $\lambda_{1,k} \in \mathcal{E}^+(4n^{-3/4}, 2 + 4n^{-3/2})$,
(iii) $\lambda_{2,j} \in \mathcal{E}^+(4m^{-3/4}, 2 + 4m^{-3/2})$.

With this lemma we will now prove the nonsingularity of A.

THEOREM 4.2. *The coefficient matrix* A *defined in* (16) *is nonsingular.*

Proof. From Lemma 4.1 we conclude that there exist nonsingular matrices V_d, $d = 1, 2$, such that

$$A = (V_2 \otimes V_1)\Lambda(V_2 \otimes V_1)^{-1},$$

where

$$\Lambda = 4I_{n\times m} + \tilde{\kappa}_1 I_m \otimes \Lambda_1 + \tilde{\kappa}_2 \Lambda_2 \otimes I_n = \operatorname{diag}(\lambda_{11},\ldots,\lambda_{nm}),$$

$$\Lambda_1 = \operatorname{diag}(\lambda_{1,1},\ldots,\lambda_{1,n}), \quad \text{and} \quad \Lambda_2 = \operatorname{diag}(\lambda_{2,1},\ldots,\lambda_{2,m}).$$

Now from Lemma 4.1 we get

$$\Re e(\lambda_{kj}) = 4 + \tilde{\kappa}_1 \Re e(\lambda_{1,k}) + \tilde{\kappa}_2 \Re e(\lambda_{2,j}) \geq 4$$

which proves the theorem. □

5. Preconditioners. Holmgren and Otto have investigated semi and block circulant preconditioners for (13) showing very good results, [14], [15]. The idea is to create blocks in the preconditioner that are circulant, and then use Fourier techniques to compute the inverse. A somewhat similar approach is presented in [11], where the blocks are Toeplitz with a certain symmetry. By a modified sine transform defined in [10], the solver for the preconditioner system again is based on Fourier techniques. In this report we employ preconditioners based on Toeplitz blocks. Circulant preconditioners have also been implemented, but since the results from these were inferior to the ones presented here, we do not further discuss such preconditioners in this report. However, we cannot conclude that circulant preconditioners do not work in a domain decomposition context. Numerical spectra indicate that for larger problems they might do, but since we are interested in all problem sizes, these results are left out.

5.1. Definition of matrices and operators. We start by defining some important matrices that we will use in what follows.

DEFINITION 5.1.

$$A_{00}(D) = I_{p-1} \otimes \hat{D},$$

where

$$\hat{D} = \begin{pmatrix} 4 & \tilde{\kappa}_{1,1} & & & \\ -\tilde{\kappa}_{1,2} & 4 & \tilde{\kappa}_{1,2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\tilde{\kappa}_{1,n-1} & 4 & \tilde{\kappa}_{1,n-1} \\ & & & -\tilde{\kappa}_{1,n} & 4 \end{pmatrix}.$$

$$A_{11}^{(q)}(\bullet, D) = \begin{pmatrix} D_{j'} & G_{j'} & & & \\ -G_{j'+1} & D_{j'+1} & G_{j'+1} & & \\ & \ddots & \ddots & \ddots & \\ & & -G_{j'+m_q-2} & D_{j'+m_q-2} & G_{j'+m_q-2} \\ & & & -G_{j'+m_q-1} & D_{j'+m_q-1} \end{pmatrix},$$

$$q = 1, \ldots, p,$$

where $j' = q + \sum_{\ell=1}^{q-1} m_\ell$, *and* D_j, G_j *are defined in* (14) *and* (15). *Finally*

$$A_{11}^{(q)}(D, D) = \begin{pmatrix} \hat{D} & G_{j'} & & & \\ -G_{j'+1} & \hat{D} & G_{j'+1} & & \\ & \ddots & \ddots & \ddots & \\ & & -G_{j'+m_q-2} & \hat{D} & G_{j'+m_q-2} \\ & & & -G_{j'+m_q-1} & \hat{D} \end{pmatrix},$$

$$q = 1, \ldots, p.$$

Next we define the matrix that will be the basis of the preconditioning matrices.

DEFINITION 5.2. *R_n is the matrix of order n defined by*

$$R_n = \begin{pmatrix} 0 & 1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & -1 & 0 \end{pmatrix}.$$

Those elements that are not written explicitly are understood to be zero.

R_n has the following eigenvalue decomposition

$$R_n = S_n \Lambda_n^R S_n^H,$$

where S_n is the modified sine matrix defined in [10], with entries

$$S_n(k, \ell) = \sqrt{\tfrac{2}{n+1}} i^{k+\ell+1} \sin\left(\tfrac{k\ell\pi}{n+1}\right), \qquad k, \ell = 1, \ldots, n.$$

The eigenvalue matrix $\Lambda_n^R = \operatorname{diag}(\lambda_1^R, \ldots, \lambda_n^R)$ is defined by

$$\lambda_k^R = 2i \cos\left(\tfrac{k\pi}{n+1}\right), \qquad k = 1, \ldots, n.$$

In the following we will define preconditioners that have a decomposition in modified sine matrices. We terminate this section by defining some operators that we will use.

DEFINITION 5.3. *In Table* 1 *we define a number of operators to be used.*

TABLE 1
Definition of operators.

Operator	Operand Q		Definition
	number of blocks	size of blocks	
c	1×1	$n \times n$	Equation (17)
$c_s^{(1)}$	$s \times s$	$n \times n$	Equation (18)
$\tilde{c}_s^{(1)}$	$s \times s$	$n \times n$	Equation (20)
$c_s^{(2)}$	$s \times s$	$n \times n$	$\tilde{c}_s^{(1)} \circ c_s^{(1)}$

$$c(Q) = \frac{1}{n}\left(\sum_{k=1}^{n} q_{k,k}\right) I_n + \frac{1}{2(n-1)}\left(\sum_{k=1}^{n-1}(q_{k,k+1} - q_{k+1,k})\right) R_n, \tag{17}$$

$$c_s^{(1)}(Q) = \begin{pmatrix} c([Q]_{11}) & c([Q]_{12}) & & \\ c([Q]_{21}) & \ddots & \ddots & \\ & \ddots & \ddots & c([Q]_{s-1,s}) \\ & & c([Q]_{s,s-1}) & c([Q]_{s,s}) \end{pmatrix}, \tag{18}$$

$$\tilde{c}_s^{(1)}(Q) = \tfrac{1}{s}\left[I_s \otimes \left(\textstyle\sum_{j=1}^{s}[Q]_{j,j}\right)\right] + \tfrac{1}{2(s-1)}\left[R_s \otimes \left(\textstyle\sum_{j=1}^{s-1}([Q]_{j,j+1} - [Q]_{j+1,j})\right)\right]. \tag{19}$$

5.2. Definition of preconditioners. In this section we define the preconditioners $\hat{C}$ to (4), and $\hat{A}_{11}$ to (6). We start with $\hat{A}_{11}$ which is defined by

$$\hat{A}_{11}(\bullet, D) = \operatorname{diag}(\hat{A}_{11}^{(1)}(\bullet, D), \ldots, \hat{A}_{11}^{(p)}(\bullet, D)), \tag{20}$$

where

$$\hat{A}_{11}^{(q)}(\bullet, D) = \tilde{c}_{m_q}^{(1)}(A_{11}^{(q)}(\bullet, D)), \qquad q = 1, \ldots, p\ .$$

We also define

$$\hat{C}(D) = \hat{A}_{00}(D) - A_{01}\hat{A}_{11}^{-1}(D, D)A_{10},$$

where

$$\hat{A}_{00}(D) = c_{p-1}^{(1)}(A_{00}(D)),$$

and

$$\hat{A}_{11}^{(q)}(D, D) = c_{m_q}^{(2)}(A_{11}^{(q)}(D, D)),$$

and

$$\hat{A}_{11}(D, D) = \operatorname{diag}(\hat{A}_{11}^{(1)}(D, D), \ldots, \hat{A}_{11}^{(p)}(D, D)).$$

5.3. Preconditioner solve.

5.3.1. Schur complement preconditioner. From the definition it is clear that $\hat{C}(D)$ has the following decomposition:

$$\hat{C}(D) = (I_{p-1} \otimes S_n) U(D) (I_{p-1} \otimes S_n^H). \tag{21}$$

The matrix $U(D)$ is sparse with only three nonzero diagonals

$$U(D) = \begin{pmatrix} \Lambda^{(1)}(D) & \Delta^{(1)}(D) & & \\ \Theta^{(2)}(D) & \ddots & \ddots & \\ & \ddots & \ddots & \Delta^{(p-2)}(D) \\ & & \Theta^{(p-1)}(D) & \Lambda^{(p-1)}(D) \end{pmatrix}, \tag{22}$$

where

$$\left.\begin{aligned} \Lambda^{(q)}(D) &= \operatorname{diag}(\lambda_1^{(q)}(D), \ldots, \lambda_n^{(q)}(D)) \\ \Delta^{(q)}(D) &= \operatorname{diag}(\delta_1^{(q)}(D), \ldots, \delta_n^{(q)}(D)) \\ \Theta^{(q)}(D) &= \operatorname{diag}(\theta_1^{(q)}(D), \ldots, \theta_n^{(q)}(D)) \end{aligned}\right\}, \qquad q = 1, \ldots, p-1.$$

Here

$$\lambda_k^{(q)}(D) = 4 + 2i\bar{\kappa}_1(D) \cos\left(\frac{k\pi}{n+1}\right) + \tilde{\kappa}_{2,\bar{\jmath}} \tilde{\kappa}_{2,\bar{\jmath}+1} \sum_{j=1}^{m_q} \frac{S_{m_q}(m_q, j) S_{m_q}^H(j, m_q)}{4 + 2i\bar{\kappa}_1(D) \cos\left(\frac{k\pi}{n+1}\right) + 2i\bar{\kappa}_2(D, q) \cos\left(\frac{j\pi}{m_q+1}\right)}$$

$$+ \tilde{\kappa}_{2,\bar{\jmath}+1} \tilde{\kappa}_{2,\bar{\jmath}+2} \sum_{j=1}^{m_{q+1}} \frac{S_{m_{q+1}}(1, j) S_{m_{q+1}}^H(j, 1)}{4 + 2i\bar{\kappa}_1(D) \cos\left(\frac{k\pi}{n+1}\right) + 2i\bar{\kappa}_2(D, q+1) \cos\left(\frac{j\pi}{m_{q+1}+1}\right)}$$

$$\delta_k^{(q)}(D) = -\tilde{\kappa}_{2,\bar{\jmath}+1} \tilde{\kappa}_{2,\bar{\jmath}+m_{q+1}+1} \sum_{j=1}^{m_{q+1}} \frac{S_{m_{q+1}}(1, j) S_{m_{q+1}}^H(j, m_{q+1})}{4 + 2i\bar{\kappa}_1(D) \cos\left(\frac{k\pi}{n+1}\right) + 2i\bar{\kappa}_2(D, q+1) \cos\left(\frac{j\pi}{m_{q+1}+1}\right)}$$

$$\theta_k^{(q)}(D) = -\tilde{\kappa}_{2,\bar{\jmath}+1} \tilde{\kappa}_{2,\bar{\jmath}-m_q+1} \sum_{j=1}^{m_q} \frac{S_{m_q}(m_q, j) S_{m_q}^H(j, 1)}{4 + 2i\bar{\kappa}_1(D) \cos\left(\frac{k\pi}{n+1}\right) + 2i\bar{\kappa}_2(D, q) \cos\left(\frac{j\pi}{m_q+1}\right)},$$

$$k = 1, \ldots, n, \qquad q = 1, \ldots, p,$$

where $\bar{\jmath} = q - 1 + \sum_{\ell=1}^{q} m_\ell$ and

$$\bar{\kappa}_1(D) = \frac{1}{2(n-1)} \sum_{k=1}^{n-1} (\tilde{\kappa}_{1,k} + \tilde{\kappa}_{1,k+1}),$$

$$\bar{\kappa}_2(D, q) = \frac{1}{2(m_q - 1)} \sum_{j=j_{\text{start}}}^{j_{\text{stop}}} (\tilde{\kappa}_{2,j} + \tilde{\kappa}_{2,j+1}), \qquad q = 1, \ldots, p,$$

$j_{\text{start}} = q + \sum_{\ell=1}^{q-1} m_\ell$, and $j_{\text{stop}} = q - 2 + \sum_{\ell=1}^{q} m_\ell$.

From (21) and the fact that S_n is unitary it is clear that the solution of

$$\hat{C}(D) x = y$$

is defined by

ALGORITHM SCHUR COMPLEMENT PRECONDITIONER SOLVE

1. $w = (I_{p-1} \otimes S_n^H)y$.
2. Solve $U(D)z = w$.
3. $x = (I_{p-1} \otimes S_n)z$.

Steps 1 and 3 are $p-1$ modified sine transforms and inverse modified sine transforms, respectively. The transforms can be computed completely in parallel. In [10] it is shown that each such transform can be accomplished using $\mathcal{O}(2.5n \log_2 n)$ operations, provided that $n+1$ is a power of 2. The system defined in Step 2 and (22), decouples into n systems of order $p-1$. These systems that are spread over the processors are solved using cyclic reduction, [13]. By symmetry only half of the number of systems must be solved for, [10].

5.3.2. Subdomain preconditioner. The subdomain preconditioner $\hat{A}_{11}$ is defined in (20). The inversion of $\hat{A}_{11}$ decouples into p independent local solves. Hence they can be performed in parallel with no communication. In [11] it is shown that

$$\hat{A}_{11}^{(q)}(\bullet, D) = (S_{m_q} \otimes I_n)T^{(q)}(\bullet, D)(S_{m_q}^H \otimes I_n),$$

where $T^{(q)}(\bullet, D)$ is sparse with only three nonzero diagonals. Hence

$$\hat{A}_{11}^{(q)} x_1^{(q)} = y_1^{(q)} \tag{23}$$

can be solved using the following algorithm.

ALGORITHM SUBDOMAIN PRECONDITIONER SOLVE

1. $w = (S_{m_q}^H \otimes I_n)y_1^{(q)}$.
2. Solve $T^{(q)}(\bullet, D)z = w$.
3. $x_1^{(q)} = (S_{m_q} \otimes I_n)z$.

By the previous section it is clear that Steps 1 and 3 are n transforms requiring $\mathcal{O}(2.5m_q \log_2 m_q)$ operations each. Step 2 yields the solution of m_q tridiagonal systems of order n. By symmetry we must only solve for half of the number of systems.

6. Theoretical analysis. Here we consider (11) with $\sigma_1 = \sigma_2 = 1$. Initially, we state a lemma that we use in the following. The proof can be found in [12].

LEMMA 6.1. *Let α be a complex number with $\Re e\,(\alpha) > 0$, and $\beta > 0$ real. Then*

$$\begin{aligned}\Phi_{k,\ell}(\alpha, \beta, N) &\equiv \frac{2}{N+1} \sum_{j=1}^{N} \frac{\sin\left(\frac{kj\pi}{N+1}\right) \sin\left(\frac{\ell j\pi}{N+1}\right)}{2\alpha + 2i\beta \cos\left(\frac{j\pi}{N+1}\right)} \\ &= \frac{i}{\beta}\left(\frac{z^{k+\ell} - z^{|k-\ell|}}{z - z^{-1}} + \frac{(z^{-k} - z^k)(z^{-\ell} - z^{\ell})}{z - z^{-1}} \cdot \frac{z^{2N+2}}{1 - z^{2N+2}}\right),\end{aligned}$$

where

$$z = i\left(\frac{\alpha}{\beta} - \sqrt{1 + \left(\frac{\alpha}{\beta}\right)^2}\right).$$

6.1. The Schur complement system for a semiperiodic problem. Here we theoretically analyze the semiperiodic problem, i.e., we impose periodic boundary conditions in one space direction. We replace (12) by

$$\frac{\partial u}{\partial x_1} \approx D_{0,x_1} u_{k,j},$$

$$\frac{\partial u}{\partial x_2} \approx D_{0,x_2} u_{k,j}, \qquad k = 1,\dots,n-1 \text{ and } j = 1,\dots,m-1,$$

$$\frac{\partial u}{\partial x_1} \approx \frac{u_{1,j} - u_{n-1,j}}{2h_1}, \qquad j = 1,\dots,m,$$

$$\frac{\partial u}{\partial x_2} \approx D_{-,x_2} u_{k,j}, \qquad k = 1,\dots,n \text{ and } j = m.$$

We consider the two-domain decomposition and restrict our formulation to the case $m_1 = m_2$. The latter condition is not necessary, it is just to simplify the notation. The coefficient matrices then become

$$A_{00} = \tilde{D}, \tag{24}$$

$$A_{11}^{(1)} = \begin{pmatrix} \tilde{D} & G & & & \\ -G & \tilde{D} & G & & \\ & \ddots & \ddots & \ddots & \\ & & -G & \tilde{D} & G \\ & & & -G & \tilde{D} \end{pmatrix}, \tag{25}$$

and

$$A_{11}^{(2)} = \begin{pmatrix} \tilde{D} & G & & & \\ -G & \tilde{D} & G & & \\ & \ddots & \ddots & \ddots & \\ & & -G & \tilde{D} & G \\ & & & -2G & \tilde{D}+2G \end{pmatrix}, \tag{26}$$

where

$$\tilde{D} = \begin{pmatrix} 4 & \kappa_1 & & -\kappa_1 \\ -\kappa_1 & \ddots & \ddots & \\ & \ddots & \ddots & \kappa_1 \\ \kappa_1 & & -\kappa_1 & 4 \end{pmatrix},$$

and $G = \kappa_2 I_n$. We also have

$$A_{01}^{(11)} = \begin{pmatrix} 0 & \cdots & 0 & -G \end{pmatrix},$$

$$A_{01}^{(12)} = \begin{pmatrix} G & 0 & \cdots & 0 \end{pmatrix},$$

$$A_{10}^{(11)} = -(A_{01}^{(11)})^T,$$

and

$$A_{10}^{(21)} = -(A_{01}^{(12)})^T.$$

We are going to study the eigenvalue decomposition of C defined in (5), when A_{11} is defined through (25) and (26). We start by considering the eigenvalue decomposition of $A_{11}^{(q)}$, $q = 1, 2$. $A_{11}^{(1)}$ defined in (25) is block Toeplitz with circulant blocks, which means that it has an eigenvalue decomposition as $A_{11}^{(1)} = (S_{m_1} \otimes F_n)\Lambda^{(1)}(S_{m_1}^H \otimes F_n^H)$, where F_n is the Fourier matrix defined by

$$F_n(k,\ell) = \sqrt{\tfrac{1}{n}} e^{\left(\frac{2\pi i(k-1)(\ell-1)}{n}\right)}, \qquad k,\ell = 1,\ldots,n.$$

The eigenvalue matrix $\Lambda^{(1)}$ is defined by $\Lambda^{(1)} = \operatorname{diag}(\Lambda_1^{(1)},\ldots,\Lambda_{m_1}^{(1)})$, where $\Lambda_j^{(1)} = \operatorname{diag}(\lambda_{1j}^{(1)},\ldots,\lambda_{nj}^{(1)})$, and

$$\lambda_{kj}^{(1)} = 4 + 2i\kappa_1 \sin\left(\tfrac{2(k-1)\pi}{n}\right) + 2i\kappa_2 \cos\left(\tfrac{j\pi}{m_1+1}\right), \qquad k = 1,\ldots,n, j = 1,\ldots,m_1.$$

Since it is the eigenvalue decomposition of C we really are interested in, we observe that the term $-A_{01}^{(11)}(A_{11}^{(1)})^{-1}A_{10}^{(11)}$ can be written as

$$\begin{aligned} &\kappa_2^2(u^T \otimes I_n)(S_{m_1} \otimes F_n)(\Lambda^{(1)})^{-1}(S_{m_1}^H \otimes F_n^H)(u \otimes I_n) \\ = \;& F_n\left(\kappa_2^2 \sum_{j=1}^{m_1} S_{m_1}(m_1,j)S_{m_1}^H(j,m_1)(\Lambda_j^{(1)})^{-1}\right) F_n^H \equiv F_n E^{(1)} F_n^H, \end{aligned}$$

where

$$u = (\,0 \quad \ldots \quad 0 \quad 1\,)^T. \tag{27}$$

$E^{(1)} = \operatorname{diag}(e_1^{(1)},\ldots,e_n^{(1)})$, where

$$e_k^{(1)} = \kappa_2^2 \sum_{j=1}^{m_1} \frac{S_{m_1}(m_1,j)S_{m_1}^H(j,m_1)}{4 + 2i\kappa_1 \sin\left(\frac{2(k-1)\pi}{n}\right) + 2i\kappa_2 \cos\left(\frac{j\pi}{m_1+1}\right)}, \qquad k = 1,\ldots,n.$$

By Lemma 6.1 we get

$$e_k^{(1)} = \kappa_2^2 \Phi_{m_1,m_1}\left(2 + i\kappa_1 \sin\left(\tfrac{2(k-1)\pi}{n}\right), \kappa_2, m_1\right) = \kappa_2 i z_k \frac{1 - z_k^{2m_1}}{1 - z_k^{2m_1+2}}, \qquad k = 1,\ldots,n,$$

$$z_k = i\left(\frac{2 + i\kappa_1 \sin\left(\frac{2(k-1)\pi}{n}\right)}{\kappa_2} - \sqrt{1 + \left(\frac{2 + i\kappa_1 \sin\left(\frac{2(k-1)\pi}{n}\right)}{\kappa_2}\right)^2}\,\right). \tag{28}$$

As $m_2 = m_1$ we express $A_{11}^{(2)}$ as $A_{11}^{(2)} = A_{11}^{(1)} + UV^T$, where

$$\begin{aligned} U &= \kappa_2(u \otimes I_n), \\ V &= v \otimes I_n, \\ v &= (\,0 \quad \ldots \quad 0 \quad -1 \quad 2\,)^T, \end{aligned}$$

and u is defined in (27). Using the Sherman–Morrison–Woodbury formula we obtain

$$(A_{11}^{(1)} + UV^T)^{-1} = (A_{11}^{(1)})^{-1} - (A_{11}^{(1)})^{-1}U(I_n + V^T(A_{11}^{(1)})^{-1}U)^{-1}V^T(A_{11}^{(1)})^{-1}. \tag{29}$$

Now

$$\begin{aligned} & I_n + V^T(A_{11}^{(1)})^{-1}U \\ &= I_n + \kappa_2(v^T \otimes I_n)(S_{m_1} \otimes F_n)(\Lambda^{(1)})^{-1}(S_{m_1}^H \otimes F_n^H)(u \otimes I_n) \\ &= I_n + \kappa_2(v^T S_{m_1} \otimes F_n)(\Lambda^{(1)})^{-1}(S_{m_1}^H u \otimes F_n^H) \\ &= F_n(I_n + \kappa_2 \sum_{j=1}^{m_1}(-S_{m_1}(m_1-1,j) + 2S_{m_1}(m_1,j))S_{m_1}^H(j,m_1)(\Lambda_j^{(1)})^{-1})F_n^H \\ &\equiv F_n L F_n^H. \end{aligned} \tag{30}$$

Here $L = \operatorname{diag}(l_1, \ldots, l_n)$ and

$$l_k = 1 + \kappa_2 \sum_{j=1}^{m_1} \frac{(-S_{m_1}(m_1-1,j) + 2S_{m_1}(m_1,j))S_{m_1}^H(j,m_1)}{4 + 2i\kappa_1 \sin\left(\frac{2(k-1)\pi}{n}\right) + 2i\kappa_2 \cos\left(\frac{j\pi}{m_1+1}\right)}, \quad k = 1, \ldots, n. \tag{31}$$

Using Lemma 6.1, (32) yields

$$\begin{aligned} l_k &= 1 + \kappa_2 \times \left(i\Phi_{m_1-1,m_1}\left(2 + i\kappa_1 \sin\left(\frac{2(k-1)\pi}{n}\right), \kappa_2, m_1\right)\right. \\ &\quad \left. + 2\Phi_{m_1,m_1}\left(2 + i\kappa_1 \sin\left(\frac{2(k-1)\pi}{n}\right), \kappa_2, m_1\right)\right) \\ &= 1 - z_k^2 \frac{1 - z_k^{2m_1-2}}{1 - z_k^{2m_1+2}} + 2iz_k \frac{1 - z_k^{2m_1}}{1 - z_k^{2m_1+2}}, \qquad k = 1, \ldots, n, \end{aligned} \tag{32}$$

where z_k is defined in (28). Inserting (30) in (29) yields

$$(A_{11}^{(1)} + UV^T)^{-1} = (S_{m_1} \otimes F_n)((\Lambda^{(1)})^{-1} - \kappa_2(\Lambda^{(1)})^{-1}W(\Lambda^{(1)})^{-1})(S_{m_1}^H \otimes F_n^H),$$

where W is defined by

$$\begin{aligned} W &= (S_{m_1}^H \otimes F_n^H)(u \otimes I_n)F_n L^{-1} F_n^H (v^T \otimes I_n)(S_{m_1} \otimes F_n) \\ &= (S_{m_1}^H u v^T S_{m_1}) \otimes L^{-1}. \end{aligned} \tag{33}$$

Now $-A_{01}^{(12)}(A_{11}^{(2)})^{-1}A_{10}^{(21)}$ can be written as $\kappa_2^2(w^T \otimes I_n)(A_{11}^{(1)} + UV^T)^{-1}(w \otimes I_n)$, where $w = (\ 1\ \ 0\ \ \ldots\ \ 0\ )^T$ which, together with (33), yields

$$-A_{01}^{(12)}(A_{11}^{(2)})^{-1}A_{10}^{(21)} = F_n E^{(2)} F_n^H,$$

where $E^{(2)}$ is defined by

$$E^{(2)} = \kappa_2^2 \left(\sum_{j=1}^{m_1} S_{m_1}(1,j)S_{m_1}^H(j,1)(\Lambda_j^{(1)})^{-1} \right.$$

$$
\begin{aligned}
&- \kappa_2 L^{-1} \sum_{j=1}^{m_1} S_{m_1}(1,j) S_{m_1}^H(j,m_1)(\Lambda_j^{(1)})^{-1} \\
&\times \sum_{j=1}^{m_1} S_{m_1}^H(j,1)(-S_{m_1}(m_1-1,j) + 2S_{m_1}(m_1,j))(\Lambda_j^{(1)})^{-1} \Bigg) \\
&\equiv \operatorname{diag}(e_1^{(2)}, \ldots, e_n^{(2)}).
\end{aligned}
$$

The sums in $e_k^{(2)}$, $k = 1, \ldots, n$, can be computed using Lemma 6.1, and by inserting (32) we get

$$
e_k^{(2)} = \kappa_2 \left(\frac{iz_k(1-z_k^{2m_1})}{1-z_k^{2m_1+2}} + \frac{iz_k^{2m_1-1}(1-z_k^2)(1-z_k^4)+2z_k^{2m_1}(1-z_k^2)^2}{(1-z_k^{2m_1+2})(1-z_k^{2m_1+2}-z_k^2(1-z_k^{2m_1-2})+2iz_k(1-z_k^{2m_1}))} \right),
$$

$$
k = 1, \ldots, n.
$$

Finally A_{00} defined in (24), has a well-known decomposition in Fourier matrices, and hence the semiperiodic problem divided in two subdomains gives a Schur complement C with an eigenvalue decomposition

$$
C = F_n \Lambda F_n^H, \tag{34}
$$

with eigenvalues

(35)

$$
\begin{aligned}
\lambda_k = 4 &+ 2i\kappa_1 \sin\left(\frac{2(k-1)\pi}{n}\right) \\
&+ \kappa_2 \Bigg(2iz_k \frac{1-z_k^{2m_1}}{1-z_k^{2m_1+2}} \\
&\quad + \frac{iz_k^{2m_1-1}(1-z_k^2)(1-z_k^4) + 2z_k^{2m_1}(1-z_k^2)^2}{(1-z_k^{2m_1+2})(1-z_k^{2m_1+2}-z_k^2(1-z_k^{2m_1-2}) + 2iz_k(1-z_k^{2m_1}))} \Bigg),
\end{aligned}
$$

where z_k is defined in (28).

From (34) we see that the solution of the Schur complement system for a semiperiodic problem can be obtained from one Fourier transform, the solution of a diagonal system of order n, and finally one inverse Fourier transform. This can be generalized to an arbitrary number of subdomains. Then we perform $p-1$ Fourier transforms in parallel, solve n tridiagonal systems of order $p-1$, and finally compute $p-1$ parallel inverse Fourier transforms. The solution of the tridiagonal systems can be parallelized using cyclic reduction for example.

From (35) we will derive a theorem concerning the asymptotic spectrum. Note that we do not require

$$
\Delta t = ch_1^{\alpha}, \qquad 0 < \alpha < 1, \qquad c > 0,
$$

which is the case in [17] and [12].

THEOREM 6.2. *Assume that*

$$
\phi \equiv \frac{n}{m} < 1, \tag{36}
$$

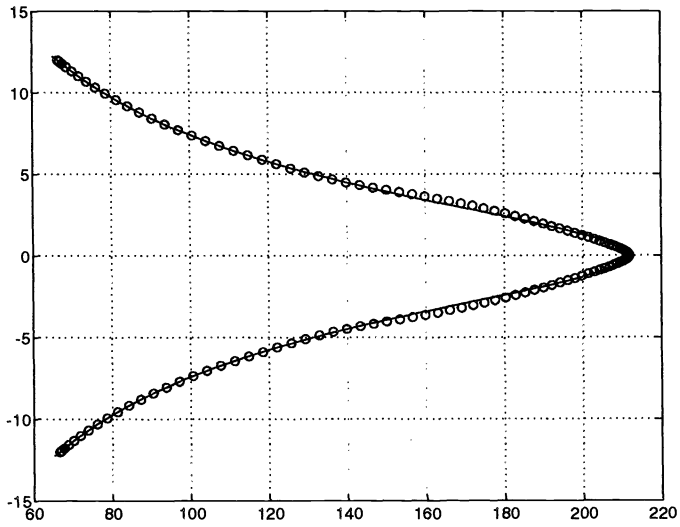


FIG. 2. *Asymptotic spectrum solid line and eigenvalues marked with circles for* $\kappa_1 = 100$, $n = 250$, *and* $m = 263$, *i.e.*, $\phi = 0.95$.

and that κ_1 *is large and fixed. In the limit* $n \to \infty$ *the eigenvalues* λ_k, $k = 1, \ldots, n$, *defined in* (35) *reside on a curve-segment* $\tilde{\lambda}(\zeta)$ *defined by*

$$\tilde{\lambda}(\zeta) = 2\kappa_1\phi^{-1}\sqrt{1-\zeta^2} + \frac{4i\zeta}{\sqrt{1-\zeta^2}} + \mathcal{O}(\kappa_1^{-1}), \qquad -\phi \le \zeta \le \phi. \tag{37}$$

Proof. Let δ_k be defined by

$$\delta_k = \phi \sin\left(\frac{2(k-1)\pi}{n}\right).$$

Then for large κ_1 we get from a Taylor expansion that

$$z_k = -\delta_k - i\sqrt{1-\delta_k^2} + \frac{2i\phi}{\kappa_1} + \frac{2\delta_k\phi}{\kappa_1\sqrt{1-\delta_k^2}} + \mathcal{O}(\kappa_1^{-2}), \qquad k = 1, \ldots, n.$$

Since

$$|z_k|^2 = 1 - \frac{4\phi}{\kappa_1\sqrt{1-\delta_k^2}} + \mathcal{O}(\kappa_1^{-2}) < 1,$$

we get $\lim_{n\to\infty} |z_k|^{2m_1} = 0$, and hence, from (35) we get (37). □

From Theorem 6.2 we see that the eigenvalues stay bounded and are well separated from the origin when the problem size increases. In Figs. 2 and 3 we show how well the asymptotic formula agrees with the eigenvalues for large problems.

We now enclose the asymptotic spectrum $\tilde{\lambda}$, defined in (37), in a semicircle $\mathcal{C}^-(c, R)$. Here $\mathcal{C}(c, R)$ denotes the circle with center c and radius R, and $\mathcal{C}^-(c, R)$ is the semicircle defined by

$$\{z | (z \in \mathcal{C}(c, R) \text{ and } \Re e(z) \le c) \text{ or } (\Re e(z) = c \text{ and } |\Im m(z)| \le R)\}.$$

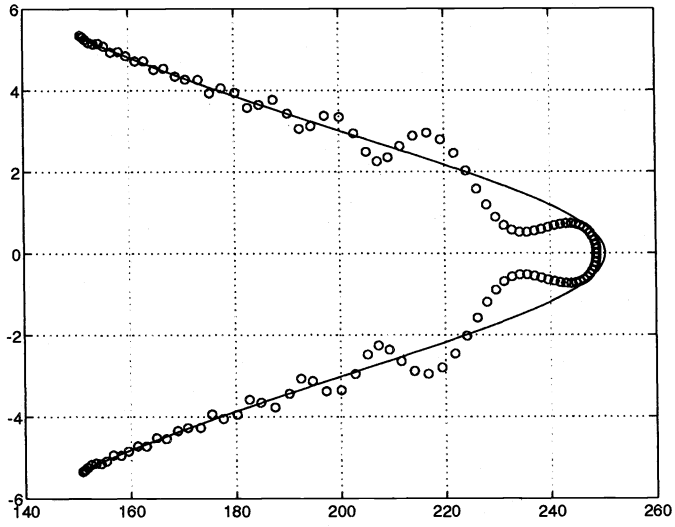


FIG. 3. *Asymptotic spectrum solid line and eigenvalues marked with circles for* $\kappa_1 = 100$, $n = 250$, *and* $m = 313$, *i.e.,* $\phi = 0.80$.

In the following we neglect $\mathcal{O}(\kappa_1^{-1})$-terms.

LEMMA 6.3. *The semicircle* $\mathcal{C}^-(c, R)$, *where* c *and* R *are defined by*

$$c = \frac{2\kappa_1}{\phi},$$

$$R = |\tilde{\lambda}(\pm\phi) - c| = \sqrt{\frac{16\phi^2}{1-\phi^2} + \frac{4\kappa_1^2(2-\phi^2-2\sqrt{1-\phi^2})}{\phi^2}},$$

encloses the asymptotic spectrum $\tilde{\lambda}$ *defined in* (37).

Proof. Define $r(\zeta)$ as the distance between c and $\tilde{\lambda}(\zeta)$. Then

$$r^2 = \frac{16\zeta^2}{1-\zeta^2} + \frac{4\kappa_1^2(1-\sqrt{1-\zeta^2})^2}{\phi^2} = \frac{16\zeta^2}{1-\zeta^2} + \frac{4\kappa_1^2(2-\zeta^2-2\sqrt{1-\zeta^2})}{\phi^2}$$

and

$$\frac{d^2r^2}{d\zeta^2} = \frac{32(1-\zeta^2)^2 - 128(\zeta^4-\zeta^2)}{(1-\zeta^2)^4} + \frac{4\kappa_1^2}{\phi^2}\left(-2 + \frac{2(1+\zeta^2-\zeta^4)}{\sqrt{1-\zeta^2}}\right) > 0.$$

Hence, r^2 obtains its maximum value at the endpoints yielding $r(\zeta) \le R$. Finally

$$\Re e\ \left(\tilde{\lambda}(\zeta)\right) = \frac{2\kappa_1\sqrt{1-\zeta^2}}{\phi} \le \frac{2\kappa_1}{\phi},$$

which proves the theorem. □

From Lemma 6.3 and the general result in [18] we derive the following theorem.

THEOREM 6.4. *The asymptotic convergence factor* ρ *defined in* (10), *fulfills*

$$\rho \le \sqrt{\frac{4\phi^4}{\kappa_1^2(1-\phi^2)} + 2 - \phi^2 - 2\sqrt{1-\phi^2}}. \tag{38}$$

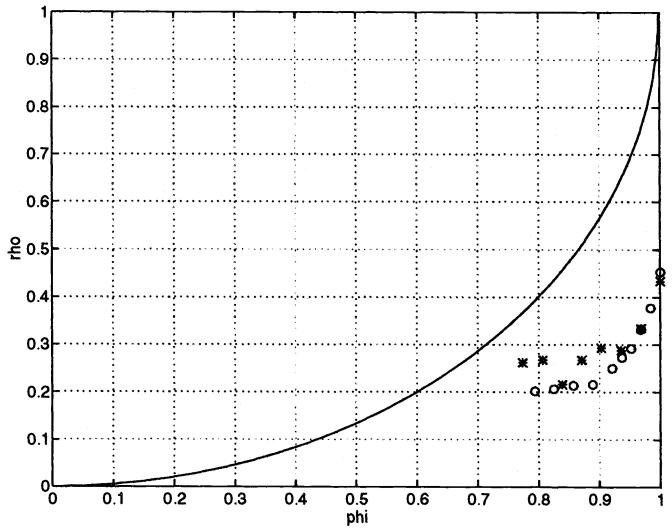


FIG. 4. *Asymptotic convergence factor solid line. The residual reduction* $\tilde{\rho}_{20}$ *is marked with asterisks for* $m = 31$ *and circles for* $m = 63$.

Proof. From [18] we get that if the spectrum is enclosed in a circle $C(c, R)$, then the asymptotic convergence factor fulfills

$$\rho \leq \frac{R}{c}.$$

With this result and Lemma 6.3 we get (38). □

Now define the *residual reduction* $\tilde{\rho}_k$ as

$$\tilde{\rho}_k \equiv \left(\frac{\|r_k\|_2}{\|r_0\|_2} \right)^{1/k}. \tag{39}$$

From (9), (34), (39), and the fact that the Fourier matrix is unitary, we conclude that

$$\tilde{\rho}_\infty \leq \rho.$$

In Fig. 4 we display the asymptotic convergence factor as a function of ϕ for $\kappa_1 = 100$. In the same figure we present $\tilde{\rho}_{20}$ for the original problem (12).

From Fig. 4 we conclude that the shape of the obtained residual reduction agrees very well with the asymptotic convergence factor, especially for the larger problem. Hence we can expect an extensive gain in time by simply letting the aspect ratio of the space discretization slightly decrease from 1.

6.2. The subdomain systems. From [12] we can draw the following conclusions for the preconditioned subdomain systems. PGCR on (7) with preconditioner

(i) $\hat{A}_{11}^{(q)}(\bullet, D)$, $q = 1, \ldots, p-1$, converges in one iteration since $\hat{A}_{11}^{(q)}(\bullet, D) = A_{11}^{(q)}$.

(ii) $\hat{A}_{11}^{(p)}(\bullet, D)$ yields an asymptotic convergence factor ρ defined in (10), which fulfills

$$\rho \leq \frac{\sqrt{2 + 3\phi^2 - 2\sqrt{1 - \phi^2}}}{2},$$

where ϕ is defined in (36) and the time step Δt is given by

$$\Delta t = ch_1^{\alpha}, \qquad 0 < \alpha < 1, \qquad c > 0.$$

The proof of this result can be found in [12].

The most time-consuming part in the solution of (4), is the local preconditioner solves. It requires

$$\sum_{k=1}^{\text{it}_{\text{outer}}} \max_q \text{it}_{\text{inner},k}^{(q)}$$

such solves to solve (4), where $\text{it}_{\text{inner},k}^{(q)}$ is the number of inner iterations required in outer iteration k in Ω_q. Hence, for constant coefficients, most of the time in the iteration will be spent by performing inner iterations in Ω_p. Thus, processor number p will be working most of the time while the others remain idle. This is not a good load balancing in a parallel method. So, no matter how much we can reduce the number of outer iterations by preconditioning the Schur complement system, it will still be the inner iterations in Ω_p that limits the performance of the method. A remedy to this problem is to only consider the grid line $j = m$ as subdomain Ω_p. Then the coefficient matrix $A_{11}^{(p)}$ is tridiagonal, and this system can be solved with a direct method using $\mathcal{O}(8n)$ operations. Note also that for $q = 1, \ldots, p-1$ we solve (23) instead of (7) which leads to a reduction in memory requirements.

7. Spectra. In this section we present the spectra of both the preconditioned system and the original problem. In all figures the parameters are $n = 23$, $m = 23$, $p = 4$, $m_q = 5$, $q = 1, \ldots, 4$, $\kappa_1 = \kappa_2 = 100$, and $\sigma_1 = \sigma_2 = 1$. We also present different properties of the spectra in each case. For that reason we define the *spectral quotient* μ as a measure of the condition of the spectrum

$$\mu = \frac{\max_i |\lambda_i|}{\min_i |\lambda_i|}.$$

Here λ_i are the eigenvalues of the coefficient matrix considered.

7.1. The Schur complement system. In Fig. 5 and Table 2 we see that the Schur complement system seems to be relatively well-conditioned even without preconditioning. The Toeplitz approximations lead to a preconditioner that yields a well-clustered spectrum.

TABLE 2
Properties of the spectra of the Schur complement system.

	Preconditioner	
	None	$\hat{C}(D)$
$\max_i \Re e(\lambda_i)$	484.69	4.56
$\min_i \Re e(\lambda_i)$	12.43	0.21
$\max_i \Im m(\lambda_i)$	258.74	1.12
$\max_i \lvert\lambda_i\rvert$	542.25	4.56
$\min_i \lvert\lambda_i\rvert$	25.38	0.21
μ	21.36	22.16
$\#(\lambda_i = 1)$	0	34

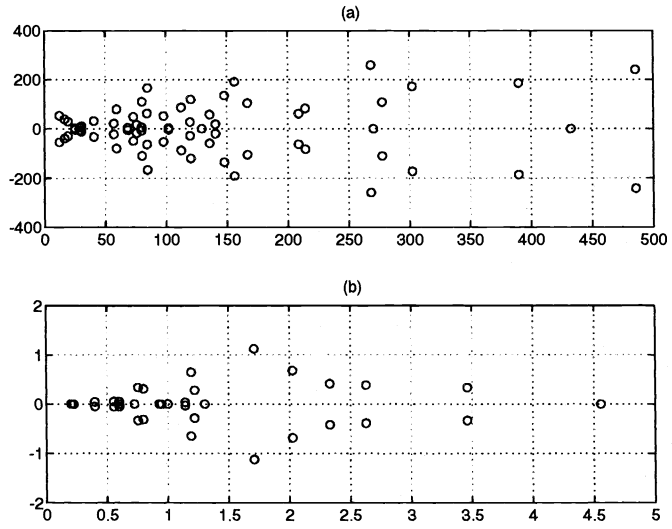


FIG. 5. *Spectra of the Schur complement system with preconditioner* (a) I *and* (b) $\hat{C}(D)$.

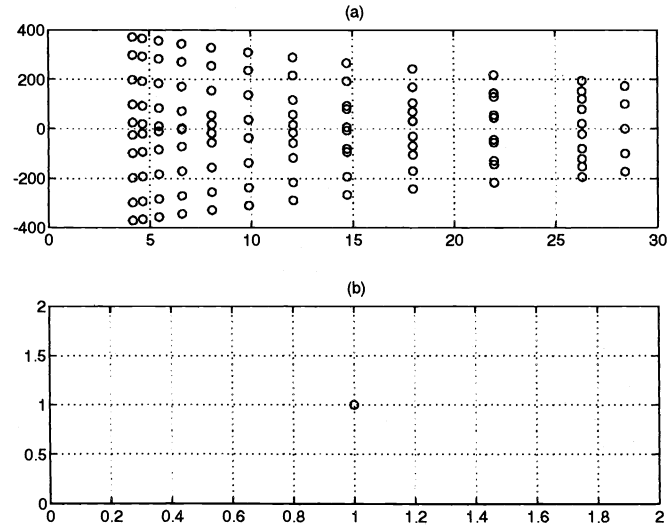


FIG. 6. *Spectra of a subdomain system with preconditioner* (a) I *and* (b) $\hat{A}_{11}^{(q)}(\bullet, D)$.

7.2. The subdomain systems. We only consider the subdomains Ω_q, $q = 1, \ldots, p-1$, since the spectra for Ω_p are shown in [11].

From Fig. 6, and Table 3 we immediately see that $\hat{A}_{11}^{(q)}(\bullet, D)$ yields the exact solver and hence we have convergence in one iteration, which is also pointed out in §6.2.

8. Results from the implementation on a parallel computer. Here we show the results obtained from a parallel implementation on an Intel Paragon. The program was originally written for the iPSC/2 at the Department of Scientific Computing, Uppsala University. It was then transferred to the Paragon located at Para//ab

TABLE 3
Properties of the spectra of a subdomain system.

	Preconditioner	
	None	$\hat{A}_{11}^{(q)}(\bullet, D)$
$\max_i \Re e(\lambda_i)$	28.41	1.00
$\min_i \Re e(\lambda_i)$	4.16	1.00
$\max_i \Im m(\lambda_i)$	371.35	0.00
$\max_i \lvert\lambda_i\rvert$	371.38	1.00
$\min_i \lvert\lambda_i\rvert$	6.93	1.00
μ	53.62	1.00
$\#(\lambda_i = 1)$	0	115

at the University of Bergen.

The coefficients σ_1 and σ_2 are chosen as

$$\sigma_d(x_d) = \cosh^2\left(s_d(2x_d - 1)\right) \frac{\tanh(s_d)}{s_d}. \tag{40}$$

This is suitable when we want to simulate stretchings in the physical grid. We then employ the coefficients (40), and perform the calculations in the equidistant computational grid. The scalar parameter s_d determines how much the grid is stretched in the x_d-direction. We consider both stretching in one space dimension, and stretchings in both space dimensions. Stretching in one space dimension is interesting, for instance, when we want to compute a two-dimensional viscous channel flow. In order to resolve the boundary layers, the grid lines are denser near the solid walls.

We will study the nonuniform decomposition of the domain discussed in §6.2. This is to avoid the difficulties introduced by the numerical outflow boundary conditions. For $s_2 = 0$, i.e., when we have constant coefficients in the x_2-direction, the subdomain solves are carried out exactly by the fast solver defined in §5.3.2 in Ω_q, $q = 1, \ldots, p-1$. In Ω_p we use a direct method. For $s_2 > 0$ we employ an iterative method to solve the subdomain problems in Ω_q, $q = 1, \ldots, p-1$. We then use the preconditioner defined in §5.3.2. Again we use a direct method in Ω_p.

The iterative method that we use to solve (4), and possibly also (6), is PGCR(6). A longer restarting length cannot be motivated due to memory restrictions. However, numerical experiments indicate that the number of iterations decreases considerably with increasing restarting length. As convergence criterion for a general system of equations (8), we use

$$\frac{\|\hat{B}^{-1}(y - Bx_k)\|_2}{\|\hat{B}^{-1}y\|_2} < 10^{-6}.$$

For all iterations we have taken the right-hand side as the initial guess to the solution.

In all cases we consider $p = \#subdomains = \#processors$.

8.1. Definitions. To determine how efficient a parallel program is, two quantities have been defined: *speedup* $S_{p,t}$ and *rate of efficiency* $E_{p,t}$ and $E_{p,s}$. The speedup shows how much faster the program executes on p nodes than on one node

$$S_{p,t} = \frac{T_1}{T_p}.$$

Here T_p denotes the execution time on p nodes. T_p could be determined in two different ways. To get the "nonscaled" speedup, T_p is measured as the time required on p nodes

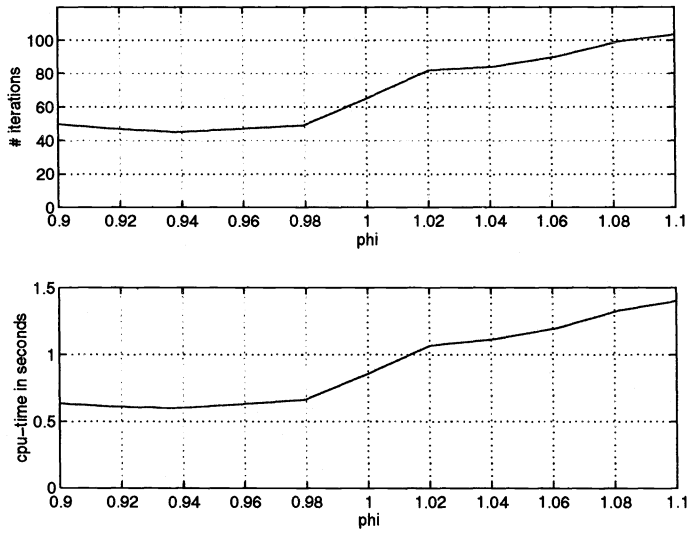


FIG. 7. *Number of iterations and* cpu *time obtained for convergence.*

when the problem in all is as big as the one on one node. To get the "scaled" speedup, the execution time is measured when each node deals with a problem as big as the one executed on one node. Then the time obtained is divided by p to get T_p.

The rate of efficiency indicates how efficient the program is. Two quantities are defined here, the rate of efficiency in time

$$E_{p,t} = \frac{S_{p,t}}{p}$$

and the rate of efficiency in arithmetic speed

$$E_{p,s} = \frac{M_p}{M_1 p}.$$

Here M_p is the arithmetic speed in floating point operations per second on p processors.

8.2. The method. In this section we study how the method depends on different parameters. We consider the nonuniform decomposition of the domain discussed in §6.2. We present here the results from iterating on (4).

The parameters used are $p = 4$, $\kappa_1 = 100$, $s_1 = s_2 = 0$, $m = 49$, $m_q = 15$, $q = 1, \ldots, p-1$, and $m_p = 1$. We present the results for varying ϕ defined in (36).

In Fig. 7 we see that the iteration count is *strongly* dependent on ϕ, i.e., the grid ratio. By decreasing ϕ from 1 to $\simeq 0.90$–0.95 we decrease the number of iterations considerably. Next we show the results from iterating on (4) for different problem sizes. The parameters used are $p = 4$, $\kappa_1 = 100$, $s_1 = s_2 = 0$, $m_q = (m-p)/(p-1)$, $q = 1, \ldots, p-1$, and $m_p = 1$. Since the grid ratio is so crucial for the performance, we have used $n \approx 0.95 \cdot m$.

From Fig. 7 and Table 4 we draw the following conclusions.

(i) The convergence depends *only* on the grid ratio and *not* on the number of unknowns.

(ii) Using only the grid line $j = m$ as Ω_p is very suitable for constant coefficients.

TABLE 4
Number of iterations and cpu *time obtained for convergence on an Intel Paragon.*

n	m	# iterations	T[s]
24	25	77	0.45
47	49	47	0.63
92	97	27	1.20
183	193	19	3.40

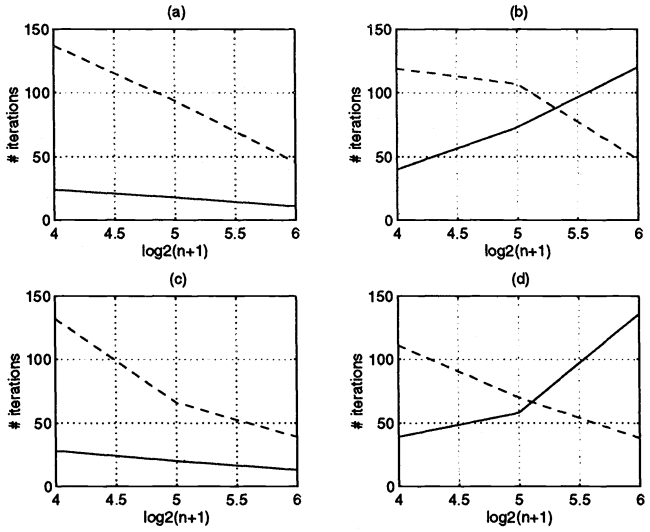


FIG. 8. *Number of iterations obtained for convergence for* (a) $s_1 = s_2 = 0.0$, (b) $s_1 = 0.5$, $s_2 = 0.0$, (c) $s_1 = 0.0$, $s_2 = 0.5$, *and* (d) $s_1 = s_2 = 0.5$. *The solid line represents the preconditioned system and the dashed line the original system.*

(iii) The fact that the number of iterations decreases with increasing problem size, is even more pronounced here compared to [14], [11].

Henceforth we will only consider this nonuniform domain decomposition. The subdomain solves are accomplished through solving (23) with $\hat{A}_{11}(\bullet, D)$ for $s_2 = 0$. For $s_2 > 0$ we solve (7) with an iterative method and preconditioner $\hat{A}_{11}(\bullet, D)$.

Now that we have an efficient method to solve (4) due to the nonuniform decomposition of the domain, we will study the influence of the preconditioner on the Schur complement system. The parameters used are $p = 4$, $\kappa_1 = 100$, $m_q = (n+1)/2 - 1$, $q = 1, \ldots, p-1$, and $m_p = 1$.

From Figs. 8 and 9 we draw the following conclusions.

(i) For $s_1 = 0$ the preconditioner reduces both the number of iterations and the cpu time.

(ii) For $s_1 = 0.5$ the preconditioner works well for small problems, but for larger problems, the iteration on the original system is preferable.

(iii) The iteration on the original system is independent of the stretching in the grid. Note however that in all four cases, the stretching is quite moderate.

(iv) For $s_2 = 0$ we can solve (23) instead of (7) which reduces the cpu time considerably.

Now we have found out how the method depends on the problem size and the stretching in the grid. In the following we consider $s_1 = s_2 = 0$. The next thing we

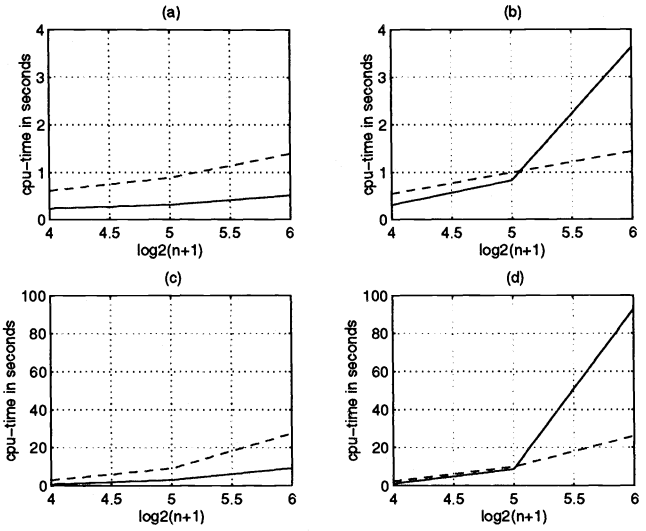


FIG. 9. Cpu *time in seconds obtained for convergence for* (a) $s_1 = s_2 = 0.0$, (b) $s_1 = 0.5$, $s_2 = 0.0$, (c) $s_1 = 0.0$, $s_2 = 0.5$, *and* (d) $s_1 = s_2 = 0.5$. *The solid line represents the preconditioned system and the dashed line the original system. The timings are from an Intel Paragon.*

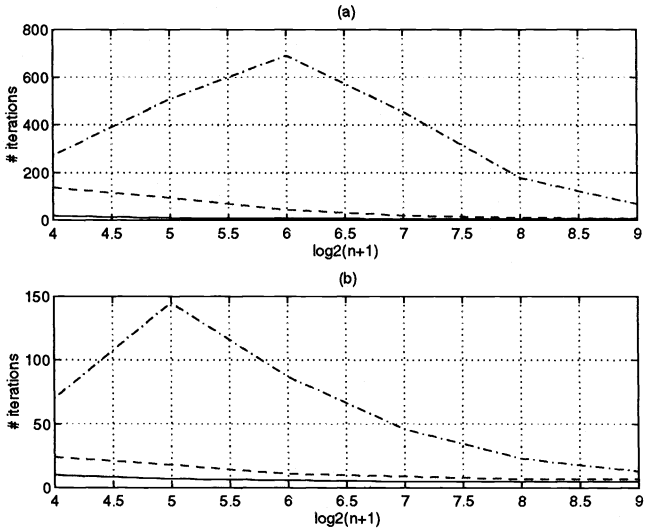


FIG. 10. *Number of iterations for* $\kappa_1 = 10$ *solid line,* $\kappa_1 = 100$ *dashed line, and* $\kappa_1 = 1000$ *dashed-dotted line.* (a) *presents the original system and* (b) *the preconditioned system.*

will study is the dependency on the time step, i.e., we vary κ_1 and κ_2. Here we have used $p = 4$ and $n + 1 = 2 \cdot ((m - p)/(p - 1) + 1)$.

In Fig. 10 we see that the number of iterations decreases with increasing problem size with very few deviations. We also see that the iteration count increases with increasing time step, but the increase is less for larger problem sizes. Note that the shapes of the curves from the original system and the preconditioned system are similar.

Finally we have examined the behavior when we increase the number of subdo-

TABLE 5
Number of iterations and cpu *time obtained for convergence on an Intel Paragon.*

$\hat{C}$	n	m	p	# iterations	T[s]
I	255	385	4	11	7.18
I	255	449	8	35	9.26
I	255	481	16	147	17.3
$\hat{C}(D)$	255	385	4	7	5.21
$\hat{C}(D)$	255	449	8	9	3.21
$\hat{C}(D)$	255	481	16	12	2.14

TABLE 6
Number of iterations, cpu *time, arithmetic speed, and rate of efficiency on an Intel Paragon.*

p	n	m	# iterations	T_{it}[s]	T_{tot}[s]	Mflops	$E_{p,t}$	$E_{p,s}$
4	255	385	7	5.21	7.84	9.33	1.00	1.00
8	255	449	9	3.21	4.95	19.1	0.79	1.02
16	255	481	12	2.14	3.17	37.2	0.62	1.00
32	255	497	16	1.68	2.36	65.8	0.42	0.88
64	255	505	26	1.70	2.21	105	0.22	0.70

TABLE 7
Number of iterations, cpu *time, and arithmetic speed on an Intel Paragon for some "large" problems.*

p	n	m	# iterations	T_{it}[s]	T_{tot}[s]	Mflops
32	511	993	11	4.20	5.96	76.8
32	1023	993	10	7.16	10.5	82.1
32	1023	1985	9	13.5	20.0	84.1
64	1023	1009	12	4.67	6.42	151
64	1023	2017	11	8.54	11.9	157
64	2047	2017	9	14.3	20.7	157

mains for the same problem, i.e., the scalability of the method. Note that we must vary the problem size slightly when we vary the number of subdomains due to the fast modified sine transforms. We have used the parameter $\kappa_1 = 100$.

As seen in Table 5 the iteration on the original system clearly isn't very scalable. The preconditioned system on the other hand seems to be quite scalable. We shall further investigate this in §8.3.

8.3. Parallel computing. In this section we consider the solution of the whole problem, i.e., Algorithm SCM. The Schur complement system is preconditioned with preconditioner $\hat{C}(D)$. Here T_{it} is the time to solve the Schur complement system and T_{tot} is the time to solve the whole problem. # iterations refers to the number of iterations to solve the Schur complement system. The parameter used is $\kappa_1 = 100$. To obtain the rate of efficiency, the reference level is set to $p = 4$ as shown in Table 6.

The reason we obtain a rate of efficiency > 1 for the arithmetic speed is that the inefficiency due to the nonuniform decomposition of the domain becomes less pronounced the more subdomains that we have. Note that the poor rate of efficiency in time is partly due to the fact that we are solving larger problems when we have more processors.

We end this section by presenting the results from some "large" problems in Table 7. Again $\kappa_1 = 100$.

9. Conclusions. In this report we have presented a domain decomposition method to solve first-order PDEs in two space dimensions. The method is analyzed theoreti-

cally and numerical results corroborating the theory have been performed.

We have defined preconditioners built on Toeplitz blocks, which have fast solvers based on a fast modified sine transform. Numerical spectra for the preconditioned systems are presented, showing highly clustered spectra.

Due to a nonuniform decomposition of the domain, the method was very successful for constant coefficients in the x_2-direction. The preconditioned system was the one that gave the best results. This method was also highly scalable. However, we have shown theoretically and numerically that the Schur complement system can be solved successfully without preconditioning, for a decomposition with only few subdomains. Hence, this method can be used for more irregular domains. The preconditioned Schur complement method on the other hand requires logically rectangular domains, due to the solver of the preconditioner.

Acknowledgments. I would like to thank my adviser Professor Bertil Gustafsson for helpful support. I also want to express my gratitude to Doctor Sverker Holmgren and Doctor Kurt Otto for sharing their wide experience of iterative methods and preconditioners for nonsymmetric problems with me. Finally I am very grateful to Para//ab at the University of Bergen for letting me use their Paragon. The system administrator Bjarne Herland gave invaluable advice.

REFERENCES

[1] O. AXELSSON, *A restarted version of a generalized preconditioned conjugate gradient method*, Comm. Appl. Numer. Meth., 4 (1988), pp. 521–530.

[2] O. AXELSSON AND G. LINDSKOG, *On the eigenvalue distribution of a class of preconditioning methods*, Numer. Math., 48 (1986), pp. 479–498.

[3] P. BJÖRSTAD AND O. WIDLUND, *Iterative methods for the solution of elliptic problems on regions partitioned into substructures*, SIAM J. Numer. Anal., 23 (1986), pp. 1097–1120.

[4] T. F. CHAN, *Analysis of preconditioners for domain decomposition*, SIAM J. Numer. Anal., 24 (1987), pp. 382–390.

[5] M. DRYJA, *A capacitance matrix method for Dirichlet problem on polygonal region*, Numer. Math., 39 (1982), pp. 51–64.

[6] H. C. ELMAN, *Iterative Methods for Large Sparse Nonsymmetric Systems of Linear Equations*, Ph. D. thesis and Report RR-229, Department of Computer Science, Yale University, New Haven, CT, 1982.

[7] R. W. FREUND, G. H. GOLUB, AND N. M. NACHTIGAL, *Iterative solution of linear systems*, Acta Numerica, 1992, pp. 57–100.

[8] J. GUERRA AND B. GUSTAFSSON, *A semi-implicit method for hyperbolic problems with different time-scales*, SIAM J. Numer. Anal., 23 (1986), pp. 734–749.

[9] B. GUSTAFSSON AND H. STOOR, *Navier–Stokes equations for almost incompressible flow*, SIAM J. Numer. Anal., 28 (1991), pp. 1523–1547.

[10] L. HEMMINGSSON, *A fast modified sine transform for solving block-tridiagonal systems with Toeplitz blocks.*, Numer. Algor., 7 (1994), pp. 375–389.

[11] ———, *Toeplitz preconditioners with block-structure for first-order* PDEs, Numer. Linear Algebra, to appear.

[12] L. HEMMINGSSON AND K. OTTO, *Analysis of semi-Toeplitz preconditioners for first-order* PDEs, SIAM J. Sci. Comput, to appear.

[13] R. W. HOCKNEY AND C. R. JESSHOPE, *Parallel Computers* 2, IOP Publishing Ltd, 1988.

[14] S. HOLMGREN AND K. OTTO, *A Comparison of Preconditioned Iterative Methods for Nonsymmetric Block-Tridiagonal Systems of Equations*, Report No. 123 (revised), Dept. of Scientific Computing, Uppsala University, Uppsala, Sweden, 1990.

[15] ———, *Iterative solution methods and preconditioners for block-tridiagonal systems of equations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 863–886.

[16] D. E. KEYES AND W. D. GROPP, *A comparison of domain decomposition techniques for elliptic partial differential equations and their parallel implementation*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 166–202.

[17] K. OTTO, *Analysis of preconditioners for hyperbolic* PDE, SIAM J. Numer. Anal., to appear.

[18] Y. SAAD, *Krylov subspace methods for solving large unsymmetric linear systems*, Math. Comp., 37 (1981), pp. 105–126.

[19] H. A. SCHWARTZ, *Gesammelte mathematische Abhandlungen*, Springer-Verlag, Berlin, 2, 1890, pp. 133–134.

[20] H. STOOR, *Numerical Solution of the Navier–Stokes Equations for Small Mach Numbers*, Ph.D. thesis, Department of Scientific Computing, Uppsala University, Uppsala, Sweden, 1990.

[21] H. A. VAN DER VORST, *Preconditioning by Incomplete Decompositions*, Ph.D. thesis, Rijksuniversiteit Utrecht, Utrecht, The Netherlands, 1982.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, pp. 1268–1286, October 1995

017

SOME PROPERTIES OF FULLY SEMIMONOTONE, Q_0-MATRICES*

G. S. R. MURTHY† AND T. PARTHASARATHY‡

Abstract. Stone [*Ph.D. thesis, Dept. of Operations Research, Stanford University, Stanford, CA*, 1981] proved that within the class of Q_0-matrices, the U-matrices are P_0-matrices and conjectured that the same must be true for fully semimonotone (E_0^f) matrices. In this paper we show that this conjecture is true for matrices of order up to 4×4 and partially resolve it for higher order matrices. This is done by establishing the result that if A is in $E_0^f \cap Q_0$ and if every proper principal minor of A is nonnegative, then A is a P_0-matrix. Using this key result we settle the conjecture for a number of special cases of matrices of general order. These special cases include E_0^f-matrices which are either symmetric or nonnegative or copositive-plus or Z-matrices or E-matrices. Also the conjecture is established for 5×5 matrices with all diagonal entries positive. While trying to settle the conjecture, we obtained a number of results on Q_0-matrices. The main among these are characterizations of nonnegative Q_0-matrices and symmetric semimonotone Q_0-matrices; results providing sufficient conditions under which, principal submatrices of order $(n-1)$ of a $n \times n$ Q_0-matrix are also in Q_0.

Key words. linear complementarity problem, matrix classes, Lemke's algorithm

AMS subject classification. 90C33

1. Introduction.

For any $A \in \boldsymbol{R}^{n \times n}$ and any $q \in \boldsymbol{R}^n$, define the sets $F(q, A)$ and $S(q, A)$ as:

$$F(q, A) = \{ z \in \boldsymbol{R}^n_+ : Az + q \geq 0 \}, \tag{1.1}$$

$$S(q, A) = \{ z \in F(q, A) : (Az + q)^t z = 0 \}. \tag{1.2}$$

The linear complementarity problem (LCP) with data A and q is to find an element of $S(q, A)$. This problem is denoted by (q, A). For $z \in F(q, A)$, let $w = Az + q$. Then $w \geq 0$. Note that if $z \in S(q, A)$, then $w_i z_i = 0$ for each i. Thus, if $z \in S(q, A)$, then w and z are *complementary* to each other. Any z of $S(q, A)$ (or equivalently the pair (w, z)) is called a solution of (q, A).

Cottle and Stone [5] introduced the classes of $\boldsymbol{U}$-matrices and fully semimonotone ($\boldsymbol{E}_0^f$) matrices and studied their properties (see Table 1 for notations and definitions). If A is a $\boldsymbol{U}$-matrix, then (q, A) has a unique solution for every q in the interior of $K(A)$. If A is in $\boldsymbol{E}_0^f$, then (q, A) has a unique solution for every q in the interior of any complementary cone. It was observed [5] that $\boldsymbol{P} \subseteq \boldsymbol{U} \subseteq \boldsymbol{E}_0^f$. Stone [21] showed that $\boldsymbol{U} \cap \boldsymbol{Q}_0 \subseteq \boldsymbol{P}_0$ and raised the following conjecture.

CONJECTURE 1.1. *Every $\boldsymbol{E}_0^f \cap \boldsymbol{Q}_0$-matrix is a $\boldsymbol{P}_0$-matrix.*

In this paper we show that this conjecture is true for matrices of order up to 4×4 and partially resolve it for higher order matrices. We first establish that if A is in $\boldsymbol{E}_0^f \cap \boldsymbol{Q}_0$ and if every proper principal minor of A is nonnegative, then A is a $\boldsymbol{P}_0$-matrix. An important corollary of this key result is that $\bar{\boldsymbol{Q}}_0 \cap \boldsymbol{E}_0^f \subseteq \boldsymbol{P}_0$, where $\bar{\boldsymbol{Q}}_0$ is the class of completely $\boldsymbol{Q}_0$-matrices. As a consequence of this corollary, we establish Conjecture 1.1 for a number of special cases. These special cases include $\boldsymbol{E}_0^f$-matrices that are either symmetric or nonnegative or copositive-plus or $\boldsymbol{Z}$-matrices or $\boldsymbol{E}$-matrices. Using the above result it is shown that interior of $\boldsymbol{R}^{n \times n} \cap \boldsymbol{E}_0^f$ is

* Received by the editors August 12, 1993; accepted for publication (in revised form) by R. Cottle November 23, 1994.

† Indian Statistical Institute, 110 Nelson Manickam Road, Madras 600029, India.

‡ Indian Statistical Institute, 7 S.J.S. Sansanwal Marg., New Delhi-110016, India (tps@isid.ernet.in).

same as $\boldsymbol{R}^{n\times n} \cap \boldsymbol{P}$. In the sequel we introduce a subclass of $\boldsymbol{E}_0^f$, the class of fully copositive matrices ($\boldsymbol{C}_0^f$), and prove that symmetric $\boldsymbol{E}_0^f$-matrices are fully copositive. Furthermore, it is shown that if A is a $\boldsymbol{C}_0^f$-matrix with at most one zero diagonal entry, then A is a $\boldsymbol{P}_0$-matrix.

Aganagic and Cottle [2] characterized $\boldsymbol{Q}_0$-matrices with nonnegative principal minors and established that Lemke's algorithm—with a suitable apparatus to resolve degeneracy—processes (q, A) whenever A is in $\boldsymbol{P}_0 \cap \boldsymbol{Q}_0$. Appealing to this result, we conclude that Lemke's algorithm processes (q, A) for all those A for which we establish Conjecture 1.1.

Our other main results of this paper are on properties of $\boldsymbol{Q}_0$-matrices. Murty [15] gave a characterization of nonnegative $\boldsymbol{Q}$-matrices. We present a characterization of nonnegative $\boldsymbol{Q}_0$-matrices. Jeter and Pye [8], [9] studied the connections of $\boldsymbol{Q}$-matrices with their principal submatrices. While extending their result to $\boldsymbol{Q}_0$-matrices, we derive conditions, in terms of principal pivotal transforms (PPTs), for an $n \times n$ $\boldsymbol{Q}_0$-matrix to have all its principal submatrices of order $(n-1)$ to be in $\boldsymbol{Q}_0$. As applications of these results, we show that matrices that are either nonnegative or symmetric semimonotone are $\boldsymbol{Q}_0$ if, and only if (iff) they are $\bar{\boldsymbol{Q}}_0$. The study of $\bar{\boldsymbol{Q}}_0$-matrices in general is a complex problem (see [3], [6]).

In §2, we present results on $\boldsymbol{Q}_0$-matrices. Section 3 is devoted to results on $\boldsymbol{E}_0^f$-matrices of general order. In §4, we establish Conjecture 1.1 for n less than or equal to 4 and prove that it is valid for 5×5 and 6×6 matrices with some additional assumptions.

With every matrix $A \in \boldsymbol{R}^{n\times n}$ there is a real number associated with it, called the (minimax) value of A and is denoted by $v(A)$. It is well known that $v(A)$ is positive (nonnegative) iff there exists a nonzero nonnegative z such that Az is positive (nonnegative) (see [22]).

Remark 1.2. It is easy to see from the definition that every $\boldsymbol{E}_0$-matrix must have all its diagonal entries nonnegative. Also if A is in $\boldsymbol{E}_0^f$, then A and all its PPTs must have all their diagonal entries nonnegative. In addition, if A is also in $\boldsymbol{N}_0$, then A must have all its diagonal entries equal to zero.

Remark 1.3. A matrix A is in $\boldsymbol{E}_0$ iff all principal submatrices of A have value nonnegative. If A is in $\boldsymbol{E}_0$, then A^t is also in $\boldsymbol{E}_0$. Furthermore, if A is a symmetric $\boldsymbol{E}_0$-matrix, then it is copositive (see [4, pp. 187,188]).

Remark 1.4. If A is a $\boldsymbol{Q}_0$-matrix and $v(A)$ is positive, then A is in $\boldsymbol{Q}$.

Remark 1.5. If A belongs to any of $\boldsymbol{E}_0$, $\boldsymbol{E}_0^f$, $\boldsymbol{P}_0$, $\boldsymbol{N}_0$, $\boldsymbol{C}_0$, and $\boldsymbol{C}_0^f$, then every principal submatrix of A is also in the same class. If A belongs to any of $\boldsymbol{E}_0$, $\boldsymbol{E}_0^f$, $\boldsymbol{P}_0$, $\boldsymbol{N}_0$, $\boldsymbol{C}_0$, $\boldsymbol{C}_0^f$, $\boldsymbol{Q}$, and $\boldsymbol{Q}_0$, then every principal rearrangement of A is also in the same class. Furthermore, if A belongs to any of $\boldsymbol{E}_0^f$, $\boldsymbol{P}_0$, $\boldsymbol{Q}$, $\boldsymbol{Q}_0$, then every PPT of A is also in the same class.

2. Some results on Q_0-matrices.

The following result is due to Jeter and Pye [9].

THEOREM 2.1. *Suppose $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{Q}$ and $\alpha = \bar{n} \setminus \{i\}$. Then either $A_{\alpha\alpha} \in \boldsymbol{Q}$ or there exists a $u \in S(e_i, A)$ such that $A_{i.}u = -1$, where e_i is ith column of the identity matrix of order n.*

We extend the above theorem to Q_0-matrices.

TABLE 1. *Notations and definitions.*

Symbol	Definition
$\bar{n}$	the set $\{1, 2, \ldots, n\}$, n is any positive integer
n^*	collection of all nonempty subsets of $\bar{n}$
$\|\alpha\|$	cardinality of the set α
α, β, γ	denote the subsets of $\bar{n}$
$\bar{\alpha}$	complement of the set α relative to $\bar{n}$
$\alpha \setminus \beta$	the set $\alpha \cap \bar{\beta}$
R^n	n-dimensional space of reals
$R^{m\times n}$	the space of $m \times n$ real matrices
R^n_+	the nonnegative orthant of R^n
$A = (a_{ij})$	a matrix with a_{ij}'s as its entries. We denote the entries of a matrix by the corresponding lower case letters, for example entries of B are denoted by b_{ij}
I	the identity matrix
det A	determinant of matrix A
$A_{\alpha\beta}$	the submatrix of A obtained by dropping rows and columns of A corresponding $\bar{\alpha}$ and $\bar{\beta}$
$A_{i\beta}$	stands for $A_{\{i\}\beta}$, $i \in \bar{n}$
$A_{\alpha.}$	stands for $A_{\alpha\bar{n}}$, $A \in R^{m\times n}$
$A_{.\beta}$	stands for $A_{\bar{m}\beta}$, $A \in R^{m\times n}$
$A_{i.}$	ith row of A
$A_{.j}$	jth column of A
$v(A)$	(minimax) value of A
supp(z)	the index set $\{i \in \bar{n} : z_i \neq 0\}$, where $z \in R^n$
$(A/A_{\alpha\alpha})$	$A_{\bar{\alpha}\bar{\alpha}} - A_{\bar{\alpha}\alpha}(A_{\alpha\alpha})^{-1}A_{\alpha\bar{\alpha}}$.
$\wp_\alpha(A)$	if det $A_{\alpha\alpha} \neq 0$, then $M = \wp_\alpha(A)$ is called the principal pivotal transform of A with respect to α, where $M_{\alpha\alpha} = (A_{\alpha\alpha})^{-1}$, $M_{\alpha\bar{\alpha}} = -(A_{\alpha\alpha})^{-1}A_{\alpha\bar{\alpha}}$, $M_{\bar{\alpha}\alpha} = A_{\bar{\alpha}\alpha}(A_{\alpha\alpha})^{-1}$, $M_{\bar{\alpha}\bar{\alpha}} = (A/A_{\alpha\alpha})$
$C_A(\alpha)$	the complementary matrix with respect to $\alpha \subseteq \bar{n}$, where $C_A(\alpha)_{.j} = -A_{.j}$ if $j \in \alpha$ and $C_A(\alpha)_{.j} = I_{.j}$ otherwise
pos $C_A(\alpha)$	$\{C_A(\alpha)z : z \in R^n_+\}$, the complementary cone with respect to α Columns of $C_A(\alpha)$ are called generators of pos $C_A(\alpha)$
$K(A)$	the set $\{q \in R^n : S(q, A) \neq \phi\}$
C_0	$\cup_n\{A \in R^{n\times n} : x^tAx \geq 0\ \forall x \in R^n_+\}$
C_0^+	$\cup_n\{A \in R^{n\times n} \cap C_0 : [x^tAx = 0,\ x \in R^n_+] \Rightarrow (A + A^t)x = 0\}$
E	$\cup_n\{A \in R^{n\times n} : \forall\ 0 \neq x \in R^n_+\ \exists\ k \in \bar{n} \ni x_k > 0$ and $(Ax)_k > 0\}$
E_0	$\cup_n\{A \in R^{n\times n} : \forall\ 0 \neq x \in R^n_+\ \exists\ k \in \bar{n} \ni x_k > 0$ and $(Ax)_k \geq 0\}$
E_0^f	$\cup_n\{A \in R^{n\times n} : \forall \alpha \in n^*, \det A_{\alpha\alpha} \neq 0 \Rightarrow \wp_\alpha(A) \in E_0\}$
N	$\cup_n\{A \in R^{n\times n} : \forall \alpha \in n^*, \det A_{\alpha\alpha} < 0\}$
N_0	$\cup_n\{A \in R^{n\times n} : \forall \alpha \in n^*, \det A_{\alpha\alpha} \leq 0\}$
P	$\cup_n\{A \in R^{n\times n} : \forall \alpha \in n^*, \det A_{\alpha\alpha} > 0\}$
P_0	$\cup_n\{A \in R^{n\times n} : \forall \alpha \in n^*, \det A_{\alpha\alpha} \geq 0\}$
Q	$\cup_n\{A \in R^{n\times n} : S(q, A) \neq \phi\ \forall\ q \in R^n\}$
Q_0	$\cup_n\{A \in R^{n\times n} : \forall q \in R^n,\ F(q, A) \neq \phi \Rightarrow S(q, A) \neq \phi\}$
$\bar{Q}$	$\cup_n\{A \in R^{n\times n} : \forall \alpha \in n^*, A_{\alpha\alpha} \in Q\}$
$\bar{Q}_0$	$\cup_n\{A \in R^{n\times n} : \forall \alpha \in n^*, A_{\alpha\alpha} \in Q_0\}$
R_0	$\cup_n\{A \in R^{n\times n} : (0, A)$ has a unique solution$\}$
U	$\cup_n\{A \in R^{n\times n} : \|S(q, A)\| = 1\ \forall\ q \in$ interior of $K(A)\}$

THEOREM 2.2. *Suppose* $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{Q}_0$ *and* $\alpha = \bar{n} \setminus \{i\}$. *Then either* $A_{\alpha\alpha} \in \boldsymbol{Q}_0$ *or there exists a* $u \in S(e_i, A)$ *such that* $A_{i.}u = -1$.

Proof. Without loss of generality take $i = n$. Suppose $A_{\alpha\alpha} \notin \boldsymbol{Q}_0$. Then there exists a $\bar{q} \in R^{n-1}$ such that $F(\bar{q}, A_{\alpha\alpha}) \neq \phi$ and $S(\bar{q}, A_{\alpha\alpha}) = \phi$. For each positive integer k, define $q^k_\alpha = \bar{q}/k$, $q^k_n = 1$. Observe that $F(q^k_\alpha, A_{\alpha\alpha}) \neq \phi$ for every $k \geq 1$. This is because, for any $z_\alpha \in F(q_\alpha, A_{\alpha\alpha})$, $\frac{1}{k}z_\alpha \in F(q^k_\alpha, A_{\alpha\alpha})$ for all $k \geq 1$. Then $F(q^k, A) \neq \phi$ for all positive integers k sufficiently large. As $A \in \boldsymbol{Q}_0$, $S(q^k, A) \neq \phi$ for all k sufficiently large. So each q^k lies in a complementary cone for

all k sufficiently large. Since there are only finitely many complementary cones, there is a complementary cone containing a subsequence $q^{k_1}, q^{k_2}, q^{k_3}, \ldots$ of $\{q^k\}$. Since q^k converges to e_n and as the complementary cones are all closed, e_n also lies in a complementary cone containing the subsequence $\{q^{k_i}\}$. Also note that for each $k \geq 1$, if $z^k \in S(q^k, A)$, then $z_n^k > 0$ as otherwise it would imply that $S(\bar{q}, A_{\alpha\alpha}) \neq \phi$. Thus, for all k sufficiently large, q^k lies in a complementary cone with $A_{.n}$ as one of its generators. Therefore, e_n lies in a complementary cone with one of its generators as $A_{.n}$ and there exists a $u \in S(e_n, A)$ such that $(Au)_n = -1$. □

It can be proved that if $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{Q}$ and $A_{i.} \geq 0$ for some i, then $A_{\alpha\alpha} \in \boldsymbol{Q}$, where $\alpha = \bar{n} \setminus \{i\}$. The following is an analog of this for $\boldsymbol{Q}_0$-matrices.

COROLLARY 2.3. *Suppose $A \in R^{n\times n} \cap Q_0$ and that $A_{i.} \geq 0$ for some i. Then $A_{\alpha\alpha} \in Q_0$, where $\alpha = \bar{n} \setminus \{i\}$.*

Proof. The proof follows from Theorem 2.2 and the fact that $(Au)_i \geq 0$ for every $u \in \boldsymbol{R}_+^n$. □

THEOREM 2.4. *Suppose $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{Q}_0$. Assume that A is nonnegative. Then A is a $\boldsymbol{Q}_0$-matrix iff A is a $\bar{\boldsymbol{Q}}_0$-matrix.*

Proof. If A is in $\bar{\boldsymbol{Q}}_0$, then obviously A is in $\boldsymbol{Q}_0$. Conversely, assume that A is in $\boldsymbol{Q}_0$. Since every row of A is nonnegative, by Corollary 2.3 every principal submatrix of A of order $(n-1)$ is also a nonnegative $\boldsymbol{Q}_0$-matrix. Repeating this argument with principal submatrices, we conclude that A is in $\bar{\boldsymbol{Q}}_0$. □

THEOREM 2.5. *Suppose $A \in \boldsymbol{R}^{n\times n}$ is a nonnegative matrix where $n \geq 2$. Then A belongs to $\boldsymbol{Q}_0$ iff the following implication is valid:*

$$\text{for every } i \in \bar{n}, \quad A_{i.} \neq 0 \Rightarrow a_{ii} > 0.$$

Proof. (Necessity) We shall prove this by induction on n. It is easy to check this when $n = 2$. Assume that the result is true for all $(n-1) \times (n-1)$ matrices. Let $A \in \boldsymbol{R}^{n\times n}$, $n \geq 3$, be a nonnegative $\boldsymbol{Q}_0$-matrix. Suppose $A_{i.} \neq 0$ for some $i \in \bar{n}$. Let j be such that $a_{ij} > 0$. If $j = i$ we are done. Suppose $j \neq i$. Choose any $k \in \overline{\{i,j\}}$ (we can do this as $n \geq 3$). Let $\alpha = \overline{\{k\}}$. By Theorem 2.4, $A_{\alpha\alpha} \in \boldsymbol{Q}_0$. By choice of k, $A_{i\alpha} \neq 0$. By induction, we must have $a_{ii} > 0$.

(Sufficiency) Assume that $a_{ii} > 0$ for every i such that $A_{i.} \neq 0$.
Let $\alpha = \{i \in \bar{n} : A_{i.} \neq 0\}$. Then

$$\begin{bmatrix} A_{\alpha\alpha} & A_{\alpha\bar{\alpha}} \\ A_{\bar{\alpha}\alpha} & A_{\bar{\alpha}\bar{\alpha}} \end{bmatrix} = \begin{bmatrix} A_{\alpha\alpha} & A_{\alpha\bar{\alpha}} \\ 0 & 0 \end{bmatrix}.$$

Suppose $q \in \boldsymbol{R}^n$ is such that $F(q, A) \neq \phi$. Then we must have $q_{\bar{\alpha}} \geq 0$. Since $A_{\alpha\alpha}$ is nonnegative with all its diagonal entries positive, $A_{\alpha\alpha} \in \boldsymbol{Q}$ (see [15], [16]). Let $z_\alpha \in S(q_\alpha, A_{\alpha\alpha})$. Then $(z_\alpha^t, 0^t)^t \in S(q, A)$. As q was arbitrary, $A \in \boldsymbol{Q}_0$. □

COROLLARY 2.6. *Suppose $A \in \boldsymbol{R}^{n\times n}$ is a nonnegative nonsingular $\boldsymbol{Q}_0$-matrix. Then A is in $\bar{\boldsymbol{Q}}$.*

Proof. Since A is nonsingular $\boldsymbol{Q}_0$-matrix, we must have $a_{ii} > 0$ for every $i \in \bar{n}$. It follows that A is in $\bar{\boldsymbol{Q}}$. □

The following examples illustrate the application of the above theorem.

Example 2.7. Let

$$A = \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 3 & 2 & 1 & -1 \\ 2 & 2 & -1 & 0 \\ -1 & 0 & 0 & 1 \\ -1 & -2 & 1 & 0 \end{bmatrix}.$$

Note that $A_{3.} \geq 0$. If $A \in \boldsymbol{Q}_0$, then $A_{\alpha\alpha} \in \boldsymbol{Q}_0$, where $\alpha = \{1,2,4\}$. But by Theorem 2.5, $A \notin \boldsymbol{Q}_0$. Hence $A \notin \boldsymbol{Q}_0$.

Let $\alpha = \{3,4\}$ and $M = \wp_\alpha(B)$. Then

$$M = \begin{bmatrix} 3 & 4 & -1 & 1 \\ 2 & 0 & 0 & -1 \\ 1 & 2 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}.$$

Suppose $M \in \boldsymbol{Q}_0$. Since last two rows of M are nonnegative, $M_{\bar{\alpha}\bar{\alpha}} \in \boldsymbol{Q}_0$. But by Theorem 2.5, $M_{\bar{\alpha}\bar{\alpha}} \notin \boldsymbol{Q}_0$. It follows that $M \notin \boldsymbol{Q}_0$.

THEOREM 2.8. *Suppose $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{E}_0 \cap \boldsymbol{Q}_0$. If A is symmetric, then A is in $\bar{\boldsymbol{Q}}_0$.*

Proof. We will show this by induction on n. If $n = 1$, there is nothing to prove. Assume that the result is true for all real square matrices of order less than or equal to $n-1$, $n > 1$. Suppose $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{E}_0 \cap \boldsymbol{Q}_0$ and A is symmetric. Let α be any subset of $\bar{n}$ such that $|\alpha| = n-1$. Without loss of generality, we may assume that $\alpha = \overline{\{n\}}$. Suppose $A_{\alpha\alpha} \notin \boldsymbol{Q}_0$. Then by Theorem 2.2, there exists a $u \in \boldsymbol{R}^n_+$ such that $Au + e_n \geq 0$, $u^tAu + u_n = 0$, and $(Au)_n = -1$. Since A is symmetric $\boldsymbol{E}_0$-matrix, A is copositive. As $u \geq 0$ and $u^tAu + u_n = 0$, $u^tAu = 0$ and $u_n = 0$. Since A is symmetric copositive matrix, it follows that $(Au) \geq 0$ (see [20, Lemma 3.1]). This contradicts $(Au)_n = -1$. Hence $A_{\alpha\alpha} \in \boldsymbol{Q}_0$. As α was arbitrary, it follows that every $(n-1) \times (n-1)$ principal submatrix of A is in $\boldsymbol{Q}_0$. By induction, it follows that $A \in \bar{\boldsymbol{Q}}_0$. □

Pang [17] proved that if A is a $\boldsymbol{E}_0 \cap \boldsymbol{Q}$-matrix, then every nontrivial solution of $(0, A)$ must have at least two nonzero coordinates. Paraphrasing, if A is a $\boldsymbol{E}_0 \cap \boldsymbol{Q}$-matrix, then A cannot have a diagonal entry zero and all other entries in the corresponding column nonnegative. We have the following results for $\boldsymbol{Q}_0$-matrices in this direction.

THEOREM 2.9 *Suppose $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{E}_0 \cap \boldsymbol{Q}_0$. Assume that for some i_0, $j_0 \in \bar{n}$, $a_{i_0i_0} = 0$ and $a_{i_0j_0} > 0$. Then there exists a $k \in \bar{n}$ such that $a_{ki_0} < 0$.*

Proof. Since $a_{i_0j_0}$ is positive, we can choose a $q \in \boldsymbol{R}^n$ such that $q_{i_0} < 0$, $q_j > 0$ for all $j \neq i_0$ and $F(q, A) \neq \phi$. Since $A \in \boldsymbol{Q}_0$, $S(q, A) \neq \phi$. Let $z \in S(q, A)$ and let $\alpha = \text{supp}(z)$. Let $\beta = \alpha \setminus \{i_0\}$. Since $a_{i_0i_0} = 0$ and $q_{i_0} < 0$, $\beta \neq \emptyset$. Since z_β is positive, we have

$$0 = A_{\beta i_0} z_{i_0} + A_{\beta\beta} z_\beta + q_\beta.$$

Note that $q_\beta > 0$. If $A_{.i_0} \geq 0$, then

$$A_{\beta\beta} z_\beta = -q_\beta - z_{i_0} A_{\beta i_0} < 0,$$

which in turn implies that $v(A^t_{\beta\beta}) < 0$. This is not possible as $A \in \boldsymbol{E}_0$. Therefore, $A_{.i_0}$ must contain a negative entry. There exists a $k \in \bar{n}$ such that $a_{ki_0} < 0$. □

COROLLARY 2.10. *Suppose $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{E}_0 \cap \boldsymbol{Q}_0$. Assume that every row of A contains a positive entry. Then every nontrivial solution of $(0, A)$ contains at least two positive coordinates.*

THEOREM 2.11. *Suppose $A \in \boldsymbol{R}^{n\times n}$, where $n \geq 3$. Assume that $a_{11} = a_{22} = 0$, $a_{12} > 0$ and $a_{21} > 0$. Let $g = (a_{13}, a_{14}, \ldots, a_{1n})$ and $h = (a_{23}, a_{24}, \ldots, a_{2n})$. Assume that A satisfies any of the following conditions:*

(a) *$g \leq 0$ and $h \geq 0$.*

(b) $g \geq 0$ *and* $h \leq 0$.

Then A is not a $\boldsymbol{Q}_0$-matrix.

Proof. Suppose A satisfies condition (a). Since $a_{12} > 0$, there exists a $q \in \boldsymbol{R}^n$ such that $q_1 < 0$, $q_j > 0$ for every $j \neq 1$, and $F(q, A) \neq \phi$. Hypothesis implies that for every $z \in F(q, A)$, $z_2 > 0$. Since $A_{2.} \geq 0$, $w_2 = (Az)_2 + q_2 > 0$. Thus (q, A) cannot have a complementary solution. Therefore, A is not a $\boldsymbol{Q}_0$-matrix. Similarly, we can show that if A satisfies (b), then A is not in $\boldsymbol{Q}_0$. □

In the above theorem, there is nothing special about the indices 1 and 2. The theorem is valid even when they are replaced by any other indices.

THEOREM 2.12. *Suppose $A \in \boldsymbol{R}^{n \times n}$. Let $k \in \bar{n}$ and let $\alpha = \overline{\{k\}}$. Assume that A satisfies the following conditions:*

(a) $A_{\alpha\alpha} \leq 0$,

(b) $a_{i_0 k} > 0$ *for some* $i_0 \in \alpha$, *and*

(c) $A_{k.} \geq 0$.

Then A does not belong to $\boldsymbol{Q}_0$.

Proof. Note that the assumptions of the theorem imply $i_0 \neq k$. Since $a_{i_0 k} > 0$, there exists a $q \in \boldsymbol{R}^n$ such that $q_{i_0} < 0$, $q_j > 0$ for every $j \in \bar{n}$, $j \neq i_0$ and $F(q, A) \neq \phi$. Let $z \in F(q, A)$. Since $A_{\alpha\alpha} \leq 0$, we must have $z_k > 0$. Note that as $k \neq i_0$, $q_k > 0$ and $w_k = (Az)_k + q_k > 0$. This implies (q, A) cannot have a solution. Therefore, A is not in $\boldsymbol{Q}_0$. □

For any $A \in \boldsymbol{R}^{n \times n}$, define the sign pattern matrix of A, denoted by $SP(A)$, as a matrix of the same order with entries as either the corresponding entries of A or their possible signs. For example, if

$$A = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 0 & 2 \end{bmatrix},$$

then

$$\begin{bmatrix} + & - & 0 \\ -1 & \oplus & \star \end{bmatrix}, \begin{bmatrix} 1 & \ominus & \oplus \\ - & 0 & + \end{bmatrix}, \text{ and } \begin{bmatrix} \oplus & - & 0 \\ \star & 0 & 2 \end{bmatrix}$$

are all sign pattern matrices of A (here $\oplus$ stands for nonnegative, $\ominus$ for nonpositive, and $\star$ for the corresponding entry).

THEOREM 2.13. *Suppose $A \in \boldsymbol{R}^{3 \times 3} \cap \boldsymbol{Q}_0$. Then $SP(A)$ cannot be equal to any of the following:*

$$\text{(a)} \begin{bmatrix} \oplus & \oplus & \oplus \\ \star & \star & \star \\ + & \ominus & 0 \end{bmatrix} \quad \text{(b)} \begin{bmatrix} \oplus & - & \oplus \\ \oplus & \oplus & \oplus \\ + & \star & 0 \end{bmatrix}.$$

Proof. Suppose $SP(A)$ is given by (a). As $a_{31} > 0$, there exists a $q \in \boldsymbol{R}^3$ such that $SP(q) = (+, +, -)^t$ and $F(q, A) \neq \phi$. Let $z \in F(q, A)$. Then $z_1 > 0$. Since $A_{1.} \geq 0$ and $q_1 > 0$, $w_1 = (Az)_1 + q_1 > 0$. This implies (q, A) cannot have a solution. This contradicts the hypothesis. Therefore, $SP(A)$ cannot be equal to the sign pattern given by (a).

Suppose $SP(A)$ is as in (b). Note that $A_{2.} \geq 0$. By Corollary 2.3, $A_{\alpha\alpha} \in \boldsymbol{Q}_0$ where $\alpha = \{1, 3\}$. Observe that

$$SP(A_{\alpha\alpha}) = \begin{bmatrix} \oplus & \oplus \\ + & 0 \end{bmatrix}.$$

By Theorem 2.5, $A_{\alpha\alpha} \notin \boldsymbol{Q}_0$. From this contradiction it follows that A cannot have the sign pattern given by (b). □

DEFINITION 2.14. *Let $A \in \boldsymbol{R}^{n\times n}$ and $q \in \boldsymbol{R}^n$. A complementary matrix B is called a complementary basis of (q, A) provided B is nonsingular and q belongs to* pos B.

A linear complementarity problem may have a complementary solution without having a complementary basis. Consider the following example due to Mohan [11].

Example 2.15. Let

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \quad \text{and} \quad q = \begin{bmatrix} 0 \\ 0 \\ -1 \\ -1 \end{bmatrix}.$$

It is clear that (q, A) does not have a complementary basis even though it has a complementary solution, namely, $z = (1, 1, 0, 0)^t$.

Thus, in general, the existence of a complementary solution need not necessarily imply the existence of a complementary basis. However, the existence of complementary bases can be asserted in some special cases.

DEFINITION 2.16. *Let $A \in \boldsymbol{R}^{n\times n}$. Say that A has* property (D) *if the following implication is valid for every $\alpha \in n^*$:*

$$\det A_{\alpha\alpha} = 0 \quad \Rightarrow \quad \textit{columns of } A_{.\alpha} \textit{ are linearly dependent.}$$

Remark 2.17. The class of matrices having property (D) is rather large. Obviously it contains column adequate matrices and all nondegenerate matrices (see [4], [7] for definitions).

THEOREM 2.18. *Suppose $A \in \boldsymbol{R}^{n\times n}$. Assume that A has property* (D). *If $q \in \boldsymbol{R}^n$ is such that (q, A) has a solution, then (q, A) has a complementary basis.*

Proof. Since $S(q, A) \neq \phi$, choose $z \in S(q, A)$. Let $\alpha = \text{supp}(z)$. If $\alpha = \phi$, then $z = 0$ is a solution of (q, A) and I is a complementary basis. Suppose $\alpha \neq \phi$. Without loss of generality, assume $\alpha = \{1, 2, \ldots, k\}$. If $\det A_{\alpha\alpha} \neq 0$, then $C_A(\alpha)$ is a complementary basis for (q, A). Suppose $\det A_{\alpha\alpha} = 0$. Then by property (D), there exists a $\bar{d} \in R^{|\alpha|}$ such that

$$A_{.\alpha}\bar{d} = 0, \quad \bar{d} \neq 0.$$

Let d be such that $d_\alpha = \bar{d}$ and $d_{\bar{\alpha}} = 0$. Since $z_\alpha > 0$, we can choose a real number λ such that $z_\alpha - \lambda d_\alpha \geq 0$ and at least one coordinate of $z_\alpha - \lambda d_\alpha$ is equal to zero. Define $\bar{z} \in \boldsymbol{R}^n_+$ by

$$\bar{z}_\alpha = z_\alpha - \lambda d_\alpha \quad \text{and} \quad \bar{z}_{\bar{\alpha}} = 0.$$

Then

$$A\bar{z} + q = A_{.\alpha}z_\alpha - \lambda A_{.\alpha}d_\alpha + q = Az + q \geq 0.$$

Let $w = Az + q$. Since $z_\alpha > 0$, $w_\alpha = 0$ and hence $d^t(Az + q) = 0$.

Since $Ad = A_{.\alpha}d_\alpha = 0$, $\bar{z}^t(A\bar{z} + q) = 0$. Thus, $\bar{z} \in S(q, A)$. Let $\beta = \text{supp}(\bar{z})$. It is clear that $|\beta| < |\alpha|$. If $\det A_{\beta\beta} \neq 0$, then $C_A(\beta)$ is a complementary basis for (q, A). Otherwise we can repeat the above process to get a new solution whose cardinality of its support is strictly less than $|\beta|$. It is clear that in a finite number of steps (at most n) repeating the above process we end up in one of the following situations:

(a) (q, A) has a solution with a complementary basis,
(b) 0 is a solution of (q, A).

In either case (q, A) has a complementary basis. □

THEOREM 2.19. *Suppose* $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{Q}_0$. *Let* $i \in \bar{n}$ *and* $\alpha = \overline{\{i\}}$. *Suppose* A *has property* (D). *Then either* $A_{\alpha\alpha} \in \boldsymbol{Q}_0$ *or there exists a* $\beta \in n^*$ *satisfying:*

(a) $i \in \beta$,
(b) $\det A_{\beta\beta} \neq 0$,
(c) $M_{.i} \leq 0$, *where* $M = \wp_\beta(A)$,
(d) $u \in S(e_i, A)$, *where* $u_\beta = -M_{\beta i}$ *and* $u_{\bar{\beta}} = 0$, *and*
(e) $(Au)_i = -1$.

Proof. Without loss of generality, take $i = n$. Suppose $A_{\alpha\alpha} \notin \boldsymbol{Q}_0$. Then there exists a $\bar{q} \in R^{n-1}$ such that $F(\bar{q}, A_{\alpha\alpha}) \neq \phi$ and $S(\bar{q}, A_{\alpha\alpha}) = \phi$. For each positive integer k, define q^k by $q^k_\alpha = \bar{q}/k$, and $q^k_n = 1$. As $F(\bar{q}, A_{\alpha\alpha}) \neq \phi$, $F(q^k, A) \neq \phi$ for all k sufficiently large. Since $A \in \boldsymbol{Q}_0$, for all k sufficiently large, there exists a solution (w^k, z^k) of (q^k, A). By Theorem 2.17, we may assume, without loss of generality, that (w^k, z^k) corresponds to a complementary basis with $\beta_k = \mathrm{supp}(z^k)$. Then $\det C_A(\beta_k) \neq 0$ for all k sufficiently large. Since $\bar{n}$ has only finitely many subsets, one of its subsets must repeat infinitely often in the sequence $\beta_1, \beta_2, \beta_3, \ldots$. Again, without loss of generality, we can assume $\beta_k = \beta$ for all k sufficiently large. Then $\det C_A(\beta) \neq 0$. Let $M = \wp_\beta(A)$. Note that for each k, $z^k_n > 0$, as otherwise it would imply that $S(\bar{q}, A_{\alpha\alpha}) \neq \phi$. Thus, $n \in \beta$. Hence we have:

$$\begin{bmatrix} I_{|\bar{\beta}|} & -A_{\bar{\beta}\beta} \\ 0 & -A_{\beta\beta} \end{bmatrix} \begin{bmatrix} w^k_{\bar{\beta}} \\ z^k_\beta \end{bmatrix} = \begin{bmatrix} q^k_{\bar{\beta}} \\ q^k_\beta \end{bmatrix} \quad \forall\ k \text{ sufficiently large.}$$

Since $\det C_A(\beta) \neq 0$,

$$\begin{bmatrix} w^k_{\bar{\beta}} \\ z^k_\beta \end{bmatrix} = \begin{bmatrix} I_{|\bar{\beta}|} & -A_{\bar{\beta}\beta}(A_{\beta\beta})^{-1} \\ 0 & -(A_{\beta\beta})^{-1} \end{bmatrix} \begin{bmatrix} q^k_{\bar{\beta}} \\ q^k_\beta \end{bmatrix} \quad \forall\ k \text{ sufficiently large.}$$

Note that as $k \to \infty$, $q^k \to e_n$, and hence $\begin{bmatrix} w^k_{\bar{\beta}} \\ z^k_\beta \end{bmatrix} \to \begin{bmatrix} v_{\bar{\beta}} \\ u_\beta \end{bmatrix}$,

where

$$\begin{bmatrix} v_{\bar{\beta}} \\ u_\beta \end{bmatrix} = \begin{bmatrix} I_{|\bar{\beta}|} & -A_{\bar{\beta}\beta}(A_{\beta\beta})^{-1} \\ 0 & -(A_{\beta\beta})^{-1} \end{bmatrix} e_n = -M_{.n}.$$

Since $v_{\bar{\beta}} \geq 0$, $u_\beta \geq 0$, $M_{.n} \leq 0$ and $u = (0^t, -M^t_{\beta n})^t \in S(e_n, A)$. Obviously $(Au)_n = -1$, as $w^k_n = 0$ for all k sufficiently large implies $v_n = 0$. This completes the proof of the theorem. □

COROLLARY 2.20. *Suppose* $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{Q}_0$. *Assume that* A *has property* (D). *If every* PPT M *of* A *is such that* $v(M^t) > 0$, *then every principal submatrix of* A *of order* $(n-1)$ *is in* $\boldsymbol{Q}_0$.

Proof. Suppose there exists an $\alpha \subseteq \bar{n}$ such that $|\alpha| = n - 1$ and $A_{\alpha\alpha} \notin \boldsymbol{Q}_0$. By Theorem 2.19, there exists a PPT M of A such that $M_{.k} \leq 0$ where $\{k\} = \bar{\alpha}$. This implies $v(M^t) \leq 0$ which contradicts the hypothesis. It follows that every principal submatrix of A of order $(n-1)$ is in $\boldsymbol{Q}_0$. □

3. Results on $E^f_0 \cap Q_0$-matrices. In this section, we consider $\boldsymbol{E}^f_0$-matrices of general order. We need the following results on $\boldsymbol{N}_0$ and almost $\boldsymbol{P}_0$-matrices.

DEFINITION 3.1. *Let $A \in \boldsymbol{R}^{n\times n}$. Say that A is an almost $\boldsymbol{P}_0$-matrix (almost $\boldsymbol{P}$-matrix) if $\det A_{\alpha\alpha} \geq 0$ ($\det A_{\alpha\alpha} > 0$) for all $\alpha \in n^*$, $\alpha \neq \bar{n}$ and $\det A < 0$.*

It is a well-known fact (see Pye [19]) that a matrix A is an almost $\boldsymbol{P}_0$-matrix (almost $\boldsymbol{P}$-matrix) iff A is nonsingular and A^{-1} is an $\boldsymbol{N}_0$-matrix ($\boldsymbol{N}$-matrix). A matrix A is said to be an $\bar{\boldsymbol{N}}$-matrix if it can be obtained as a limit of a sequence of $\boldsymbol{N}$-matrices. If $A \in \boldsymbol{R}^{n\times n}$ is an $\bar{\boldsymbol{N}}$-matrix, then there exists a nonempty subset α of $\bar{n}$ such that $A_{\alpha\alpha}$ and $A_{\bar{\alpha}\bar{\alpha}}$ are nonpositive, and $A_{\alpha\bar{\alpha}}$ and $A_{\bar{\alpha}\alpha}$ are nonnegative [12] (see also [13], [18]).

THEOREM 3.2. *Suppose $A \in \boldsymbol{R}^{n\times n}$ is an almost $\boldsymbol{P}_0$-matrix. Let $B = A^{-1}$. Then there exists a nonempty subset α of $\bar{n}$ satisfying:*

$$B_{\alpha\alpha} \leq 0, \quad B_{\bar{\alpha}\bar{\alpha}} \leq 0, \quad B_{\alpha\bar{\alpha}} \geq 0 \quad \text{and} \quad B_{\bar{\alpha}\alpha} \geq 0.$$

Proof. It suffices to show that B is an $\bar{\boldsymbol{N}}$-matrix. It is easy to show that for all positive ε sufficiently small, $A + \varepsilon I$ is an almost $\boldsymbol{P}$-matrix. Therefore, $(A + \varepsilon I)^{-1}$ is an $\boldsymbol{N}$-matrix for all positive ε sufficiently small. Note that $(A + \varepsilon I)^{-1}$ converges to B as ε converges to 0. □

LEMMA 3.3. *Let $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{E}_0 \cap \boldsymbol{N}_0$. Suppose $A \leq 0$. Then there exists a principal rearrangement M of A such that M is a strict upper triangular matrix, that is, $m_{ij} = 0$ for all $i, j \in \bar{n}$ such that $i \geq j$. In other words, there exists a permutation matrix $P \in \boldsymbol{R}^{n\times n}$ such that PAP^t is a strict upper triangular matrix.*

Proof. We shall prove this by induction on n. If $n = 1$, the result is trivially true. So assume that the lemma is valid for all matrices of order up to $(n-1) \times (n-1)$, $n > 1$. Now assume $A \in \boldsymbol{R}^{n\times n}$ satisfies the hypothesis of the lemma.

If every column of A has a negative entry, then, as $A \leq 0$, we have

$$e^t A < 0, \text{ where } e = (1, 1, \ldots, 1)^t \in \boldsymbol{R}^n.$$

This implies that value of A is negative. This contradicts the hypothesis that $A \in \boldsymbol{E}_0$. Hence A must have a zero column. Suppose $A_{\cdot j} = 0$. Then interchange the first column and jth column and then the first row and jth row. In the resulting matrix the first column will be zero. Since both $\boldsymbol{E}_0$ and $\boldsymbol{N}_0$ properties are invariant under principal rearrangements, the new matrix is also in $\boldsymbol{E}_0 \cap \boldsymbol{N}_0$. Hence assume, without loss of generality, that $A_{\cdot 1} = 0$. Let $\alpha = \{2, 3, \ldots, n\}$. Then $A_{\alpha\alpha} \in R^{(n-1)\times(n-1)} \cap \boldsymbol{E}_0 \cap \boldsymbol{N}_0$. Also $A_{\alpha\alpha} \leq 0$. By induction hypothesis, there exists a permutation matrix $\bar{P} \in R^{(n-1)\times(n-1)}$ such that $\bar{P}A_{\alpha\alpha}\bar{P}^t$ is a strict upper triangular matrix. Let

$$P = \begin{bmatrix} 1 & 0 \\ 0 & \bar{P} \end{bmatrix}.$$

Then

$$PAP^t = \begin{bmatrix} 0 & A_{1\alpha}\bar{P}^t \\ 0 & \bar{P}A_{\alpha\alpha}\bar{P}^t \end{bmatrix}.$$

Since $\bar{P}A_{\alpha\alpha}\bar{P}^t$ is a strict upper triangular matrix so is PAP^t. □

THEOREM 3.4. *Suppose $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{E}_0 \cap \boldsymbol{N}_0$. Assume that A is nonsingular. Then there exists a principal rearrangement*

$$\begin{bmatrix} A_{\alpha\alpha} & A_{\alpha\bar{\alpha}} \\ A_{\bar{\alpha}\alpha} & A_{\bar{\alpha}\bar{\alpha}} \end{bmatrix}$$

of A such that $\alpha \neq \phi$, $\alpha \neq \bar{n}$, $A_{\alpha\alpha}$, and $A_{\bar{\alpha}\bar{\alpha}}$ are nonpositive strict upper triangular matrices, and $A_{\alpha\bar{\alpha}}$, and $A_{\bar{\alpha}\alpha}$ are nonnegative matrices.

Proof. Since A is a nonsingular $\boldsymbol{N}_0$-matrix, A^{-1} is an almost $\boldsymbol{P}_0$-matrix. By Theorem 3.2, there exists a nonempty subset α of $\bar{n}$ such that $A_{\alpha\alpha}$ and $A_{\bar{\alpha}\bar{\alpha}}$ are nonpositive, and $A_{\alpha\bar{\alpha}}$ and $A_{\bar{\alpha}\alpha}$ are nonnegative matrices. Since A is nonsingular, $\alpha \neq \bar{n}$. By Lemma 3.3, there exist permutation matrices $M \in R^{|\alpha|\times|\alpha|}$ and $L \in R^{|\bar{\alpha}|\times|\bar{\alpha}|}$ such that $MA_{\alpha\alpha}M^t$ and $LA_{\bar{\alpha}\bar{\alpha}}L^t$ are strict upper triangular matrices. Let

$$P = \begin{bmatrix} M & 0 \\ 0 & L \end{bmatrix}.$$

Then

$$PAP^t = \begin{bmatrix} MA_{\alpha\alpha}M^t & MA_{\alpha\bar{\alpha}}L^t \\ LA_{\bar{\alpha}\alpha}M^t & LA_{\bar{\alpha}\bar{\alpha}}L^t \end{bmatrix}.$$

Since $A_{\bar{\alpha}\alpha}$, $A_{\alpha\bar{\alpha}}$, M, and L are all nonnegative, we have $LA_{\bar{\alpha}\alpha}M^t \geq 0$ and $MA_{\alpha\bar{\alpha}}L^t \geq 0$. This completes the proof. □

THEOREM 3.5. *Suppose $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{E}_0^f \cap \boldsymbol{Q}_0$. Assume that every proper principal minor of A is nonnegative. Then A belongs to $\boldsymbol{P}_0$.*

Proof. It suffices to show that $\det A \geq 0$. Suppose $\det A < 0$. Then A is an almost $\boldsymbol{P}_0$-matrix and hence $A^{-1} \in \boldsymbol{N}_0$. Since A^{-1} is a PPT of A, $A^{-1} \in \boldsymbol{E}_0^f \cap \boldsymbol{N}_0 \cap \boldsymbol{Q}_0$. Let $B = A^{-1}$. Then by Theorem 3.4, there exists a principal rearrangement of B such that $B_{\alpha\alpha}$ and $B_{\bar{\alpha}\bar{\alpha}}$ are nonpositive strict upper triangular matrices, and $B_{\alpha\bar{\alpha}}$ and $B_{\bar{\alpha}\alpha}$ are nonnegative matrices for some $\alpha \subseteq \bar{n}$ with $\alpha \neq \phi$ and $\alpha \neq \bar{n}$. For simplicity, we assume $\alpha = \{1, 2, \ldots, k\}$, $k < n$. Observe that

$$b_{ij} \geq 0 \ \forall\ i,\ j \in \bar{n} \text{ such that } i \geq j. \tag{3.1}$$

In particular, $B_{n.} \geq 0$. By Corollary 2.3, $B_{\beta\beta} \in \boldsymbol{Q}_0$, where $\beta = \overline{\{n\}}$. Note that, from the above observation (3.1), the last row of $B_{\beta\beta}$ is nonnegative. By Corollary 2.3, $B_{\gamma\gamma} \in \boldsymbol{Q}_0$, where $\gamma = \{1, 2, \ldots, n-2\}$. Thus it can be seen that all the leading principal submatrices of B are in $\boldsymbol{Q}_0$.

We will now show that $B_{\alpha(k+1)} = 0$ which will in turn imply that $B_{.(k+1)} = 0$ leading to the contradiction that B is singular.

Let

$$M = \begin{bmatrix} B_{\alpha\alpha} & B_{\alpha(k+1)} \\ B_{(k+1)\alpha} & 0 \end{bmatrix}.$$

M is the leading principal submatrix of B of order $(k+1)$. From the above argument, $M \in \boldsymbol{Q}_0$. If $B_{\alpha(k+1)}$ has a positive entry, then by Theorem 2.12, $M \notin \boldsymbol{Q}_0$. Hence $B_{\alpha(k+1)} = 0$. It follows that A belongs to $\boldsymbol{P}_0$. □

We shall now identify a number of subclasses of $\boldsymbol{E}_0^f$ for which Conjecture 1.1 is valid.

COROLLARY 3.6. *Suppose $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{E}_0^f \cap \bar{\boldsymbol{Q}}_0$. Then A belongs to $\boldsymbol{P}_0$.*

Proof. We prove this by induction on n. If $n = 1$, the result is obviously true. Hence assume that the result is true for all real square matrices of order less than or equal to $n-1$, $n > 1$. Suppose $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{E}_0^f \cap \bar{\boldsymbol{Q}}_0$. Then $A_{\alpha\alpha} \in R^{(n-1)\times(n-1)} \cap \boldsymbol{E}_0^f \cap \bar{\boldsymbol{Q}}_0$ for all $\alpha \in n^*$. By induction hypothesis, $A_{\alpha\alpha} \in \boldsymbol{P}_0$ for all $\alpha \in n^*$ with $|\alpha| < n$. By Theorem 3.5, A belongs to $\boldsymbol{P}_0$. □

COROLLARY 3.7. *Suppose $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{E}_0^f$. Assume that A satisfies any one of the following conditions:*

(a) *A is nonnegative $\boldsymbol{Q}_0$-matrix.*

(b) *A is a copositive-plus matrix.*

(c) *A is a $\boldsymbol{Z}$-matrix.*
(d) *A is an $\boldsymbol{E}$-matrix.*
(e) *A is symmetric $\boldsymbol{Q}_0$-matrix.*

Then A belongs to $\boldsymbol{P}_0$.

Proof. This is a direct consequence of Corollary 3.7 and the fact that if A satisfies any of the conditions (a)–(e), then A is in $\bar{\boldsymbol{Q}}_0$; see [4, pp. 181, 196, 201]. □

It is a well-known fact that the set of $\boldsymbol{P}$-matrices is an open set. Intuitively one feels that the interior of the set of $\boldsymbol{E}_0^f$-matrices in $\boldsymbol{R}^{n\times n}$ should coincide with $\boldsymbol{P}$-matrices of $\boldsymbol{R}^{n\times n}$. Our aim, here, is to show that this is indeed the case. We are not aware of a specific mention of this result in the literature. Our main interest here is to establish this as an application of our Theorem 3.5. The following results are fairly well known [3], [4].

THEOREM 3.8. *Suppose $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{E}_0$. Then for every positive ε, $A+\varepsilon I$ is in $\boldsymbol{E}$ and hence in $\bar{\boldsymbol{Q}}$.*

Proof. One can easily check that if $A \in \boldsymbol{E}_0$, then $A + \varepsilon I$ is in $\boldsymbol{E}$. The second assertion follows from Cottle's result [3], $\boldsymbol{E} = \bar{\boldsymbol{Q}}$. □

LEMMA 3.9. *Closure* $(\boldsymbol{R}^{n\times n} \cap \boldsymbol{P}) = \boldsymbol{R}^{n\times n} \cap \boldsymbol{P}_0$.

Proof. The proof follows from the fact that if $A \in \boldsymbol{P}_0$, then $A + \varepsilon I \in \boldsymbol{P}$ for all $\varepsilon > 0$. □

THEOREM 3.10. *Let $T = \{A \in \boldsymbol{R}^{n\times n} : A \in \boldsymbol{E}_0^f\}$. Then*

$$\text{interior}(T) = \boldsymbol{R}^{n\times n} \cap \boldsymbol{P}.$$

Proof. In the light of Lemma 3.9 and the fact that $\boldsymbol{R}^{n\times n} \cap \boldsymbol{P}$ is an open set, it is sufficient to show that if M is in the interior(T), then M belongs to the interior of $\boldsymbol{R}^{n\times n} \cap \boldsymbol{P}_0$.

Since $M \in$ interior$(\ T\ )$, there exists a $\delta > 0$ such that

$$B_\delta(M) = \{A \in \boldsymbol{R}^{n\times n} : ||M - A|| < \delta\} \subseteq T,$$

where $||.||$ is any norm on $\boldsymbol{R}^{n\times n}$. We will show that if $A \in B_\delta(M)$, then $A \in \boldsymbol{P}_0$. This will then imply that M is an interior point of $\{A \in \boldsymbol{R}^{n\times n} : A \in \boldsymbol{P}_0\}$. Observe that $A \in B_\delta(M)$ implies $A + \varepsilon I \in B_\delta(M)$ for all positive ε sufficiently small. Also $A + \varepsilon I \in \boldsymbol{E}_0^f$ for all positive ε sufficiently small. By Theorem 3.8, $A + \varepsilon I \in \bar{\boldsymbol{Q}}$ for all positive ε sufficiently small. By Corollary 3.6, $A+\varepsilon I \in \boldsymbol{P}_0$ for all positive ε sufficiently small. It follows that $A \in \boldsymbol{P}_0$ as $\boldsymbol{R}^{n\times n} \cap \boldsymbol{P}_0$ is a closed set. Therefore, M is an interior point of $\boldsymbol{R}^{n\times n} \cap \boldsymbol{P}_0$. Since interior of $\boldsymbol{R}^{n\times n} \cap \boldsymbol{P}_0$ is $\boldsymbol{P}$, the theorem follows. □

THEOREM 3.11. *Suppose $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{E}_0^f$, where $n \leq 3$. If all the diagonal entries of A are positive, then A is a $\boldsymbol{P}_0$-matrix.*

Proof. If A is a 2×2 $\boldsymbol{E}_0^f$-matrix, then it is easy to check that even with one diagonal entry positive, A must be a $\boldsymbol{P}_0$-matrix. Now suppose A is a 3×3 matrix satisfying hypothesis of the theorem. Assume, to the contrary, that A is not a $\boldsymbol{P}_0$-matrix. Then A is an almost $\boldsymbol{P}_0$-matrix with $\det A < 0$. Let B be the inverse of A. Then B is an $\boldsymbol{N}_0 \cap \boldsymbol{E}_0^f$-matrix and must have all diagonal entries zero. Also $\det B < 0$. Since all diagonal entries of A are positive and B is in $\boldsymbol{E}_0^f$, b_{ij} and b_{ji} must be positive for all $i \neq j$. But this implies $\det B > 0$, which is a contradiction. It follows that A is a $\boldsymbol{P}_0$-matrix. □

From the above theorem a logical question that can arise is that : If A is in $\boldsymbol{R}^{n\times n} \cap \boldsymbol{E}_0^f$ and $a_{ii} > 0$ for all i, then is it true that A belongs to $\boldsymbol{P}_0$? Our investigation for $n = 4$ proved that this is not true for $n \geq 4$. Consider the following example.

Example 3.12. Let

$$A = \begin{bmatrix} 2 & -1 & 1 & 2 \\ -2 & 1 & -1 & 1 \\ -1 & 2 & 1 & -1 \\ 2 & -1 & -2 & 2 \end{bmatrix}.$$

It can be checked that A is an almost $\boldsymbol{P}_0$-matrix and that all the PPTs of A are $\boldsymbol{E}_0$-matrices. Thus, A is an $\boldsymbol{E}_0^f$-matrix with all diagonal entries positive but $A \notin \boldsymbol{P}_0$.

DEFINITION 3.13. *Let $A \in \boldsymbol{R}^{n\times n}$. Say that A is a* fully copositive *matrix if A and all its* PPTs *are all copositive matrices. This class will be denoted by $\boldsymbol{C}_0^f$.*

Remark 3.14. As $\boldsymbol{C}_0 \subseteq \boldsymbol{E}_0$, it is obvious from the definition that $\boldsymbol{C}_0^f \subseteq \boldsymbol{E}_0^f$. Observe that positive semidefinite matrices and permutation matrices are $\boldsymbol{C}_0^f$-matrices. Furthermore, if $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{C}_0^f$, then $A_{\alpha\alpha} \in \boldsymbol{C}_0^f$ for every $\alpha \in n^*$.

Example 3.15. Let

$$A = \begin{bmatrix} 1 & 5 \\ -1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \text{and } C = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix}.$$

Note that A is a $\boldsymbol{P}$-matrix (and hence $\boldsymbol{E}_0^f$) but not a fully copositive matrix (A^{-1} does not belong to $\boldsymbol{C}_0$). B is a fully copositive matrix but not a positive semidefinite matrix. Last, C is an $\boldsymbol{E}_0^f$-matrix but not a fully copositive matrix.

The following theorem establishes that within the class of symmetric matrices there is no difference between $\boldsymbol{C}_0^f$ and $\boldsymbol{E}_0^f$.

THEOREM 3.16. *Suppose $A \in \boldsymbol{R}^{n\times n}$. Assume that A is symmetric. Then A belongs to $\boldsymbol{C}_0^f$ iff A belongs to $\boldsymbol{E}_0^f$.*

Proof. Since $\boldsymbol{C}_0^f \subseteq \boldsymbol{E}_0^f$, we need to show that if $A \in \boldsymbol{E}_0^f$, then $A \in \boldsymbol{C}_0^f$. Suppose $A \in \boldsymbol{E}_0^f$. Since A is symmetric, $A \in \boldsymbol{C}_0$. Let $\alpha \in n^*$ be such that $\det A_{\alpha\alpha} \neq 0$. We will show that $\wp_\alpha(A) \in \boldsymbol{C}_0$. Let $B = \wp_\alpha(A)$. Since A is symmetric $(A_{\bar{\alpha}\alpha})^t = A_{\alpha\bar{\alpha}}$ and $(A_{\alpha\bar{\alpha}})^t = A_{\bar{\alpha}\alpha}$. Therefore,

$$\frac{1}{2}(B + B^t) = \begin{bmatrix} (A_{\alpha\alpha})^{-1} & 0 \\ 0 & (\ A/A_{\alpha\alpha}\ ) \end{bmatrix}.$$

Since $A \in \boldsymbol{E}_0^f$, $(A_{\alpha\alpha})^{-1}$, and $(\ A/A_{\alpha\alpha}\ )$ are both $\boldsymbol{E}_0$-matrices. Observe, also, that both $(A_{\alpha\alpha})^{-1}$, and $(\ A/A_{\alpha\alpha}\ )$ are symmetric. Therefore, $(A_{\alpha\alpha})^{-1}$, and $(\ A/A_{\alpha\alpha}\ )$ are copositive matrices. Hence $\frac{1}{2}(B + B^t)$ is copositive. Since $x^t Bx = \frac{1}{2}x^t(B + B^t)x$, it follows that $B \in \boldsymbol{C}_0$. Since α was arbitrary, it follows that $A \in \boldsymbol{C}_0^f$. □

Recall that for $n \leq 3$, if $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{E}_0^f$ and if $a_{ii} > 0$ for all i, then $A \in \boldsymbol{P}_0$. In Example 3.12, it was shown that this assertion does not hold good for $n = 4$. However, for A in $\boldsymbol{C}_0^f$ we have the following results.

THEOREM 3.17. *Suppose $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{C}_0^f$. Assume that $a_{ii} > 0$ for all $i \in \bar{n}$. Then $A \in \boldsymbol{P}_0$.*

Proof. We prove this by induction on n. If $n = 1$, then the result is trivially true. Assume that the result is true for all $(n-1) \times (n-1)$ real matrices. Suppose $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{C}_0^f$ and $a_{ii} > 0$ for every $i \in \bar{n}$. Observe that for all $\alpha \subseteq \bar{n}$ with $|\alpha| = n-1$, $A_{\alpha\alpha}$ satisfies the assumptions of the theorem (see Remark 1.5). Suppose $A \notin \boldsymbol{P}_0$. By induction hypothesis, A is an almost $\boldsymbol{P}_0$-matrix. Then $\det A < 0$ and $A^{-1} \in \boldsymbol{C}_0^f \cap \boldsymbol{N}_0$. Let $B = A^{-1}$. By Theorem 3.4, there exists a subset α of $\bar{n}$ such that

$$\phi \neq \alpha \neq \bar{n},\ B_{\alpha\alpha} \leq 0,\ B_{\bar{\alpha}\bar{\alpha}} \leq 0,\ B_{\alpha\bar{\alpha}} \geq 0 \quad \text{and} \quad B_{\bar{\alpha}\alpha} \geq 0.$$

Since $B \in \boldsymbol{C}_0$, we must have $B_{\alpha\alpha} = 0$ and $B_{\bar{\alpha}\bar{\alpha}} = 0$. Without loss of generality, we may assume

$$B = \begin{bmatrix} 0 & B_{\alpha\bar{\alpha}} \\ B_{\bar{\alpha}\alpha} & 0 \end{bmatrix}.$$

Let $k = |\alpha|$. Since B is nonsingular, we must have $|\alpha| = |\bar{\alpha}| = k$. It is easy to see, from the structure of B, that if we drop any row and the corresponding column from B, then the resulting $(n-1) \times (n-1)$ principal submatrix of B must be singular. Let $\beta = \overline{\{1\}}$. Then

$$a_{11} = \frac{\det B_{\beta\beta}}{\det B} = 0,$$

which contradicts the hypothesis. It follows that A belongs to $\boldsymbol{P}_0$. □

COROLLARY 3.18. *Suppose $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{C}_0^f$. Assume that A has at most one zero diagonal entry. Then A belongs to $\boldsymbol{P}_0$.*

Proof. The proof is exactly similar to the proof of the above theorem. Note that induction hypothesis works because if A has the property that it has at most one zero diagonal entry, then every principal submatrix of A also has this property. □

4. $E_0^f \cap Q_0$-matrices of order less than 7. Before establishing Conjecture 1.1 for 4×4 matrices we prove some theorems on sign patterns of $\boldsymbol{E}_0^f$-matrices which will be needed in the sequel.

THEOREM 4.1. *Suppose $A \in R^{2\times 2} \cap \boldsymbol{E}_0^f$. Then $SP(A)$ cannot be equal to any of the following:*

$$\text{(a)} \begin{bmatrix} \oplus & - \\ - & 0 \end{bmatrix} \quad \text{(b)} \begin{bmatrix} 0 & - \\ - & \oplus \end{bmatrix} \quad \text{(c)} \begin{bmatrix} + & + \\ + & 0 \end{bmatrix} \quad \text{(d)} \begin{bmatrix} 0 & + \\ + & + \end{bmatrix}.$$

Proof. It is easy to check that value of any matrix having the above sign pattern given by (a) or (b) is negative and hence cannot be an $\boldsymbol{E}_0$-matrix. If A has the sign pattern as in (c), then the second diagonal entry of $\wp_\alpha(A)$ is negative, where $\alpha = \{1\}$. Hence A cannot have the sign pattern given by (c). Similarly we can show that A cannot have the sign pattern given by (d). □

THEOREM 4.2. *Suppose $A \in \boldsymbol{R}^{3\times 3} \cap \boldsymbol{E}_0^f$. Then $SP(A)$ cannot be equal to any of the following:*

$$\text{(a)} \begin{bmatrix} \star & - & \star \\ + & 0 & + \\ \star & + & 0 \end{bmatrix} \quad \text{(b)} \begin{bmatrix} \oplus & \star & - \\ \star & 0 & + \\ + & + & 0 \end{bmatrix} \quad \text{(c)} \begin{bmatrix} \star & - & \star \\ \star & 0 & - \\ - & 0 & 0 \end{bmatrix} \quad \text{(d)} \begin{bmatrix} \star & \star & - \\ - & \star & 0 \\ 0 & - & 0 \end{bmatrix}.$$

Proof. Suppose

$$SP(A) = \begin{bmatrix} \oplus & - & \star \\ + & 0 & + \\ \star & + & 0 \end{bmatrix}.$$

Let M be a PPT of A with respect to $\alpha = \{2,3\}$. Then note that $SP(M_{\alpha\alpha}) = SP((A_{\alpha\alpha})^{-1}) = \begin{bmatrix} 0 & + \\ + & 0 \end{bmatrix}$. Furthermore,

$$\begin{aligned} SP(M_{\alpha\bar{\alpha}}) &= SP(-(A_{\alpha\alpha})^{-1}A_{\alpha\bar{\alpha}}) \\ &= SP(-(A_{\alpha\alpha})^{-1})SP(A_{\alpha\bar{\alpha}}) \\ &= \begin{bmatrix} 0 & - \\ - & 0 \end{bmatrix} \begin{bmatrix} + \\ \star \end{bmatrix} = \begin{bmatrix} \star \\ - \end{bmatrix}. \end{aligned}$$

Similarly, $SP(M_{\bar{\alpha}\alpha}) = (\star, -)$. Since M is a PPT of an $\boldsymbol{E}_0^f$-matrix, $m_{11} \geq 0$. Thus

$$SP(M) = \begin{bmatrix} \oplus & \star & - \\ \star & 0 & + \\ - & + & 0 \end{bmatrix}.$$

Note that $SP(M_{\beta\beta}) = \begin{bmatrix} \oplus & - \\ - & 0 \end{bmatrix}$, where $\beta = \{1,3\}$. This implies that $M \notin \boldsymbol{E}_0^f$. This shows that $SP(A)$ cannot be equal to the one given by (a). Similarly we can show that $SP(A)$ cannot be equal to the one given by (b). Suppose $SP(A)$ is given by (c). It is clear that an $x > 0$ can be found so that $Ax < 0$. This contradicts the fact that $A \in \boldsymbol{E}_0$. A similar argument will show that $SP(A)$ cannot be equal to the one given by (d). □

COROLLARY 4.3. *Suppose $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{E}_0^f$. Then no principal submatrix of A or any of its principal rearrangements can have any of the sign patterns listed in Theorems* 4.1 *and* 4.2.

Proof. The proof follows from the fact that every principal submatrix of $\boldsymbol{E}_0^f$-matrix is also in $\boldsymbol{E}_0^f$. □

THEOREM 4.4. *Suppose $A \in \boldsymbol{R}^{3\times 3}$. Assume that $SP(A)$ is equal to one of the following sign patterns:*

$$\text{(a)} \begin{bmatrix} \oplus & \oplus & \oplus \\ + & 0 & + \\ + & + & 0 \end{bmatrix} \quad \text{(b)} \begin{bmatrix} \oplus & \oplus & \oplus \\ - & 0 & + \\ - & + & 0 \end{bmatrix}.$$

If $a_{12} + a_{13} > 0$, then $A \notin \boldsymbol{E}_0^f$.

Proof. Suppose $A \in \boldsymbol{E}_0^f$ and $SP(A)$ is given by (a). Since $a_{12} + a_{13} > 0$, we may assume, without loss of generality, that $a_{12} > 0$. If $a_{11} > 0$, then, by Corollary 4.3, $A \notin \boldsymbol{E}_0^f$. So $a_{11} = 0$. But then the first diagonal entry of $\wp_\alpha(A)$ is negative, where $\alpha = \{2,3\}$. This contradicts our supposition that $A \in \boldsymbol{E}_0^f$. It follows that $A \notin \boldsymbol{E}_0^f$. Suppose $SP(A)$ is given by (b). Let $B = \wp_\alpha(A)$, where $\alpha = \{2,3\}$. Then

$$SP(B) = \begin{bmatrix} \oplus & \oplus & \oplus \\ + & 0 & + \\ + & + & 0 \end{bmatrix}.$$

Since $a_{12} + a_{13} > 0$, it can be seen that $b_{12} + b_{13} > 0$. From the earlier argument $B \notin \boldsymbol{E}_0^f$. □

THEOREM 4.5. *Suppose $A \in \boldsymbol{R}^{3\times 3} \cap \boldsymbol{E}_0^f \cap \boldsymbol{Q}_0$. Then $A \in \boldsymbol{P}_0$.*

Proof. In view of Theorem 3.5, it is sufficient to show that $A_{\alpha\alpha} \in \boldsymbol{P}_0$ for all $\alpha \subseteq \{1,2,3\}$ such that $|\alpha| \leq 2$. Since $A \in \boldsymbol{E}_0^f$, $a_{ii} \geq 0$ for all i. Suppose there exists an $\alpha \subseteq \{1,2,3\}$ such that $|\alpha| = 2$ and $A_{\alpha\alpha} \notin \boldsymbol{P}_0$. Since $\boldsymbol{E}_0^f \cap \boldsymbol{Q}_0$ property is invariant under principal rearrangements, we may assume, without loss of generality, that $\alpha = \{2,3\}$. Since $A_{\alpha\alpha} \in \boldsymbol{R}^{2\times 2} \cap \boldsymbol{E}_0^f$ and $A_{\alpha\alpha} \notin \boldsymbol{P}_0$ (Theorem 4.1),

$$SP(A_{\alpha\alpha}) = \begin{bmatrix} 0 & + \\ + & 0 \end{bmatrix}.$$

By Theorem 2.11, we must have either $(a_{21}, a_{31}) < 0$ or $(a_{21}, a_{31}) > 0$.

Suppose $(a_{21}, a_{31}) > 0$. If $a_{12} < 0$ or $a_{13} < 0$, then $A \notin \boldsymbol{E}_0^f$ (Corollary 4.3). But this contradicts Theorem 2.9.

Suppose $(a_{21}, a_{31}) < 0$. By Corollary 4.3, a_{12} and a_{13} must be nonnegative. But this contradicts Theorem 2.9. From this contradiction it follows that A belongs to $\boldsymbol{P}_0$. □

We now establish Conjecture 1.1 for 4×4 matrices. The outline of the proof is as follows. We first show that every 2×2 principal submatrix of a $\boldsymbol{R}^{4\times4} \cap \boldsymbol{E}_0^f \cap \boldsymbol{Q}_0$-matrix is in $\boldsymbol{P}_0$ (see Lemma 4.7) and then show that every 3×3 principal submatrix of an $\boldsymbol{R}^{4\times4} \cap \boldsymbol{E}_0^f \cap \boldsymbol{Q}_0$-matrix is in $\boldsymbol{P}_0$. Then invoking Theorem 3.5, we conclude the result.

LEMMA 4.6. *Suppose* $A \in \boldsymbol{R}^{4\times4} \cap \boldsymbol{Q}_0$. *Assume that* $a_{33} = a_{44} = 0$, *and* a_{34} *and* a_{43} *are positive. Then there exists a* PPT B *of* A *such that, subject to principal rearrangement,*

$$b_{33} = b_{44} = 0,\ b_{31} > 0,\ b_{34} > 0 \text{ and } b_{43} > 0. \tag{4.1}$$

Proof. Let $\alpha = \{3, 4\}$. From Theorem 2.11, it follows that $A_{\alpha\bar{\alpha}} \neq 0$. If $A_{\alpha\bar{\alpha}}$ contains a positive entry, then it is easy to see that A or a principal rearrangement B of it will satisfy (4.1). If $A_{\alpha\bar{\alpha}}$ has no positive entry, then it must have a negative entry. Let $M = \wp_\alpha(A)$. Then $m_{33} = m_{44} = 0$, and m_{34} and m_{43} are positive. Also $M_{\alpha\bar{\alpha}}$ will have a positive entry. Then a principal rearrangement B of M will satisfy (4.1). □

LEMMA 4.7. *Suppose* $A \in \boldsymbol{R}^{4\times4} \cap \boldsymbol{E}_0^f \cap \boldsymbol{Q}_0$. *Assume that* $a_{33} = a_{44} = 0, a_{34} > 0$ *and* $a_{43} > 0$. *Then* $A_{1.}$ *and* $A_{2.}$ *both must have negative entries.*

Proof. Let $\alpha = \{2, 3, 4\}$ and $\beta = \{1, 3, 4\}$. From the hypothesis, $A_{\alpha\alpha}$ and $A_{\beta\beta}$ are not in $\boldsymbol{P}_0$. If $A_{1.}$ is nonnegative, then by Corollary 2.3, $A_{\alpha\alpha} \in \boldsymbol{R}^{3\times3} \cap \boldsymbol{E}_0^f \cap \boldsymbol{Q}_0$, and by Theorem 4.5, $A_{\alpha\alpha} \in \boldsymbol{P}_0$. This contradiction implies that $A_{1.}$ must have a negative entry. Similar argument shows that $A_{2.}$ must contain a negative entry. □

LEMMA 4.8. *Suppose* $A \in \boldsymbol{R}^{4\times4} \cap \boldsymbol{E}_0^f \cap \boldsymbol{Q}_0$. *Then every* 2×2 *principal submatrix of* A *is in* $\boldsymbol{P}_0$.

Proof. Suppose A has a 2×2 principal submatrix which is not in $\boldsymbol{P}_0$. Let $\alpha = \{3, 4\}$. Without loss of generality assume that $A_{\alpha\alpha} \notin \boldsymbol{P}_0$. Then we must have $a_{33} = a_{44} = 0$, and a_{34} and a_{43} are positive. In view of Lemma 4.6, we may assume, without loss of generality, that a_{31} is positive (see Remark 1.5). By Theorem 4.2 and Corollary 4.3, we must have $a_{13} \geq 0$. Since $A \in \boldsymbol{E}_0^f$, a_{11} and a_{22} are nonnegative. By Theorem 2.9, $a_{23} < 0$ and this in turn implies $a_{32} = 0$ (Corollary 4.3). By Theorem 2.11, we must have

$$\text{either } a_{41} > 0 \text{ or } a_{42} > 0.$$

Suppose $a_{41} > 0$. Then by Corollary 4.3, $a_{14} \geq 0$ and by Theorem 2.9, $a_{24} < 0$. Since $a_{24} < 0$, $a_{42} = 0$ (by Theorem 4.1 and Theorem 4.2). Thus,

$$SP(A) = \begin{bmatrix} \oplus & - & \oplus & \oplus \\ \star & \oplus & - & - \\ + & 0 & 0 & + \\ + & 0 & + & 0 \end{bmatrix}.$$

Note that $a_{12} < 0$, as otherwise it would contradict Lemma 4.7. By Theorem 4.4, $a_{13} = a_{14} = 0$. Then

$$SP(\wp_\alpha(A)) = \begin{bmatrix} \oplus & - & 0 & 0 \\ \star & \oplus & - & - \\ - & 0 & 0 & + \\ - & 0 & + & 0 \end{bmatrix}.$$

This contradicts Corollary 4.3. So we must have $a_{41} \leq 0$ and $a_{42} > 0$. This in turn implies $a_{24} \geq 0$ (Corollary 4.3), $a_{14} < 0$ (Theorem 2.9) and $a_{41} = 0$ (Corollary 4.3). Let $\beta = \{1,2,3\}$ and $\gamma = \{1,2,4\}$. Observe that $A_{3.}$ and $A_{4.}$ are nonnegative. This implies that $A_{\beta\beta}$ and $A_{\gamma\gamma}$ are in $\boldsymbol{P}_0$ (Corollary 2.3 and Theorem 4.5), which in turn implies that $a_{13} = a_{24} = 0$. Thus,

$$SP(A) = \begin{bmatrix} \oplus & \star & 0 & - \\ \star & \oplus & - & 0 \\ + & 0 & 0 & + \\ 0 & + & + & 0 \end{bmatrix}.$$

As $A_{\beta\beta} \in \boldsymbol{Q}_0$, by Theorem 2.13, we must have $a_{12} < 0$. Observe that

$$SP(\wp_\alpha(A)) = \begin{bmatrix} + & - & - & 0 \\ \star & + & 0 & - \\ 0 & - & 0 & + \\ - & 0 & + & 0 \end{bmatrix}.$$

This contradicts Corollary 4.3. Hence every 2×2 principal submatrix of a $\boldsymbol{R}^{4\times 4} \cap \boldsymbol{E}_0^f \cap \boldsymbol{Q}_0$-matrix must be in $\boldsymbol{P}_0$. □

THEOREM 4.9. *Suppose $A \in \boldsymbol{R}^{4\times 4} \cap \boldsymbol{E}_0^f \cap \boldsymbol{Q}_0$. Then A belongs to $\boldsymbol{P}_0$.*

Proof. By Lemma 4.8, every 2×2 principal submatrix of A is in $\boldsymbol{P}_0$. If every 3×3 principal submatrix of A is also in $\boldsymbol{P}_0$, then by Theorem 3.5, $A \in \boldsymbol{P}_0$. Suppose there exists an $\alpha \subseteq \{1,2,3,4\}$ such that $|\alpha| = 3$ and $A_{\alpha\alpha} \notin \boldsymbol{P}_0$. Since $\boldsymbol{E}_0^f \cap \boldsymbol{Q}_0$ property is invariant under principal rearrangements, we may assume, without loss of generality, that $\alpha = \{2,3,4\}$. Since every 2×2 principal submatrix of A is in $\boldsymbol{P}_0$, we must have $\det A_{\alpha\alpha} < 0$ and $(A_{\alpha\alpha})^{-1} \in \boldsymbol{E}_0^f \cap \boldsymbol{N}_0$. Let $B = \wp_\alpha(A)$. Then $B \in \boldsymbol{E}_0^f \cap \boldsymbol{Q}_0$. Note that $B_{\alpha\alpha} = (A_{\alpha\alpha})^{-1} \in \boldsymbol{E}_0 \cap \boldsymbol{N}_0$. By Theorem 3.4 and Remark 1.5, we can assume, without loss of generality, that

$$SP(B_{\alpha\alpha}) = \begin{bmatrix} 0 & \ominus & \oplus \\ 0 & 0 & \oplus \\ \oplus & \oplus & 0 \end{bmatrix}.$$

Since $B_{\alpha\alpha}$ is nonsingular, $b_{42} > 0$, $b_{34} > 0$, and $b_{23} < 0$. Since $B \in \boldsymbol{E}_0^f \cap \boldsymbol{Q}_0$, all its 2×2 principal submatrices are in $\boldsymbol{P}_0$. This implies $b_{24} = b_{43} = 0$. Thus

$$SP(B) = \begin{bmatrix} \oplus & \star & \star & \star \\ \star & 0 & - & 0 \\ \star & 0 & 0 & + \\ \star & + & 0 & 0 \end{bmatrix}.$$

From Theorem 2.9, b_{12} and b_{14} must both be negative (to see that $b_{12} < 0$ examine the sign pattern of A, observe that $a_{14} < 0$ (Theorem 2.9) and compute the sign of b_{12}). This in turn implies that b_{21} and b_{41} are both nonnegative. If $b_{31} \leq 0$, then we can choose a $q \in \boldsymbol{R}^{4\times 4}$ with $SP(q) = (+,+,-,+)^t$ satisfying $F(q,A) \neq \phi$ and $S(q,A) = \phi$. So b_{31} must be positive. But then

$$SP(A) = \begin{bmatrix} \oplus & \star & \star & - \\ \star & 0 & 0 & + \\ \star & - & 0 & 0 \\ - & 0 & + & 0 \end{bmatrix}.$$

This contradicts Corollary 4.3. It follows that every 3×3 principal submatrix of A is in $\boldsymbol{P}_0$. Invoking Theorem 3.5, we conclude that A belongs to $\boldsymbol{P}_0$. □

The following result is due to Jeter and Pye [10]. We give an alternative proof of this using Theorem 4.9 and a result due to Aganagic and Cottle [1].

THEOREM 4.10. *Suppose $A \in \boldsymbol{R}^{4\times 4} \cap \boldsymbol{E}_0^f \cap \boldsymbol{Q}$. Then A belongs to $\boldsymbol{R}_0$.*

Proof. Since $\boldsymbol{Q} \subseteq \boldsymbol{Q}_0$, by Theorem 4.9, $A \in \boldsymbol{P}_0$. Aganagic and Cottle [1] showed that, if $A \in \boldsymbol{P}_0$, then $A \in \boldsymbol{Q}$ iff $A \in \boldsymbol{R}_0$. Hence A belongs to $\boldsymbol{R}_0$. □

In [14], it was shown that if $A \in \boldsymbol{R}^{4\times 4} \cap \boldsymbol{E}_0^f \cap \boldsymbol{Q}$ and $a_{ii} > 0$ for all i, then $A \in \boldsymbol{P}_0$. In this direction we have the following result for $A \in \boldsymbol{R}^{5\times 5}$.

THEOREM 4.11. *Suppose $A \in \boldsymbol{R}^{5\times 5} \cap \boldsymbol{E}_0^f \cap \boldsymbol{Q}_0$. If $a_{ii} > 0$ for all $i \in \{1, 2, \ldots, 5\}$, then A belongs to $\boldsymbol{P}_0$.*

Proof. Since $A \in \boldsymbol{E}_0^f$, and $a_{ii} > 0$ for all i, by Theorem 3.12, every 3×3 principal submatrix of A is a $\boldsymbol{P}_0$-matrix. If every 4×4 principal submatrix of A is in $\boldsymbol{P}_0$, then by Theorem 3.5, $A \in \boldsymbol{P}_0$. Suppose A has a 4×4 principal submatrix which is not a $\boldsymbol{P}_0$-matrix. Without loss of generality, assume $A_{\alpha\alpha} \notin \boldsymbol{P}_0$, where $\alpha = \{1, 2, 3, 4\}$. Then, by the above observation, $\det A_{\alpha\alpha}$ is negative and $(A_{\alpha\alpha})^{-1} \in \boldsymbol{E}_0^f \cap \boldsymbol{N}_0$. Let $B = (A_{\alpha\alpha})^{-1}$. By Theorem 3.4, there exists a principal rearrangement of B whose sign pattern is

$$\text{either (a)} \quad \begin{bmatrix} 0 & \ominus & \oplus & \oplus \\ 0 & 0 & \oplus & \oplus \\ \oplus & \oplus & 0 & \ominus \\ \oplus & \oplus & 0 & 0 \end{bmatrix} \quad \text{or (b)} \quad \begin{bmatrix} 0 & \ominus & \ominus & \oplus \\ 0 & 0 & \ominus & \oplus \\ 0 & 0 & 0 & \oplus \\ \oplus & \oplus & \oplus & 0 \end{bmatrix}.$$

We may assume, without loss of generality, that $SP(B)$ itself is given by either (a) or (b). Suppose $SP(B)$ is as in (a). Note that $\det B < 0$.
Since $a_{11} = \frac{\det B_{\beta\beta}}{\det B}$, where $\beta = \{2, 3, 4\}$,

$$a_{11} > 0 \Rightarrow b_{42} > 0,\ b_{34} < 0 \text{ and } b_{23} > 0.$$

Similarly,

$$a_{22} > 0 \Rightarrow b_{41} > 0,\ b_{13} > 0.$$

$$a_{33} > 0 \Rightarrow b_{24} > 0.$$

$$a_{44} > 0 \Rightarrow b_{31} > 0,\ b_{23} > 0, b_{12} < 0.$$

Then

$$SP(B) = \begin{bmatrix} 0 & - & + & \oplus \\ 0 & 0 & + & + \\ + & \oplus & 0 & - \\ + & + & 0 & 0 \end{bmatrix}.$$

Let $D = \wp_\alpha(A)$. Then

$$SP(D) = \begin{bmatrix} 0 & - & + & \oplus & \star \\ 0 & 0 & + & + & \star \\ + & \oplus & 0 & - & \star \\ + & + & 0 & 0 & \star \\ \star & \star & \star & \star & \oplus \end{bmatrix}.$$

Write $D = (d_{ij})$. If $d_{25} \geq 0$, then $D_{2.} \geq 0$, and $D_{\beta\beta} \in \boldsymbol{E}_0^f \cap \boldsymbol{Q}_0$, where $\beta = \{1,3,4,5\}$. But $D_{\beta\beta}$ has a principal submatrix with determinant negative. This contradicts Theorem 4.9. Hence $d_{25} < 0$. By Theorem 2.9, $a_{51} < 0$. But then $v(A_{\gamma\gamma})$ is negative (observe $SP(A_{\gamma\gamma})$), where $\gamma = \{1,2,5\}$, which contradicts the hypothesis that A is in $\boldsymbol{E}_0^f$. Thus B cannot have the sign pattern given by (a). So B must be given by (b). But this sign pattern is ruled out because it implies that $a_{44} = 0$ which contradicts the hypothesis. It follows that A is in $\boldsymbol{P}_0$. □

COROLLARY 4.12. *Suppose* $A \in \boldsymbol{R}^{5\times 5} \cap \boldsymbol{E}_0^f \cap \boldsymbol{Q}$. *Assume that all the diagonal entries of* A (*or any of its* PPTs) *are positive. Then* A *belongs to* $\boldsymbol{R}_0$.

Proof. If A (or any of its PPTs) has all diagonal entries positive, then A belongs to $\boldsymbol{P}_0$. Since A is in $\boldsymbol{Q}$, A belongs to $\boldsymbol{R}_0$. □

THEOREM 4.13. *Suppose* $A \in \boldsymbol{R}^{6\times 6} \cap \boldsymbol{E}_0^f \cap \boldsymbol{Q}_0$. *Suppose* A *satisfies the following conditions:*

(a) $a_{ii} > 0$ *for every* $i \in \{1,2,\ldots,6\}$,
(b) A *has property* (D),
(c) *for every* PPT M *of* A, $v(M^t) > 0$.

Then A *belongs to* $\boldsymbol{P}_0$.

Proof. Note that Corollary 2.20 implies $A_{\alpha\alpha} \in \boldsymbol{Q}_0$ for all $\alpha \subseteq \{1,2,\ldots,6\}$ with $|\alpha| = 5$. By Theorem 4.11, $A_{\alpha\alpha} \in \boldsymbol{P}_0$ for every $\alpha \subseteq \{1,2,\ldots,6\}$ such that $|\alpha| = 5$. By Theorem 3.5, A belongs to $\boldsymbol{P}_0$. □

5. Concluding remarks. Aganagic and Cottle [1] gave a constructive characterization of $\boldsymbol{P}_0 \cap \boldsymbol{Q}_0$ and showed that Lemke's algorithm processes (q, A) when A is in this class. Hence for all the cases for which we have established Conjecture 1.1 this result will apply. We believe that Conjecture 1.1 can be established even in the case of 5×5 matrices using proof techniques employed in §4. It can be shown, using sign patterns, that if $A \in \boldsymbol{E}_0^f \cap \boldsymbol{Q}_0$ and every principal submatrix of A of order $(n-2)$ is in $\boldsymbol{P}$, then A belongs to $\boldsymbol{P}_0$.

Acknowledgments. We are grateful to Professor Gowda, Professor Mohan, Dr. Ravindran, and Dr. Sridhar for several useful discussions. We express our sincere thanks to the two anonymous referees and Professor Cottle for various suggestions and comments that removed some obscurities and improved the presentation of the paper. Our special thanks are due to Shri Bal Menon for his kind cooperation in getting a draft of this paper printed.

REFERENCES

[1] M. AGANAGIC AND R. W. COTTLE, *A note on Q-matrices*, Math. Programming, 16 (1979), pp. 374–377.
[2] ———, *A constructive characterization of Q_0-matrices with nonnegative principal minors*, Math. Programming, 37 (1987), pp. 223–231.
[3] R. W. COTTLE, *A note on completely Q-matrices*, Math. Programming, 19 (1980), pp. 347–351.
[4] R.W. COTTLE, J.S. PANG, AND R.E. STONE, *The Linear Complementarity Problem*, Academic Press, Inc., New York, 1992.
[5] R. W. COTTLE AND R. E. STONE, *On the uniqueness of solutions to linear complementarity problems*, Math. Programming, 27 (1983), pp. 191–213.
[6] J. T. FREDRICKSEN, L. T. WATSON, AND K. G.MURTY, *A finite characterization of K-matrices in dimension less than four*, Math. Programming, 35 (1986), pp. 17–31.
[7] A. W. INGLETON, *The linear complementarity problem*, J. London Math. Soc. (2), 2 (1970), pp. 330–336.
[8] M. W. JETER AND W. C. PYE, *Some properties of Q-matrices*, Linear Algebra Appl., 57 (1984), pp. 169–180.

[9] M. W. JETER AND W. C. PYE, *Some remarks on copositive Q-matrices and on a conjecture of Pang*, J. Indust. Math. Soc., 35 (1985), pp. 75–80.
[10] ———, *An example of a nonregular semimonotone Q-matrix*, Math. Programming, 44 (1989), pp. 351–356.
[11] S. R. MOHAN, *Degeneracy in linear complimentarity problems: A Survey*, Ann. Oper. Res., 46 (1993), pp. 179–194.
[12] S. R. MOHAN, T. PARTHASARATHY, AND R. SRIDHAR, *$\bar{N}$-Matrices and the class Q*, Lecture Notes in Economics and Mathematical Systems, 389 (1992), pp. 24–36.
[13] S. R. MOHAN AND R. SRIDHAR, *On characterizing N-matrices using linear complementarity*, Linear Algebra Appl., 160 (1992), pp. 231–245.
[14] G. S. R. MURTHY, T. PARTHASARATHY, AND G. RAVINDRAN, *On copositive semimonotone Q-matrices*, Math. Programming, 68 (1995), pp. 187–203.
[15] K. G. MURTY, *On the number of solutions to the complementarity problem and spanning properties of complementary cones*, Linear Algebra Appl., 5 (1972), pp. 65–108.
[16] ———, *Linear Complementarity, Linear and Nonlinear Programming*, Heldermann-Verlag, Berlin, 1988.
[17] J. S. PANG, *On Q-matrices*, Math. Programming, 17 (1979), pp. 243–247.
[18] T. PARTHASARATHY AND G. RAVINDRAN, *N-matrices*, Linear Algebra Appl., 139 (1990), pp. 89–102.
[19] W. C. PYE, *Almost P_0-matrices and the class Q*, Math. Programming, 57 (1992), pp. 439–444.
[20] SONG XU, *Notes on sufficient matrices*, Linear Algebra Appl., 191 (1993), pp. 1–13.
[21] R. E. STONE, *Geometric Aspects of Linear Complementarity Problem*, Ph.D. thesis, Department of Operations Research, Stanford University, Stanford, CA, 1981.
[22] H. A. TAHA, *Operations Research*, 3rd ed., Macmillan Publishing Company, New York, 1982.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, pp. 1287–1307, October 1995

018

STABILITY OF LINEAR EQUATIONS SOLVERS IN INTERIOR-POINT METHODS*

STEPHEN J. WRIGHT†

Abstract. Primal-dual interior-point methods for linear complementarity and linear programming problems solve a linear system of equations to obtain a modified Newton step at each iteration. These linear systems become increasingly ill-conditioned in the later stages of the algorithm, but the computed steps are often sufficiently accurate to be useful. We use error analysis techniques tailored to the special structure of these linear systems to explain this observation and examine how theoretically superlinear convergence of a path-following algorithm is affected by the roundoff errors.

Key words. primal-dual interior-point methods, error analysis, stability

AMS subject classifications. 65G05, 65F05, 90C33

1. Introduction.

The monotone linear complementarity problem (LCP) is the problem of finding a vector pair $(x, y) \in \mathbf{R}^n \times \mathbf{R}^n$ such that

$$y = Mx + q, \qquad (x, y) \geq 0, \qquad x^T y = 0, \tag{1}$$

where M (a real, $n \times n$ positive semidefinite matrix) and q (a real vector with n elements) are given. Note that M need not be symmetric. It is well known that (1) includes the linear programming problem as a special case. Specifically, for the linear programming formulation

$$\min_z c^T z \text{ subject to } Az \geq b, \quad z \geq 0, \tag{2}$$

where $A \in \mathbf{R}^{m \times p}$, we can introduce the dual variable $\lambda \in \mathbf{R}^m$ for the constraint $Az \geq b$ and obtain the following necessary and sufficient conditions for optimality of the primal-dual pair (z, λ):

$$\begin{bmatrix} 0 & -A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} z \\ \lambda \end{bmatrix} + \begin{bmatrix} c \\ -b \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \begin{bmatrix} z \\ \lambda \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{3a}$$

$$z^T(c - A^T\lambda) + \lambda^T(Az - b) = 0. \tag{3b}$$

For appropriate definitions of M and q, (3) has the form (1). Little is lost from either the practical or theoretical point of view by applying interior-point algorithms for (1) to the special cases of linear and convex quadratic programming, provided that the special structure of each problem is exploited in the solution of the linear systems at each iteration.

Interior-point methods for (1) generate a sequence of iterates (x^k, y^k) that are strictly positive. Many such methods require a linear system of the form

$$\begin{bmatrix} M & -I \\ Y & X \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} r \\ -XYe + \sigma\mu e \end{bmatrix}, \tag{4}$$

* Received by the editors December 21, 1993; accepted for publication (in revised form) by N. Higham December 13, 1994. This work was based on research supported by the Office of Scientific Computing, U.S. Department of Energy, Contract W-31-109-Eng-38.

† Mathematics and Computer Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, Illinois 60439 (wright@mcs.anl.gov).

where

$$X = \operatorname{diag}(x_1, x_2, \ldots, x_n), \quad Y = \operatorname{diag}(y_1, y_2, \ldots, y_n), \quad e = (1, 1, \ldots, 1)^T,$$
$$\mu = x^T y/n, \qquad r = y - Mx - q, \qquad \sigma \in [0, 1],$$

to be solved for a search direction (u, v) at each iteration. Affine-scaling methods solve (4) with $\sigma = 0$ to find a search direction, then step a fraction of the distance along this direction to the boundary of the nonnegative orthant defined by $(x, y) \geq 0$. Affine-scaling steps (u, v) are simply Newton steps for the system of nonlinear equations

$$F(x, y) = \begin{bmatrix} Mx + q - y \\ XYe \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{5}$$

Path-following methods (see, for example, Monteiro and Adler [10], Zhang [19], Wright [14]) generate steps by using generally positive values of σ in (4). (As we see later, the algorithm of [14] allows $\sigma = 0$ on some iterates in an attempt to attain the rapid local convergence associated with Newton's method.) Potential-reduction methods (see, for example, Kojima, Mizuno, and Yoshise [6], Kojima, Kurita, and Mizuno [5]) also determine search directions by solving systems like (4), but they refer to a logarithmic potential function to decide how far to move along the computed direction. Predictor-corrector methods (see, for example, Ye and Anstreicher [18], Ji, Potra, and Huang [4], Potra [12]) take steps with either $\sigma = 0$ or $\sigma = 1$.

The system (4) is highly structured; since the diagonals of X and Y are strictly positive, we can rearrange the system to obtain

$$(M + X^{-1}Y)u = r - y + \sigma\mu X^{-1}e, \tag{6a}$$
$$v = -X^{-1}Yu - y + \sigma\mu X^{-1}e. \tag{6b}$$

In the case of linear programming (2), equation (6a) contains even more structure. Its form is

$$\begin{bmatrix} Z^{-1}Y_z & -A^T \\ A & \Lambda^{-1}Y_\lambda \end{bmatrix} \begin{bmatrix} u_z \\ u_\lambda \end{bmatrix} = \begin{bmatrix} A^T\lambda - c + \sigma\mu Z^{-1}e \\ -Az + b + \sigma\mu\Lambda^{-1}e \end{bmatrix}, \tag{7}$$

where Y_z and Y_λ are positive diagonal matrices and

$$Z = \operatorname{diag}(z_1, \ldots, z_p), \qquad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m).$$

The matrix in (7) can be made symmetric indefinite by multiplying the first block row by -1. This system can be reduced even further by eliminating either u_z or u_λ. For instance, if u_z is eliminated, we obtain

$$(AY_z^{-1}ZA^T + \Lambda^{-1}Y_\lambda)u_\lambda = (b - Az + \sigma\mu\Lambda^{-1}e) - AY_z^{-1}Z(A^T\lambda - c + \sigma\mu Z^{-1}e). \tag{8}$$

Some interior-point codes for linear programming use the formulation (8), with modifications for handling dense columns in A and for dealing with nonstandard linear programming formulations (see Lustig, Marsten, and Shanno [7], [8] and Xu, Hung, and Ye [17]). Other codes, notably those of Fourer and Mehrotra [1] and Vanderbei [13], handle the formulation (7). Analysis of algorithms for these formulations are discussed in another preprint [16]. In this paper, we focus on the system arising from general monotone LCP (6), and analyze the behavior of Gaussian elimination with pivoting applied to this system.

Since for any index $i = 1, \ldots, n$, at least one of x_i and y_i is zero at the solution, we would expect some of the diagonal elements of $X^{-1}Y$ to approach zero and some to approach $+\infty$ as the solution set is approached. Hence the coefficient matrix in (6a) tends to become increasingly ill-conditioned during the later stages of the algorithm. From the standard error analysis of linear systems, we might therefore expect that rounding errors in the step (u, v) make it useless in advancing the algorithm towards convergence. In this paper, we show that while the theoretical superlinear properties suggested by the exact analysis are not generally observed, implementations of the algorithms can still exhibit rapid convergence if the parameters are set to appropriate values. For a particular path-following infeasible-interior-point algorithm with strong theoretical convergence properties, these conclusions are presented in §4 and confirmed by computational experiments in §5. Section 3 lays the groundwork by deriving bounds on the rounding errors in the computed values of (u, v), for a wide class of algorithms that includes the algorithm of §§4 and 5. Section 2 presents the assumptions and a fundamental result from error analysis.

Linear systems that arise in logarithmic barrier methods for constrained optimization methods were analyzed by Ponceleón [11]. The Newton equations for each logarithmic subproblem are similar to (6a) in that the large elements occur only on the diagonal. Despite the apparent ill-conditioning of these systems, Ponceleón showed that their sensitivity to structured perturbations from a certain class is governed by the conditioning of the underlying problem and does not depend on the current value of the barrier parameter. Ponceleón's analysis is somewhat different from that of §3 — she looks at the relative error in the components of the solution, rather than starting with the absolute error — but her conclusions are consistent with those obtained in §3.

In subsequent sections, subscripts denote components of a vector, while iteration indices (usually k) appear as subscripts on scalars and as superscripts on vectors and matrices. The sets B and N form a partition of the index set $\{1, 2, \ldots, n\}$ defined in Assumption 1 below. If $x \in \mathbf{R}^n$, then

$$x_B = [x_i]_{i \in B}, \qquad X = \operatorname{diag}(x) = \operatorname{diag}(x_1, x_2, \ldots, x_n), \qquad X_B = \operatorname{diag}(x_B),$$

and so on. For the matrix $M \in \mathbf{R}^{n \times n}$, we have

$$M_{BN} = [M_{ij}]_{i \in B, j \in N},$$

and similarly for M_{BB}, M_{NB}, and M_{NN}. Given any matrix $H = [h_{ij}]$, we define $|H| = [|h_{ij}|]$ and denote the jth column of H by $H_{\cdot,j}$. The notation $\|\cdot\|$ denotes the 1-, 2-, or ∞-norm of a vector or matrix, while $\kappa(\cdot)$ denotes the corresponding condition number.

For any two nonnegative numbers ν and χ, we write $\nu = O(\chi)$ if there is a moderate constant τ such that $\nu \leq \tau\chi$. When W is a matrix or vector, we write $W = O(\chi)$ to denote $\|W\| = O(\chi)$. We use $\nu = \Omega(\chi)$ to indicate that both $\nu = O(\chi)$ and $\chi = O(\nu)$.

We use $\mathbf{u}$ to denote unit roundoff, which we define implicitly by the statement that when x and y are any two floating point numbers, op denotes $+, -, \times, /$, and $fl(z)$ denotes the floating point representation of any real number z, we have

$$fl(x \operatorname{op} y) = (x \operatorname{op} y)(1 + \delta), \qquad |\delta| \leq \mathbf{u}. \tag{9}$$

(See Golub and Van Loan [2, §2.4.2].) We assume throughout that $\mathbf{u}$ is small enough that $O(\mathbf{u}) \ll 1$, where $O(\cdot)$ is the order notation defined above.

2. Assumptions and basic results. In the remainder of the paper, we focus on path-following interior-point methods. "Infeasible" variants of these methods are among the most widely used practical algorithms for linear programming and LCPs [7], [8] and have also been the subject of extensive theoretical investigations, which have shown that they can have strong convergence properties under weak assumptions [19], [15], [14]. The path-following infeasible-interior-point framework stated below also includes the class of predictor-corrector methods, for appropriate choices of the initial point and parameters.

Path-following algorithms restrict their iterates (x^k, y^k) to neighborhoods of the form

$$\mathcal{N}(\gamma) = \{(x,y) > 0 \,|\, x_i y_i \geq \gamma(x^T y/n)\} \subset \mathbf{R}^n_+ \times \mathbf{R}^n_+, \tag{10}$$

where $\mathbf{R}^n_+$ denotes the nonnegative orthant in $\mathbf{R}^n$. All iterates generated by the algorithms lie in $\mathcal{N}(\gamma_{\min})$, where $\gamma_{\min} \in (0, 1/2)$ is a constant. Other quantities needed to define the general algorithm include

$$\mu_k = (x^k)^T y^k/n, \qquad r^k = y^k - Mx^k - q, \qquad \gamma_{\max} \in (\gamma_{\min}, 1/2],$$
$$X^k = \operatorname{diag}(x^k), \qquad Y^k = \operatorname{diag}(y^k).$$

We define the algorithmic framework in what follows.

ALGORITHM PFI
given $\gamma_0 \in [\gamma_{\min}, \gamma_{\max}]$ and $(x^0, y^0) \in \mathcal{N}(\gamma_0)$

for $k = 0, 1, 2, \ldots$
choose $\sigma_k \in [0, 1]$ and find (u^k, v^k) that satisfies

$$\begin{bmatrix} M & -I \\ Y^k & X^k \end{bmatrix} \begin{bmatrix} u^k \\ v^k \end{bmatrix} = \begin{bmatrix} r^k \\ -X^k Y^k e + \sigma_k \mu_k e \end{bmatrix}; \tag{11}$$

choose $\gamma_{k+1} \in [\gamma_{\min}, \gamma_{\max}]$ and $\alpha_k > 0$ such that

$$(x^{k+1}, y^{k+1}) = (x^k, y^k) + \alpha_k (u^k, v^k) \in \mathcal{N}(\gamma_{k+1})$$

and $\prod_{j=0}^{k}(1 - \alpha_j) \leq K\mu_{k+1}/\mu_0$ when $r^0 \neq 0$, for some constant $K > 0$;
end (for)

The decrease in $\|r^k\|$ at each iteration is linear — in fact, $r^{k+1} = \prod_{j=0}^{k}(1 - \alpha_j) r^0$ — so the last condition in Algorithm PFI is equivalent to

$$\|r^{k+1}\|/\|r^0\| \leq K\mu_{k+1}/\mu_0. \tag{12}$$

Hence, $\|r_k\| = O(\mu_k)$ for all k, so the infeasibility is always bounded by a multiple of the complementarity gap μ.

When the initial point is feasible ($r^0 = 0$), predictor-corrector algorithms such as that of Ji, Potra, and Huang [4] are special cases of Algorithm PFI. This framework also includes the infeasible-interior-point algorithms of Zhang [19] and Wright [15], [14]. These algorithms choose γ_{k+1} and σ_k so that a step α_k of nontrivial length can always be taken without violating the required conditions.

In practical implementations of interior-point methods, the framework of Algorithm PFI is usually modified slightly. In linear programming codes, different step

lengths are usually chosen for the primal and dual components of x, as experience has shown that this strategy tends to reduce the number of iterations slightly. Moreover, explicit membership of the neighborhood (10) is usually not enforced. (A more common strategy, for which there is no supporting theory, is to find the largest value of α in $[0,1]$ that keeps $(x^k, y^k) + \alpha(u^k, v^k)$ in the nonnegative orthant and then choose α_k to be a fixed fraction of this length.) The predictor-corrector strategy of Mehrotra [9], used also in the codes of Lustig, Marsten, and Shanno [8], Vanderbei [13], and Xu, Hung, and Ye [17], adds extra terms to the lower part of the right-hand side on "corrector" iterations. Nevertheless, the coefficient matrices used in these practical algorithms are the same as in (11), and our conclusions about the accuracy of the computed steps continue to hold, with minor modifications to the analysis of §3.

For most of our analysis, we make the following assumptions about the data for problem (1) and its solution set.

Assumption 1.

(a) Problem (1) has a unique solution (x^*, y^*) such that $x^* + y^* > 0$ (i.e. *strict complementarity* holds). We can define an associated partition B, N of the index set $\{1, \ldots, n\}$ such that $x_i^* > 0$ for all $i \in B$ and $y_i^* > 0$ for all $i \in N$.

(b) The quantities

$$\|M\|, \qquad \|M_{BB}^{-1}\|, \qquad \|X_B^*\|, \qquad \|(X_B^*)^{-1}\|, \qquad \|Y_N^*\|, \qquad \|(Y_N^*)^{-1}\|$$

are all moderate in size.

Assumption 1 implies that the coefficient matrix in (11) approaches a nonsingular limit, since there are $2n \times 2n$ permutation matrices P and Π such that

$$P \begin{bmatrix} M & -I \\ Y^* & X^* \end{bmatrix} \Pi = \begin{bmatrix} X_B^* & 0 & 0 & 0 \\ 0 & Y_N^* & 0 & 0 \\ -I & M_{BN} & M_{BB} & 0 \\ 0 & M_{NN} & M_{NB} & -I \end{bmatrix}, \tag{13}$$

and each of the submatrices on the diagonal of (13) is nonsingular.

When the problem (1) is derived from a linear program as in (3), existence of a solution implies existence of a strictly complementary solution. However, for both this special case and the general case of M symmetric positive semidefinite, uniqueness of the solution and well-conditioning of M_{BB} are often not satisfied in practice, so Assumption 1 is quite strong. As we see, however, this assumption plays an important role in showing that the errors in the computed solutions are not disastrous for the interior-point algorithm, just as well-conditioning of the square coefficient matrix A in a linear system $Az = b$ is needed to ensure that the relative errors in the computed version of z are not too large. Our computational experience (§5) tends to indicate that Assumption 1 is necessary as well as sufficient for rapid local convergence of the algorithm.

We make one further assumption on the iterates generated by the basic algorithm.

Assumption 2. The iterates generated by Algorithm PFI satisfy

$$\lim_{k \to \infty} (x^k, y^k) = (x^*, y^*).$$

Of course, it is not necessary to make this assumption for any reasonable instance of Algorithm PFI, since convergence to a solution should be one of the properties implied by the particular schemes for choosing σ_k, α_k, and γ_k. We make this assumption here

merely to divorce the error estimates of the next section from any particular variant of Algorithm PFI.

By the implicit function theorem, the nonsingularity of the matrix (13), equation (11), and $\|r^k\| = O(\mu_k)$, we have

$$\|(u^k, v^k)\| = O(\mu_k), \qquad \|(x^k, y^k) - (x^*, y^*)\| = O(\mu_k). \tag{14}$$

We also have the following simple result.

LEMMA 2.1. *There are positive constants C_1 and C_2 such that for all k sufficiently large, we have*

$$C_1\mu_k \le x_i^k \le C_2\mu_k \qquad \forall i \in N, \tag{15a}$$

$$C_1\mu_k \le y_i^k \le C_2\mu_k \qquad \forall i \in B. \tag{15b}$$

Proof. Because of Assumption 2, we can define an index K and positive constants $\bar{C}_1$ and $\bar{C}_2$ such that

$$x_i^k \in [\bar{C}_2, \bar{C}_1] \qquad \forall k \ge K \ \ \forall i \in B,$$
$$y_i^k \in [\bar{C}_2, \bar{C}_1] \qquad \forall k \ge K \ \ \forall i \in N.$$

Therefore, since $(x^k, y^k) > 0$, we have for $k \ge K$, $i \in N$, that

$$x_i^k y_i^k < (x^k)^T y^k = n\mu_k \ \Rightarrow \ x_i^k < \frac{n\mu_k}{y_i^k} \le \frac{n\mu_k}{\bar{C}_2}.$$

Also,

$$x_i^k y_i^k \ge \gamma_k \mu_k \ge \gamma_{\min}\mu_k \ \Rightarrow \ x_i^k \ge \frac{\gamma_{\min}\mu_k}{y_i^k} \ge \frac{\gamma_{\min}\mu_k}{\bar{C}_1}.$$

Therefore (15a) holds with $C_1 = \gamma_{\min}/\bar{C}_1$ and $C_2 = n/\bar{C}_2$. The proof of (15b) is similar. □

Finally, we state a result from the roundoff error analysis of Gaussian elimination, for reference in the next section.

THEOREM 2.2. *Suppose that the $m \times m$ linear system $Az = b$ is solved by using Gaussian elimination, possibly with row and/or column pivoting. Let us denote the row permutation matrix by P, the column permutation matrices by Π, the computed unit lower triangular factor by $\hat{L}$, and the computed upper triangular factor by $\hat{U}$. Then the computed solution $\hat{z}$ solves the perturbed system $(A + H)\hat{z} = b$, where*

$$|PH\Pi| \le \epsilon_m(2 + \epsilon_m)|\hat{L}||\hat{U}|, \tag{16}$$

and $\epsilon_m = m\mathbf{u}/(1 - m\mathbf{u}) = O(\mathbf{u})$.

Proof. The proof follows immediately from Theorem 6.4 of Higham [3]. □

During Gaussian elimination, the size of the largest element in each column of the remaining submatrix may grow as multiples of the pivot rows are added to later rows in the matrix. We quantify this growth by the growth factor ρ, defined as the smallest positive number such that

$$\max_{i=1,\ldots,m} |\hat{U}_{ij}| \le \rho \max_{i=1,\ldots,m} |(PA\Pi)_{ij}| \qquad \forall j = 1, 2, \ldots, m. \tag{17}$$

We then have the following simple corollary of Theorem 2.2.

COROLLARY 2.3. *Let the system $Az = b$ be as in Theorem 2.2, and suppose the pivots $\hat{U}_{jj}$ are chosen so that $|\hat{L}_{ij}| \leq 1$ for all $i = 2, \ldots, m$, $j = 1, \ldots, i-1$. Then*

$$\|H_{\cdot,j}\| \leq m\epsilon_m(2+\epsilon_m)\rho\|A_{\cdot,j}\|\mathbf{u}. \tag{18}$$

3. Error bounds for the steps. In this section, we derive estimates for the difference between the step actually computed by solving (6), which we denote by $(\hat{u}, \hat{v})$, and the corresponding exact values, denoted by (u, v). We treat the cases in which $(\hat{u}, \hat{v})$ is determined by Gaussian elimination with row partial pivoting and with complete pivoting.

We start with a purely technical result.

LEMMA 3.1. *Let G be a square matrix partitioned as*

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix},$$

where G_{11} and G_{22} are also square. Suppose that G_{11} and $G_{22} - G_{21}G_{11}^{-1}G_{12}$ are nonsingular. Then G is nonsingular and G^{-1} has the form

$$\begin{bmatrix} G_{11}^{-1} + G_{11}^{-1}G_{12}(G_{22} - G_{21}G_{11}^{-1}G_{12})^{-1}G_{21}G_{11}^{-1} & -G_{11}^{-1}G_{12}(G_{22} - G_{21}G_{11}^{-1}G_{12})^{-1} \\ -(G_{22} - G_{21}G_{11}^{-1}G_{12})^{-1}G_{21}G_{11}^{-1} & (G_{22} - G_{21}G_{11}^{-1}G_{12})^{-1} \end{bmatrix}.$$

Our first main result concerning components of $\hat{u}$ is the following.

THEOREM 3.2. *Let $\hat{u}$ be computed by applying Gaussian elimination with row partial pivoting to* (6a)*, and suppose that the growth factor ρ is not too large. Then for all μ sufficiently small, we have*

$$\|\hat{u}_B\| = O(\mathbf{u} + \mu), \qquad \|\hat{u}_N\| = O(\mu).$$

Proof. We assume that μ is much smaller than all the quantities in Assumption 1(b), so that $\mu \ll 1$. We retain only the lowest-order terms in $\mathbf{u}$ and μ in the analysis since, by our assumptions, higher order terms are small enough to be absorbed into lower-order terms with minor perturbations of the coefficients.

From Theorem 2.2 and (18), we have, by permuting the rows and columns of (6a), that

$$\begin{bmatrix} M_{BB} + X_B^{-1}Y_B + E_{BB} & M_{BN} + E_{BN} \\ M_{NB} + E_{NB} & M_{NN} + X_N^{-1}Y_N + E_{NN} \end{bmatrix} \begin{bmatrix} \hat{u}_B \\ \hat{u}_N \end{bmatrix} = \begin{bmatrix} r_B - y_B - \sigma\mu X_B^{-1}e \\ r_N - y_N - \sigma\mu X_N^{-1}e \end{bmatrix}, \tag{19}$$

where

$$\|E_{BB}\| \leq \rho O(\|M_{BB} + X_B^{-1}Y_B\| + \|M_{NB}\|)\mathbf{u}, \tag{20a}$$

$$\|E_{NB}\| \leq \rho O(\|M_{BB} + X_B^{-1}Y_B\| + \|M_{NB}\|)\mathbf{u}, \tag{20b}$$

$$\|E_{BN}\| \leq \rho O(\|M_{NN} + X_N^{-1}Y_N\| + \|M_{BN}\|)\mathbf{u}, \tag{20c}$$

$$\|E_{NN}\| \leq \rho O(\|M_{NN} + X_N^{-1}Y_N\| + \|M_{BN}\|)\mathbf{u}. \tag{20d}$$

Now from Assumption 1(b) and Lemma 2.1, we have

$$\|X_B^{-1}Y_B\| = O(\mu), \qquad \|X_N^{-1}Y_N\| = O(\mu^{-1}),$$

and $\|M_{BB}\|$, $\|M_{BN}\|$, $\|M_{NB}\|$, and $\|M_{NN}\|$ are all $O(1)$. Combining these observations with (20), we obtain

$$\|E_{BB}\| = O(\mathbf{u}), \qquad \|E_{NB}\| = O(\mathbf{u}), \qquad \|E_{BN}\| = O(\mu^{-1}\mathbf{u}), \qquad \|E_{NN}\| = O(\mu^{-1}\mathbf{u}).$$

Therefore (19) can be rewritten as

$$\begin{bmatrix} M_{BB} + \bar{E}_{BB} & M_{BN} + \bar{E}_{BN} \\ M_{NB} + \bar{E}_{NB} & M_{NN} + X_N^{-1}Y_N + \bar{E}_{NN} \end{bmatrix} \begin{bmatrix} \hat{u}_B \\ \hat{u}_N \end{bmatrix} = \begin{bmatrix} b_B \\ b_N \end{bmatrix}, \tag{21}$$

where

$$\begin{aligned} \|\bar{E}_{BB}\| &= O(\mu + \mathbf{u}), & \|\bar{E}_{NB}\| &= O(\mathbf{u}), \\ \|\bar{E}_{BN}\| &= O(\mu^{-1}\mathbf{u}), & \|\bar{E}_{NN}\| &= O(\mu^{-1}\mathbf{u}), \end{aligned} \tag{22}$$

and

$$b_B = r_B - y_B - \sigma\mu X_B^{-1}e, \qquad b_N = r_N - y_N - \sigma\mu X_N^{-1}e.$$

If we denote the coefficient matrix in (21) by G, with $G_{11} = M_{BB} + \bar{E}_{BB}$ and so on, we have from Assumption 1(b) and Lemma 3.1 that

$$G_{11}^{-1} = [I + M_{BB}^{-1}\bar{E}_{BB}]^{-1}M_{BB}^{-1} = M_{BB}^{-1} + O(\mu + \mathbf{u}).$$

Since

$$\|G_{21}\| = O(1), \qquad \|G_{12}\| = O(1 + \mu^{-1}\mathbf{u}), \qquad \|G_{11}^{-1}\| = O(1),$$

and since $\|X_N Y_N^{-1}\| = O(\mu)$ from (15a), we have

$$\begin{aligned} G_{22} - G_{21}G_{11}^{-1}G_{12} &= X_N^{-1}Y_N \left[I + X_N Y_N^{-1}(M_{NN} + \bar{E}_{NN}) - X_N Y_N^{-1}G_{21}G_{11}^{-1}G_{12}\right] \\ &= X_N^{-1}Y_N \left[I + O(\mu)O(1 + \mu^{-1}\mathbf{u}) + O(\mu)O(1)O(1)O(1 + \mu^{-1}\mathbf{u})\right] \\ &= X_N^{-1}Y_N[I + O(\mu + \mathbf{u})]. \end{aligned}$$

Hence

$$(G_{22} - G_{21}G_{11}^{-1}G_{12})^{-1} = (I + O(\mu + \mathbf{u}))\, Y_N^{-1}X_N = O(\mu).$$

Substitution in Lemma 3.1, together with Assumption 1(b) and some manipulation, yields

$$\begin{aligned} (G^{-1})_{11} &= M_{BB}^{-1} + O(\mu + \mathbf{u}), & (G^{-1})_{12} &= O(\mu + \mathbf{u}), \\ (G^{-1})_{21} &= O(\mu), & (G^{-1})_{22} &= O(\mu). \end{aligned}$$

We have $\|r\| = O(\mu)$ in exact arithmetic, but roundoff errors in the calculation of $r = y - Mx - q$ restrict us to assuming that $\|r\| = O(\mu + \mathbf{u})$. Using this fact together with Assumption 1(b), formula (14), and Lemma 2.1, we have

$$\|b_B\| = O(\mu + \mathbf{u}), \qquad \|b_N\| = O(1).$$

Therefore, from (21), we have

$$\begin{aligned} \hat{u}_B &= (G^{-1})_{11}b_B + (G^{-1})_{12}b_N = O(\mu + \mathbf{u}), \\ \hat{u}_N &= (G^{-1})_{21}b_B + (G^{-1})_{22}b_N = O(\mu), \end{aligned}$$

as required. □

Our next result bounds the difference between u and $\hat{u}$.

THEOREM 3.3. *Suppose the assumptions of Theorem* 3.2 *hold. Then for all* μ *sufficiently small, we have*

$$\|\hat{u}_B - u_B\| = O(\mu + \mathbf{u}), \tag{23a}$$

$$\|\hat{u}_N - u_N\| = O(\mu(\mu + \mathbf{u})), \tag{23b}$$

and, for all $i \in N$,

$$u_i/x_i = -1 + O(\sigma + \mu), \tag{24a}$$

$$\hat{u}_i/x_i = -1 + O(\sigma + \mu + \mathbf{u}). \tag{24b}$$

Proof. The expression (23a) follows immediately from (14) and Theorem 3.2, since

$$\|\hat{u}_B - u_B\| \le \|\hat{u}_B\| + \|u_B\| = O(\mu + \mathbf{u}).$$

For (23b), note first from (6) and (21) that

$$M_{NB}u_B + (M_{NN} + X_N^{-1}Y_N)u_N = b_N = (M_{NB} + \bar{E}_{NB})\hat{u}_B + (M_{NN} + X_N^{-1}Y_N + \bar{E}_{NN})\hat{u}_N.$$

Hence, from (14), (22), (23a), Assumption 1(b), and Theorem 3.2, we have

$$\begin{aligned} X_N^{-1}Y_N(u_N - \hat{u}_N) &= M_{NB}(\hat{u}_B - u_B) + M_{NN}(\hat{u}_N - u_N) + \bar{E}_{NB}\hat{u}_B + \bar{E}_{NN}\hat{u}_N \\ &= O(\mu + \mathbf{u}) + O(\mu) + O(\mu\mathbf{u}) + O(\mathbf{u}) = O(\mu + \mathbf{u}). \end{aligned}$$

From (15a), we have $X_N Y_N^{-1} = O(\mu)$ and so (23b) is proved.

From (4), we have

$$y_i u_i + x_i v_i = -x_i y_i + \sigma\mu$$

and therefore

$$\frac{u_i}{x_i} = -1 - \frac{v_i}{y_i} + \frac{\sigma\mu}{x_i y_i}. \tag{25}$$

Because $x_i y_i \ge \gamma_{\min}\mu$, we have $\mu/(x_i y_i) = O(1)$. Also from Assumption 2 and (14), we have for $i \in N$ that $v_i/y_i = O(\mu)$. Hence, (24a) is obtained by using these estimates in (25). For (24b), we have from (15a), (23b), and (24a) that for all $i \in N$,

$$\frac{\hat{u}_i}{x_i} = \frac{u_i}{x_i} + \frac{1}{x_i}O(\mu(\mu + \mathbf{u})) = -1 + O(\mu + \sigma) + O(\mu + \mathbf{u}). \qquad \square$$

Note that in Theorems 3.2 and 3.3, we have ignored possible errors that are introduced into the computation during the formation of the right-hand side b from the vectors r, x, and y, and the scalars σ and μ. Since the formation process introduces a relative perturbation of $O(\mathbf{u})$ into each component of b, we lose nothing by ignoring the perturbations.

We now turn to recovery of the step $\hat{v}$. From the exact formula (6b), we have

$$v_i = -y_i(1 + u_i/x_i) + \sigma\mu/x_i. \tag{26}$$

In the actual computation of $\hat{v}$, we have only the computed value $\hat{u}$ available to us. Moreover, errors are introduced when each of the five or six floating-point operations on the right-hand side of (26) are performed. The exact nature of these errors will depend on the order in which the operations in (26) are performed. Two possibilities are suggested by the parentheses in the expressions

$$-y_i[1 + (u_i/x_i)] + (\sigma\mu)/x_i, \qquad [-y_i - (y_i u_i)/x_i] + \sigma(\mu/x_i).$$

We can, however, perform an analysis that takes all the possibilities into account, as we show in the following theorem.

THEOREM 3.4. *Suppose the assumptions of Theorem* 3.3 *hold and that* $\hat{v}$ *is computed from the formula* (6b) (*equivalently,* (26)) *with* $\hat{u}$ *replacing* u. *Then we have*

$$\|\hat{v}_B - v_B\| = O(\mu(\mu + \mathbf{u})), \tag{27a}$$

$$\|\hat{v}_N - v_N\| = O(\mu + \mathbf{u}), \tag{27b}$$

and, for all $i \in B$,

$$v_i/y_i = -1 + O(\sigma + \mu), \tag{28a}$$

$$\hat{v}_i/y_i = -1 + O(\sigma + \mu + \mathbf{u}). \tag{28b}$$

Proof. In all the formulae of this proof, we use the notation δ_j, $j = 1, 2, \ldots$, to represent scalar quantities of order $\mathbf{u}$ (We certainly have $\delta_j \leq 10\mathbf{u}$ throughout the proof.) Recall that relative errors of $O(\mathbf{u})$ are incurred whenever a real number is approximated by a floating-point number and when an arithmetic operation involving two floating-point numbers is performed (cf. (9)). Therefore, regardless of the order in which the operations required to recover $\hat{v}_i$ are performed, we have

$$\hat{v}_i = \left[-y_i\left(1 + \delta_1 + \frac{\hat{u}_i}{x_i}(1 + \delta_2)\right)(1 + \delta_3) + \frac{\sigma\mu}{x_i}(1 + \delta_4)\right](1 + \delta_5). \tag{29}$$

By rearranging and combining terms in (29), we obtain

$$\begin{aligned}\hat{v}_i &= -y_i(1 + \delta_6) - \frac{y_i\hat{u}_i}{x_i}(1 + \delta_7) + \frac{\sigma\mu}{x_i}(1 + \delta_8) \\ &= -y_i - \frac{y_i u_i}{x_i} + \frac{\sigma\mu}{x_i} + \frac{y_i}{x_i}(u_i - \hat{u}_i) - \delta_6 y_i - \delta_7\frac{y_i\hat{u}_i}{x_i} - \delta_8\frac{\sigma\mu}{x_i}.\end{aligned}$$

By substituting from (26), we obtain

$$|v_i - \hat{v}_i| = \left|\frac{y_i}{x_i}\right||u_i - \hat{u}_i| + O(\mathbf{u})\left[|y_i| + \left|\frac{y_i\hat{u}_i}{x_i}\right| + \left|\frac{\sigma\mu}{x_i}\right|\right]. \tag{30}$$

Consider first the case of $i \in B$. From (30) together with Assumption 2, Theorem 3.2, expressions (15) and (23a), and $\delta_j = O(\mathbf{u})$, we have

$$|v_i - \hat{v}_i| = O(\mu(\mu + \mathbf{u})) + O(\mathbf{u})\,[O(\mu) + O(\mu(\mu + \mathbf{u})) + O(\sigma\mu)] = O(\mu(\mu + \mathbf{u})),$$

proving (27a). For $i \in N$, we have from Assumption 2, Theorem 3.2, and expressions (15) and (23b) that

$$\left|\frac{y_i}{x_i}\right| = O(\mu^{-1}), \qquad |u_i - \hat{u}_i| = O(\mu(\mu + \mathbf{u})), \qquad \left|\frac{y_i\hat{u}_i}{x_i}\right| = O(1), \qquad \left|\frac{\sigma\mu}{x_i}\right| = O(1).$$

Therefore we obtain from (30) that

$$|v_i - \hat{v}_i| = O(\mu + \mathbf{u}) \quad \forall i \in N,$$

as required.

The inequalities (28) are proved in the same way as (24). □

Gaussian elimination with complete pivoting is possibly more relevant to practical algorithms, since sparse elimination algorithms rearrange both rows and columns and hence can be regarded as approximations to the complete pivoting strategy. The main error results for complete pivoting are the same as those for partial pivoting. To justify this claim, we note first that the nonbasic indices will eventually be used as pivots before any of the basic indices are used, because of the large sizes of y_i/x_i, $i \in N$. Moreover, the error matrices E_{BN} and E_{NN} are $O(\mathbf{u})$ rather than $O(\mu^{-1}\mathbf{u})$, because the elements y_i/x_i, $i \in N$ cannot appear in a pivot row except on the diagonal, so they cannot "contaminate" other elements in the nonbasic columns of $M + X^{-1}Y$. In other words, $\hat{u}$ actually solves the system

$$\begin{bmatrix} M_{BB} + X_B^{-1}Y_B + E_{BB} & M_{BN} + E_{BN} \\ M_{NB} + E_{NB} & M_{NN} + X_N^{-1}Y_N + E_{NN} \end{bmatrix} \begin{bmatrix} \hat{u}_B \\ \hat{u}_N \end{bmatrix} = \begin{bmatrix} r_B - y_B - \sigma\mu X_B^{-1}e \\ r_N - y_N - \sigma\mu X_N^{-1}e \end{bmatrix}, \tag{31}$$

where

$$\|E_{BB}\| \le \rho\, O(\|M_{BB} + X_B^{-1}Y_B\| + \|M_{NB}\|)\mathbf{u} = O(\mathbf{u}), \tag{32a}$$

$$\|E_{NB}\| \le \rho\, O(\|M_{BB} + X_B^{-1}Y_B\| + \|M_{NB}\|)\mathbf{u} = O(\mathbf{u}), \tag{32b}$$

$$\|E_{BN}\| \le \rho\, O(\|M_{NN}\| + \|M_{BN}\|)\mathbf{u} = O(\mathbf{u}), \tag{32c}$$

$$\|E_{NN}\| \le \rho\, O(\|M_{NN}\| + \|M_{BN}\|)\mathbf{u} = O(\mathbf{u}). \tag{32d}$$

By defining G as the coefficient matrix in (31) and partitioning as before, we obtain after some manipulation that

$$(G^{-1})_{11} = M_{BB}^{-1} + O(\mu + \mathbf{u}), \qquad (G^{-1})_{12} = O(\mu),$$
$$(G^{-1})_{21} = O(\mu), \qquad (G^{-1})_{22} = O(\mu).$$

Therefore, using $\|b_B\| = O(\mu + \mathbf{u})$ and $\|b_N\| = O(1)$, we have

$$\hat{u}_B = (G^{-1})_{11}b_B + (G^{-1})_{12}b_N = O(\mu + \mathbf{u}),$$
$$\hat{u}_N = (G^{-1})_{21}b_B + (G^{-1})_{22}b_N = O(\mu),$$

which is the same error result as the one obtained for partial pivoting in Theorem 3.2. The other results in the section also hold for complete pivoting, with minor modifications to the proofs.

4. Effect of roundoff error on local convergence. We now consider the algorithm in [14], which can be described as follows. Given parameters $\gamma_{k+1} \in (\gamma_{\min}, \gamma_k]$ and $\beta_k \in [0, 1)$, the step α_k is chosen as

$$\alpha_k = \arg\min_{\alpha \in (0,1]} \mu_k(\alpha) \triangleq (x^k + \alpha u^k)^T (y^k + \alpha v^k)/n, \tag{33}$$

subject to

$$(x^k, y^k) + \alpha(u^k, v^k) \in \mathcal{N}(\gamma_{k+1}), \tag{34a}$$

$$\mu_k(\alpha) \geq (1-\alpha)(1-\beta_k)\mu_k, \qquad \text{if } r^k \neq 0 \tag{34b}$$

for all $\alpha \in [0, \alpha_k]$. The choices of σ_k, γ_{k+1}, and β_k at each iteration are made according to the following scheme.

given $\bar{\gamma} \in (0,1)$, $\gamma_{\min}$, $\gamma_{\max}$ with $0 < \gamma_{\min} < \gamma_{\max} < 1/2$, $\bar{\sigma} \in (0, 1/2)$,
$\rho \in (0, \bar{\gamma})$, and (x^0, y^0) with $x_i^0 y_i^0 \geq \gamma_{\max}\mu_0 > 0$;

$t_0 \leftarrow 1$, $\gamma_0 \leftarrow \gamma_{\max}$;

for $k = 0, 1, 2, \ldots$

if $\mu_k = 0$ **then** stop;
Find (u^k, v^k) and α_k from (11), (33), and (34) with
$\sigma_k = 0$, $\beta_k = \bar{\gamma}^{t_k}$, $\gamma_{k+1} = \gamma_{\min} + \bar{\gamma}^{t_k}(\gamma_{\max} - \gamma_{\min})$;
if $\mu_k(\alpha_k) \leq \rho\mu_k$
then accept this step; $t_{k+1} \leftarrow t_k + 1$; go to next k;
end if
Find (u^k, v^k) and α_k from (11), (33), and (34) with
$\sigma_k \in [\bar{\sigma}, 1/2]$, $\beta_k = 0$, $\gamma_{k+1} = \gamma_k$;
accept this step; $t_{k+1} \leftarrow t_k$; go to next k;

end for.

This algorithm takes two types of steps — "safe" steps, for which $\sigma_k \geq \bar{\sigma}$, and "fast" steps, for which $\sigma_k = 0$. Theoretically, the safe steps ensure good global convergence properties and complexity, while the fast steps ensure asymptotic superlinear convergence. The counter t_k keeps track of the number of fast steps taken prior to iteration k.

The choice of step length α_k ensures that $\|r^{k+1}\| = O(\mu_{k+1})$. To see this, note from condition (34b) that

$$\mu_{k+1} = (1-\alpha_k)(1-\beta_k)\mu_k = \left[\prod_{j=0}^{k}(1-\alpha_j)(1-\beta_j)\right]\mu_0.$$

Since

$$\prod_{j=0}^{k}(1-\beta_j) \geq \prod_{j=0}^{\infty}(1-\bar{\gamma}^j) = \hat{\beta} > 0, \qquad r^{k+1} = \prod_{j=0}^{k}(1-\alpha_j)r^0,$$

we have

$$\mu_{k+1}/\mu_0 \geq \hat{\beta}\|r^{k+1}\|/\|r^0\|,$$

so condition (12) holds with $K = \hat{\beta}^{-1}$.

We focus on this algorithm because of its strong theoretical properties, namely, global convergence from any positive starting point (x^0, y^0), polynomial complexity

when properly initialized, and superlinear local convergence. Also, the method performs well in computational tests and is quite similar (at least in its "nonsuperlinear" phase where $\sigma_k \geq \bar{\sigma}$) to the algorithm implemented by Lustig, Marsten, and Shanno [8]. We assume throughout that finite termination does not occur, that is, the algorithm generates an infinite sequence of strictly positive iterates (x^k, y^k).

In this section, we examine how the behavior of this algorithm is affected when the computed steps $(\hat{u}^k, \hat{v}^k)$ are used in place of the exact steps (u^k, v^k). We start by showing that near-unit steplengths can eventually be taken by this algorithm without violating the positivity condition $(x^k, y^k) > 0$. Consequently, there exists the possibility of rapid convergence of the sequence of complementarity gaps μ_k to zero, even in the presence of roundoff error. We refine the results to show that for the safe steps $(\sigma_k \geq \bar{\sigma})$, we actually have $\alpha_k = 1$ when μ_k is sufficiently small.

In all the analysis below, our convention is to use the iteration index k in the statement of each result, but omit it in the proofs.

LEMMA 4.1. *For all k sufficiently large, we have*

$$(x^k + \alpha \hat{u}^k, y^k + \alpha \hat{v}^k) \geq 0$$

for all $\alpha \in [0, \bar{\alpha}_k]$, where

$$|1 - \bar{\alpha}_k| = O(\sigma_k + \mu_k + \mathbf{u}). \tag{35}$$

Proof. We consider first the indices $i \in N$. From (27b), we have

$$|\hat{v}_i| = |v_i| + O(\mu + \mathbf{u}) = O(\mu + \mathbf{u}),$$

while from Assumption 2 we have for large k that $y_i^k \approx y_i^* > 0$. Hence $y_i^k + \alpha \hat{v}_i^k > 0$ for all $\alpha \in [0, 1]$ and all k sufficiently large. On the other hand, we have from (24b) that

$$x_i + \alpha \hat{u}_i = x_i + \alpha x_i(-1 + O(\sigma + \mu + \mathbf{u})).$$

Therefore, if $x_i + \alpha_i \hat{u}_i = 0$ for some index i, then we must have

$$1 - \alpha + \alpha O(\sigma + \mu + \mathbf{u}) = 0 \;\Rightarrow\; |1 - \alpha| = O(\sigma + \mu + \mathbf{u}).$$

Hence $x_i + \alpha \hat{u}_i \geq 0$ for all $\alpha \in [0, \bar{\alpha}]$, for $\bar{\alpha}$ satisfying (35).

The case of $i \in B$ is proved in a similar way by using (28b). □

We now show that near-unit steps produce fast linear convergence of the complementarity gap to zero.

THEOREM 4.2. *If k is sufficiently large, then*

$$(x^k + \alpha_k \hat{u}^k)^T (y^k + \alpha_k \hat{v}^k) = [1 - \alpha_k(1 - \sigma_k) + O(\mu_k + \mathbf{u})](x^k)^T y^k.$$

Proof. For any $i = 1, \ldots, n$, we have from the second part of (4) that

$$\begin{aligned} &(x_i + \alpha \hat{u}_i)(y_i + \alpha \hat{v}_i) \\ &= x_i y_i + \alpha y_i u_i + \alpha x_i v_i + \alpha y_i(\hat{u}_i - u_i) + \alpha x_i(\hat{v}_i - v_i) + \alpha^2 \hat{u}_i \hat{v}_i \\ &= (1 - \alpha) x_i y_i + \alpha \sigma \mu + \alpha y_i(\hat{u}_i - u_i) + \alpha x_i(\hat{v}_i - v_i) + \alpha^2 \hat{u}_i \hat{v}_i. \end{aligned} \tag{36}$$

Now, by Assumption 2 and relations (15a) and (27), we have

$$\begin{aligned} i \in N &\Rightarrow |x_i(\hat{v}_i - v_i)| = O(\mu) O(\mu + \mathbf{u}), \\ i \in B &\Rightarrow |x_i(\hat{v}_i - v_i)| = O(1) O(\mu(\mu + \mathbf{u})). \end{aligned}$$

A similar result holds for $|y_i(\hat{u}_i - u_i)|$. For the last term in (36), we have from (14), (27), and Theorem 3.2 that

$$i \in B \Rightarrow |\hat{u}_i \hat{v}_i| \leq O(\mu + \mathbf{u})(|v_i| + |\hat{v}_i - v_i|) = O(\mu(\mu + \mathbf{u})),$$
$$i \in N \Rightarrow |\hat{u}_i \hat{v}_i| = O(\mu)(|v_i| + |\hat{v}_i - v_i|) = O(\mu(\mu + \mathbf{u})).$$

We also have $\mu/(x_i y_i) \leq 1/\gamma_{\min} = O(1)$. Using these estimates in (36), we obtain

$$\begin{aligned}(x_i + \alpha\hat{u}_i)(y_i + \alpha\hat{v}_i) &= x_i y_i(1-\alpha) + \alpha\sigma\mu + \alpha O(\mu(\mu+\mathbf{u})) \\ &= x_i y_i(1-\alpha) + \alpha\mu\left[\sigma + O(\mu+\mathbf{u})\right]. \end{aligned} \tag{37}$$

By summing over i, we obtain

$$(x + \alpha\hat{u})^T(y + \alpha\hat{v}) = x^T y\left[(1 - \alpha + \alpha\sigma) + \alpha O(\mu + \mathbf{u})\right], \tag{38}$$

which yields the desired result. □

We now examine the safe steps, for which $\sigma_k \geq \bar{\sigma}$, and show that $\alpha_k = 1$ satisfies the criteria (33) and (34) for large enough k, even when the computed search direction $(\hat{u}^k, \hat{v}^k)$ is used in place of the exact direction (u^k, v^k). That is, a unit step is taken.

THEOREM 4.3. *Suppose that $\bar{\sigma}$ is substantially larger than $\mathbf{u}$, in a sense to be defined below. Then for all sufficiently large k, if a safe step (with $\sigma_k \geq \bar{\sigma}$, $\gamma_{k+1} = \gamma_k$, and $\beta_k = 0$) is computed, the step length parameter satisfying* (33) *and* (34) *will be* $\alpha_k = 1$.

Proof. From (38), we have

$$\frac{(x + \alpha\hat{u})^T(y + \alpha\hat{v})}{x^T y} = (1 - \alpha) + \alpha\left[\sigma + O(\mu + \mathbf{u})\right]. \tag{39}$$

Therefore (34b) will hold for all $\alpha \in [0,1]$ (with (u,v) replaced by $(\hat{u},\hat{v})$ and $\beta_k = 0$), provided that the term in square brackets in (39) is nonnegative. But nonnegativity is guaranteed for $\mathbf{u} \ll \bar{\sigma}$ and μ sufficiently small, so $\alpha_k = 1$ satisfies this inequality.

Consider now (34a), with $\gamma_{k+1} = \gamma_k$. From (37) and (38), we have

$$\begin{aligned}\frac{(x_i + \alpha\hat{u}_i)(y_i + \alpha\hat{v}_i)}{(x + \alpha\hat{u})^T(y + \alpha\hat{v})/n} &= \frac{(1-\alpha)x_i y_i + \alpha\mu[\sigma + O(\mu+\mathbf{u})]}{(1-\alpha)\mu + \alpha\mu[\sigma + O(\mu+\mathbf{u})]} \\ &\geq \frac{x_i y_i/\mu + [\sigma - C_{10}(\mu+\mathbf{u})]\alpha/(1-\alpha)}{1 + [\sigma + C_{11}(\mu+\mathbf{u})]\alpha/(1-\alpha)} \end{aligned} \tag{40}$$

for some positive constants C_{10} and C_{11}. Since $x_i y_i \geq \gamma_k \mu$ for all $i = 1, \ldots, n$, we find that (34a) is satisfied if

$$\frac{\sigma - C_{10}(\mu + \mathbf{u})}{\sigma + C_{11}(\mu + \mathbf{u})} \geq \gamma_{\max},$$

or, equivalently,

$$\frac{\mu + \mathbf{u}}{\sigma} \leq \frac{1 - \gamma_{\max}}{C_{10} + C_{11}}.$$

This last inequality, and therefore (34a), holds provided that μ and $\mathbf{u}$ are small enough with respect to $\bar{\sigma}$, as we have assumed.

Finally, we show that

$$\hat{\mu}(\alpha) \triangleq (x + \alpha\hat{u})^T(y + \alpha\hat{v})/n$$

is decreasing on the interval $\alpha \in [0, 1]$. Since from (38), we have

$$\frac{\hat{\mu}(\alpha)}{\mu} = 1 - \alpha + \alpha\sigma + \alpha O(\mu + \mathbf{u}),$$

the derivative $\hat{\mu}'(\alpha)$ is nonpositive provided that

$$-1 + \sigma + O(\mu + \mathbf{u}) \leq 0. \tag{41}$$

Since $\bar{\sigma} \leq \sigma \leq 1/2$ and, by our assumptions in the first part of this proof, $\bar{\sigma}$ dominates the $O(\mu + \mathbf{u})$ term, we have that (41) holds.

We have shown that all $\alpha \in [0, 1]$ satisfy the conditions (34) with (u, v) replaced by $(\hat{u}, \hat{v})$. Moreover, the function in (33) is decreasing over this interval. We conclude that $\alpha_k = 1$ is the step chosen by the line search procedure, giving the result. □

We turn now to fast step, for which $\sigma_k = 0$, $\beta_k = \bar{\gamma}^{t_k}$, and $\gamma_{k+1} = \gamma_{\min} + \bar{\gamma}^{t_k}(\gamma_{\max} - \gamma_{\min})$. The (exact) analysis in Wright [14] shows that fast steps are eventually always taken by the algorithm. Note that if the fast step is accepted, we have

$$\gamma_k - \gamma_{k+1} = (\bar{\gamma}^{t_k - 1} - \bar{\gamma}^{t_k})(\gamma_{\max} - \gamma_{\min}) = \bar{\gamma}^{t_k}(\bar{\gamma}^{-1} - 1)(\gamma_{\max} - \gamma_{\min}) = \Omega(\beta_k). \tag{42}$$

The following theorem gives an estimate for the length of a fast step.

THEOREM 4.4. *If a fast step is attempted at iteration k, where k is sufficiently large and $\mathbf{u} \ll 1$, then*

$$\alpha_k \geq \left[1 + \eta_k \frac{\mu_k + \mathbf{u}}{\bar{\gamma}^{t_k}}\right]^{-1},$$

where η_k satisfies $0 \leq \eta_k \leq O(1)$.

Proof. As in (37), we have for $\sigma_k = 0$ that

$$(x_i + \alpha\hat{u}_i)(y_i + \alpha\hat{v}_i) = x_i y_i(1 - \alpha) + \mu\alpha O(\mu + \mathbf{u}) \geq \gamma_k\mu(1 - \alpha) - \mu\alpha(\mu + \mathbf{u})\eta_k^{(1)}, \tag{43}$$

where $0 \leq \eta_k^{(1)} \leq O(1)$. From (38) we have

$$(x + \alpha\hat{u})^T(y + \alpha\hat{v})/n = \mu\left[1 - \alpha + \alpha O(\mu + \mathbf{u})\right] \leq \mu\left[1 - \alpha + \alpha(\mu + \mathbf{u})\eta_k^{(2)}\right], \tag{44}$$

where $0 \leq \eta_k^{(2)} \leq O(1)$. Putting (43) and (44) together, we deduce that (34a) holds provided that

$$\gamma_k(1 - \alpha) - \alpha(\mu + \mathbf{u})\eta_k^{(1)} \geq \gamma_{k+1}[1 - \alpha + \alpha(\mu + \mathbf{u})\eta_k^{(2)}],$$

which in turn is true if

$$(\gamma_k - \gamma_{k+1})(1 - \alpha) \geq \alpha(\mu + \mathbf{u})(\eta_k^{(1)} + \eta_k^{(2)}).$$

This last inequality is implied by the following bound on α:

$$\alpha \leq \left[1 + \frac{(\mu + \mathbf{u})(\eta_k^{(1)} + \eta_k^{(2)})}{\gamma_k - \gamma_{k+1}}\right]^{-1} = \left[1 + \frac{(\mu + \mathbf{u})(\eta_k^{(1)} + \eta_k^{(2)})}{\bar{\gamma}^{t_k}(\bar{\gamma}^{-1} - 1)(\gamma_{\max} - \gamma_{\min})}\right]^{-1}, \tag{45}$$

where we have used (42) to derive the final inequality.

Using (44) again, we have

$$(x+\alpha\hat{u})^T(y+\alpha\hat{v})/n = \mu\left[1-\alpha+\alpha O(\mu+\mathbf{u})\right] \geq \mu\left[1-\alpha-\alpha(\mu+\mathbf{u})\eta_k^{(3)}\right],$$

where $0 \leq \eta_k^{(3)} \leq O(1)$, so the inequality (34b) is satisfied when

$$1-\alpha-\alpha(\mu+\mathbf{u})\eta_k^{(3)} \geq (1-\alpha)(1-\beta_k),$$

that is, when

$$\alpha \leq \left[1+\frac{(\mu+\mathbf{u})\eta_k^{(3)}}{\beta_k}\right]^{-1} = \left[1+\frac{(\mu+\mathbf{u})\eta_k^{(3)}}{\bar{\gamma}^{t_k}}\right]^{-1}. \tag{46}$$

Providing we can show that $\hat{\mu}_k(\alpha)$ is decreasing on $[0,1]$, we have from (45) and (46) that the result holds for η_k defined by

$$\eta_k = \max\left(\eta_k^{(3)}, \frac{\eta_k^{(1)}+\eta_k^{(2)}}{(\bar{\gamma}^{-1}-1)(\gamma_{\max}-\gamma_{\min})}\right).$$

However,

$$\hat{\mu}_k'(\alpha) \leq -\mu\left[1-O(\mu+\mathbf{u})\right] < 0,$$

so $\hat{\mu}_k(\alpha)$ is certainly decreasing on $[0,1]$, and the result is proved. □

This result accurately indicates the behavior of fast steps on later iterations of the algorithm. The quantity η_k is typically either extremely small or else quite significant (that is, $\eta_k = \Omega(1)$), depending on the sign of certain products such as $\hat{u}^T\hat{v}$, $\hat{u}_i\hat{v}_i$, and so on. When η_k is tiny and $\mu_k + \mathbf{u} \ll \bar{\gamma}^{t_k}$, the value of α_k is very close to 1, and the fast step is accepted with a large reduction in μ. When η_k is larger, or when $\bar{\gamma}^{t_k} = O(\mathbf{u})$, the fast step may not lead to a very large decrease in μ and may even be rejected in favor of a safe step.

We summarize the results of this section in the following theorem.

THEOREM 4.5. *Suppose that* $\mathbf{u}$ *is much smaller than* $\bar{\sigma}$, *in the sense of Theorem 4.3. Then for all sufficiently large* k *we have that either*

(i) *a fast step is taken, and*

$$\mu_{k+1} \leq \rho\mu_k, \tag{47}$$

with

$$\mu_{k+1} = O\left(\frac{\mu_k+\mathbf{u}}{\bar{\gamma}^{t_k}}\right)\mu_k, \tag{48}$$

or

(ii) *a safe step is taken, with*

$$\mu_{k+1} = [\sigma_k + O(\mu_k+\mathbf{u})]\mu_k. \tag{49}$$

Proof. The condition for fast-step acceptance yields (47). The estimate (48) follows from Theorems 4.2 and 4.4 and the identity $\sigma_k = 0$. The safe-step estimate (49) follows from Theorem 4.2 when we use $\alpha_k = 1$. □

TABLE 1
Convergence of the algorithm: Problem type (i), $n = n_r = 20$.

k	$\log_{10}\mu_k$	$\log_{10}\|r^k\|_1$	Fast step?
0	4.0	4.7	
1	3.7	4.3	
2	3.2	3.3	
3	2.7	-11.7	
⋮	⋮	⋮	
15	-3.7	-13.4	
16	-4.8	-13.3	
17	-5.7	-13.2	*
18	-7.0	-13.1	*
19	-9.5	-13.2	*
20	-14.5	-13.5	*
21	-23.8	-13.4	terminate

5. Computational results. The algorithm of §4 was implemented in double-precision Fortran, using the LAPACK routines dgetrf and dgetrs to solve the linear system (6a). Our test problems are of two types.

(i) The matrix M has the form $M = ADA^T$, where A is $n \times n_r$ dense with all elements drawn from a uniform distribution on $[-1, 1]$, while D is diagonal with diagonal elements of the form 10^τ, where τ is drawn from a uniform distribution on $[0, 1]$. Since we choose $n_r \le n$, the rank of M is n_r. (Rank-deficiency of M is not an artificial feature; in certain applications of (1), including (3), M is structurally rank-deficient.) The solutions x^* and y^* are chosen so that the even-numbered components of x^* and the odd-numbered components of y^* are zero, while the remaining components are uniform on $[0, 1]$.

(ii) The matrix M has the form (3a), where the matrix A is dense with elements of the form $\tau_1 10^{\tau_2}$, where τ_1 and τ_2 are drawn from uniform distributions on $[-\frac{1}{2}, \frac{1}{2}]$ and $[0, 1]$, respectively. If p and m denote the dimensions of z and λ, respectively, we choose the even-numbered components $2, 4, \ldots, \min(p, m)$ of both z^* and λ^* to be nonzero. (It is a consequence of nondegeneracy of the solution of (2) that the same number of components of z^* and λ^* be nonzero. This requirement is also necessary for nonsingularity of M_{BB}.) The nonzero components of z^*, λ^* and the complementary vector pair in (3) are all drawn from $[0, 1]$.

We use the following values for the constants:

$$\gamma_{\min} = 10^{-5}, \qquad \gamma_{\max} = 10^{-2}, \qquad \rho = \bar{\gamma}/2, \qquad \bar{\gamma} = 10^{-1}. \tag{50}$$

We choose the centering parameter σ_k at each safe iteration according to the formula

$$\sigma_k = \text{mid}(\bar{\sigma}, \mu_k/\sqrt{n}, .2), \qquad \text{where } \bar{\sigma} = .01. \tag{51}$$

Though we made no special effort to tune these constants to their optimal values, our experience indicates that the choices (50), (51) are efficient for these and other types of problems.

In Tables 1–4 we tabulate the behavior of the algorithm of §4. Many of the uninteresting safe iterates are omitted. An asterisk in the last column indicates that a fast step was taken from this iterate. We terminate the algorithm when μ falls below 10^{-20}.

These tables indicate rapid convergence of the algorithm during its final stages. Typically, the algorithm takes only fast steps after it has decreased μ below a certain

TABLE 2
Convergence of the algorithm: Problem type (i), $n = n_r = 100$.

k	$\log_{10} \mu_k$	$\log_{10} \|r^k\|_1$	Fast step?
0	4.9	6.5	
1	4.6	6.2	
2	4.1	5.5	
3	3.8	4.9	
4	3.3	-9.8	
⋮	⋮	⋮	
21	-5.8	-11.6	
22	-6.7	-11.6	
23	-7.8	-11.6	
24	-9.0	-11.6	*
25	-10.9	-11.5	*
26	-14.7	-11.6	*
27	-22.1	-11.8	terminate

TABLE 3
Convergence of the algorithm: Problem type (i), $n = 100$, $n_r = 60$.

k	$\log_{10} \mu_k$	$\log_{10} \|r^k\|_1$	Fast step?
0	4.5	6.1	
1	4.3	5.8	
2	4.1	5.5	
3	3.7	5.0	
4	3.2	3.9	
5	2.8	-9.5	
⋮	⋮	⋮	
24	-2.8	-11.8	
25	-3.5	-11.7	
26	-4.4	-11.7	
27	-5.6	-11.7	
28	-6.7	-11.7	*
29	-8.1	-11.7	*
30	-10.1	-11.8	*
31	-13.8	-11.8	*
32	-20.8	-11.8	terminate

threshold. (The experience of the author and others indicates that this threshold is quite small for linear and quadratic problems, so that superlinear convergence does not set in until quite late in the process. Preliminary experience with nonlinear problems indicates that fast steps are typically taken at an earlier stage, that is, the threshold is not so small.)

The behavior observed in Tables 1–4 certainly confirms the efficacy of Gaussian elimination with partial pivoting in the context of this interior-point method. The linear algebra continues to produce good steps even when μ is extremely small. The convergence of μ to zero appears to be superlinear in each case (even quadratic, in the case of Table 4). These tables do not, however, show the asymptotic behavior suggested by Theorem 4.5. To see it, we must continue to run the algorithm past the point of convergence. Table 5 shows what happens when we continue to iterate on the problem of Table 3 until μ is reduced below 10^{-100}. (The late asymptotic convergence was qualitatively similar on all the problems we tried, so we report just this one instance.) Note that fast steps are taken on each iteration with decrease

TABLE 4
Convergence of the algorithm: Problem type (ii), $n = 200$, *matrix* A *is* 40×160.

k	$\log_{10} \mu_k$	$\log_{10} \|r^k\|_1$	Fast step?
0	4.0	5.2	
1	3.8	4.9	
2	3.5	4.7	
3	3.1	4.2	
4	2.1	3.1	
5	1.7	2.6	
6	1.4	2.2	
7	1.0	1.7	
8	0.5	1.2	
9	0.2	0.8	
10	-0.7	-0.3	
11	-2.1	-1.8	*
12	-4.4	-4.1	*
13	-8.3	-7.9	*
14	-15.6	-10.2	*
15	-28.2	-10.2	terminate

TABLE 5
Later iterates on the problem of Table 3.

k	$\log_{10} \mu_k$	$\log_{10} \|r^k\|_1$	Fast step?
⋮	⋮	⋮	
32	-20.8	-11.8	*
33	-31.4	-11.9	*
34	-42.1	-11.9	*
35	-46.8	-11.9	*
36	-50.9	-11.8	*
37	-61.7	-11.9	*
38	-71.9	-11.9	
39	-73.9	-11.9	*
40	-78.4	-11.9	*
41	-90.5	-11.8	*
42	-102.4	-11.9	terminate

factors between 10^{-4} and 10^{-12}, except for one iteration — the 38th — on which a safe step is taken with a decrease ratio of almost exactly $\sigma_k = 10^{-2}$. The existence of these two kinds of steps and their effects on μ_k are in close accord with the predictions of Theorem 4.5.

Note that in all the tables the residual norm $\|r_k\|$ decreases to $O(\mathbf{u})$ but no further. As discussed in the proof of theorem 3.2, this behavior is due to roundoff error in the calculation of r^k via the formula $r^k = y^k - Mx^k - q$.

We experimented with a version of the code in which a modified complete pivoting strategy was used for solving (6a). The columns of the coefficient matrix were ordered by decreasing value of $\|\cdot\|_\infty$ before Gaussian elimination with partial pivoting was applied. Asymptotically, this strategy has the effect of ordering the nonbasic columns first, so the analysis at the end of §3 still applies. As predicted in that analysis, this version of the algorithm behaves only slightly differently from the partial pivoting version described above.

The assumption that M_{BB} is nonsingular (indeed, well conditioned) plays an important role in the analysis of §§3 and 4. Theoretically, the algorithm of §4 is

TABLE 6
Convergence in the case of M_{BB} rank-deficient: Problem type (i), $n = 100$, $n_r = 25$.

k	$\log_{10} \mu_k$	$\log_{10} \|r^k\|_1$	Fast step?
0	4.2	5.9	
1	4.0	5.5	
2	3.6	5.1	
3	2.7	4.2	
4	2.1	3.4	
⋮	⋮	⋮	
11	-0.2	0.2	
12	-1.2	-0.9	*
13	-3.7	-3.5	*
14	-7.3	-7.0	
15	-7.4	-7.1	
16	-7.4	-7.2	
17	-7.6	-7.4	
⋮	⋮	⋮	
98	-8.4	-8.2	
99	-8.5	-8.2	
100	-8.5	-8.3	
⋮	⋮	⋮	

known to have fast local convergence even when M_{BB} is singular and the solution is not unique. We tested to see whether fast convergence was attainable in practice by forming a problem from the class (i) with $n_r = 25$. Since B contains 50 indices, the submatrix M_{BB} is certainly rank deficient. The result of this run is summarized in Table 6. It is clear that the behavior indicated in Theorem 4.5 does not occur. After taking two fast steps and converging to $\mu_k \approx 10^{-7}$ by iteration 14, the algorithm stalls and makes very little progress from that point on. This and other similar examples suggest that the assumption of M_{BB} nonsingular probably cannot be relaxed.

Acknowledgment. I thank the editor and referees for their insightful comments on an earlier draft.

REFERENCES

[1] R. FOURER AND S. MEHROTRA, *Solving symmetric indefinite systems in an interior-point method for linear programming*, Math. Programming, 62 (1993), pp. 15–39.

[2] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, 1989.

[3] N. J. HIGHAM, *Accuracy and Stability of Numerical Algorithms (provisional title)*, 1994, manuscript.

[4] J. JI, F. A. POTRA, AND S. HUANG, *A predictor-corrector method for linear complementarity problems with polynomial complexity and superlinear convergence*, Tech. Report 18, Department of Mathematics, University of Iowa, Iowa City, August 1991.

[5] M. KOJIMA, Y. KURITA, AND S. MIZUNO, *Large-step interior point algorithms for linear complementarity problems*, SIAM J. Optim., 3 (1993), pp. 398–412.

[6] M. KOJIMA, S. MIZUNO, AND A. YOSHISE, *An $O(\sqrt{n}L)$ iteration potential reduction algorithm for linear complementarity problems*, Math. Programming, 50 (1991), pp. 331–342.

[7] I. J. LUSTIG, R. E. MARSTEN, AND D. F. SHANNO, *Computational experience with a primal-dual interior point method for linear programming*, Linear Algebra Appl., 152 (1991), pp. 191–222.

[8] ———, *Computational experience with a globally convergent primal-dual predictor-corrector algorithm for linear programming*, Tech. Report SOR 92–10, Program in Statistics and

Operations Research, Princeton University, Princeton, NJ, 1992.

[9] S. MEHROTRA, *On the implementation of a primal-dual interior point method*, SIAM J. Optim., 2 (1992), pp. 575–601.

[10] R. D. C. MONTEIRO AND I. ADLER, *Interior path-following primal-dual algorithms. part* II: *Convex quadratic programming*, Math. Programming, 44 (1989), pp. 43–66.

[11] D. B. PONCELEÓN, *Barrier methods for large-scale quadratic programming*, Ph.D. thesis, Stanford University, Stanford, CA, 1990.

[12] F. A. POTRA, *An o(nl) infeasible-interior-point algorithm for* LCP *with quadratic convergence*, Report on Computational Mathematics 50, Department of Mathematics, University of Iowa, Iowa City, January 1994.

[13] R. J. VANDERBEI, *LOQO User's Manual*, Tech. Report SOR 92–5, Program in Statistics and Operations Research, Princeton University, Princeton, NJ, 1992.

[14] S. J. WRIGHT, *A path-following interior-point algorithm for linear and quadratic optimization problems*, Preprint MCS–P401–1293, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, December 1993; Ann. Oper. Res., to appear.

[15] ———, *An infeasible-interior-point algorithm for linear complementarity problems*, Math. Programming, 67 (1994), pp. 29–52.

[16] ———, *Stability of linear algebra computations in interior-point methods for linear programming*, Preprint MCS–P446–0694, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, June 1994.

[17] X. XU, P. HUNG, AND Y. YE, *A simplified homogeneous and self-dual linear programming algorithm and its implementation*, September 1993, manuscript.

[18] Y. YE AND K. ANSTREICHER, *On quadratic and* $O(\sqrt{n}L)$ *convergence of a predictor-corrector algorithm for* LCP, Math. Programming, Series A, 62 (1993), pp. 537–551.

[19] Y. ZHANG, *On the convergence of a class of infeasible-interior-point methods for the horizontal linear complementarity problem*, SIAM J. Optim., 4 (1994), pp. 208–227.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, pp. 1308 1327, October 1995

019

THE ALGEBRAIC RICCATI EQUATION AND INEQUALITY FOR SYSTEMS WITH UNCONTROLLABLE MODES ON THE IMAGINARY AXIS *

CARSTEN W. SCHERER†

Abstract. If $(\mathcal{A}, \mathcal{B})$ is stabilizable, one pretty well knows algebraic conditions for the solvability and for the existence of largest solutions of the algebraic Riccati equation and inequality

$$\mathcal{A}^*\mathcal{X} + \mathcal{X}\mathcal{A} - \mathcal{X}\mathcal{B}\mathcal{B}^*\mathcal{X} + \mathcal{Q} = 0 \quad \text{and} \quad \mathcal{A}^*\mathcal{X} + \mathcal{X}\mathcal{A} - \mathcal{X}\mathcal{B}\mathcal{B}^*\mathcal{X} + \mathcal{Q} \geq 0,$$

which leads to immediate existence results for positive definite solutions. In this paper we work out how far these properties may be generalized if $(\mathcal{A}, \mathcal{B})$ could have *uncontrollable modes on the imaginary axis.* Since the relations of the equation and inequality are not as tight any more, we provide separate conditions for the existence of Hermitian or positive definite solutions and give a detailed discussion how to verify them. As auxiliary steps we discuss various new aspects for the corresponding Lyapunov equation/inequality and a complete solvability test for the quadratic equation

$$X^*RX + SX + (SX)^* + T = 0$$

with Hermitian R and T. Finally, we briefly point out the consequences of our results for the general state-feedback H_∞-control problem at optimality.

Key words. algebraic Riccati equation, algebraic Riccati inequality, positive definite solutions, ordering of solutions, state-feedback H_∞-optimal control

AMS subject classifications. 15A06, 15A24, 15A39, 15A45, 93C05, 93C45

1. Introduction.

We study the algebraic Riccati equation (ARE)

$$\mathcal{A}^*\mathcal{X} + \mathcal{X}\mathcal{A} - \mathcal{X}\mathcal{B}\mathcal{B}^*\mathcal{X} + \mathcal{Q} = 0 \tag{1}$$

and the related nonstrict algebraic Riccati inequality (ARI)

$$\mathcal{A}^*\mathcal{X} + \mathcal{X}\mathcal{A} - \mathcal{X}\mathcal{B}\mathcal{B}^*\mathcal{X} + \mathcal{Q} \geq 0, \tag{2}$$

where $\mathcal{A}$ is an $n \times n$, $\mathcal{B}$ an $n \times m$, and $\mathcal{Q}$ an $n \times n$ Hermitian complex matrix. We call $\mathcal{X}$ a *solution* of the ARE or ARI if it is a complex *Hermitian* $n \times n$-matrix satisfying (1) or (2).

Let us briefly recall the relation of and the algebraic solvability criteria for the ARE and ARI if $(\mathcal{A}, \mathcal{B})$ is stabilizable [1], [3], [4], [7], [10], [11], [13], [15], [24], [25], [27].

(a) If the ARI has a solution $\mathcal{X}$ then there exists a *unique* solution $\tilde{\mathcal{X}}$ of the ARE such that $\mathcal{A} - \mathcal{B}\mathcal{B}^*\tilde{\mathcal{X}}$ has all its eigenvalues in the *closed left half-plane.*

* Received by the editors July 23, 1992; accepted for publication (in revised form) by G. Cybenko September 15, 1994.

†Mechanical Engineering Systems and Control Group, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands (scherer@tudw03.tudelft.nl). This work was supported by Deutsche Forschungsgemeinschaft Sche 402/1-1. This paper was conducted while the author was affiliated with the Mathematical Institute of the University of Würzburg, Germany. The work was performed while the author was visiting the EECS Department of the University of Michigan, Ann Arbor, Michigan.

(b) If $\mathcal{X}$ satisfies the ARE or the ARI then $\mathcal{X} \leq \tilde{\mathcal{X}}$.

(c) The ARE or the ARI have solutions if and only if (iff) the Jordan blocks corresponding to the eigenvalues of the Hamiltonian matrix

$$\mathcal{H} := \begin{pmatrix} \mathcal{A} & -\mathcal{B}\mathcal{B}^* \\ -\mathcal{Q} & -\mathcal{A}^* \end{pmatrix}$$

on the imaginary axis have *even size*. If existent, the solution $\tilde{\mathcal{X}}$ can be computed using well-defined generalized eigenvectors of $\mathcal{H}$.

Item (a) implies that the solvability of the ARE and the ARI is equivalent. Moreover, in case of solvability, there exists a unique particular solution of the ARE satisfying a certain spectral condition. The existence of this solution can be checked by identifying the Jordan structure of the Hamiltonian matrix $\mathcal{H}$, and then computing it algebraically. Finally, this particular solution has an additional distinguishing property of being the *largest* element of the solution set of both the ARE and the ARI. These relations reveal that the ARE or the ARI has a positive definite solution iff the largest solution $\tilde{\mathcal{X}}$ of (1) exists and is positive definite and this condition is verifiable.

We recall what is known under weaker assumptions. If $(\mathcal{A}, \mathcal{B})$ is sign-controllable (i.e., the set of uncontrollable modes and its reflection on the imaginary axis are disjoint) then (c) persists to hold [3], [4], [25], [26]. If $(\mathcal{A}, \mathcal{B})$ has no uncontrollable modes on the imaginary axis, then the same algebraic condition characterizes the existence of a Hermitian solution of the ARI [5]. In this generality, the existence of positive definite solutions is not characterized in the literature.

Obviously, all these hypotheses explicitly exclude uncontrollable modes of $(\mathcal{A}, \mathcal{B})$ on the imaginary axis. Our main interest in this paper concerns the problem of how far the above results can be generalized to systems $(\mathcal{A}, \mathcal{B})$ that may have uncontrollable modes on the imaginary axis. More precisely, we assume that

$(\mathcal{A}, \mathcal{B})$ has all its uncontrollable modes in the *closed left half-plane*,

which is, of course, *necessary* for the existence of a solution of the ARE as in (a).

Under this assumption, the solvability of the ARE and the ARI is not as tightly related as for a stabilizable system. If we choose $\mathcal{A} = 0, \mathcal{B} = 0$ then (2) is solvable for any $\mathcal{Q} \geq 0$ whereas (1) is not solvable if $\mathcal{Q}$ is nonzero. Hence the solvability of the ARE and the ARI is generally not equivalent, which leads us to consider the ARE and the ARI separately.

In §2 we provide generalizations of (a) and (b) for the ARE and give a detailed discussion how to use this nontrivial result to verify the existence of Hermitian and positive definite solutions. For the verification procedure it is required to characterize the existence of a positive definite solution of the Lyapunov equation $\mathcal{A}^*\mathcal{X} + \mathcal{X}\mathcal{A} + \mathcal{Q} = 0$, where $\mathcal{A}$ has all its eigenvalues on the imaginary axis. Moreover, we need to check the existence of a not necessarily Hermitian X satisfying the quadratic equation $X^*RX + SX + X^*S^* + T = 0$ (R, T Hermitian). Complete solvability criteria for this equation and the corresponding inequality are developed in §3. As a preliminary step for the ARI, we discuss in §4 new necessary and sufficient conditions for the existence of arbitrarily large solutions of the nonstrict Lyapunov inequality $\mathcal{A}^*\mathcal{X} + \mathcal{X}\mathcal{A} + \mathcal{Q} \geq 0$, again assuming that $\mathcal{A}$ has all its eigenvalues on the imaginary axis. Finally, §5 contains separate necessary and sufficient conditions for the existence of (positive definite) solutions of the ARI together with a detailed investigation of verifiability and a discussion of when the gap between necessity and sufficiency disappears. Two technical proofs are deferred to the Appendix.

Apart from the interest in providing further steps towards solvability criteria for the general Riccati equation or inequality without any restriction on $(\mathcal{A}, \mathcal{B})$, our results do have immediate applications to the general state-feedback H_∞-control problem at optimality. We have proved in [16] and [19] that the optimal value is attained iff the ARI (2) has a positive definite solution where $\mathcal{A}, \mathcal{B}$, and $\mathcal{Q}$ may be directly determined from the data matrices and the optimal value. We stress that, by construction, the pair $(\mathcal{A}, \mathcal{B})$ precisely satisfies our hypothesis and, in general, no stronger ones. Therefore, the most general conditions for the existence of a positive definite solution of (2) or of (1) provide the best sufficient conditions for the optimum to be attained. To decide whether the optimal value is *not* attained, we can falsify the existence of a positive definite solution of (2) by applying the necessary conditions. In any case, this paper contains the most general available necessary and sufficient conditions for the optimal value being attained.

We finally mention that the existence of Hermitian or positive definite solutions of the strict ARI

$$\mathcal{A}^*\mathcal{X} + \mathcal{X}\mathcal{A} - \mathcal{X}\mathcal{B}\mathcal{B}^*\mathcal{X} + \mathcal{Q} > 0 \tag{3}$$

found a complete algebraic characterization in our work [16], [18], where $(\mathcal{A}, \mathcal{B})$ is *in no way restricted.* For a recent exhaustive discussion of the case $\mathcal{Q} \geq 0$, we refer to [28].

Notation. The notation is standard. $\mathcal{R}$ are the real, $\mathcal{C} = \mathcal{C}^- \cup \mathcal{C}^0 \cup \mathcal{C}^+$ the complex numbers partitioned into the open left half-plane, the imaginary axis and the open right half-plane. Any matrix, space, or subspace in this paper is complex. The matrix A is called *basis matrix* of the subspace $U \subset \mathcal{C}^n$ if the columns of A are linearly independent and span U. A^+ denotes the Moore–Penrose inverse, A^* the complex conjugate transpose, and $\mathcal{H}^{n\times n} := \{A \in \mathcal{C}^{n\times n} | A = A^*\}$. $A > B$ $(A \geq B)$ means that $A, B \in \mathcal{H}^{n\times n}$ (for some n) and $A - B$ is positive (semi)definite. For $A \in \mathcal{H}^{n\times n}, i_+(A), i_-(A), i_0(A)$ denote the numbers of eigenvalues of A in $\mathcal{C}^+, \mathcal{C}^-, \mathcal{C}^0$ and $i(A) := (i_+(A), i_-(A), i_0(A))$ is the inertia. If U is any subspace of $\mathcal{C}^n$, we slightly generalize this notion to the quadratic form $Q : U \ni x \to x^*Ax \in \mathcal{R}$: $i(Q) := i(B^*AB)$ with any basis matrix B of U. Finally, any system $\dot{x} = Ax + Bu$ or pair (A, B) is identified with the pencil $(A - sI\ B)$ and we denote the zeros of this pencil (which are the uncontrollable modes) by $\sigma(A - sI\ B)$ [6].

2. Solvability criteria for the ARE. Suppose that the ARE (1) has a solution. If we are seeking for $\mathcal{X}$ with

$$\mathcal{A}^*\mathcal{X} + \mathcal{X}\mathcal{A} - \mathcal{X}\mathcal{B}\mathcal{B}^*\mathcal{X} + \mathcal{Q} = 0, \qquad \sigma(\mathcal{A} - \mathcal{B}\mathcal{B}^*\mathcal{X}) \subset \mathcal{C}^- \cup \mathcal{C}^0, \tag{4}$$

an obvious *necessary* condition is $\sigma(\mathcal{A} - sI\ \mathcal{B}) \subset \mathcal{C}^- \cup \mathcal{C}^0$. The following by no means obvious result states that this condition is even *sufficient* for the existence of $\mathcal{X}$ with (4). If $\sigma(\mathcal{A} - sI\ \mathcal{B}) \cap \mathcal{C}^0 \neq \varnothing$, there exist infinitely many $\mathcal{X}$ satisfying (4). Nevertheless, one can always select one out of the multitude of solutions satisfying (4) which *overbounds* an arbitrary solution of (1).

THEOREM 1. *Suppose the uncontrollable modes of* $(\mathcal{A} - sI\ \mathcal{B})$ *are contained in* $\mathcal{C}^- \cup \mathcal{C}^0$. *If the* ARE (1) *has the solution* $\mathcal{X}$, *then there exists a Hermitian* $\tilde{\mathcal{X}}$ *which satisfies* (4) *and*

$$\mathcal{X} \leq \tilde{\mathcal{X}}. \tag{5}$$

The set of all $\mathcal{X}$ with (4) *contains a linear manifold. This set reduces to one point iff* $(\mathcal{A} - sI\ \mathcal{B})$ *is stabilizable.*

An immediate corollary is important for verifying the existence of positive definite solutions.

COROLLARY 2. *If* (1) *has a positive definite solution then there exists a Hermitian positive definite $\mathcal{X}$ which satisfies* (4).

To test the existence of Hermitian or positive definite solutions it hence suffices to check the existence of Hermitian or positive definite solutions that satisfy the more restricting conditions (4), and this simplifies the validation problem considerably.

To be more explicit, we now display the $\mathcal{C}^0$-zero structure of $(\mathcal{A} - sI\ \mathcal{B})$ by performing a suitable coordinate change. Indeed, it is well known how to construct a transformation $\mathcal{T}$ such that $\mathcal{A}_\mathcal{T} := \mathcal{T}^{-1}\mathcal{A}\mathcal{T}, \mathcal{B}_\mathcal{T} := \mathcal{T}^{-1}\mathcal{B}$, and $\mathcal{Q}_\mathcal{T} := \mathcal{T}^*\mathcal{Q}\mathcal{T}$ admit the special structures

$$\mathcal{A}_\mathcal{T} = \begin{pmatrix} A & F \\ 0 & M \end{pmatrix}, \quad \mathcal{B}_\mathcal{T} = \begin{pmatrix} B \\ 0 \end{pmatrix}, \quad \mathcal{Q}_\mathcal{T} = \begin{pmatrix} Q & R \\ R^* & S \end{pmatrix}, \tag{6}$$

where

$$(A - sI\ B) \text{ is stabilizable and } \sigma(M) \subset \mathcal{C}_0.$$

Then $\mathcal{X}$ satisfies (1) iff $\mathcal{X}_\mathcal{T} = \mathcal{T}^*\mathcal{X}\mathcal{T}$ satisfies $\mathcal{A}_\mathcal{T}^*\mathcal{X}_\mathcal{T} + \mathcal{X}_\mathcal{T}\mathcal{A}_\mathcal{T} - \mathcal{X}_\mathcal{T}\mathcal{B}_\mathcal{T}\mathcal{B}_\mathcal{T}^*\mathcal{X}_\mathcal{T} + \mathcal{Q}_\mathcal{T} = 0$. If partitioning $\mathcal{X}_\mathcal{T}$ as

$$\begin{pmatrix} X & Y \\ Y^* & Z \end{pmatrix}, \tag{7}$$

this latter ARE is equivalent to the coupled system of equations

$$A^*X + XA - XBB^*X + Q = 0, \tag{8}$$

$$(A - BB^*X)^*Y + YM + XF + R = 0, \tag{9}$$

$$M^*Z + ZM + F^*Y + Y^*F - Y^*BB^*Y + S = 0. \tag{10}$$

Moreover, $\mathcal{A}_\mathcal{T} - \mathcal{B}_\mathcal{T}\mathcal{B}_\mathcal{T}^*\mathcal{X}_\mathcal{T} = \mathcal{T}^{-1}(\mathcal{A} - \mathcal{B}\mathcal{B}^*\mathcal{X})\mathcal{T}$ implies $\sigma(\mathcal{A}_\mathcal{T} - \mathcal{B}_\mathcal{T}\mathcal{B}_\mathcal{T}^*\mathcal{X}_\mathcal{T}) = \sigma(\mathcal{A} - \mathcal{B}\mathcal{B}^*\mathcal{X})$. With (7) and by observing that the diagonal blocks of $\mathcal{A}_\mathcal{T} - \mathcal{B}_\mathcal{T}\mathcal{B}_\mathcal{T}^*\mathcal{X}_\mathcal{T}$ are $A - BB^*X$ and $M, \sigma(\mathcal{A} - \mathcal{B}\mathcal{B}^*\mathcal{X}) \subset \mathcal{C}^- \mathcal{C}^0$ is equivalent to

$$\sigma(A - BB^*X) \subset \mathcal{C}^- \cup \mathcal{C}^0. \tag{11}$$

Whenever necessary we can and do assume without loss of generality $\mathcal{T} = I$, i.e., $\mathcal{A}, \mathcal{B}$, and $\mathcal{Q}$ themselves already have the structures as in (6). Let us finally introduce the notation $\{i\omega_1, \ldots, i\omega_l\} := \sigma(M)$ and let L_j, R_j be basis matrices of $\ker(M - i\omega_j I)^*, \ker(M - i\omega_j I)$ for $j = 1, \ldots, l$.

We arrive at the following existence characterizations, which are an immediate consequence of our main result.

COROLLARY 3. *The* ARE (1) *has a solution iff the* (*unique*) *Hermitian solution X of* (8) *with* (11) *exists and* (9) *has a solution Y such that* (10) *is solvable. The* ARE (1) *has a positive definite solution iff the unique X with* (8) *and* (11) *exists, if it is positive definite, and if there exists a solution Y of* (9) *such that* (10) *has a solution Z with $Z > Y^*X^{-1}Y$.*

For the structure of the whole solution set of (1) it is interesting to prove a little more which may be done without additional effort.

THEOREM 4. *Suppose that* (1) *is solvable. Then for any Hermitian solution* X *of* (8) *the linear equation* (9) *has a solution* Y *such that* (10) *is solvable.*

The proofs of Theorems 1 and 4 are given in the Appendix.

The consequences of these results are striking if the data matrices satisfy a certain regularity condition. Let us introduce for *any* solution X of (8)

$$\Omega := \sigma(A - BB^*X) \cap \sigma(M).$$

We first clarify that Ω only depends on the data $\mathcal{A}, \mathcal{B}, \mathcal{Q}$ and *not* on X or the transformation with $\mathcal{T}$ performed above. For this reason we recall

$$(12)\quad \det \begin{pmatrix} A - sI & -BB^* \\ -Q & -A^* - sI \end{pmatrix} = \det(A - BB^*X - sI) \det(-(A - BB^*X)^* - sI)$$

for any $X = X^*$ satisfying (8) and

$$\det(\mathcal{H} - sI) = \det(M - sI) \det(-M^* - sI) \det \begin{pmatrix} A - sI & -BB^* \\ -Q & -A^* - sI \end{pmatrix}.$$

Both relations allow us to define Ω in terms of the given data matrices as follows.

> Ω consists of the set of all $i\omega \in \sigma(\mathcal{A} - sI\ \mathcal{B})$ whose algebraic multiplicity viewed as an eigenvalue of $\mathcal{H}$ is *larger* than twice its multiplicity viewed as a zero of $(\mathcal{A} - sI\ \mathcal{B})$.

The problem is called *regular* if $\Omega = \varnothing$; otherwise it is called *nonregular* [7].

2.1. Solvability test. We first check whether (8) has a solution. Since $(A - sI\ B)$ is stabilizable, this may be done by looking at the Jordan structure of the corresponding Hamiltonian matrix. If no solution exists we can stop since (1) cannot have a solution either. If (1) is solvable, we can proceed and compute the unique solution $\tilde{X}$ of (8) and (11). Throughout this section we assume $\tilde{X}$ to exist.

Now we verify whether the linear equation (9) with X replaced by $\tilde{X}$ is solvable, which can be accomplished by well-known techniques [9]. If no solution exists we can stop since (1) is not solvable. To proceed we hence assume that a solution exists and we denote it by $\tilde{Y}$.

After defining the subspace

$$\mathcal{Y} := \{Y | (A - BB^*\tilde{X})^*Y + YM = 0\},$$

we must finally test whether there exists a Y in the linear manifold $\tilde{Y} + \mathcal{Y}$ for which (10) is solvable. With $\tilde{S} := S + F^*\tilde{Y} + \tilde{Y}F - \tilde{Y}BB^*\tilde{Y}, \tilde{F} := F - BB^*\tilde{Y}$, we must hence check

$$(13)\qquad \exists Y \in \mathcal{Y}\quad \exists Z\colon\ M^*Z + ZM + \tilde{F}^*Y + Y^*\tilde{F} - Y^*BB^*Y + \tilde{S} = 0.$$

In the general case, we therefore must find a $Y \in \mathcal{Y}$ such that $\tilde{F}^*Y + Y^*\tilde{F} - Y^*BB^*Y + \tilde{S}$, depending quadratically on Y, is contained in a certain subspace. This problem is easily recast into a convex optimization problem and is hence amenable to powerful numerical techniques. At the moment, however, no direct and complete algebraic approach is available for this validation problem.

Hence we solve this problem only under a mild technical hypothesis. Suppose that Ω is given, without restriction, by $\{i\omega_1, \ldots, i\omega_k\}$. Let

$$K_j \text{ be a basis matrix of } \ker((A - BB^*\tilde{X} - i\omega_j I)^*), \qquad j = 1, \ldots, k.$$

Now suppose that $Y \in \mathcal{Y}$ and Z satisfy (13). Then $(A - BB^*\tilde{X} - i\omega_j I)^*Y + Y(M - i\omega_j I) = 0$ holds. If we multiply from the right with R_j we infer $Y_j R_j = K_j Y_j$ for some Y_j. Similarly, (13) leads to $(M - i\omega_j I)^* Z + Z(M - i\omega_j I) + \tilde{F}^* Y + Y^* \tilde{F} - Y^* BB^* Y + \tilde{S} = 0$. By multiplying R_j^* and $-R_j$ from the left and the right, we arrive at

$$Y_j^*(K_j^* BB^* K_j)Y_j - Y_j^*(K_j^* \tilde{F} R_j) - (K_j^* \tilde{F} R_j)^* Y_j - R_j^* \tilde{S} R_j = 0. \tag{14}$$

Hence each Y_j satisfies a quadratic equation for which we develop a complete solvability characterization in the following section. Since $(A - sI\ B)$ is stabilizable, it easily seen that $K_j^* BB^* K_j$ is positive definite. If we now apply Theorem 8 to (14), we infer that (14) is constructively solvable iff

$$R_j^*[\tilde{F}^* K_j (K_j^* BB^* K_j)^{-1} K_j^* \tilde{F} + \tilde{S}] R_j \geq 0 \text{ with rank } \leq \text{ rank } (K_j^* BB^* K_j). \tag{15}$$

For $j > k$ we similarly obtain $YR_j = 0$ and (14) reduces to $R_j^* \tilde{S} R_j = 0$. If K_j is taken to be an empty matrix [23] and if the rank of an empty matrix is defined as zero, the characterization (15) persists to hold for these indices. We have shown that (13) implies (15) for all $j = 1, \ldots, l$.

To reverse these arguments we need to assume that

$$\text{the } \Omega\text{-zero structure of } \sigma(\mathcal{A} - sI\ \mathcal{B}) \text{ is diagonizable} \tag{16}$$

or, in special coordinates and equivalently,

all the Jordan blocks of M associated to $i\omega \in \Omega$ are diagonal.

The importance of this property is stated in the following auxiliary result.

LEMMA 5. *Suppose* (16) *holds. Given arbitrary* $Y_1, \ldots, Y_k$ *of compatible size, there exists a unique* $Y \in \mathcal{Y}$ *with* $YR_j = K_j Y_j$ *for* $j = 1, \ldots, k$.

Proof. We can assume $M = \operatorname{diag}(M_1\ M_2)$ with $\sigma(M_1) = \Omega, \sigma(M_2) \cap \Omega = \varnothing$, which yields $R_j = (\hat{R}_j^*\ 0)^*$ for $j = 1, \ldots, k$. The assumption (16) implies that M_1 is diagonizable and, hence, $\hat{R} := (\hat{R}_1 \cdots \hat{R}_k)$ is square and nonsingular. If we partition Y as $(\hat{Y}\ ^*)$ then $Y \in \mathcal{Y}$ is equivalent to $(A - BB^*\tilde{X})^*\hat{Y} + \hat{Y}M_1 = 0$ and $* = 0$ (by $\sigma((A - BB^*\tilde{X})^*) \cap \sigma(-M_2) = \varnothing$). Hence the only freedom is left in $\hat{Y}$. However, the requirement $\hat{Y}\hat{R} = (K_1 Y_1 \cdots K_k Y_k)$ uniquely determines $\hat{Y}$ and, obviously, $Y := (\hat{Y}\ 0)$ is contained in $\mathcal{Y}$. This proves the existence and uniqueness of Y. □

Now we are ready to formulate and prove the main result of this section.

THEOREM 6. *The existence of* $\tilde{X}, \tilde{Y}$, *and* (15) *for* $j = 1, \ldots, l$ *is necessary for the solvability of* (1). *Suppose that these necessary conditions and, in addition,* (16) *are valid. If* Y_j *denote arbitrary solutions of* (14), $j = 1, \ldots, k$ *and if we choose the unique solution* Y *of* (9) *with* $(Y - \tilde{Y})R_j = K_j Y_j$, *the following equivalences hold.*

(a) *The* ARE (1) *has a Hermitian solution iff the Lyapunov equation*

$$M^* Z + ZM + F^* Y + Y^* F - Y^* BB^* Y + S = 0 \tag{17}$$

is solvable.

(b) *There exists a positive definite solution of* (1) *iff* $\tilde{X}$ *is positive definite and* (17) *has a solution* Z *with* $Z - Y^* \tilde{X}^{-1} Y > 0$.

Proof. By Corollary 3, the sufficiency parts of (a) and (b) need no proof. We must prove necessity *no matter how we choose* Y_j *as solutions of* (14). Suppose Y^1

is a solution of (9) such that (17) (where Y is replaced with Y^1) is solvable. Let Y_j^2 be arbitrary solutions of (14) and let Y^2 denote the corresponding solution of (9). With the abbreviation $P(Y) := \tilde{S} + \tilde{F}^*Y + Y^*\tilde{F} - Y^*BB^*Y$, we hence know that $M^*Z^1 + Z^1M + P(Y^1 - \tilde{Y}) = 0$ for some Z^1 and we must prove the existence of Z^2 with $M^*Z^2 + Z^2M + P(Y^2 - \tilde{Y}) = 0$.

For this reason let us use the same notations as in the proof of Lemma 5 and partition all square matrices as M. Then $M^*Z + ZM + P = 0$ is equivalent to $M_1^*Z_1 + Z_1M_1 + P_1 = 0, M_1^*Z_{12} + Z_{12}M_2 + P_{12} = 0$, and $M_2^*Z_2 + Z_2M_2 + P_2 = 0$.

We first observe that $Y^j - \tilde{Y}$ have the structure $(^* \ 0)$ for $j = 1, 2$. Therefore, $P_2(Y^1 - \tilde{Y}) = P_2(Y^2 - \tilde{Y})$. Hence $M_2^*Z_2^1 + Z_2^1M_2 + P_2(Y^1 - \tilde{Y}) = 0$ implies $M_2^*Z_2^2 + Z_2^2M_2 + P_2(Y^2 - \tilde{Y}) = 0$ for $Z_2^2 := Z_2^1$. By $\sigma(-M_1^*) \cap \sigma(M_2) = \varnothing$, the equation $M_1^*Z_{12}^2 + Z_{12}^2M_2 + P_{12}(Y^2 - \tilde{Y}) = 0$ has a solution Z_{12}^2. Finally, since Y_j^2 satisfies (14), we get $R_j^*P(Y^2 - \tilde{Y})R_j = \hat{R}_j^*P_1(Y^2 - \tilde{Y})\hat{R}_j = 0$ for all $j = 1, \ldots, k$. Since M_1 is diagonizable, we can apply Theorem 12 to infer that $M_1^*Z_1^2 + Z_1^2M_1 + P_1(Y^2 - \tilde{Y}) = 0$ has a solution Z_1^2. This finishes the proof of (a).

For proving necessity in (b) we can assume, in addition, $Z^1 - (Y^1)^*\tilde{X}^{-1}Y^1 > 0$. We try to adjust Z^2 such that

$$\begin{pmatrix} Z_1^2 & Z_{12}^2 \\ (Z_{12}^2)^* & Z_2^2 \end{pmatrix} - (Y^2)^*\tilde{X}^{-1}(Y^2) > 0. \tag{18}$$

Since $Z_2^2 = Z_2^1$ and by $Y^1 - Y^2 = (^* \ 0)$, the (2, 2) block of this matrix coincides with that of $Z^1 - (Y^1)^*\tilde{X}^{-1}Y^1 > 0$ and is hence positive definite. By Theorem 12, $M_1^*Z_1^2 + Z_1^2M_1 + P_1(Y^2 - \tilde{Y}) = 0$ actually has arbitrarily large solutions. Hence we can choose Z_1^2 large enough to ensure (18). □

Again the solvability of (17) is checked by standard techniques [9]. The existence of solutions that satisfy an additional inequality, as required in (b), seems not to appear in the literature. Let us finally turn to this problem.

If defining $\hat{S} := M^*(Y^*\tilde{X}^{-1}Y) + (Y^*\tilde{X}^{-1}Y)M + \tilde{F}^*Y + Y^*\tilde{F} - Y^*BB^*Y + \tilde{S}$, we simply must check the existence of a positive definite solution of

$$M^*Z + ZM + \hat{S} = 0. \tag{19}$$

The last result of this section gives a solution.

THEOREM 7. *Suppose $\sigma(M) \subset \mathcal{C}^0$ and let $T^{-1}MT = \operatorname{diag}(J_1 \cdots J_p)$ be a Jordan normal form of M with Jordan blocks J_j of size ν_j. Introduce*

$$\kappa_j := \left\{ \begin{matrix} \frac{\nu_j}{2} & \text{if } \nu_j \text{ is even} \\ \frac{\nu_j - 1}{2} & \text{if } \nu_j \text{ is odd} \end{matrix} \right\} \quad \text{for } j = 1, \ldots, p$$

and partition $T = (E_1 \ F_1 \cdots E_p \ F_p)$, where E_j, F_j have $\kappa_j, \nu_j - \kappa_j$ columns, respectively.

Let Z denote any arbitrary solution of (19). *Then* (19) *has a positive definite solution iff the matrix $(E_\alpha^* Z E_\beta)_{\alpha,\beta=1,\ldots,p}$* is positive definite.

Proof. Without restriction we assume $T = I$. With

$$Z_{\alpha\beta} = \begin{pmatrix} Z_{\alpha\beta}^E & Z_{\alpha\beta}^{EF} \\ Z_{\alpha\beta}^{FE} & A_{\alpha\beta}^F \end{pmatrix} := (E_\alpha \ F_\alpha)^* Z (E_\beta \ F_\beta),$$

$M^*Z + ZM = 0$ is equivalent to

$$J_\alpha^* Z_{\alpha\beta} + Z_{\alpha\beta} J_\beta = 0 \tag{20}$$

for $\alpha, \beta \in \{1, \dots, p\}$.

Necessity now follows, by linearity, from the simple observation $M^*Z + ZM = 0 \Rightarrow Z^E_{\alpha\beta} = 0$. In the case of $\sigma(J_\alpha) \cap \sigma(J_\beta) = \varnothing$ this is obvious from (20). Hence suppose that J_α and J_β have the same eigenvalue $i\omega$. Then (20) implies

$$(J_\alpha^* + i\omega I)^{\kappa_\alpha} Z_{\alpha\beta} + (-1)^{\kappa_\alpha + 1} Z_{\alpha\beta} (J_\beta - i\omega I)^{\kappa_\alpha} = 0. \tag{21}$$

If ν_α is even we have

$$(J_\alpha^* + i\omega I)^{\kappa_\alpha} = \begin{pmatrix} 0_{\kappa_\alpha} & 0 \\ I_{\kappa_\alpha} & 0 \end{pmatrix} \quad \text{and} \quad (J_\beta - i\omega I)^{\kappa_\alpha} = \begin{pmatrix} 0_{\kappa_\alpha} & * \\ 0 & 0 \end{pmatrix}.$$

Hence the (2, 1) block of (21) yields $Z^E_{\alpha\beta} = 0$. A similar argument applies if ν_α is odd.

For proving sufficiency we first choose any Hermitian Z satisfying (19) and then adjust the diagonal blocks Z_α by adding D_α with $J_\alpha^* D_\alpha + D_\alpha J_\alpha = 0$ such that $Z + \operatorname{diag}(D_1 \cdots D_p)$ is positive definite.

By assumption, $(Z^E_{\alpha\beta})_{\alpha,\beta=1,\dots,p}$ is positive definite.

Let us prove the central induction step for $p = 3$ which considerably simplifies the notations but captures the general features. We assume that Z_1 has been adjusted such that the $\nu_1 + \kappa_2 + \kappa_3$ dimensional submatrix

$$\left(\begin{array}{c|c|c|c} Z_1^E & Z_1^{EF} & Z_{12}^E & Z_{13}^E \\ Z_1^{FE} & Z_1^F & Z_{12}^{FE} & Z_{13}^{FE} \\ \hline Z_{21}^E & Z_{21}^{EF} & Z_2^E & Z_{23}^E \\ \hline Z_{31}^E & Z_{31}^{EF} & Z_{32}^E & Z_3^E \end{array}\right)$$

of Z is positive definite.

Suppose ν_2 is odd. It is easily seen that

$$D(\gamma) = \left(\begin{array}{ccc|c|ccc} 0 & \dots & 0 & 0 & 0 & \dots & \pm\gamma \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & 0 & -\gamma & \dots & 0 \\ \hline 0 & \dots & 0 & \gamma & 0 & \dots & 0 \\ \hline 0 & \dots & -\gamma & 0 & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \pm\gamma & \dots & 0 & 0 & 0 & \dots & 0 \end{array}\right) \in \mathcal{R}^{(\kappa_2+1+\kappa_2)\times(\kappa_2+1+\kappa_2)}$$

satisfies $J_2^* D(\gamma) + D(\gamma) J_2 = 0$ for any $\gamma \in \mathcal{R}$. Let c_1, c_2, c_3, c_4 denote the first columns of $Z_{12}^{EF}, Z_{12}^F, Z_2^{EF}, Z_{32}^{EF}$ and γ_0 be the (1, 1) element of Z_2^F. Consider the $\nu_1 + (\kappa_2 + 1) + \kappa_3$ dimensional submatrix

$$\left(\begin{array}{c|c|c|c|c} Z_1^E & Z_1^{EF} & Z_{12}^E & c_1 & Z_{13}^E \\ Z_1^{FE} & Z_1^F & Z_{12}^{FE} & c_2 & Z_{13}^{FE} \\ \hline Z_{21}^E & Z_{21}^{EF} & Z_2^E & c_3 & Z_{23}^E \\ c_1^* & c_2^* & c_3^* & \gamma_0 & c_4^* \\ \hline Z_{31}^E & Z_{31}^{EF} & Z_{32}^E & c_4 & Z_3^E \end{array}\right) \tag{22}$$

of Z. By the shape of $D(\gamma)$, the same submatrix of $Z + \operatorname{diag}(0_{\nu_1}\ D(\gamma)\ 0_{\nu_3})$ differs from (22) only in the $\nu_1+\kappa_2+1$ diagonal element which equals $\gamma_0+\gamma$. This central structural property allows us to choose a suitably large positive γ such that this submatrix of $Z+\operatorname{diag}(0\ D(\gamma)\ 0)$ is positive definite. Moreover, these steps can be repeated κ_2 times (the number of the remaining diagonal elements of Z_2^F) using solutions D of $J_2^* D + DJ_2 = 0$, whose lower antidiagonal vectors are given by $(\pm\gamma, \dots, -\gamma, \gamma, -\gamma, \dots, \pm\gamma)$ of length $\nu_2 - 2, \nu_2 - 4, \dots, 1$. Again similar arguments apply if ν_2 is even.

In any case this gives a successive procedure to determine $Z_2^{EF}, Z_2^{FE}, Z_2^F$ such that

$$\left(\begin{array}{cc|cc|c} Z_1^E & Z_1^{EF} & Z_{12}^E & Z_{12}^{EF} & Z_{13}^E \\ Z_1^{EF} & Z_1^F & Z_{12}^{FE} & Z_{12}^F & Z_{13}^{FE} \\ \hline Z_{21}^E & Z_{21}^{EF} & Z_2^E & Z_2^{EF} & Z_{23}^E \\ Z_{21}^{FE} & Z_{21}^F & Z_2^{FE} & Z_2^F & Z_{23}^{FE} \\ \hline Z_{31}^E & Z_{31}^{EF} & Z_{32}^E & Z_{32}^{EF} & Z_3^E \end{array}\right)$$

is positive definite and this finishes the proof of the induction step. □

2.2. The regular case. In the regular case the situation is much simpler. $\Omega = \varnothing$ implies that $Y = \tilde{Y}$ is the unique solution of (9) and we can directly apply Theorem 6.

The purpose of this short section is to reveal the structure of the solution set if the problem is regular and (1) is solvable. By regularity, any solution X of (8) *uniquely* determines $Y(X)$ satisfying (9). According to Theorem 4, the solution set of (10) is then necessarily nonempty and actually a linear manifold $L(X)$ whose determining subspace is just $\{Z | M^* Z + ZM = 0\}$ and hence independent of X. This shows that the solution set of (1) equals

$$\left\{ \begin{pmatrix} X & Y(X) \\ Y(X)^* & Z \end{pmatrix} | X \text{ satisfies (8)}, Z \in L(X) \right\}.$$

By the stabilizability of $(A - sI\ B)$, we can apply to (8) the parametrization results obtained in [17] that provide a detailed picture of the solution set of the ARE (1). Finally, we note that the set of all Hermitian $\mathcal{X}$ satisfying (4) not only contains but actually *is* a linear manifold.

3. A general quadratic equation and inequality. The explicit test for the solvability of (1) in the last section required a criterion for the existence of an X satisfying $X^* R X + SX + (SX)^* + T = 0$ where R is positive definite. This section provides a complete discussion for unrestricted matrices.

More precisely, we intend to characterize the existence of $X \in \mathcal{C}^{n\times m}$, which solves the *quadratic* equation

$$X^* R X + SX + X^* S^* + T = 0 \tag{23}$$

or the corresponding inequality

$$X^* R X + SX + X^* S^* + T \geq 0 \tag{24}$$

with $R \in \mathcal{H}^{n\times n}, S \in \mathcal{C}^{m\times n}$, and $T \in \mathcal{H}^{m\times m}$. Note that X is generally rectangular and even for $n = m$ it is not required to be symmetric. If $n = m$, the search for symmetric

solutions would result in discussing the very general ARE $XRX + SX + XS^* + T = 0$ or ARI $XRX + SX + XS^* + T \geq 0$, which is far less understood.

It is not difficult to provide a complete algebraic solvability test for both the equation and the inequality. The proof is *constructive* and provides a recipe how to compute solutions.

THEOREM 8. *Fix $R \in \mathcal{H}^{n\times n}, S \in \mathcal{C}^{m\times n}, T \in \mathcal{H}^{m\times m}$ and define the quadratic form*

$$Q : [S\ker(R)]^\perp \ni x \longrightarrow x^*(SR^+S^* - T)x.$$

Then the equation $X^*RX + SX + X^*S^* + T = 0$ has a solution $X \in \mathcal{C}^{n\times m}$ iff

$$i_+(Q) \leq i_+(R) \quad \text{and} \quad i_-(Q) \leq i_-(R).$$

There exists a $X \in \mathcal{C}^{n\times m}$ with $X^*RX + SX + X^*S^* + T \geq 0$ iff

$$i_+(Q) \leq i_+(R).$$

If the data matrices satisfy $\ker(R) \subset \ker(S)$, which holds in particular if R is nonsingular or $\left(\begin{smallmatrix} R & S^* \\ S & T \end{smallmatrix}\right)$ is positive/negative semidefinite, we infer $S\ker(R) = \{0\}$ and our characterizations admit the nice form

$$i_+(SR^+S^* - T) \leq i_+(R) \quad \text{and} \quad i_-(SR^+S^* - T) \leq i_-(R)$$

directly in terms of the given matrices.

Obviously, the solvability of

$$X^*RX + SX + X^*S^* + T \leq 0 \tag{25}$$

is characterized by $i_-(Q) \leq i_-(R)$. Note that this leads to the interesting conclusion that the solvability of both *inequalities* (24) and (25) imply the solvability of the *equation* (23).

Finally, the equation $X^*RX + SX - X^*S^* + T = 0$ for skew-Hermitian R and T is easily recast to the present situation by noting

$$i[X^*RX + SX - X^*S^* + T] = (iX)^*(iR)(iX) + S(iX) + (iX)^*S^* + (iT) = 0.$$

Proof. Define $(r_+, r_-, r_0) = i(R)$, choose a unitary U with $\tilde{R} = \operatorname{diag}(\Sigma_+ \; -\Sigma_- \; 0) = URU^*, \Sigma_+, \Sigma_- > 0$, and introduce $\tilde{S} = (S_+ \; S_- \; S_0) = SU^*$ with a column partition corresponding to that of $\tilde{R}$. Then X satisfies (23) iff $(Y_+^* \; Y_-^* \; Y_0^*)^* = UX$ satisfies $Y_+^*\Sigma_+Y_+ - Y_-^*\Sigma_-Y_- + S_+Y_+ + S_-Y_- + S_0Y_0 + (S_+Y_+ + S_-Y_- + S_0Y_0)^* + T = 0$, which can be rearranged (by completion of the squares) to

(26)

$$\begin{aligned}[\Sigma_+^{1/2}Y_+ + \Sigma_+^{-1/2}S_+^*]^*[\Sigma_+^{1/2}Y_+ + \Sigma_+^{-1/2}S_+^*] - [\Sigma_-^{1/2}Y_- + \Sigma_-^{-1/2}S_-^*]^*[\Sigma_-^{1/2}Y_- + \Sigma_-^{-1/2}S_-^*] \\ = [S_+\Sigma_+^{-1}S_+^* - S_-\Sigma_-^{-1}S_-^* - T] - S_0Y_0 - (S_0Y_0)^* \\ = [SR^+S^* - T] - S_0Y_0 - Y_0^*S_0^*,\end{aligned}$$

where the last equality follows by computation.

We want to decide the solvability of this equation in Y_+, Y_- for any fixed Y_0. For this reason we first prove the following auxiliary result which is intuitively clear.

LEMMA 9. *For a given $P \in \mathcal{H}^{p\times p}$, the equation $V^*V - W^*W = P$ in $V \in \mathcal{C}^{v\times p}, W \in \mathcal{C}^{w\times p}$ is solvable iff $i_+(P) \leq v$ and $i_-(P) \leq w$.*

Proof. The "only if" part follows from $v \geq i_+(V^*V) = i_+(P + W^*W) \geq i_+(P)$ and similarly $w \geq i_+(W^*W) = i_+(V^*V - P) = i_-(P - V^*V) \geq i_-(P)$. The "if" part can be proved constructively by observing that neither the solvability problem nor our characterization are affected by a congruence transformation on P. Hence we can assume $P = \operatorname{diag}(P_v \ -P_w \ 0)$, where both P_v, P_w are positive semidefinite and of dimension v, w, respectively. Then $V = (\sqrt{P_v}\ 0\ 0)$ and $W = (0\ \sqrt{P_w}\ 0)$ are solutions. □

Let us return to (26). Since Σ_+ and Σ_- are nonsingular, (26) has, for any fixed Y_0, solutions $Y_+ \in \mathcal{C}^{r_+ \times m}$ and $Y_- \in \mathcal{C}^{r_- \times m}$ iff

$$i_j((SR^+S^* - T) - S_0Y_0 - Y_0^*S_0^*) \leq r_j \quad \text{for} \quad j = +, - . \tag{27}$$

How far can these inequalities be enforced by varying Y_0? To decide this question, let us choose any basis matrix K of the kernel of S_0^*. Then there exist L, M such that $M, (L\ K)$ are nonsingular and satisfy $MS_0^*(L\ K) = \left(\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right)$. Clearly, Y_0 yields (27) iff $\left(\begin{smallmatrix} Z_1 & Z_{12} \\ Z_{21} & Z_2 \end{smallmatrix}\right) := M^{-*}Y_0(L\ K)$ leads to

$$i_j\left(\begin{pmatrix} * & * \\ * & K^*[SR^+S^* - T]K \end{pmatrix} - \begin{pmatrix} Z_1 + Z_1^* & Z_{12} \\ Z_{12}^* & 0 \end{pmatrix}\right) \leq r_j$$

for $j = +, -$. Now the left-hand side of this inequality is bounded below by $i_j(K^*[SR^+S^* - T]K)$. Since we can obviously achieve this bound by choosing suitable Z_1, Z_{12}, we infer that there exists a Y_0 with (27) iff

$$i_+(K^*[SR^+S^* - T]K) \leq r_+ \quad \text{and} \quad i_-(K^*[SR^+S^* - T]K) \leq r_- .$$

Since K is an arbitrary basis matrix of $\ker(S_0^*) = \operatorname{im}(S_0)^\perp$, the result follows by noticing $\operatorname{im}(S_0) = \tilde{S}\ker(\tilde{R}) = SU^*\ker(\tilde{R}) = S\ker(R)$.

In the same way one proves the algebraic characterization for the solvability of (24). □

4. The Lyapunov inequality. This section serves to investigate the solvability of the Lyapunov inequality

$$M^*X + XM + S \geq 0, \tag{28}$$

where M is only restricted to having all its eigenvalues in $\mathcal{C}^0$ and $S = S^*$ is possibly indefinite. We will give necessary and sufficient conditions for the existence of arbitrarily large Hermitian solutions of (28) in the following sense.

For all Hermitian X_0 there exists a Hermitian solution X of (28) with $X > X_0$.

Recall the following complete result for the strict Lyapunov inequality [16], [18].

THEOREM 10. *Suppose $\sigma(M) \subset \mathcal{C}^0$ and $S = S^*$. Then there exists a Hermitian solution X of*

$$M^*X + XM + S > 0 \tag{29}$$

iff any eigenvector x of M satisfies $x^*Sx > 0$. If (29) is solvable then it has arbitrarily large solutions.

Simple examples show that (28) may be solvable without having arbitrarily large solutions, which reveals an essential difference between both inequalities. The general solvability problem for (28) without any requirements on the solutions seems very hard, and our results provide the weakest general sufficient conditions for the solvability of (28) that are available.

An obvious necessary condition for the solvability of (28) is

$$x^*Sx \geq 0 \text{ for each eigenvector } x \text{ of } M: \tag{30}$$

just note that $x^*(M^*X + XM + S)x = x^*Sx$. If M is diagonizable, it was observed in [16] that (30) implies the existence of a Hermitian solution of (28). Moreover, as for the strict inequality, the solution can be chosen arbitrarily large. However, if M is not diagonizable, (30) is generally much too weak to imply the solvability of (28) and far from sufficient, again, in contrast to the strict inequality.

Let us now assume that (28) has arbitrarily large solutions. As expected from the above discussion, we should derive further necessary conditions by choosing an x with $(M - i\omega I)x = 0$ for which x^*Sx vanishes. We obtain $0 = x^*Sx = x^*((M - i\omega I)^*X + X(M - i\omega I) + S)x = 0$ and, since (28) is equivalent to $(M - i\omega I)^*X + X(M - i\omega I) + S \geq 0$, we infer $(M - i\omega I)^*Xx + Sx = 0$. Due to the fact that X can be chosen arbitrarily large, we claim that x cannot be the starting vector of a Jordan chain of M: There is no y with $(M - i\omega I)y = x$. Suppose there existed a solution y to this equation. Then we could infer $x^*Xx + y^*Sx = 0$, which shows that x^*Xx is a fixed number and this contradicts our assumption. Hence we conclude the existence of a y with $y^*(M - i\omega I) = 0$ such that y^*x does not vanish.

As for the strict inequality, it is surprising that these conditions in terms of eigenvectors turn out to be sufficient and there is no need to consider further Jordan chain vectors.

THEOREM 11. *If $\sigma(M) \subset \mathcal{C}^0$ and $S = S^*$, the following conditions are equivalent.*

(a) *The Lyapunov inequality $M^*X + XM + S \geq 0$ has arbitrarily large solutions.*

(b) *Any eigenvector of M satisfies $x^*Sx \geq 0$ and if $x^*Sx = 0$ then there exists a y with $y^*(M - i\omega I) = 0$ and $y^*x \neq 0$.*

(c) *$R_j^*SR_j \geq 0$ and $\ker(R_j^*SR_j) \cap \ker(L_j^*R_j) = \{0\}$ for all $j = 1, \ldots, l$.*

Proof. The equivalence of (b) and (c) is easy to establish and (a) $\Rightarrow$ (b) has been proved above. Hence we only show (b) $\Rightarrow$ (a) by induction. This requires the following two observations. If (M, S) satisfies (b) then the same is true for $(T^{-1}MT, T^*ST)$ and any nonsingular matrix T. Moreover, suppose that M and S are structured as

$$\begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix} \text{ and } \begin{pmatrix} S_1 & S_{21}^* \\ S_{21} & S_2 \end{pmatrix}.$$

Since any eigenvector x_1 of M_1 can be trivially extended to an eigenvector $x := (x_1^* \; 0)^*$ of M with $x^*Sx = x_1^*S_1x_1$, it is obvious that (M_1, S_1) and, by symmetry, (M_2, S_2) satisfy (b) as well.

First we convince ourselves that we only need to prove the theorem under the assumption that M is nilpotent. Without restriction we can assume M given as $\operatorname{diag}(M_1 \cdots M_q)$, where $\sigma(M_\alpha)$ and $\sigma(M_\beta)$ are singletons having no intersection for $\alpha \neq \beta$. We partition $X = (X_{\alpha\beta})$ and $S = (S_{\alpha\beta})$ accordingly and observe that $(M_\alpha, S_{\alpha\alpha})$ satisfies (b). Clearly, the (α, β) block of $M^*X + XM + S$ equals

$$M_\alpha^*X_{\alpha\beta} + X_{\alpha\beta}M_\beta + S_{\alpha\beta}. \tag{31}$$

For $\alpha \neq \beta$ we can choose, by $\sigma(-M_\alpha^*) \cap \sigma(M_\beta) = \varnothing, X_{\alpha\beta}$ such that (31) vanishes. This fixes the nondiagonal blocks of X with $X_{\alpha\beta}^* = X_{\beta\alpha}$. If the diagonal blocks $X_{\alpha\alpha}$ satisfy

$$M_\alpha^* X_{\alpha\alpha} + X_{\alpha\alpha} M_\alpha + S_{\alpha\alpha} \geq 0 \tag{32}$$

then X is a solution of (28) and if $X_{\alpha\alpha}$ may be chosen arbitrarily large then X may be taken arbitrarily large as a solution of (28). Hence we only need to prove the result for (32) or, equivalently, for $(M_\alpha - i\omega I)^* X + X(M_\alpha - i\omega I) + S_{\alpha\alpha} \geq 0$, where $\sigma(M_\alpha) = \{i\omega\}$. Therefore we assume $\sigma(M) = \{0\}$. If all kernel vectors of M satisfy $x^* S x > 0$ then Theorem 10 finishes the proof. Hence suppose there exists an $x \neq 0$ with $Mx = 0$ and $x^* S x = 0$. Without restriction x equals e_1, the first standard unit vector, and then we have

$$M = \begin{pmatrix} 0 & M_{12} \\ 0 & M_2 \end{pmatrix} \quad \text{and} \quad S = \begin{pmatrix} 0 & S_{12} \\ S_{12}^* & S_2 \end{pmatrix}.$$

By (b) there exists a y with $y^* M = 0$ and $y^* x \neq 0$. Since the first coefficient of y does not vanish, the row M_{12} is actually linear dependent on the rows of M_2. If we choose a z with $z^* M_2 = M_{12}$ and define $T = \left(\begin{smallmatrix} 1 & z^* \\ 0 & I \end{smallmatrix}\right)$, we infer that $T^{-1} M T$ and $T^* S T$ are given as

$$\begin{pmatrix} 0 & 0 \\ 0 & M_2 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & S_{12} \\ S_{12}^* & \tilde{S}_2 \end{pmatrix}.$$

We can again assume that M and S themselves are already given in this form and partition X accordingly. Then the (1, 1) block of $M^* X + X M + S$ vanishes whereas the (2, 1) and the (2, 2) blocks are given by

$$M_2^* X_{21} + S_{21} \text{ and } M_2^* X_2 + X_2 M_2 + S_2.$$

We now prove that there exists a X_{21} with $M_2^* X_{21} + S_{12} = 0$: Any y_2 with $M_2 y_2 = 0$ can be extended as $y = (0\ y_2^*)^*$ such that $My = 0$ and we actually show $y^* S x = 0$ which leads, recalling $x = e_1$, to $y_2^* S_{12} = 0$. If E denotes a basis matrix of the kernel of M we must show $E^* S x = 0$. Since x is a kernel vector of M there exists a ξ with $x = E\xi$ and we must prove $E^* S E \xi = 0$. Now we just recall $E^* S E \geq 0$ and $\xi^* E^* S E \xi = x^* S x = 0$ which gives the desired result.

Let us hence fix X_{12} with $M_2^* X_{21} + S_{12} = 0$. If we use induction on the dimension of M we can finish the proof as follows. Since (M_2, S_2) satisfies (b), the induction hypothesis implies that $M_2^* X_2 + X_2 M_2 + S_2 \geq 0$ has arbitrarily large solutions. The block $X_1 \in \mathcal{R}$ is free and hence we can choose X_1 and X_2 such that X becomes as large as desired and solves (28). □

It is important to note that the proof gives us a recipe to actually compute solutions. For any eigenvector x of M with $x^* S x = 0$, it is demonstrated how to reduce the dimension of M and S by one. The sequential application of this procedure leads us to a problem where M and S are one dimensional (which is trivial) or where any eigenvector x of M satisfies $x^* S x > 0$. Note that it is not required to check (b) a priori, but one may just start the iteration on (M, S) and verify whether all the steps can be performed as described. This gives a test for verifying (b). Moreover, the final structure of the matrices determines the space on which $M^* X + X M + S$ can be rendered positive and on which it necessarily vanishes.

If M is diagonizable, (b) clearly reduces to (30), and (c) to $R_j^* S R_j \geq 0$. For reasons of comparison, we finally formulate the analogous result for the Lyapunov equation, whose proof is now a simple exercise.

THEOREM 12. *The Lyapunov equation $M^*X + XM + S = 0$ has arbitrarily large solutions iff M is diagonizable and $x^*Sx = 0$ holds for each eigenvector x of M.*

5. Solvability criteria for the ARI. Our solvability tests for the nonstrict ARI are separated into necessary and sufficient conditions that are formulated in terms of the transformed versions (6) of $\mathcal{A}, \mathcal{B}, \mathcal{Q}$, and the basis matrices L_j, R_j for the left-kernel, right-kernel of $(M - i\omega_j I)$ as introduced in §2.

THEOREM 13. *If* (2) *has a Hermitian* (*positive definite*) *solution then there exist a Hermitian* (*positive definite*) X *and a* Y *with*

$$\sigma(A - BB^*X) \subset \mathcal{C}^- \cup \mathcal{C}^0, A^*X + XA - XBB^*X + Q = 0, \tag{33}$$

$$(A - BB^*X)^*Y + YM + XF + R = 0, \tag{34}$$

$$\forall j = 1, \ldots, l: R_j^*[Y^*F + F^*Y - Y^*BB^*Y + S]R_j \geq 0. \tag{35}$$

The proof of this result is given in the Appendix.

THEOREM 14. *Suppose that there exists a Hermitian* X *with* (33) *and a solution* Y *of* (34) *such that* (35) *and, in addition,*

$$\forall j = 1, \ldots, l: \ker(R_j^*[Y^*F + F^*Y - Y^*BB^*Y + S]R_j) \cap \ker(L_j^* R_j) = \{0\} \tag{36}$$

hold true. Then (2) *is solvable. If* X *is, in addition, positive definite, then* (2) *has a positive definite solution.*

Proof. By Theorem 11, $M^*Z + ZM + Y^*F + F^*Y - Y^*BB^*Y + S \geq 0$ has arbitrarily large solutions. For any solution Z of this inequality, (7) obviously satisfies (2). If X is positive definite we can choose Z large enough to render (7) positive definite. □

Remark. For reasons of comparison we recall the complete solvability test for the strict ARI (3) provided in [18]. The inequality (3) has a (positive definite) solution iff the unique X with (33) exists and satisfies in fact $\sigma(A - BB^*X) \subset \mathcal{C}^-$ (as well as $X > 0$) and the hence unique solution Y of (34) satisfies $R_j^*[Y^*F + F^*Y - Y^*BB^*Y + S]R_j > 0$ for all $j = 1, \ldots, l$.

For the discussion on how to verify the conditions in both theorems, we again assume that X with (33) exists and that (34) has the solution $\tilde{Y}$.

If the problem is regular, $\tilde{Y}$ is unique and we just need to check (35) or (36). Hence we concentrate on the nonregular case $\Omega = \{i\omega_1, \ldots, i\omega_k\} \neq \varnothing$ and introduce $\mathcal{Y}, \tilde{S}, \tilde{F}, K_j$ as in §2.

If we let, for any $Y \in \mathcal{Y}, Y_j$ satisfy $YR_j = K_j Y_j$ and define

$$Z_j := Y_j - (K_j^* BB^* K_j)^{-1} K_j^* \tilde{F} R_j,$$

one easily computes

$$\begin{aligned} R_j^*[(\tilde{Y} + Y)^*F + F^*(\tilde{Y} + Y) - (\tilde{Y} + Y)^*BB^*(\tilde{Y} + Y) + S]R_j \\ = R_j^*[\tilde{F}^* K_j (K_j^* BB^* K_j)^{-1} K_j^* \tilde{F} + \tilde{S}]R_j - Z_j^*(K_j^* BB^* K)_j^{-1} Z_j. \end{aligned} \tag{37}$$

Hence if there exists a Y with (34) and (35) then

$$\forall j = 1, \ldots, l: R_j^*[\tilde{F}^* K_j (K_j^* BB^* K_j)^{-1} K_j^* \tilde{F} + \tilde{S}]R_j \geq 0. \tag{38}$$

Moreover, if Y satisfies in addition (36), we infer

$$(39)\quad \forall j = 1,\dots,l\colon\ \ker(R_j^*\{\tilde{F}^*K_j(K_j^*BB^*K_j)^{-1}K_j^*\tilde{F}+\tilde{S}\}R_j)\cap\ker(L_jR_j)=\{0\}.$$

To reverse the above reasoning, we again need to assume (16). If (38) (and (39)) are valid we define (motivated by (37) and the definition of Z_j) $Y_j := (K_j^*BB^*K_j)^{-1}K_j^*\tilde{F}R_j$. Lemma 5 allows to construct a $Y\in\mathcal{Y}$ with $YR_j = K_jY_j$ and then, recalling (37), $\tilde{Y}+Y$ satisfies (35) (and (36)).

We have proved the following result which reveals how far we are presently able to verify the necessary (Theorem 13) and sufficient (Theorem 14) conditions for the existence of (positive definite) solutions of (2) in an algebraic manner.

THEOREM 15. *If Y solves* (34) *then*: (35) (*and* (36)) $\Rightarrow$ (38) (*and* (39)). *If the Ω-zero structure of $(\mathcal{A}-sI\ \mathcal{B})$ is diagonizable, there exists a unique solution Y of* (34) *with $YR_j = \tilde{Y}R_j + K_j(K_j^*BB^*K_j)^{-1}K_j^*\tilde{F}R_j, j=1,\dots,l$. Then* (38) (*and* (39)) $\Rightarrow$ (35) (*and* (36)).

If the $\mathcal{C}^0$-zero structure of $(\mathcal{A}-sI\ \mathcal{B})$ or, equivalently, M is diagonizable, then the matrices $L_j^*R_j$ are nonsingular and the above necessary and sufficient conditions coincide—our characterization is complete.

COROLLARY 16. *Suppose that the $\mathcal{C}^0$-zero structure of $(\mathcal{A}-sI\ \mathcal{B})$ is diagonizable. Then the* ARI (2) *has a* (*positive definite*) *solution iff the unique X with* (33) *exists* (*and is positive definite*), (34) *has a solution $\tilde{Y}$, and* (38) *holds.*

6. Conclusions. The results of this paper can be summarized as follows. If the Ω-zero structure of $(\mathcal{A}-sI\ \mathcal{B})$ is diagonizable (with Ω encompassing all points in $\mathcal{C}^0$ whose algebraic multiplicity viewed as an eigenvalue of $\mathcal{H}$ is larger than twice their multiplicity viewed as a zero of $(\mathcal{A}-sI\ \mathcal{B})$), we obtained a full verifiable characterization for the existence of (positive definite) solutions of the ARE (1), and necessary and slightly stronger sufficient conditions for the existence of (positive definite) solutions of the ARI (2).

As auxiliary results of independent interest, we provided a new structural property for the solution set of the ARE (1) (Theorems 1 and 4) with potentials for future applications, and a complete algebraic solvability tests for the quadratic equation $X^*RX+SX+(SX)^*+T=0$ and inequality $X^*RX+SX+(SX)^*+T\geq 0$ with Hermitian R and T.

Under the assumption $\sigma(\mathcal{A})\subset\mathcal{C}^0$ and with a general indefinite $\mathcal{Q}$, we developed complete algebraic characterizations for the existence of positive definite or arbitrarily large solutions of the Lyapunov equation $\mathcal{A}^*\mathcal{X}+\mathcal{X}\mathcal{A}+\mathcal{Q}=0$ and a complete algebraic characterization for the existence of arbitrarily large solutions of the Lyapunov inequality $\mathcal{A}^*\mathcal{X}+\mathcal{X}\mathcal{A}+\mathcal{Q}\geq 0$.

Appendix.

Proof of Theorems 1 *and* 4. The essential structural property is provided by the following auxiliary result.

LEMMA 17. *Suppose that $\mathcal{X}=\left(\begin{smallmatrix}X & Y\\ Y^* & Z\end{smallmatrix}\right)$ satisfies* (1) *and let $\tilde{X}$ be any solution of* (8). *Then there exists a solution $\tilde{\mathcal{X}}$ of* (1) *which satisfies the relation*

$$(40)\qquad \tilde{\mathcal{X}}-\mathcal{X}=\begin{pmatrix}I\\ V\end{pmatrix}(\tilde{X}-X)\begin{pmatrix}I\\ V\end{pmatrix}^*$$

for some V.

Proof. Let us start by transforming $A - BB^*X$ with a suitable similarity transformation $T := (T_1\ T_2)$ to obtain

$$T^{-1}(A - BB^*X)T = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}, \qquad T^{-1}B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix},$$

where

$$\sigma(A_1) \cap \mathcal{C}^0 = \varnothing \quad \text{and} \quad \sigma(A_2) \subset \mathcal{C}^0.$$

It is easily seen that $(A-BB^*X)^*(\tilde{X}-X)+(\tilde{X}-X)(A-BB^*X)-(\tilde{X}-X)BB^*(\tilde{X}-X) = 0$ and we proved in [17, p. 111] that the $\mathcal{C}^0$-root subspace of $A - BB^*X$ is contained in the kernel of $\tilde{X} - X$. Hence $\Delta := T^*(\tilde{X} - X)T$ satisfies

$$\begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}^*\Delta + \Delta\begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} - \Delta\begin{pmatrix} B_1 \\ B_2 \end{pmatrix}\begin{pmatrix} B_1 \\ B_2 \end{pmatrix}^*\Delta = 0$$

and, due to our particular coordinates, admits the structure

$$\Delta = \begin{pmatrix} \Delta_1 & 0 \\ 0 & 0 \end{pmatrix}.$$

We arrive at

$$A_1^*\Delta_1 + \Delta_1 A_1 - \Delta_1 B_1 B_1^*\Delta_1 = 0. \tag{41}$$

Let us now consider

$$\begin{pmatrix} T_1 & T_2 & 0 \\ 0 & 0 & I \end{pmatrix}^{-1}(\mathcal{A} - \mathcal{B}\mathcal{B}^*\mathcal{X})\begin{pmatrix} T_1 & T_2 & 0 \\ 0 & 0 & I \end{pmatrix} = \begin{pmatrix} A_1 & 0 & F_1 \\ 0 & A_2 & F_2 \\ 0 & 0 & M \end{pmatrix}.$$

Since A_1 and M do not have common eigenvalues, we can clearly find an R with $A_1R - RM + F_1 = 0$. If we define the similarity transformation

$$\mathcal{T} := \begin{pmatrix} T_1 & T_2 & 0 \\ 0 & 0 & I \end{pmatrix}\begin{pmatrix} I & 0 & R \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix}$$

we infer after a simple computation

$$\tilde{\mathcal{A}} := \mathcal{T}^{-1}(\mathcal{A} - \mathcal{B}\mathcal{B}^*\mathcal{X})\mathcal{T} = \begin{pmatrix} A_1 & 0 & 0 \\ 0 & A_2 & F_2 \\ 0 & 0 & M \end{pmatrix}, \qquad \tilde{\mathcal{B}} := \mathcal{T}^{-1}\mathcal{B} = \begin{pmatrix} B_1 \\ B_2 \\ 0 \end{pmatrix}.$$

Now remember that $\tilde{\mathcal{X}}$ satisfies (1) iff $(\mathcal{A} - \mathcal{B}\mathcal{B}^*\mathcal{X})^*(\tilde{\mathcal{X}} - \mathcal{X}) + (\tilde{\mathcal{X}} - \mathcal{X})(\mathcal{A} - \mathcal{B}\mathcal{B}^*\mathcal{X}) - (\tilde{\mathcal{X}} - \mathcal{X})\mathcal{B}\mathcal{B}^*(\tilde{\mathcal{X}} - \mathcal{X}) = 0$ iff $\mathcal{D} = \mathcal{T}^*(\tilde{\mathcal{X}} - \mathcal{X})T$ satisfies

$$\tilde{\mathcal{A}}^*\mathcal{D} + \mathcal{D}\tilde{\mathcal{A}} - \mathcal{D}\tilde{\mathcal{B}}\tilde{\mathcal{B}}^*\mathcal{D} = 0. \tag{42}$$

Due to the fact that the (1, 3) block in $\tilde{\mathcal{A}}$ vanishes, the choice

$$\mathcal{D} := \begin{pmatrix} \Delta_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

leads to

$$\tilde{\mathcal{A}}^*\mathcal{D} + \mathcal{D}\tilde{\mathcal{A}} - \mathcal{D}\tilde{\mathcal{B}}\tilde{\mathcal{B}}^*\mathcal{D} = \begin{pmatrix} A_1^*\Delta_1 + \Delta_1 A_1 - \Delta_1 B_1 B_1^* \Delta_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

By (41), $\mathcal{D}$ is a solution of (42).

We can conclude that (8) has a solution $\tilde{\mathcal{X}}$ with $\tilde{\mathcal{X}} - \mathcal{X} = \mathcal{T}^{-*}\mathcal{D}\mathcal{T}^{-1}$. With the notation

$$T^{-1} = \begin{pmatrix} S_1 \\ S_2 \end{pmatrix} \text{ we infer } \mathcal{T}^{-1} = \begin{pmatrix} S_1 & -R \\ S_2 & 0 \\ 0 & I \end{pmatrix}$$

and thus

$$\tilde{\mathcal{X}} - \mathcal{X} = \mathcal{T}^{-*}\mathcal{D}\mathcal{T}^{-1} = \begin{pmatrix} S_1 & -R \\ S_2 & 0 \end{pmatrix}^* \begin{pmatrix} \Delta_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} S_1 & -R \\ S_2 & 0 \end{pmatrix}.$$

Since $\begin{pmatrix} S_1 & -R \\ S_2 & 0 \end{pmatrix} = \begin{pmatrix} S_1 \\ S_2 \end{pmatrix} (I \quad V^*) = T^{-1} \begin{pmatrix} I \\ V \end{pmatrix}^*$ for some V, we arrive at

$$\tilde{\mathcal{X}} - \mathcal{X} = \begin{pmatrix} I \\ V \end{pmatrix} T^{-*} \begin{pmatrix} \Delta_1 & 0 \\ 0 & 0 \end{pmatrix} T^{-1} \begin{pmatrix} I \\ V \end{pmatrix}^* = \begin{pmatrix} I \\ V \end{pmatrix} (\tilde{X} - X) \begin{pmatrix} I \\ V \end{pmatrix}^*$$

by the definition of Δ. □

Now let $\mathcal{X}$ be an arbitrary solution of (1). For an arbitrary $\tilde{X}$ satisfying (8), Lemma 17 implies the existence of a solution $\tilde{\mathcal{X}}$ of (1) whose left upper block is given by $(\tilde{X} - X) + X = \tilde{X}$. This proves Theorem 4.

Let us now choose $\tilde{X}$ as the unique solution of (8) with $\sigma(A - BB^*\tilde{X}) \subset \mathcal{C}^- \cup \mathcal{C}^0$ and construct $\tilde{\mathcal{X}}$ as in Lemma 17. Again, the left upper block of $\tilde{\mathcal{X}}$ is just $\tilde{X}$ and, therefore, $\sigma(\mathcal{A} - \mathcal{B}\mathcal{B}^*\tilde{\mathcal{X}}) \subset \mathcal{C}^- \cup \mathcal{C}^0$. Moreover, since $\tilde{X}$ is the largest solution of (8), we get $\tilde{X} - X \geq 0$. Again by Lemma 17 we arrive at $\tilde{\mathcal{X}} - \mathcal{X} \geq 0$. This proves Theorem 1.

Proof of Theorem 13. We assume

$$\begin{pmatrix} U & V \\ V^* & W \end{pmatrix} = \mathcal{A}^*\mathcal{X} + \mathcal{X}\mathcal{A} - \mathcal{X}\mathcal{B}\mathcal{B}^*\mathcal{X} + \mathcal{Q} \geq 0 \tag{43}$$

and evaluate the equality blockwise as

$$A^*X + XA - XBB^*X + Q = U, \tag{44}$$

$$(A - BB^*X)^*Y + YM + XF + R = V, \tag{45}$$

$$M^*Z + ZM + F^*Y + Y^*F - Y^*BB^*Y + S = W. \tag{46}$$

The inequality $A^*X + XA - XBB^*X + Q \geq 0$ implies the existence $\tilde{X}$ satisfying (33). If $\mathcal{X}$ is positive definite, we infer $X > 0$ and hence $\tilde{X} > 0$. We intend to prove the existence of a $\tilde{Y}$ with

$$(A - BB^*\tilde{X})^*\tilde{Y} + \tilde{Y}M + \tilde{X}F + R = 0 \tag{47}$$

such that

$$R_j^*[F^*\tilde{Y} + \tilde{Y}^*F - \tilde{Y}^*BB^*\tilde{Y} + S]R_j \geq 0 \tag{48}$$

for all $j = 1, \dots, l$.

Let us first discuss the regular case $\Omega = \varnothing$. Though we could give a direct algebraic proof, we prefer the following quicker perturbation approach using the known results for the strict ARI. After choosing some $\alpha > 0$ with $\alpha I > \mathcal{Q}$, define

$$\mathcal{Q}(\mu) := \mathcal{Q} - \mu(\alpha I - \mathcal{Q}) \quad \text{for} \quad \mu \leq 0$$

and partition it as $\mathcal{Q}$. By $\mathcal{Q}(\mu) > \mathcal{Q}$ for $\mu < 0$, $\mathcal{X}$ satisfies the *strict* ARI $\mathcal{A}^*\mathcal{X} + \mathcal{X}\mathcal{A} - \mathcal{X}\mathcal{B}\mathcal{B}^*\mathcal{X} + \mathcal{Q}(\mu) > 0$ and hence (see the remark in §5) there exist $X(\mu)$ and $Y(\mu)$ with

$$\begin{aligned} &\sigma(A - BB^*X(\mu)) \subset \mathcal{C}^-, A^*X(\mu) + X(\mu)A - X(\mu)BB^*X(\mu) + Q(\mu) = 0, \\ &\qquad (A - BB^*X(\mu))^*Y(\mu) + Y(\mu)M + X(\mu)F + R(\mu) = 0, \\ &\forall_j = 1, \dots, l\colon R_j^*[F^*Y(\mu) + Y(\mu)^*F - Y(\mu)^*BB^*Y(\mu) + S]R_j > 0. \end{aligned} \tag{49}$$

Let us now take the limit $\mu \to 0$. By $Q(\mu) \to 0$, a standard result shows $X(\mu) \to \tilde{X}$ [14], [18]. Regularity and $R(\mu) \to R$ imply that $Y(\mu)$ converges to the unique solution $\tilde{Y}$ of $(A - BB^*\tilde{X})^*\tilde{Y} + \tilde{Y}M + \tilde{X}F + R = 0$. Hence (49) leads by $S(\mu) \to S$ to the desired inequality.

In the nonregular case $A - BB^*\tilde{X}$ has eigenvalues in C^0 and we can assume without loss of generality

$$\tilde{A} := A - BB^*\tilde{X} = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} \quad \text{with} \quad \sigma(A_1) \subset \mathcal{C}^-, \sigma(A_2) \subset \mathcal{C}^0 \tag{50}$$

and partition all other matrices similarly. With $\Delta := X - \tilde{X}$, it is easily seen that (44)–(46) can be rewritten as

$$\tilde{A}^*\Delta + \Delta\tilde{A} - \Delta BB^*\Delta = U, \tag{51}$$

$$(\tilde{A} - BB^*\Delta)^*Y + YM + \Delta F + (\tilde{X}F + R) = V, \tag{52}$$

$$M^*Z + ZM + F^*Y + Y^*F - Y^*BB^*Y + S = W. \tag{53}$$

As in the proof of Lemma 17 we can again invoke [17, p. 111] to infer from (50) the structure

$$\Delta = \begin{pmatrix} \Delta_1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Consequently, $U = \begin{pmatrix} U_1 & 0 \\ 0 & 0 \end{pmatrix}$. By $\begin{pmatrix} U & V \\ V^* & W \end{pmatrix} \geq 0$ we get $V = \begin{pmatrix} V_1 \\ 0 \end{pmatrix}$. Using this information we further rewrite (51)–(53) to

$$A_1^*\Delta_1 + \Delta_1 A_1 - \Delta_1 B_1 B_1^* \Delta_1 = U_1, \tag{54}$$

$$(A_1 - B_1B_1^*\Delta_1)^*Y_1 + Y_1M + \Delta_1(F_1 - B_1B_2^*Y_2) + (\tilde{X}F + R)_1 = V_1, \tag{55}$$

$$A_2^*Y_2 + Y_2M + (\tilde{X}F + R)_2 = 0, \tag{56}$$

$$M^*Z + ZM + (F_1 - B_1B_2^*Y_2)^*Y_1 + Y_1^*(F_1 - B_1B_2^*Y_2) - Y_1^*B_1B_1^*Y_1 + \tilde{S} = W, \tag{57}$$

with $\tilde{S} := F_2^*Y_2 + Y_2^*F_2 - Y_2^*B_2B_2^*Y_2 + S$ being independent of Y_1.

If we fix Y_2 and if we concentrate on (54), (55), and (57), we infer that they belong to an ARI for *regular* data. Hence we can apply what we have proved in the first step: Since $\sigma(A_1) \subset \mathcal{C}^-$ and $(A_1 - sI\ B_1)$ is stabilizable, the (unique) solution of (54) with $\sigma(A_1 - B_1B^*{}_1\Delta) \subset \mathcal{C}^- \cup \mathcal{C}^0$ is $\tilde{\Delta}_1 = 0$. As shown above, the unique

solution of $A_1^*\tilde{Y}_1 + \tilde{Y}_1 M + (\tilde{X}F + R)_1 = 0$ yields $R_j^*[(F_1 - B_1B_2^*Y_2)^*\tilde{Y}_1 + \tilde{Y}_1^*(F_1 - B_1B_2^*Y_2) - \tilde{Y}_1^*B_1B_1^*\tilde{Y}_1 + \tilde{S}]R_j \geq 0$ for all $j = 1, \ldots, l$. Due to (56), it is easily seen that $\tilde{Y} = (\tilde{Y}_1^* \; Y_2^*)^*$ is as required. □

Acknowledgments. I would like to thank Professor Khargonekar and the other members of the EECS Department at the University of Michigan in Ann Arbor for their excellent hospitality. I would also like to express my gratitude to Professor Wimmer from the University of Würzburg, Germany, who pointed out the route to a short and elegant proof of Theorem 4.

REFERENCES

[1] T. ANDO, *Matrix Quadratic Equations*, Lecture Notes, Hokkaido University, Sapporo, Japan, 1988.

[2] W. A. COPPEL, *Matrix quadratic equations*, Bull. Austral. Math. Soc., 10 (1974), pp. 377–401.

[3] A. N. ČURILOV, *On the solutions of quadratic matrix equations*, Nonlinear Vibrations and Control Theory, 2 (1978), pp. 24–33.

[4] L. E. FAIBUSOVICH, *Algebraic Riccati equation and symplectic algebra*, Internat. J. Control, 43 (1986), pp. 781–792.

[5] ———, *Matrix Riccati inequality: existence of solutions*, Systems Control Lett., 9 (1987), pp. 59–64.

[6] F. R. GANTMACHER, *Matrizentheorie*, Springer-Verlag, Berlin, 1986.

[7] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *On Hermitian solutions of the symmetric algebraic Riccati equation*, SIAM J. Control Optim., 24 (1986), pp. 1323–1334.

[8] ———, *Invariant Subspaces of Matrices with Applications*, John Wiley & Sons, New York, 1986.

[9] V. KUČERA, *The matrix equation* $AX + XB = C^*$, SIAM J. Appl. Math., 26 (1974), pp. 15–25.

[10] P. LANCASTER AND L. RODMAN, *Existence and uniqueness theorems for the algebraic Riccati equation*, Internat. J. Control, 32 (1980), pp. 285–309.

[11] B. P. MOLINARI, *The stabilizing solution of the algebraic Riccati equation*, SIAM J. Control Optim., 11 (1973), pp. 262–271.

[12] ———, *Equivalence relations for the algebraic Riccati equation*, SIAM J. Control Optim., 11 (1973), pp. 272–285.

[13] ———, *The time-invariant Linear-Quadratic optimal control problem*, Automatica, 13 (1977), pp. 347–357.

[14] A. C. M. RAN AND L. RODMAN, *On parameter dependence of solutions of algebraic Riccati equations*, Math. Control Signals Systems, 1 (1988), pp. 269–284.

[15] A. C. M. RAN AND R. VREUGDENHIL, *Existence and comparison theorems for algebraic Riccati equations for continuous- and discrete-time systems*, Linear Algebra Appl., 99 (1988), pp. 63–83.

[16] C. W. SCHERER, *The Riccati Inequality and State-Space* H_∞*-Optimal Control*, Ph.D. thesis, Universität Würzburg, Germany, 1990.

[17] ———, *The solution set of the algebraic Riccati equation and the algebraic Riccati inequality*, Linear Algebra Appl., 153 (1991), pp. 99–122.

[18] ———, H_∞*-control by state-feedback for plants with zeros on the imaginary axis*, SIAM J. Control Optim., 30 (1992), pp. 123–142.

[19] ———, *The state-feedback* H_∞*-problem at optimality*, Automatica, 30 (1994), pp. 293–305.

[20] M. A. SHAYMAN, *Geometry of the algebraic Riccati equation., Part* I, SIAM J. Control Optim., 21 (1983), pp. 375–394.

[21] ———, *Geometry of the algebraic Riccati equation., Part* II, SIAM J. Control Optim., 21 (1983), pp. 395–412.

[22] J. SNYDERS AND M. ZAKAI, *On nonnegative solutions of the equation* $AD + DA' = -C$, SIAM J. Appl. Math., 18 (1970), pp. 704–714.

[23] J. STOER AND C. WITZGALL, *Convexity and Optimization in Finite Dimensions* I, Springer-Verlag, Berlin, 1970.

[24] J. C. WILLEMS, *Least-squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 21 (1971), pp. 319–338.

[25] H. K. WIMMER, *The algebraic Riccati equation without complete controllability*, SIAM J. Discrete Algebraic Meth., 3 (1982), pp. 1–12.

[26] ———, *The algebraic Riccati equation: Conditions for the existence and uniqueness of solutions*, Linear Algebra Appl., 58 (1984), pp. 441–452.

[27] ———, *Monotonicity of maximal solutions of algebraic Riccati equations*, Systems Control Lett., 5 (1985), pp. 317–319.

[28] ———, *Decomposition and parametrization of semidefinite solutions of the continuous-time algebraic Riccati equation*, SIAM J. Control Optim., 32 (1994), pp. 995–1007.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, pp. 1328–1340, October 1995

020

PERTURBATION BOUNDS FOR THE GENERALIZED SCHUR DECOMPOSITION*

JI-GUANG SUN†

Abstract. This paper uses a technique described by M. M. Konstantinov, P. Hr. Petkov, and N. D. Christov [*SIAM J. Matrix Anal. Appl.*, 15 (1994), pp. 383–392] to derive perturbation bounds for the generalized Schur decomposition of a regular matrix pair with distinct eigenvalues.

Key words. generalized Schur decomposition, perturbation bound, condition number

AMS subject classifications. 15A21, 65F15, 93B35

1. Introduction.

Let A, B be complex $n \times n$ matrices, and let (A, B) be a regular matrix pair, i.e., A and B satisfy $\det(A - \lambda B) \not\equiv 0$. The generalized Schur decomposition of (A, B) is a decomposition of the form

$$A = UTV^{\mathrm{H}}, \quad B = URV^{\mathrm{H}}, \tag{1.1}$$

where T and R are $n \times n$ upper triangular matrices, U and V are $n \times n$ unitary matrices, V^{H} denotes the conjugate transpose of V (see [4]). The matrix pair (T, R) is called the generalized Schur form of the matrix pair (A, B). The diagonal elements of $T = (t_{ij})$ and $R = (r_{ij})$ reveal the eigenvalues of the matrix pair (A, B) and they are the pairs (t_{ii}, r_{ii}), or equivalently, $\lambda_i = t_{ii}/r_{ii}$ (if $r_{ii} \neq 0$) and $\lambda_i = \infty$ (if $r_{ii} = 0$). It is known that the generalized Schur decomposition is an important tool in linear algebra and control theory (e.g., [1], [2], [7]).

The generalized Schur decomposition of a regular matrix pair (A, B) is obviously a generalization of the Schur decomposition of a square matrix A : $A = UTU^{\mathrm{H}}$, where U is a unitary matrix, and T is an upper triangular matrix (see, e.g., [1], [7]). Recently, M. M. Konstantinov, P. Hr. Petkov, and N. D. Christov [3] presented a perturbation analysis of the Schur decomposition. After that, [8] derives new perturbation bounds of the Schur factors U and T by using the technique described in [3]. The new results are qualitatively the same, but somewhat simpler than the corresponding results of [3]. Extending the new results to regular matrix pairs, we get perturbation bounds for the generalized Schur factors U, V, T, R of a regular matrix pair (A, B). This paper, as a rewrite and extension of [8, §§5–6], presents the perturbation bounds and illustrates them by some numerical examples.

In §2 we derive perturbation equations. In §3 we discuss basic properties of the operator $\mathbf{L}$ (defined below by (2.9)) and the function $l(T, R)$ (defined below by (3.9)) that are important for studying perturbation bounds for the generalized Schur decomposition. In §4, a perturbation theorem for the generalized Schur decomposition is proved. Numerical results are given in §5.

Throughout this paper the symbol $\mathcal{C}^{m\times n}$ denotes the set of complex $m \times n$ matrices, and $\mathcal{C}^n = \mathcal{C}^{n\times 1}$. The matrix A^{T} is the transpose of A. I is the identity matrix, and 0 is the null matrix. $\mathcal{U}^{n\times n}(\mathcal{U}_{\mathrm{s}}^{n\times n})$ denotes the set of $n \times n$ upper (strictly upper) triangular matrices, $\mathcal{L}_{\mathrm{s}}^{n\times n}$ the set of $n\times n$ strictly lower triangular matrices, and $\mathcal{D}^{n\times n}$

* Received by the editors December 31, 1992; accepted for publication (in revised form) by G. Golub December 13, 1994.

† Department of Computing Science, Umeå University, S-901 87 Umeå, Sweden. This work was supported by the Swedish Natural Science Research Council contract F-FU 6952-300 and the Department of Computing Science, Umeå University (jisun@cs.umu.se).

the set of $n \times n$ diagonal matrices. $\| \ \|_2$ denotes the Euclidean vector norm and the spectral matrix norm, and $\| \ \|_F$ the Frobenius matrix norm.

It is evident that any $X \in \mathcal{C}^{n\times n}$ can be split uniquely as

$$X = X_L + X_D + X_U, \quad X_L \in \mathcal{L}_s^{n\times n}, \ X_D \in \mathcal{D}^{n\times n}, \ X_U \in \mathcal{U}_s^{n\times n}. \tag{1.2}$$

As in [3], the matrices X_L, X_D, X_U of (1.2) will be denoted by

$$X_L = \text{low}(X), \ X_D = \text{diag}(X), \ X_U = \text{up}(X). \tag{1.3}$$

The relation (1.3) gives the definition of the operators low( ), diag( ), and up( ) defined on $\mathcal{C}^{n\times n}$.

2. Perturbation equations. Let (A, B) and $(\tilde{A}, \tilde{B})$ be two regular matrix pairs of order n. Let (1.1) be the generalized Schur decomposition of (A, B), and let

$$\tilde{A} = \tilde{U}\tilde{T}\tilde{V}^H, \quad \tilde{B} = \tilde{U}\tilde{R}\tilde{V}^H \tag{2.1}$$

be the generalized Schur decomposition of $(\tilde{A}, \tilde{B})$. Write

$$E = \tilde{A} - A, \ F = \tilde{B} - B, \ W = \tilde{U} - U, \ Z = \tilde{V} - V, \ G = \tilde{T} - T, \ H = \tilde{R} - R. \tag{2.2}$$

Then the perturbation matrices W, Z, G, H satisfy the equations

$$E\tilde{V} + AZ = WT + \tilde{U}G, \quad F\tilde{V} + BZ = WR + \tilde{U}H. \tag{2.3}$$

Let

$$\tilde{E} = \tilde{U}^H E\tilde{V}, \ \tilde{F} = \tilde{U}^H F\tilde{V}, \ X = U^H W, \ Y = V^H Z. \tag{2.4}$$

Then from (2.3) we get

$$G = \tilde{E} + \tilde{U}^H U(TY - XT), \quad H = \tilde{F} + \tilde{U}^H U(RY - XR) \tag{2.5}$$

and

$$\|G\|_F \le \|E\|_F + \theta_T \|(W, Z)\|_F, \quad \|H\|_F \le \|F\|_F + \theta_R \|(W, Z)\|_F,$$

where

$$\theta_T = \|(I \otimes T, \ T^T \otimes I)\|_2, \quad \theta_R = \|(I \otimes R, \ R^T \otimes I)\|_2, \tag{2.6}$$

in which $A \otimes B \equiv (\alpha_{ij}B)$ is a Kronecker product.

Thus, the problem is reduced to investigating perturbation bounds of the unitary factors U and V of the generalized Schur decomposition (1.1) of (A, B), i.e., to seek upper bounds of $\|(W, Z)\|_F = \|(X, Y)\|_F$.

Combining the relation $\tilde{U}^H U = (I + X)^H$ with (2.5) we get

$$\begin{aligned} TY - XT &= G + X^H XT - X^H TY - \tilde{E}, \\ RY - XR &= H + X^H XR - X^H RY - \tilde{F}. \end{aligned} \tag{2.7}$$

Moreover, the matrices X, Y satisfy

$$X + X^H + X^H X = 0, \quad Y + Y^H + Y^H Y = 0. \tag{2.8}$$

Let (X, Y) be a solution to (2.7)–(2.8). Define the matrices $X_L, Y_L, X_D, Y_D, X_U, Y_U$ by (1.2)–(1.3), and define the operator $\mathbf{L} : \mathcal{L}_s^{n\times n} \times \mathcal{L}_s^{n\times n} \to \mathcal{L}_s^{n\times n} \times \mathcal{L}_s^{n\times n}$ by

$$\mathbf{L}(X_L, Y_L) = (\mathrm{low}(TY_L - X_L T),\ \mathrm{low}(RY_L - X_L R)). \tag{2.9}$$

Then from (2.7)

$$\begin{aligned}\mathbf{L}(X_L, Y_L) = &(\mathrm{low}(X^H XT - X^H TY),\ \mathrm{low}(X^H XR - X^H RY)) \\ &- (\mathrm{low}(\tilde{E}),\ \mathrm{low}(\tilde{F})).\end{aligned} \tag{2.10}$$

Choose a generalized Schur decomposition of $(\tilde{A}, \tilde{B})$ expressed by (2.1) so that the diagonal elements of $U^H\tilde{U}$ and $V^H\tilde{V}$ are real. Then from (2.8)

$$(X_D,\ Y_D) = -\frac{1}{2}(\mathrm{diag}(X^H X),\ \mathrm{diag}(Y^H Y)). \tag{2.11}$$

Furthermore, the relation (2.8) gives

$$(X_U,\ Y_U) = -(X_L^H,\ Y_L^H) - (\mathrm{up}(X^H X),\ \mathrm{up}(Y^H Y)). \tag{2.12}$$

In §4 we use the perturbation equations (2.10)–(2.12) to seek an upper bound of $\|(X, Y)\|_F$.

3. The operator L and function $l(T, R)$. Before we go on to derive perturbation bounds of the unitary factors U, V in the generalized Schur decomposition of (A, B) from (2.10)–(2.12), it will be necessary to determine when the operator $\mathbf{L}$ defined by (2.9) is nonsingular.

THEOREM 3.1. *Let* $\mathbf{L}$ *be the operator defined by* (2.9), *where*

$$T = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ 0 & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t_{nn} \end{pmatrix}, \quad R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{nn} \end{pmatrix},$$

and (T, R) *is a regular matrix pair. Then* $\mathbf{L}$ *is nonsingular if and only if all the eigenvalues* (t_{ii}, r_{ii}) *of the matrix pair* (T, R) *are simple, i.e.,*

$$\begin{vmatrix} t_{ii} & t_{jj} \\ r_{ii} & r_{jj} \end{vmatrix} \neq 0 \quad \forall i \neq j, \quad i, j = 1, \ldots, n. \tag{3.1}$$

Proof. First suppose that the condition (3.1) is satisfied. We must show that for any fixed $P = (p_1, \ldots, p_n), Q = (q_1, \ldots, q_n) \in \mathcal{L}_s^{n\times n}$ the system

$$\mathrm{low}(TY_L - X_L T) = P, \quad \mathrm{low}(RY_L - X_L R) = Q \tag{3.2}$$

has a unique solution $X_L = (x_1, \ldots, x_n), Y_L = (y_1, \ldots, y_n) \in \mathcal{L}_s^{n\times n}$.

For any $v = (\nu_1, \ldots, \nu_n)^T \in \mathcal{C}^n$, define the column vector

$$\mathrm{low}_k(v) = \begin{pmatrix} 0 \\ v^{(k)} \end{pmatrix} \in \mathcal{C}^n, \quad v^{(k)} = (\nu_{k+1}, \ldots, \nu_n)^T \in \mathcal{C}^{n-k}, \quad k = 1, \ldots, n-1.$$

Then the kth columns p_k, q_k, x_k, y_k of $P, Q, X_{\mathrm{L}}, Y_{\mathrm{L}}$ can be expressed by

$$p_k = \begin{pmatrix} 0 \\ p_k^{(k)} \end{pmatrix}, \quad q_k = \begin{pmatrix} 0 \\ q_k^{(k)} \end{pmatrix}, \quad x_k = \begin{pmatrix} 0 \\ x_k^{(k)} \end{pmatrix}, \quad y_k = \begin{pmatrix} 0 \\ y_k^{(k)} \end{pmatrix},$$

where $p_k^{(k)}, q_k^{(k)}, x_k^{(k)}, y_k^{(k)} \in \mathcal{C}^{n-k}$, $k = 1, \ldots, n-1$. Moreover, let

$$T^{(k)} = \begin{pmatrix} t_{k+1,k+1} & t_{k+1,k+2} & \cdots & t_{k+1,n} \\ 0 & t_{k+2,k+2} & \cdots & t_{k+2,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t_{nn} \end{pmatrix},$$

$$R^{(k)} = \begin{pmatrix} r_{k+1,k+1} & r_{k+1,k+2} & \cdots & r_{k+1,n} \\ 0 & r_{k+2,k+2} & \cdots & r_{k+2,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{nn} \end{pmatrix}$$

for $k = 1, \ldots, n-1$. Then the first columns of the two equations of the system (3.2) can be expressed by

$$\mathrm{low}_1(Ty_1 - t_{11}x_1) = p_1, \quad \mathrm{low}_1(Ry_1 - r_{11}x_1) = q_1, \tag{3.3}$$

which is equivalent to

$$T^{(1)}y_1^{(1)} - t_{11}x_1^{(1)} = p_1^{(1)}, \quad R^{(1)}y_1^{(1)} - r_{11}x_1^{(1)} = q_1^{(1)}.$$

Since $\begin{vmatrix} t_{ii} & t_{11} \\ r_{ii} & r_{11} \end{vmatrix} \neq 0 \quad \forall i > 1$, the matrix $\begin{pmatrix} T^{(1)} & -t_{11}I \\ R^{(1)} & -r_{11}I \end{pmatrix}$ is nonsingular. Hence the system (3.3) has a unique solution x_1, y_1.

Now suppose that $x_1, y_1, \ldots, x_{k-1}, y_{k-1}$ are uniquely determined. From (3.2)

$$\mathrm{low}_k\left(Ty_k - \sum_{l=1}^{k} t_{lk}x_l\right) = p_k, \quad \mathrm{low}_k\left(Ry_k - \sum_{l=1}^{k} r_{lk}x_l\right) = q_k, \tag{3.4}$$

which is equivalent to

$$T^{(k)}y_k^{(k)} - t_{kk}x_k^{(k)} = p_k^{(k)} + \textstyle\sum_{l=1}^{k-1} t_{lk}x_l^{(k)},$$

$$R^{(k)}y_k^{(k)} - r_{kk}x_k^{(k)} = q_k^{(k)} + \textstyle\sum_{l=1}^{k-1} r_{lk}x_l^{(k)}.$$

Since $\begin{vmatrix} t_{ii} & t_{kk} \\ r_{ii} & r_{kk} \end{vmatrix} \neq 0$ for all $i > k$, the matrix $\begin{pmatrix} T^{(k)} & -t_{kk}I \\ R^{(k)} & -r_{kk}I \end{pmatrix}$ is nonsingular. Hence the system (3.4) has a unique solution x_k, y_k. Thus, we have proved that the system (3.2) has a unique solution $X_{\mathrm{L}}, Y_{\mathrm{L}} \in \mathcal{L}_{\mathrm{s}}^{n\times n}$ for any fixed $P, Q \in \mathcal{L}_{\mathrm{s}}^{n\times n}$.

Conversely, we can prove that the operator $\mathbf{L}$ is singular without the condition (3.1).

Let (j, i) be an index-pair satisfying the following conditions: (i) $1 \leq j < i \leq n$,

(ii) $\begin{vmatrix} t_{jj} & t_{ii} \\ r_{jj} & r_{ii} \end{vmatrix} = 0$, (iii) there is no any another index-pair (l,k) so that $j \le l < k \le i$, $(l,k) \ne (j,i)$, and $\begin{vmatrix} t_{ll} & t_{kk} \\ r_{ll} & r_{kk} \end{vmatrix} = 0$. Now we are going to prove that there exist matrices $X_{\mathrm{L}}, Y_{\mathrm{L}} \in \mathcal{L}_{\mathrm{s}}^{n\times n}$ with the form

$$X_{\mathrm{L}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & X_{\mathrm{L}}^{(j,i)} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad Y_{\mathrm{L}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & Y_{\mathrm{L}}^{(j,i)} & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

where

$$X_{\mathrm{L}}^{(j,i)} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ \xi_{j+1,j} & 0 & \cdots & 0 & 0 \\ \xi_{j+2,j} & \xi_{j+2,j+1} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \xi_{ij} & \xi_{i,j+1} & \cdots & \xi_{i,i-1} & 0 \end{pmatrix},$$

$$Y_{\mathrm{L}}^{(j,i)} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ \eta_{j+1,j} & 0 & \cdots & 0 & 0 \\ \eta_{j+2,j} & \eta_{j+2,j+1} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \eta_{ij} & \eta_{i,j+1} & \cdots & \eta_{i,i-1} & 0 \end{pmatrix},$$

and $(X_{\mathrm{L}}^{(j,i)}, Y_{\mathrm{L}}^{(j,i)}) \ne (0,0)$ such that

$$\mathrm{low}(TY_{\mathrm{L}} - X_{\mathrm{L}}T) = 0, \quad \mathrm{low}(RY_{\mathrm{L}} - X_{\mathrm{L}}R) = 0. \tag{3.5}$$

Write

$$T = \begin{pmatrix} * & * & * \\ 0 & T^{(j,i)} & * \\ 0 & 0 & * \end{pmatrix}, \quad R = \begin{pmatrix} * & * & * \\ 0 & R^{(j,i)} & * \\ 0 & 0 & * \end{pmatrix},$$

where

$$T^{(j,i)} = \begin{pmatrix} t_{jj} & t_{j,j+1} & \cdots & t_{ji} \\ 0 & t_{j+1,j+1} & \cdots & t_{j+1,i} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t_{ii} \end{pmatrix}, \quad R^{(j,i)} = \begin{pmatrix} r_{jj} & r_{j,j+1} & \cdots & r_{ji} \\ 0 & r_{j+1,j+1} & \cdots & r_{j+1,i} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{ii} \end{pmatrix}.$$

Then the system (3.5) is equivalent to

$$\begin{aligned} &\mathrm{low}(T^{(j,i)}Y_{\mathrm{L}}^{(j,i)} - X_{\mathrm{L}}^{(j,i)}T^{(j,i)}) = 0, \\ &\mathrm{low}(R^{(j,i)}Y_{\mathrm{L}}^{(j,i)} - X_{\mathrm{L}}^{(j,i)}R^{(j,i)}) = 0. \end{aligned} \tag{3.6}$$

Observing the $(i-j+1,1)$-elements of the two matrix equations of (3.6), we have

$$t_{ii}\eta_{ij} - t_{jj}\xi_{ij} = 0, \quad r_{ii}\eta_{ij} - r_{jj}\xi_{ij} = 0. \tag{3.7}$$

Since $\begin{vmatrix} t_{jj} & t_{ii} \\ r_{jj} & r_{ii} \end{vmatrix} = 0$, we can take a nonzero solution (ξ_{ij}, η_{ij}) to (3.7). Furthermore, observing the $(i-j+1,2)$-elements of the two matrix equations of (3.6), we have

$$\begin{aligned} t_{ii}\eta_{i,j+1} - t_{j+1,j+1}\xi_{i,j+1} &= t_{j,j+1}\xi_{ij}, \\ r_{ii}\eta_{i,j+1} - r_{j+1,j+1}\xi_{i,j+1} &= r_{j,j+1}\xi_{ij}. \end{aligned} \tag{3.8}$$

Since $\begin{vmatrix} t_{j+1,j+1} & t_{ii} \\ r_{j+1,j+1} & r_{ii} \end{vmatrix} \neq 0$, we get a solution $(\xi_{i,j+1}, \eta_{i,j+1})$ to (3.8). After that we can get

$$\xi_{i,j+2}, \eta_{i,j+2}, \ldots, \xi_{i,i-1}, \eta_{i,i-1}, \xi_{i-1,j}, \eta_{i-1,j}, \ldots, \xi_{i-1,i-2}, \eta_{i-1,i-2}, \ldots, \xi_{j+1,j}, \eta_{j+1,j},$$

successively. This means that without the condition (3.1) there exists a nonzero solution $(X_{\mathrm{L}}, Y_{\mathrm{L}})$ to (3.5), and so the operator $\mathbf{L}$ is singular. □

Let $T, R \in \mathcal{U}^{n\times n}$, and let $\mathbf{L}$ be the operator defined by (2.9). Now we define the function $l(T,R)$ by

$$l(T,R) = \min_{\substack{X_{\mathrm{L}}, Y_{\mathrm{L}} \in \mathcal{L}_{\mathrm{s}}^{n\times n} \\ \|(X_{\mathrm{L}}, Y_{\mathrm{L}})\|_{\mathrm{F}} = 1}} \|\mathbf{L}(X_{\mathrm{L}}, Y_{\mathrm{L}})\|_{\mathrm{F}}. \tag{3.9}$$

It is easy to verify that if $\mathbf{L}$ is nonsingular then

$$l(T,R) = \|\mathbf{L}^{-1}\|^{-1}, \tag{3.10}$$

where $\|\mathbf{L}^{-1}\|$ is defined by

$$\|\mathbf{L}^{-1}\| = \max_{\substack{X_{\mathrm{L}}, Y_{\mathrm{L}} \in \mathcal{L}_{\mathrm{s}}^{n\times n} \\ \|(X_{\mathrm{L}}, Y_{\mathrm{L}})\|_{\mathrm{F}} = 1}} \|\mathbf{L}^{-1}(X_{\mathrm{L}}, Y_{\mathrm{L}})\|_{\mathrm{F}}. \tag{3.11}$$

It is worthwhile to point out that if the generalized Schur form $(T,R) = ((t_{ij}),(r_{ij}))$ of a regular matrix pair (A,B) is known, then the function $l(T,R)$ is computable. Now we show how to compute $l(T,R)$. Let $X = (\xi_{ij}), Y = (\eta_{ij}) \in \mathcal{C}^{n\times n}$, and let

$$(P,Q) = \mathbf{L}(X_{\mathrm{L}}, Y_{\mathrm{L}}), \tag{3.12}$$

where $\mathbf{L}$ is the operator defined by (2.9), and $P = (p_{ij}), Q = (q_{ij}) \in \mathcal{L}_{\mathrm{s}}^{n\times n}$. Let

$$x_j^{(\mathrm{L})} = \begin{pmatrix} \xi_{j+1,j} \\ \vdots \\ \xi_{nj} \end{pmatrix}, \quad y_j^{(\mathrm{L})} = \begin{pmatrix} \eta_{j+1,j} \\ \vdots \\ \eta_{nj} \end{pmatrix}, \quad p_j^{(\mathrm{L})} = \begin{pmatrix} p_{j+1,j} \\ \vdots \\ p_{nj} \end{pmatrix}, \quad q_j^{(\mathrm{L})} = \begin{pmatrix} q_{j+1,j} \\ \vdots \\ q_{nj} \end{pmatrix}$$

and

$$x^{(\mathrm{L})} = \begin{pmatrix} x_1^{(\mathrm{L})} \\ \vdots \\ x_{n-1}^{(\mathrm{L})} \end{pmatrix}, \quad y^{(\mathrm{L})} = \begin{pmatrix} y_1^{(\mathrm{L})} \\ \vdots \\ y_{n-1}^{(\mathrm{L})} \end{pmatrix}, \quad p^{(\mathrm{L})} = \begin{pmatrix} p_1^{(\mathrm{L})} \\ \vdots \\ p_{n-1}^{(\mathrm{L})} \end{pmatrix}, \quad q^{(\mathrm{L})} = \begin{pmatrix} q_1^{(\mathrm{L})} \\ \vdots \\ q_{n-1}^{(\mathrm{L})} \end{pmatrix}.$$

Then from (3.12) and (2.9)

$$\begin{pmatrix} p^{(\mathrm{L})} \\ q^{(\mathrm{L})} \end{pmatrix} = L \begin{pmatrix} x^{(\mathrm{L})} \\ y^{(\mathrm{L})} \end{pmatrix} \quad \text{with} \quad L = \begin{pmatrix} -L_{P,X} & L_{P,Y} \\ -L_{Q,X} & L_{Q,Y} \end{pmatrix}, \tag{3.13}$$

where $L_{P,X}$ and $L_{Q,X}$ are block lower triangular matrices, $L_{P,Y}$ and $L_{Q,Y}$ are block diagonal matrices. For example, for $n = 5$ the matrices $L_{P,X}$ and $L_{P,Y}$ have the forms

$$L_{P,X} = \begin{pmatrix} t_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\ 0 & t_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & t_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & t_{11} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & t_{12} & 0 & 0 & t_{22} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & t_{12} & 0 & 0 & t_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & t_{12} & 0 & 0 & t_{22} & 0 & 0 & 0 \\ 0 & 0 & t_{13} & 0 & 0 & t_{23} & 0 & t_{33} & 0 & 0 \\ 0 & 0 & 0 & t_{13} & 0 & 0 & t_{23} & 0 & t_{33} & 0 \\ 0 & 0 & 0 & t_{14} & 0 & 0 & t_{24} & 0 & t_{34} & t_{44} \end{pmatrix}$$

and

$$L_{P,Y} = \begin{pmatrix} t_{22} & t_{23} & t_{24} & t_{25} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & t_{33} & t_{34} & t_{35} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & t_{44} & t_{45} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & t_{55} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & t_{33} & t_{34} & t_{35} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & t_{44} & t_{45} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & t_{55} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & t_{44} & t_{45} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & t_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & t_{55} \end{pmatrix}$$

The matrices $L_{Q,X}$ and $L_{Q,Y}$ have the same forms as $L_{P,X}$ and $L_{P,Y}$. Moreover, the elements t_{ij} are replaced by r_{ij}. The relation (3.13) shows that L is the matrix representation of the operator $\mathbf{L}$. Combining (3.13) with (3.10) and (3.11) we know that the function $l(T, R)$ can be computed by

$$l(T, R) = \|L^{-1}\|_2^{-1}. \tag{3.14}$$

The following result shows that the function $l(T, R)$ is insensitive with respect to perturbations of T, R. The proof is similar to those of [6, Thm. 4.6] and [8, Thm. 3.4].

THEOREM 3.2. *Let $T, R, M, K \in \mathcal{U}^{n\times n}$. Then*

$$l(T, R) - \theta_{\mathrm{M,K}} \le l(T + M, R + K) \le l(T, R) + \theta_{\mathrm{M,K}}, \tag{3.15}$$

where

$$\theta_{\mathrm{M,K}} = \left\| \begin{pmatrix} I \otimes M & M^{\mathrm{T}} \otimes I \\ I \otimes K & K^{\mathrm{T}} \otimes I \end{pmatrix} \right\|_2 \le \sqrt{2}(\|M\|_2^2 + \|K\|_2^2)^{1/2}.$$

Proof. By the definition (3.9), we have

$$\begin{aligned} l(T, R) &= \min\{\|\mathbf{L}(X_{\mathrm{L}}, Y_{\mathrm{L}})\|_{\mathrm{F}} \ : \ X_{\mathrm{L}}, Y_{\mathrm{L}} \in \mathcal{L}_{\mathrm{s}}^{n\times n}, \ \|(X_{\mathrm{L}}, Y_{\mathrm{L}})\|_{\mathrm{F}} = 1\} \\ &= \|(\mathrm{low}(TY_{\mathrm{L}}^* - X_{\mathrm{L}}^* T), \ \mathrm{low}(RY_{\mathrm{L}}^* - X_{\mathrm{L}}^* R))\|_{\mathrm{F}}, \\ &\quad X_{\mathrm{L}}^*, Y_{\mathrm{L}}^* \in \mathcal{L}_{\mathrm{s}}^{n\times n}, \ \ \|(X_{\mathrm{L}}^*, Y_{\mathrm{L}}^*)\|_{\mathrm{F}} = 1 \end{aligned}$$

and

$$
\begin{aligned}
& l(T+M, R+K) \\
& = \min\{\|(\operatorname{low}((T+M)Y_{\mathrm{L}} - X_{\mathrm{L}}(T+M)),\ \operatorname{low}((R+K)Y_{\mathrm{L}} - X_{\mathrm{L}}(R+K)))\|_{\mathrm{F}} : \\
& \qquad X_{\mathrm{L}}, Y_{\mathrm{L}} \in \mathcal{L}_{\mathrm{s}}^{n\times n},\ \|(X_{\mathrm{L}}, Y_{\mathrm{L}})\|_{\mathrm{F}} = 1\} \\
& \le \|(\operatorname{low}((T+M)Y_{\mathrm{L}}^* - X_{\mathrm{L}}^*(T+M)),\ \operatorname{low}((R+K)Y_{\mathrm{L}}^* - X_{\mathrm{L}}^*(R+K)))\|_{\mathrm{F}} \\
& \le \|(\operatorname{low}(TY_{\mathrm{L}}^* - X_{\mathrm{L}}^*T),\ \operatorname{low}(RY_{\mathrm{L}}^* - X_{\mathrm{L}}^*R))\|_{\mathrm{F}} \\
& \quad + \|(\operatorname{low}(MY_{\mathrm{L}}^* - X_{\mathrm{L}}^*M),\ \operatorname{low}(KY_{\mathrm{L}}^* - X_{\mathrm{L}}^*K))\|_{\mathrm{F}} \\
& \le l(T,R) + \theta_{\mathrm{M,K}}\|(X_{\mathrm{L}}^*, Y_{\mathrm{L}}^*)\|_{\mathrm{F}} \\
& = l(T,R) + \theta_{\mathrm{M,K}}.
\end{aligned}
$$

Similarly, we can prove the first inequality of (3.15). □

Remark 3.3. Let (A, B) be a regular matrix pair. Note that the value of the function $l(T, R)$ is dependent on the choice of the generalized Schur form (T, R) of (A, B). For example, we consider a matrix pair

$$
(A,B) = \left(\begin{pmatrix} 3 & 0 & 0 \\ 1 & 2 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right).
$$

Take two different generalized Schur decompositions of (A, B) : $A = U_j T_j V_j^{\mathrm{T}}$, $B = U_j R_j V_j^{\mathrm{T}}$, $j = 1, 2$, where

$$
U_1 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad V_1 = U_1, \quad (T_1, R_1) = \left(\begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)
$$

and

$$
U_2 = \begin{pmatrix} 0 & 0 & 1 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \end{pmatrix}, \quad V_2 = U_2, \quad (T_2, R_2) = \left(\begin{pmatrix} 2 & 1 & \frac{1}{\sqrt{2}} \\ 0 & 1 & \frac{1}{\sqrt{2}} \\ 0 & 0 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right).
$$

By (3.14) we have

$$
l(T_1, R_1) = \left\| \begin{pmatrix} -1 & 0 & 0 & 2 & 1 & 0 \\ 0 & -1 & 0 & 0 & 3 & 0 \\ 0 & -1 & 2 & 0 & 0 & 3 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 \end{pmatrix}^{-1} \right\|_2^{-1} \approx 0.218
$$

and

$$l(T_2,R_2)=\left\|\begin{pmatrix} -2 & 0 & 0 & 1 & \frac{1}{\sqrt{2}} & 0\\ 0 & -2 & 0 & 0 & 3 & 0\\ 0 & -1 & -1 & 0 & 0 & 3\\ -1 & 0 & 0 & 1 & 0 & 0\\ 0 & -1 & 0 & 0 & 1 & 0\\ 0 & 0 & -1 & 0 & 0 & 1 \end{pmatrix}^{-1}\right\|_2^{-1}\approx 0.192.$$

Obviously, $l(T_1,R_1)\neq l(T_2,R_2)$.

4. Perturbation theorem. Now we are going to derive upper bounds of the solution (X,Y) to (2.10)–(2.12) under the assumption that all the eigenvalues of (T,R) are simple.

By Theorem 3.1, the operator $\mathbf{L}$ defined by (2.9) is nonsingular. Thus, (2.10)–(2.12) can be rewritten as a continuous mapping $\Phi : \mathcal{C}^{n\times n}\times\mathcal{C}^{n\times n}\to\mathcal{C}^{n\times n}\times\mathcal{C}^{n\times n}$ expressed by

$$\begin{aligned}(X_{\mathrm{L}},Y_{\mathrm{L}}) &= \mathbf{L}^{-1}(\mathrm{low}(X^{\mathrm{H}}XT-X^{\mathrm{H}}TY),\mathrm{low}(X^{\mathrm{H}}XR-X^{\mathrm{H}}RY))\\ &\quad -\mathbf{L}^{-1}(\mathrm{low}(\tilde{E}),\mathrm{low}(\tilde{F})),\\ (X_{\mathrm{D}},Y_{\mathrm{D}}) &= -\frac{1}{2}(\mathrm{diag}(X^{\mathrm{H}}X),\mathrm{diag}(Y^{\mathrm{H}}Y)),\\ (X_{\mathrm{U}},Y_{\mathrm{U}}) &= -(X_{\mathrm{L}}^{\mathrm{H}},Y_{\mathrm{L}}^{\mathrm{H}})-(\mathrm{up}(X^{\mathrm{H}}X),\mathrm{up}(Y^{\mathrm{H}}Y)).\end{aligned}\tag{4.1}$$

Observe the following facts:

1. By (3.10), $\|\mathbf{L}^{-1}\|=1/l(T,R)$.

2. From

$$\|(\mathrm{low}(X^{\mathrm{H}}TY),\mathrm{low}(X^{\mathrm{H}}RY))\|_{\mathrm{F}}\le\frac{1}{2}\sqrt{\|A\|_2^2+\|B\|_2^2}\,\|(X,Y)\|_{\mathrm{F}}^2$$

and

$$\|(\mathrm{low}(X^{\mathrm{H}}XT),\mathrm{low}(X^{\mathrm{H}}XR))\|_{\mathrm{F}}\le\sqrt{\frac{n-1}{2n}(\|A\|_2^2+\|B\|_2^2)}\,\|(X,Y)\|_{\mathrm{F}}^2,$$

we get

$$\begin{aligned}&\|(\mathrm{low}(X^{\mathrm{H}}TY-X^{\mathrm{H}}XT),\mathrm{low}(X^{\mathrm{H}}RY-X^{\mathrm{H}}XR))\|_{\mathrm{F}}\\ &\le\left(\frac{1}{2}+\sqrt{\frac{n-1}{2n}}\right)\sqrt{\|A\|_2^2+\|B\|_2^2}\,\|(X,Y)\|_{\mathrm{F}}^2.\end{aligned}$$

3. From (2.4)

$$\|(\mathrm{low}(\tilde{E}),\mathrm{low}(\tilde{F}))\|_{\mathrm{F}}\le\|(E,F)\|_{\mathrm{F}}.$$

4. We have

$$\|(\mathrm{diag}(X^{\mathrm{H}}X), \mathrm{diag}(Y^{\mathrm{H}}Y))\|_{\mathrm{F}} \le \|(X,Y)\|_{\mathrm{F}}^2,$$

$$\|(\mathrm{up}(X^{\mathrm{H}}X), \mathrm{up}(Y^{\mathrm{H}}Y))\|_{\mathrm{F}} \le \sqrt{\tfrac{n-1}{2n}}\|(X,Y)\|_{\mathrm{F}}^2.$$

Hence, if we let

$$\epsilon = \frac{\|(E,F)\|_{\mathrm{F}}}{l(T,R)}, \quad \mu_n = \sqrt{\frac{n-1}{2n}}, \quad \alpha = \left(\frac{1}{2} + \mu_n\right)\frac{\sqrt{\|A\|_2^2 + \|B\|_2^2}}{l(T,R)}, \tag{4.2}$$

then the mapping Φ expressed by (4.1) satisfies

$$\begin{aligned} &\|(X_{\mathrm{L}}, Y_{\mathrm{L}})\|_{\mathrm{F}} \le \alpha\|(X,Y)\|_{\mathrm{F}}^2 + \epsilon, \\ &\|(X_{\mathrm{D}}, Y_{\mathrm{D}})\|_{\mathrm{F}} \le \tfrac{1}{2}\|(X,Y)\|_{\mathrm{F}}^2, \\ &\|(X_{\mathrm{U}}, Y_{\mathrm{U}})\|_{\mathrm{F}} \le \|(X_{\mathrm{L}}, Y_{\mathrm{L}})\|_{\mathrm{F}} + \mu_n\|(X,Y)\|_{\mathrm{F}}^2. \end{aligned} \tag{4.3}$$

Let $z = (\zeta_1, \zeta_2, \zeta_3)^{\mathrm{T}} \in \mathcal{C}^3$. Consider the system

$$\begin{aligned} \zeta_1 &= \alpha(\zeta_1^2 + \zeta_2^2 + \zeta_3^2) + \epsilon, \\ \zeta_2 &= \tfrac{1}{2}(\zeta_1^2 + \zeta_2^2 + \zeta_3^2), \\ \zeta_3 &= \zeta_1 + \mu_n(\zeta_1^2 + \zeta_2^2 + \zeta_3^2). \end{aligned} \tag{4.4}$$

From the first two equations of (4.4)

$$\zeta_1 = 2\alpha\zeta_2 + \epsilon, \tag{4.5}$$

and from the last two equations of (4.4)

$$\zeta_3 = \zeta_1 + 2\mu_n\zeta_2. \tag{4.6}$$

Substituting (4.5)–(4.6) into the second equation of (4.4) we know that ζ_2 satisfies the equation

$$\phi\zeta_2^2 - \psi(\epsilon)\zeta_2 + \epsilon^2 = 0, \tag{4.7}$$

where

$$\phi = \frac{1}{2}[2(2\alpha + \mu_n)^2 + 2\mu_n^2 + 1], \quad \psi(\epsilon) = 1 - 2\epsilon(2\alpha + \mu_n). \tag{4.8}$$

Let

$$\delta(\epsilon) = (\psi(\epsilon))^2 - 4\phi\epsilon^2. \tag{4.9}$$

If $\delta(\epsilon) \ge 0$, then

$$\zeta_2^* = \frac{\psi(\epsilon) - \sqrt{\delta(\epsilon)}}{2\phi} \tag{4.10}$$

is a solution to (4.7). Thus from (4.10) and (4.5)–(4.6) we get a solution $z^* = (\zeta_1^*, \zeta_2^*, \zeta_3^*)^{\mathrm{T}}$ to the system (4.4).

Let

$$\mathcal{S}_{z^*} = \{X : X \in \mathcal{C}^{n\times n},\ \|X_{\mathrm{L}}\|_{\mathrm{F}} \le \zeta_1^*,\ \|X_{\mathrm{D}}\|_{\mathrm{F}} \le \zeta_2^*,\ \|X_{\mathrm{U}}\|_{\mathrm{F}} \le \zeta_3^*\}.$$

Obviously, $\mathcal{S}_{z^*}$ is a bounded closed convex set of $\mathcal{C}^{n\times n}$, and the relation (4.3) shows that the continuous mapping Φ maps $\mathcal{S}_{z^*}$ into $\mathcal{S}_{z^*}$. By the Brouwer fixed-point theorem (see, e.g., [5, p. 161]), the mapping Φ has a fixed point $X^* \in \mathcal{S}_{z^*}$, i.e., Φ has a fixed point X^* satisfying

$$\|X^*\|_{\mathrm{F}} \le \|z^*\|_2 = \sqrt{2\zeta_2^*}.$$

Moreover, observe that the function $\delta(\epsilon)$ defined by (4.9) can be expressed by

$$\delta(\epsilon) = 1 - 4(2\alpha + \mu_n)\epsilon - 2(2\mu_n^2 + 1)\epsilon^2,$$

and $\delta(\epsilon) \ge 0$ is equivalent to

$$\epsilon \le \frac{1}{2(2\alpha + \mu_n) + \sqrt{4(2\alpha + \mu_n)^2 + 2(2\mu_n^2 + 1)}} \equiv \bar{\epsilon}. \tag{4.11}$$

Hence, we have proved the following result.

THEOREM 4.1. *Let (A, B) be a regular matrix pair of order n with distinct eigenvalues, and let $A = UTV^{\mathrm{H}}, B = URV^{\mathrm{H}}$ be the generalized Schur decomposition of (A, B). Moreover, let $\tilde{A} = A + E, \tilde{B} = B + F$, and let ϵ, μ_n and α be defined by (4.2). If ϵ satisfies (4.11), then $(\tilde{A}, \tilde{B})$ has a generalized Schur decomposition $\tilde{A} = \tilde{U}\tilde{T}\tilde{V}^{\mathrm{H}}$, $\tilde{B} = \tilde{U}\tilde{R}\tilde{V}^{\mathrm{H}}$ such that*

$$\begin{aligned} &\|(\tilde{U} - U, \tilde{V} - V)\|_{\mathrm{F}} \\ &\le \frac{2\epsilon}{\sqrt{1 - 2(2\alpha + \mu_n)\epsilon + \sqrt{1 - 4(2\alpha + \mu_n)\epsilon - 2(2\mu_n^2 + 1)\epsilon^2}}} \\ &\equiv b_{\mathrm{U,V}}(\epsilon), \end{aligned} \tag{4.12}$$

and

$$\begin{aligned} \|\tilde{T} - T\|_{\mathrm{F}} &\le \|E\|_{\mathrm{F}} + \theta_{\mathrm{T}} b_{\mathrm{U,V}}(\epsilon) \equiv b_{\mathrm{T}}(\epsilon), \\ \|\tilde{R} - R\|_{\mathrm{F}} &\le \|F\|_{\mathrm{F}} + \theta_{\mathrm{R}} b_{\mathrm{U,V}}(\epsilon) \equiv b_{\mathrm{R}}(\epsilon), \end{aligned} \tag{4.13}$$

where $\theta_{\mathrm{T}}, \theta_{\mathrm{R}}$ are defined by (2.6).

Remark 4.2. For small ϵ the upper bound $b_{\mathrm{U,V}}(\epsilon)$ defined by (4.12) has the Taylor expansion

$$b_{\mathrm{U,V}}(\epsilon) = \sqrt{2}\epsilon + \sqrt{2}(2\alpha + \mu_n)\epsilon^2 + O(\epsilon^3), \quad \epsilon \to 0. \tag{4.14}$$

Combining (4.14) with (4.12)–(4.13) and (4.2) we see that the quantity $1/l(T, R)$ can be regarded as a condition number of the generalized Schur decomposition of (A, B). Note that by Remark 3.3 the condition number is dependent on the choice of the generalized Schur form (T, R).

Let $\tilde{U}, \tilde{V}, \tilde{T}, \tilde{R}$ be the computed generalized Schur factors of a regular matrix pair (A, B) by the QZ algorithm [4]. (Note. MATLAB supplies a function **qz** to compute the generalized Schur factors. The **qz** function is an implementation of the QZ algorithm.) It is known that the computed factors $\tilde{U}, \tilde{V}, \tilde{T}, \tilde{R}$ are exactly the generalized Schur factors of a slightly perturbed regular matrix pair $(A + E, B + F)$ [4]. Consequently, from (4.12)–(4.14) we see that the condition number $1/l(T, R)$ enables us to estimate the accuracy of the computed generalized Schur factors.

5. Numerical examples.

Example 5.1. Let

$$A = \begin{pmatrix} -20 & -0.1 & 0 & 0 & 0 \\ 0 & -10 & -0.1 & 0 & 0 \\ 0 & 0 & 0 & -0.1 & 0 \\ 0 & 0 & 0 & 10 & -0.1 \\ 0 & 0 & 0 & 0 & 20 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0.01 & 0 & 0 & 0 \\ 0 & 1 & 0.01 & 0 & 0 \\ 0 & 0 & 1 & 0.01 & 0 \\ 0 & 0 & 0 & 1 & 0.01 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$E_0 = \begin{pmatrix} 3 & -2 & 6 & 4 & -1 \\ 1 & 8 & -5 & 9 & 2 \\ -4 & 7 & -3 & 5 & 1 \\ -1 & 0 & 3 & -4 & 7 \\ 7 & 8 & 9 & 0 & -2 \end{pmatrix}, \quad F_0 = \begin{pmatrix} 1 & 2 & 3 & -2 & 7 \\ -2 & 9 & -8 & 3 & 0 \\ 6 & -2 & 4 & 7 & 1 \\ -6 & 8 & -3 & 1 & 2 \\ 0 & -5 & 3 & -1 & 1 \end{pmatrix},$$

and let $\tilde{A} = A + \tau E_0$, $\tilde{B} = B + \tau F_0$, where τ is a real parameter. The eigenvalues of the matrix pair (A, B) are $-20, -10, 0, 10$, and 20. We take $T = A, R = B$, and $U = V = I$ in the decomposition (1.1). Computation gives

$$1/l(T, R) \approx 2.24, \quad \bar{\epsilon} \approx 2.44\text{e} - 03,$$

where $\bar{\epsilon}$ is defined by (4.11). By (4.2), we have

$$\epsilon = \|(\tau E_0, \tau F_0)\|_{\mathrm{F}} / l(T, R).$$

Consequently, from the restriction (4.11) in Theorem 4.1 it follows that in order to apply the estimates (4.12)–(4.13) the parameter τ must satisfy

$$|\tau| \leq \bar{\epsilon} \cdot l(T, R) / \|(E_0, F_0)\|_{\mathrm{F}} \approx 3.30\text{e} - 05.$$

By using MATLAB we get the generalized Schur decomposition $\tilde{A} = \tilde{U}\tilde{T}\tilde{V}^{\mathrm{H}}, \tilde{B} = \tilde{U}\tilde{R}\tilde{V}^{\mathrm{H}}$, and the upper bounds in Theorem 4.1. Some numerical results are listed in Table 1, where $\epsilon, b_{\mathrm{U,V}}(\epsilon), b_{\mathrm{T}}(\epsilon)$, and $b_{\mathrm{R}}(\epsilon)$ are defined by (4.2), (4.12), and (4.13), respectively.

Example 5.2. Let A, B, U, V, T, R, and τ be as in Example 5.1 but the $(5, 5)$-element of A is changed from 20 to 9.999. The eigenvalues of the matrix pair (A, B) are $-20, -10, 0, 10$, and 9.999. Computation gives

$$1/l(T, R) \approx 14215 \gg 1, \quad \bar{\epsilon} \approx 3.88\text{e} - 07,$$

where $\bar{\epsilon}$ is defined by (4.11). In order to apply the estimates (4.12)–(4.13) the parameter τ must satisfy

$$|\tau| \leq \bar{\epsilon} \cdot l(T, R) / \|E_0, F_0)\|_{\mathrm{F}} \approx 8.24\text{e} - 13.$$

Taking $\tau = 1.00\text{e-}13$ and by using MATLAB we get the generalized Schur decomposition $\tilde{A} = \tilde{U}\tilde{T}\tilde{V}^{\mathrm{H}}, \tilde{B} = \tilde{U}\tilde{R}\tilde{V}^{\mathrm{H}}$, and

$$\|(\tilde{U}-U, \tilde{V}-V)\|_{\mathrm{F}} \approx 2.09\text{e}-09, \quad \|\tilde{T}-T\|_{\mathrm{F}} \approx 1.83\text{e}-10, \quad \|\tilde{R}-R\|_{\mathrm{F}} \approx 1.80\text{e}-11$$

$$b_{\mathrm{U,V}}(\epsilon) \approx 6.87\text{e}-08, \qquad b_{\mathrm{T}}(\epsilon) \approx 1.94\text{e}-06, \qquad b_{\mathrm{R}}(\epsilon) \approx 9.80\text{e}-08.$$

The numerical results show that the sharpness of the upper bounds of (4.12)–(4.13) may be weakened in the ill-conditioned case (i.e., in the case $1/l(T,R) \gg 1$). This is due to the fact that the condition number $1/l(T,R)$, as the condition numbers of usual matrix computation problems [1], [7], is defined by the "worst case" perturbations.

TABLE 1.

τ	1.00e-05	1.00e-07	1.00e-09	1.00e-11	1.00e-13
$\|(\tilde{U}-U, \tilde{V}-V)\|_{\mathrm{F}}$	1.71e-0.4	1.71e-06	1.71e-08	1.71e-10	1.71e-12
$b_{\mathrm{U,V}}(\epsilon)$	1.14e-03	1.05e-05	1.05e-07	1.05e-09	1.05e-11
$\|\tilde{T}-T\|_{\mathrm{F}}$	1.66e-03	1.66e-05	1.66e-07	1.66e-09	1.66e-11
$b_{\mathrm{T}}(\epsilon)$	3.26e-02	2.99e-04	2.99e-06	2.99e-08	2.99e-10
$\|\tilde{R}-R\|_{\mathrm{F}}$	2.39e-04	2.39e-06	2.39e-08	2.39e-10	2.39e-12
$b_{\mathrm{R}}(\epsilon)$	1.85e-03	1.72e-05	1.71e-07	1.71e-09	1.71e-11

Acknowledgments. I would like to thank the referee for valuable suggestions. I am also grateful to Professor Bo Kågström for his helpful comments on an earlier version of this paper.

REFERENCES

[1] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD 1989.

[2] B. KÅGSTRÖM AND L. WESTIN, *Generalized Schur methods with condition estimators for solving the generalized Sylvester equation*, IEEE Trans. Autom. Control, 34(1989), pp. 745–751.

[3] M. M. KONSTANTINOV, P. HR. PETKOV, AND N. D. CHRISTOV, *Non-local perturbation analysis of the Schur system of a matrix*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 383–392.

[4] C. MOLER AND G. W. STEWART, *An algorithm for generalized matrix eigenvalue problems*, SIAM J. Numer. Anal., 10(1973), pp. 241–256.

[5] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York 1970.

[6] G. W. STEWART, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.

[7] G. W. STEWART AND J.-G. SUN, *Matrix Perturbation Theory*, Academic Press, Boston 1990.

[8] J.-G. SUN, *Perturbation bounds for the Schur decomposition*, Report UMINF-92.20, ISSN-0348-0542, Institute of Information Processing, University of Umeå, Sweden, 1992.

SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 4, pp. 1341–1369, October 1995

021

APPLICATION OF VECTOR-VALUED RATIONAL APPROXIMATIONS TO THE MATRIX EIGENVALUE PROBLEM AND CONNECTIONS WITH KRYLOV SUBSPACE METHODS *

AVRAM SIDI†

Abstract. Let $F(z)$ be a vector-valued function $F : \mathbf{C} \to \mathbf{C}^N$, which is analytic at $z = 0$ and meromorphic in a neighborhood of $z = 0$, and let its Maclaurin series be given. In a recent work [*J. Approx. Theory*, 76 (1994), pp. 89–111] by the author, vector-valued rational approximation procedures for $F(z)$ that are based on its Maclaurin series, were developed, and some of their convergence properties were analyzed in detail. In particular, a Koenig-type theorem concerning their poles and a de Montessus-type theorem concerning their uniform convergence in the complex plane were given. With the help of these theorems it was shown how optimal approximations to the poles of $F(z)$ and the principal parts of the corresponding Laurent series expansions can be obtained. In this work we use these rational approximation procedures in conjunction with power iterations to develop bona fide generalizations of the power method for an arbitrary $N \times N$ matrix that may or may not be diagonalizable. These generalizations can be used to obtain simultaneously several of the largest distinct eigenvalues and corresponding eigenvectors and other vectors in the invariant subspaces. We provide interesting constructions for both nondefective and defective eigenvalues and the corresponding invariant subspaces, and present a detailed convergence theory for them. This is made possible by the observation that vectors obtained by power iterations with a matrix are actually coefficients of the Maclaurin series of a vector-valued rational function, whose poles are the reciprocals of some or all of the nonzero eigenvalues of the matrix being considered, while the coefficients in the principal parts of the Laurent expansions of this rational function are vectors in the corresponding invariant subspaces. In addition, it is shown that the generalized power methods of this work are equivalent to some Krylov subspace methods, among them the methods of Arnoldi and Lanczos. Thus, the theory of the present work provides a set of completely new results and constructions for these Krylov subspace methods. At the same time this theory suggests a new mode of usage for these Krylov subspace methods that has been observed to possess computational advantages over their common mode of usage in some cases. We illustrate some of the theory and conclusions derived from it with numerical examples.

Key words. Krylov subspace methods, method of Arnoldi, method of Lanczos, power iterations, generalized power methods, diagonalizable matrices, defective matrices, eigenvalues, invariant subspaces, vector-valued rational approximations

AMS subject classifications. 30E10, 41A20, 65F15, 65F30, 65F50

1. Introduction.

Let $F(z)$ be a vector-valued function, $F : \mathbf{C} \to \mathbf{C}^N$, which is analytic at $z = 0$ and meromorphic in a neighborhood of $z = 0$, and let its Maclaurin series be given as

$$F(z) = \sum_{m=0}^{\infty} u_m z^m, \tag{1.1}$$

where u_m are fixed vectors in $\mathbf{C}^N$.

*Received by the editors December 31, 1992; accepted for publication (in revised form) by A. Berman September 1, 1994. The results of this paper were presented at the International Meeting on Approximation, Interpolation, and Summability, Tel-Aviv, June 1990, the International Congress on Extrapolation and Rational Approximation, Tenerife, January 1992, and the Lanczos International Centenary Conference, Raleigh, North Carolina, December 1993.

†Computer Science Department, Technion-Israel Institute of Technology, Haifa 32000, Israel and Institute for Computational Mechanics in Propulsion, NASA Lewis Research Center, Cleveland, Ohio 44135 (asidi@cs.technion.ac.il).

In a recent work by the author [Si6] three types of vector-valued rational approximation procedures, entirely based on the expansion in (1.1), were proposed. For each of these procedures the rational approximations have two indices, n and k, attached to them, and thus form a two-dimensional table akin to the Padé table or the Walsh array. Let us denote the (n,k) entry of this table by $F_{n,k}(z)$. Then $F_{n,k}(z)$, if it exists, is defined to be of the form

$$F_{n,k}(z)=\frac{\sum_{j=0}^{k}c_j^{(n,k)}z^{k-j}F_{n+\nu+j}(z)}{\sum_{j=0}^{k}c_j^{(n,k)}z^{k-j}}\equiv\frac{P_{n,k}(z)}{Q_{n,k}(z)}\ \text{with}\ c_k^{(n,k)}=Q_{n,k}(0)=1, \tag{1.2}$$

where ν is an arbitrary but otherwise fixed integer ≥ -1, and

$$F_m(z)=\sum_{i=0}^{m}u_iz^i,\ m=0,1,2,\ldots;\qquad F_m(z)\equiv 0\quad\text{for } m<0, \tag{1.3}$$

and the $c_j^{(n,k)}$ are scalars that depend on the approximation procedure being used.

If we denote the three approximation procedures by SMPE, SMMPE, and STEA, then the $c_j^{(n,k)}\equiv c_j$ for each of the three procedures, are defined such that they satisfy a linear system of equations of the form

$$\sum_{j=0}^{k-1}u_{ij}c_j=-u_{ik},\quad 0\leq i\leq k-1;\qquad c_k=1, \tag{1.4}$$

where u_{ij} are scalars defined as

$$u_{ij}=\begin{cases}(u_{n+i},u_{n+j}) & \text{for SMPE,}\\ (q_{i+1},u_{n+j}) & \text{for SMMPE,}\\ (q,u_{n+i+j}) & \text{for STEA.}\end{cases} \tag{1.5}$$

Here $(\cdot,\cdot)$ is an inner product--not necessarily the standard Euclidean inner product—whose homogeneity property is such that $(\alpha x,\beta y)=\bar{\alpha}\beta(x,y)$ for x,y in $\mathbf{C}^N$ and α,β in $\mathbf{C}$. The vectors $q_1,q_2,\ldots,$ form a linearly independent set, and the vector q is nonzero. Obviously, $F_{n,k}(z)$ exists if the linear system in (1.4) has a solution for $c_0,c_1,\ldots,c_{k-1}$.

It is easy to verify that for SMPE the equations in (1.4) involving $c_0,c_1,\ldots,c_{k-1}$ are the normal equations for the least squares problem

$$\min_{c_0,c_1,\ldots,c_{k-1}}\left\|\sum_{j=0}^{k-1}c_ju_{n+j}+u_{n+k}\right\|, \tag{1.6}$$

where the norm $\|\cdot\|$ is that induced by the inner product $(\cdot,\cdot)$, namely, $\|x\|=\sqrt{(x,x)}$.

As is clear from (1.2) and (1.3), the numerator of $F_{n,k}(z)$ is a vector-valued polynomial of degree at most $n+\nu+k$, whereas its denominator is a scalar polynomial of degree at most k.

As can be seen from (1.4) and (1.5), the denominator polynomial $Q_{n,k}(z)$ is constructed from $u_n,u_{n+1},\ldots,u_{n+k}$ for SMPE and SMMPE, and from $u_n,u_{n+1},\ldots,u_{n+2k-1}$ for STEA. Once the denominators have been determined, the numerators involve $u_0,u_1,\ldots,u_{n+\nu+k}$ for all three approximation procedures.

The approximation procedures above are very closely related to some vector extrapolation methods. In fact, as is stated in Theorem 2.3 in Section 2 of [Si6], $F_{n,k}(z)$ for SMPE, SMMPE, and STEA are obtained by applying some generalized versions of the minimal polynomial extrapolation (MPE), the modified minimal polynomial extrapolation (MMPE), and the topological epsilon algorithm (TEA), respectively, to the vector sequence $F_m(z), m = 0, 1, 2, \ldots$. For early references pertaining to these methods and their description, see the survey paper of Smith, Ford, and Sidi [SmFSi], and for recent developments pertaining to their convergence, stability, implementation, and other additional properties, see the papers by Sidi [Si1], [Si2], [Si5], Sidi and Bridger [SiB], Sidi, Ford, and Smith [SiFSm], and Ford and Sidi [FSi]. The above mentioned generalizations of vector extrapolation methods are given in [SiB, (1.16) and (1.17)].

A detailed convergence analysis for the approximations $F_{n,k}(z)$ as $n \to \infty$ was given in [Si6], whose main results can be verbally summarized as follows: (i) Under certain conditions the denominators $Q_{n,k}(z)$ converge, and their zeros, k in number, tend to the k poles of $F(z)$ that are closest to the origin. This is a Koenig-type result and is proved in Theorems 4.1 and 4.5 of [Si6], where the precise rates of convergence are also given for both simple and multiple poles of $F(z)$, and optimal approximations to multiple poles are constructed in a simple way. (ii) Under the same conditions $F_{n,k}(z)$ converges to $F(z)$ uniformly in any compact subset of the circle containing the above-mentioned k poles of $F(z)$ with these poles excluded. This is a de Montessus-type result and is proved in Theorem 4.2 of [Si6]. (iii) The principal parts of the Laurent expansions of $F(z)$ about its poles, simple or multiple, can be constructed from $F_{n,k}(z)$ only. This construction, along with its convergence theory, is provided in Theorem 4.3 of [Si6].

It turns out that the denominator polynomials $Q_{n,k}(z)$ are very closely related to some recent extensions of the power method for the matrix eigenvalue problem, see [SiB, §6] and [Si3]. Specifically, if the vectors u_m of (1.1) are obtained from $u_m = Au_{m-1}, m = 1, 2, \ldots$, with u_0 arbitrary, and A being a complex $N \times N$ and, in general, nondiagonalizable matrix, then the reciprocals of the zeros of the polynomial $Q_{n,k}(z)$ are approximations to the k largest distinct and, in general, defective eigenvalues of A, counted according to their multiplicities, under certain conditions. In §3 of this work we provide precise error bounds for these approximations for $n \to \infty$ that are based on the results of Theorems 4.1 and 4.5 of [Si6]. While the approximations to nondefective eigenvalues have optimal accuracy in some sense, those that correspond to defective eigenvalues do not. In this paper we also show how approximations of optimal accuracy to defective eigenvalues can be constructed solely from $Q_{n,k}(z)$, providing their convergence theory for $n \to \infty$ at the same time. We then extend the treatment of [SiB] and [Si3] to cover the corresponding invariant subspaces in general, and the corresponding eigenvectors in particular. For example, we actually show how the eigenvectors corresponding to the largest distinct eigenvalues, whether these are defective or not, can be approximated solely in terms of the vectors u_j, and provide precise rates of convergence for them. The key to these results is the observation that the vector-valued power series $\sum_{m=0}^{\infty} u_m z^m$ actually represents a vector-valued *rational* function $F(z)$ whose poles are the reciprocals of some or all of the nonzero eigenvalues of A, depending on the spectral decomposition of u_0, and that corresponding eigenvectors (and certain combinations of eigenvectors and principal vectors) are related to corresponding principal parts of the Laurent expansions of the function $F(z)$. The main results of §3 pertaining to eigenvalues are given in Theorem 3.1, while those pertaining to eigenvectors and invariant subspaces are given in Theorem 3.2

and the subsequent paragraphs. A detailed description of the properties of the power iterations $u_m = Au_{m-1}, m = 1, 2, \ldots$, is provided in §2.

In §4 we present a short review of general projection methods and Krylov subspace methods for the matrix eigenvalue problem. Of particular interest to us are the methods of Arnoldi [A] and Lanczos [L], which are described in this section.

In §5 we show that the extensions of the power method developed and analyzed in §3 are very closely related to Krylov subspace methods. In particular, we show that the reciprocals of the k poles and the corresponding residues of the rational approximations $F_{n,k}(z)$ (with $\nu = -1$) obtained from the SMPE, SMMPE, and STEA procedures are the Ritz values and the Ritz vectors, respectively, of certain Krylov subspace methods of order k for the matrix A starting with the power iteration u_n. Specifically, the methods of Arnoldi and Lanczos are related to the $F_{n,k}(z)$ obtained from the SMPE and STEA procedures, respectively, precisely in this sense when $(\cdot, \cdot)$ in (1.5) is the standard Euclidean inner product. The main results of §5 concerning this are summarized in Theorem 5.4 and Corollary 5.5. In addition, Theorem 5.6 gives some optimality properties of the Arnoldi method.

Now the Ritz values and Ritz vectors obtained from Krylov subspace methods are normally used as approximations to nondefective eigenpairs. They are not very effective for defective eigenpairs. Since we know that the generalized power methods based on the SMPE, SMMPE, and STEA procedures are related to Krylov subspace methods, the constructions for approximating defective eigenvalues and their corresponding invariant subspaces that originate from generalized power methods and that are given in §3 are entirely new as far as Krylov subspace methods are concerned. Similarly, all of the convergence results of §3, whether they pertain to defective or nondefective eigenvalues and their corresponding invariant subspaces, are new and totally different from the known analyses provided by Kaniel [K], Paige [Pai], and Saad [Sa1], [Sa2]. Some of these analyses can also be found in Parlett [Par2] and Golub and Van Loan [GV]. The last two references also give a very thorough treatment of the computational aspects of Krylov subspace methods.

In §6 we show how the Ritz values and Ritz vectors obtained in a stable way from the common implementations of the Arnoldi and Lanczos methods can be used in constructing the approximations to the defective eigenvalues and their corresponding invariant subspaces in general and eigenvectors in particular.

In §7 we illustrate some of the theoretical results and claims of the paper with numerical examples.

In view of the connection between (1) the Krylov subspace methods and (2) the vector-valued rational approximations of [Si6] and the corresponding generalized power methods of the present work, we now summarize the main contributions of this paper.

(i) It is shown that Krylov subspace methods for the matrix eigenvalue problem are completely equivalent to methods founded on analytic function theory and rational approximations in the complex plane.

(ii) A mode of usage of Krylov subspace methods akin to the power method, in which one first iterates on an arbitrary initial vector many times and only then applies Krylov subspace methods, is proposed. This mode produces approximations only to the largest eigenvalues and their corresponding invariant subspaces.

(iii) The output from Krylov subspace methods, namely, the Ritz values and Ritz vectors, are used in constructing optimal approximations to defective eigenvalues and the corresponding eigenvectors and invariant subspaces. (The Ritz values and Ritz vectors by themselves are not good approximations to defective eigenvalues and

corresponding eivenvectors and invariant subspaces.)

(iv) A *complete* convergence theory for the generalized power methods is provided.

(v) This author's numerical experience suggests that at least in some cases the mode of usage proposed in this work and mentioned in (ii) above may produce the accuracy that is achieved by applying the Arnoldi method in the commonly known way using less storage and less computational work when the matrix being treated is large and sparse.

Before closing this section we note that the eigenvalue problem for defective matrices has received some attention in the literature. The problem of approximating the largest eigenvalue of a matrix when this eigenvalue is defective has been considered by Ostrowski [O], who proposes an extension of the Rayleigh quotient and inverse iteration and gives a thorough analysis for this extension. Parlett and Poole [ParPo] consider the properties of a wide range of projection methods within the framework of defective matrices. The convergence of the QR method for defective Hessenberg matrices has been analyzed in detail by Parlett [Par1]. The problem of determining the Jordan canonical form of defective matrices has been treated in Golub and Wilkinson [GW]. The use of power iterations in approximating defective eigenvalues is also treated to some extent in Wilkinson [W, Chap. 7] and Householder [H, Chap. 7].

Finally, we mention that the results of [Si6], as well as the application of vector-valued rational approximations to the matrix eigenvalue problem, were motivated by the developments in a recent work by the author [Si4] on the classical Padé approximants.

2. Properties of power iterations. Let A be an $N \times N$ matrix, which, in general, is complex and nondiagonalizable. Let u_0 be a given arbitrary vector in $\mathbf{C}^N$, and generate the vectors $u_1, u_2, \ldots$, according to

$$u_{j+1} = Au_j, \quad j \geq 0. \tag{2.1}$$

Denote by s the index of A, i.e., the size of the largest Jordan block of A with zero eigenvalue. Then u_m is of the form

$$u_m = \sum_{j=1}^{M} \left[\sum_{l=0}^{p_j} \tilde{a}_{jl} \binom{m}{l} \right] \lambda_j^m \quad \text{for } m \geq s, \tag{2.2}$$

where λ_j are some or all of the *distinct nonzero* eigenvalues of A, which we choose to order such that

$$|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_M| > 0, \tag{2.3}$$

$p_j + 1 \equiv \omega_j$ are positive integers less than or equal to the dimension of the invariant subspace of A belonging to the eigenvalue λ_j, and $\tilde{a}_{jl}, 0 \leq l \leq p_j$, are linearly independent vectors in this invariant subspace. It turns out that the vector $\tilde{a}_{jp_j}$ is an *eigenvector* of A corresponding to λ_j, while the vectors $\tilde{a}_{ji}, i = 0, 1, \ldots, p_j - 1$, are combinations of eigenvectors *and* principal vectors of A corresponding to the eigenvalue λ_j. What is more, the subspaces

$$Y_i = \operatorname{span}\{\tilde{a}_{jl}, i \leq l \leq p_j\}, \qquad i = 0, 1, \ldots, p_j,$$

are invariant subspaces of A corresponding to the eigenvalue λ_j, and satisfy $Y_0 \supset Y_1 \supset \cdots \supset Y_{p_j}$.

Whether all distinct nonzero eigenvalues are present among $\lambda_1, \lambda_2, \ldots, \lambda_M$, the exact values of the ω_j, and the precise composition of the vectors $\tilde{a}_{jl}$, all depend on the spectral decomposition of the initial vector u_0. For a detailed derivation of the above see [SiB, §2].

Before we go on, we will only mention how to determine the maximum value that ω_j can assume. Suppose that the Jordan canonical form of A has several Jordan blocks whose eigenvalues are all equal to λ_j. Then the largest value that ω_j can assume is the size of the largest of these blocks. In general, for a randomly chosen vector u_0, ω_j will take on its maximum value. In cases where ω_j is theoretically less than this maximum value, rounding errors on a computer will ultimately force ω_j to take on its maximum value.

It is obvious from the above that

$$k_0 \equiv \sum_{j=1}^{M}(p_j + 1) = \sum_{j=1}^{M} \omega_j \leq N \tag{2.4}$$

and

$$\tilde{a}_{ji},\ 0 \leq i \leq p_j, 1 \leq j \leq M, \quad \text{are linearly independent.} \tag{2.5}$$

Also the minimal polynomial of the matrix A with respect to the vector u_s has degree $k_0 = \sum_{j=1}^{M} \omega_j$, i.e.,

$$k_0 = \min\left\{k\colon \left(\sum_{i=0}^{k} \beta_i A^i\right) u_s = 0, \beta_k = 1\right\}.$$

If defined as a monic polynomial, this polynomial is unique and divides the minimal polynomial of A, which, in turn, divides the characteristic polynomial of A. Furthermore, the minimal polynomial of A with respect to u_s is also the minimal polynomial of A with respect to u_m for all $m \geq s$. Consequently, any set of vectors $\{u_m, u_{m+1}, \ldots, u_{m+k}\}$ is linearly independent for $m \geq s$ provided $k < k_0$.

Now applying Lemma 3.1 of [Si6] in conjunction with (2.2), we conclude that the vector-valued power series $\sum_{m=0}^{\infty} u_m z^m$ represents the vector-valued *rational* function

$$F(z) = (I - zA)^{-1} u_0 = \sum_{j=1}^{M} \sum_{i=0}^{p_j} \frac{a_{ji}}{(1 - \lambda_j z)^{i+1}} + G(z), \tag{2.6}$$

in which the vectors a_{ji} are uniquely determined in terms of the $\tilde{a}_{jl}$ from

$$\tilde{a}_{jl} = \sum_{i=l}^{p_j} a_{ji} \binom{i}{i-l}, \quad 0 \leq l \leq p_j, \quad 1 \leq j \leq M, \tag{2.7}$$

and hence form a linearly independent set, and $G(z)$ is a vector-valued polynomial of degree at most $s - 1$. In fact, $G(z)$ is in the invariant subspace of A corresponding to the zero eigenvalue. Also, $a_{jp_j} = \tilde{a}_{jp_j}$, i.e., a_{jp_j} is an eigenvector of A corresponding to the eigenvalue λ_j, while for each $i, 0 \leq i \leq p_j - 1, a_{ji}$ is some other vector in the invariant subspace Y_i corresponding to the eigenvalue λ_j, and involves principal vectors as well as eigenvectors.

When the matrix A is diagonalizable, $p_j = 0$ for all j in (2.2) and hence in (2.6). If, in addition, A is normal, then its eigenvectors form an orthogonal set with respect to the standard Euclidean inner product, namely, $(x, y) = x^*y$, where x^* stands for the hermitian conjugate of x. Consequently, the vectors $\tilde{a}_{j0} = a_{j0}$ in (2.2) and (2.6) are orthogonal with respect to this inner product when A is normal.

Now that we have shown that the power series $\sum_{m=0}^{\infty} u_m z^m$ represents a rational function $F(z)$ that is analytic at $z = 0$ and has poles $z_j = \lambda_j^{-1}$ of respective multiplicities $\omega_j = p_j + 1, j = 1, 2, \ldots, M$, we can apply any one of the approximation procedures SMPE, SMMPE, or STEA to the power series $\sum_{m=0}^{\infty} u_m z^m$ to obtain the vector-valued rational approximations $F_{n,k}(z)$ to $F(z)$. We can then apply the theorems of §§4 and 5 of [Si6] to construct approximations to the eigenvalues λ_j and the vectors a_{ji} in (2.6) and (2.7).

It is important to note that the linear independence of the vectors a_{jl} is an important condition for the convergence of the SMPE and SMMPE procedures, but is not needed for the STEA procedure. In addition, we assume throughout that

$$\text{(2.8)} \qquad \begin{vmatrix} (q_1, a_{10}) & \cdots & (q_1, a_{1p_1}) & \cdots & (q_1, a_{t0}) & \cdots & (q_1, a_{tp_t}) \\ \vdots & & \vdots & & \vdots & & \vdots \\ (q_k, a_{10}) & \cdots & (q_k, a_{1p_1}) & \cdots & (q_k, a_{t0}) & \cdots & (q_k, a_{tp_t}) \end{vmatrix} \neq 0 \text{ for SMMPE,}$$

where $k = \sum_{j=1}^{t} w_j$, and that

$$\text{(2.9)} \qquad \prod_{j=1}^{t} (q, a_{jp_j}) \neq 0 \text{ for STEA.}$$

No additional assumption is needed for SMPE.

In order for (2.8) to hold it is necessary (but not sufficient) that the two sets of vectors $\{a_{ji} : 0 \leq i \leq p_j, 1 \leq j \leq t\}$ and $\{q_1, \ldots, q_k\}$, each be linearly independent, as has already been assumed.

3. Theoretical development of generalized power methods. In light of the developments of §2 and Theorems 4.1, 4.3, and 4.5 of [Si6] and the developments of §5 in the same paper, we approach the matrix eigenvalue problem as follows.

Given the vector u_0 that is picked arbitrarily, we generate the vectors $u_1, u_2, \ldots$, according to (2.1). We then fix the integers n and k, and determine the coefficients $c_j^{(n,k)}, j = 0, 1, \ldots, k$, of the denominator polynomial of $F_{n,k}(z)$ for one of the procedures SMPE, SMMPE, and STEA, by using $u_n, u_{n+1}, \ldots, u_{n+k}$ for SMPE and SMMPE, and $u_n, u_{n+1}, \ldots, u_{n+2k-1}$, for STEA. By Theorem 4.1 of [Si6] the zeros of the polynomial $\hat{Q}_{n,k}(\lambda) \equiv \lambda^{-k} Q_{n,k}(\lambda^{-1}) = \sum_{j=0}^{k} c_j^{(n,k)} \lambda^j$ are approximations to the k largest λ_j in (2.2), counted according to their multiplicities ω_j, provided the conditions stated in this theorem are satisfied. In case the matrix A is normal, the zeros of the polynomial $\hat{Q}_{n,k}(\lambda)$, obtained from SMPE and STEA with the standard Euclidean inner product, are even better approximations to the eigenvalues λ_j of A as follows from Theorem 4.5 of [Si6].

3.1. Treatment of eigenvalue approximations. Theorem 3.1 below, which is of constructive nature, summarizes all the relevant results concerning the approximations to the λ_j. The corresponding approximations to eigenvectors and other vectors in the invariant subspaces are subsequently obtained with the help of the developments

in §5 of [Si6], and the relevant results for this problem are summarized in Theorem 3.2 below.

We note that in this section we have adopted all of the notation of the previous sections.

THEOREM 3.1. *Let the matrix A and the vector sequence $u_m, m = 0, 1, 2, \ldots$, be as described in the preceding section. Let the positive integers t and k be such that*

$$|\lambda_t| > |\lambda_{t+1}| \text{ and } k = \sum_{j=1}^{t} (p_j + 1) = \sum_{j=1}^{t} \omega_j. \tag{3.1}$$

Determine the coefficients $c_j^{(n,k)}, j = 0, 1, \ldots, k$, for one of the procedures SMPE, SMMPE, *and* STEA, *by utilizing $u_n, u_{n+1}, \ldots$, as described in* (1.4) *and* (1.5). *Then, under the additional conditions given in* (2.8) *and* (2.9),

$$\hat{Q}_{n,k}(\lambda) \equiv \sum_{j=0}^{k} c_j^{(n,k)} \lambda^j = \prod_{j=1}^{t} (\lambda - \lambda_j)^{\omega_j} + O(\varepsilon(n)) \text{ as } n \to \infty, \tag{3.2}$$

where

$$\varepsilon(n) = n^{\alpha} \left| \frac{\lambda_{t+1}}{\lambda_t} \right|^n, \tag{3.3}$$

α being some nonnegative integer. In fact, if the λ_j whose moduli are $|\lambda_t|$ are simple, then $\alpha = \bar{p}$, where $\bar{p} = \max\{p_j : |\lambda_j| = |\lambda_{t+1}|\}$. Consequently, the polynomial $\hat{Q}_{n,k}(\lambda)$ for $n \to \infty$, has ω_j zeros $\lambda_{jl}(n), 1 \le l \le \omega_j$, that tend to $\lambda_j, j = 1, 2, \ldots, t$. For each j and l we have

$$\lambda_{jl}(n) - \lambda_j = O(\delta_j(n)^{1/\omega_j}) \text{ as } n \to \infty, \tag{3.4}$$

where

$$\delta_j(n) = n^{\bar{p}} \left| \frac{\lambda_{t+1}}{\lambda_j} \right|^n. \tag{3.5}$$

Let us denote

$$\hat{\lambda}_j(n) = \frac{1}{\omega_j} \sum_{l=1}^{\omega_j} \lambda_{jl}(n) \text{ or } \hat{\lambda}_j(n) = \left[\frac{1}{\omega_j} \sum_{l=1}^{\omega_j} \lambda_{jl}(n)^{-1} \right]^{-1}. \tag{3.6}$$

Then

$$\hat{\lambda}_j(n) - \lambda_j = O(\delta_j(n)) \text{ as } n \to \infty. \tag{3.7}$$

Also, the p_jth derivative of $\hat{Q}_{n,k}(\lambda)$ has exactly one zero $\tilde{\lambda}_j(n)$ that tends to λ_j and satisfies

$$\tilde{\lambda}_j(n) - \lambda_j = O(\delta_j(n)) \text{ as } n \to \infty. \tag{3.8}$$

Let the matrix A be normal, i.e., $AA^* = A^*A$. Then $p_j = 0$ hence $\omega_j = 1$ for all j. If the $c_j^{(n,k)}$ are determined through the procedures SMPE *and* STEA *with the standard Euclidean inner product, and k is such that*

$$|\lambda_k| > |\lambda_{k+1}|, \tag{3.9}$$

and provided $q = u_n$ for STEA, *then* (3.2) *and* (3.4) *are substantially improved to read, respectively,*

$$\hat{Q}_{n,k}(\lambda) = \prod_{j=1}^{k}(\lambda - \lambda_j) + O\left(\left|\frac{\lambda_{k+1}}{\lambda_k}\right|^{2n}\right) \quad \text{as } n \to \infty, \tag{3.10}$$

and, for $j = 1, \ldots, k$,

$$\lambda_j(n) - \lambda_j = O\left(\left|\frac{\lambda_{k+1}}{\lambda_j}\right|^{2n}\right) \quad \text{as } n \to \infty, \tag{3.11}$$

where $\lambda_j(n)$ is the unique zero of $\hat{Q}_{n,k}(\lambda)$ that tends to λ_j.

We note again that the result in (3.2) and (3.3) was originally given in [SiB, §6, Thm. 6.1], and those in (3.10) and (3.11) were originally given for SMPE in [Si3]. The rest of Theorem 3.1 is new in that it has appeared only recently in [Si6].

One important aspect of Theorem 3.1 is the construction of *optimal* approximations to *defective* eigenvalues through (3.6) and (3.7). From (3.4) it is clear that when $p_j = 0$ hence $\omega_j = 1$, which occurs automatically if λ_j is a nondefective eigenvalue, the rate of convergence of the approximation corresponding to λ_j is optimal. In case that λ_j is a defective eigenvalue and $p_j > 0$, the rate of convergence of each of its ω_j corresponding approximations is $1/\omega_j$ of the optimal rate. For this case (3.6) and (3.7) show how the poor approximations $\lambda_{jl}(n)$ can be combined in a simple way to give an optimal approximation, namely $\hat{\lambda}_j(n)$. Similarly, (3.8) shows that $\tilde{\lambda}_j(n)$, the zero of the p_jth derivative of $\hat{Q}_{n,k}(\lambda)$ that tends to λ_j, has the same optimal convergence rate as $\hat{\lambda}_j(n)$. The results in (3.10) and (3.11) show that the approximations obtained from SMPE and STEA for a normal matrix converge twice as fast as those obtained for a nonnormal diagonalizable matrix having the same spectrum.

Another important aspect of Theorem 3.1 is that it shows clearly that the quality of the approximations to $\lambda_1, \lambda_2, \ldots,$ is better when k is larger. To see this let us consider the two different cases in which $(k,t) = (k',t')$ and $(k,t) = (k'',t'')$ in (3.1) of Theorem 3.1, where $t' < t''$. Obviously, $|\lambda_{t'}| > |\lambda_{t''}|$, and also $|\lambda_{t'+1}| > |\lambda_{t''+1}|$. Consequently, $|\lambda_{t''+1}/\lambda_j| < |\lambda_{t'+1}/\lambda_j|$ for $j = 1, 2, \ldots$. The validity of our claim now follows by comparing the outcomes of (3.2)–(3.11) with $(k,t) = (k',t')$ and $(k,t) = (k'',t'')$.

Finally, as has already been mentioned in [SiB], the methods contained in Theorem 3.1 reduce precisely to the classical power methods when $k = 1$. Specifically, solving (1.4) with $k = 1$, we have $\hat{Q}_{n,1}(\lambda) = \lambda - u_{01}/u_{00}$, from which there follows $\rho(n) = u_{01}/u_{00}$ as the approximation to the largest eigenvalue of A. Now $\rho(n) = (u_n, u_{n+1})/(u_n, u_n) = (u_n, Au_n)/(u_n, u_n)$ for SMPE procedure and this is simply the Rayleigh quotient for u_n. Similarly, $\rho(n) = (q_1, Au_n)/(q_1, u_n)$ and $\rho(n) = (q, Au_n)/(q, u_n)$, respectively, for SMMPE and STEA procedures, and this is how the standard power method is defined.

3.2. Treatment of invariant subspace approximations. For the treatment of the eigenvectors and invariant subspaces we need some preliminary work.

Let us rewrite (2.6) in the form

$$F(z) = \sum_{j=1}^{M} \sum_{i=0}^{p_j} \frac{d_{ji}}{(z - z_j)^{i+1}} + G(z), \tag{3.12}$$

where

$$z_j = \lambda_j^{-1} \text{ and } d_{ji} = (-z_j)^{i+1} a_{ji} \text{ for all } j, i. \tag{3.13}$$

Thus the d_{ji} are the coefficients of the principal part of the Laurent expansion of $F(z)$ about the pole $z_j, j = 1, \ldots, M$.

Consider the rational function

$$\hat{F}(z) = \frac{F(z) - F_{n+\nu}(z)}{z^{n+\nu+1}}, \tag{3.14}$$

which is analytic at $z = 0$ and has the Maclaurin series expansion

$$\hat{F}(z) = \sum_{i=0}^{\infty} u_{n+\nu+i+1} z^i. \tag{3.15}$$

By (3.12) we can write

$$\hat{F}(z) = \sum_{i=0}^{p_j} \frac{\hat{d}_{ji}}{(z - z_j)^{i+1}} + \hat{G}_j(z), \tag{3.16}$$

where

$$\hat{d}_{ji} = z_j^{-n-\nu-1} \sum_{l=i}^{p_j} \binom{-n-\nu-1}{l-i} z_j^{-l+i} d_{jl}, \tag{3.17}$$

and $\hat{G}_j(z)$ is analytic at z_j, i.e., as above, the $\hat{d}_{ji}$ are coefficients of the principal part of the Laurent expansion of $\hat{F}(z)$ about the pole $z_j, j = 1, \ldots, M$. Unlike before, both $\hat{F}(z)$ and the $\hat{d}_{ji}$ depend on n, in addition. The vector $\hat{d}_{jp_j}$, being a scalar multiple of the constant vector d_{jp_j}, is an eigenvector of A corresponding to the eigenvalue λ_j. For $i \neq p_j$, the vector $\hat{d}_{ji}$, being a linear combination of the constant vectors $d_{jl}, i \leq l \leq p_j$, is in the invariant subspace Y_i, and, as is obvious from (3.17), the coefficients of the d_{jl} in this linear combination are polynomials in n, up to the common multiplicative factor $z_j^{-n-\nu-1}$.

Following now the developments in §5 of [Si6], we obtain the following constructive result for the $\hat{d}_{ji}$.

THEOREM 3.2. *With the notation and conditions of Theorem* 3.1, *let us define, for* $1 \leq j \leq t$,

$$\zeta_j(n) = 1/\hat{\lambda}_j(n) \text{ or } \zeta_j(n) = 1/\tilde{\lambda}_j(n), \tag{3.18}$$

and, for $0 \le i \le p_j$ *and* $1 \le l \le \omega_j$,

$$\hat{d}_{ji,l}(n) = (z - \zeta_j(n))^i \left. \frac{\sum_{r=1}^{k} c_r^{(n,k)} z^{k-r} \sum_{m=1}^{r} u_{n+\nu+m} z^{m-1}}{\sum_{r=0}^{k} c_r^{(n,k)} (k-r) z^{k-r-1}} \right|_{z=1/\lambda_{jl}(n)} \tag{3.19}$$

and

$$\hat{d}_{ji}(n) = \sum_{l=1}^{\omega_j} \hat{d}_{ji,l}(n). \tag{3.20}$$

Then, for $0 \le i \le p_j$, $\hat{d}_{ji}(n)$ *is an approximation to* $\hat{d}_{ji}$ *in* (3.17) *in the sense*

$$\limsup_{n\to\infty} |\hat{d}_{ji}(n) - \hat{d}_{ji}|^{1/n} \le |\lambda_{t+1}|. \tag{3.21}$$

We note that Theorem 3.2 actually contains the basic ingredients of a potentially bona fide numerical method for approximating the eigenvectors and other vectors in invariant subspaces corresponding to largest eigenvalues of A. The resulting method, which is described below, (i) makes use of only $u_n, u_{n+1}, \ldots$, disregarding $u_0, u_1, \ldots, u_{n-1}$ entirely, and (ii) enables us to construct optimal approximations to the vectors $a_{ji}, 0 \le i \le p_j$, for $p_j = 0$ as well as $p_j > 0$. We now turn to these constructions.

3.2.1. Approximation of the eigenvector a_{jp_j}. Let us first specialize the result of Theorem 3.2 to the case $i = p_j$. We have

$$\hat{d}_{jp_j} = \lambda_j^{n+\nu+1} d_{jp_j}, \tag{3.22}$$

so that (3.21) can also be written as

$$\limsup_{n\to\infty} |\lambda_j^{-n-\nu-1} \hat{d}_{jp_j}(n) - d_{jp_j}|^{1/n} \le \left| \frac{\lambda_{t+1}}{\lambda_j} \right|. \tag{3.23}$$

This clearly shows that the vector $\hat{d}_{jp_j}(n)$, as $n \to \infty$, aligns itself with the *constant* vector d_{jp_j}, which is proportional to the eigenvector a_{jp_j}, practically at the rate of $|\lambda_{t+1}/\lambda_j|^n$. It is thus sufficient to compute the vectors $\hat{d}_{ji,l}(n), 1 \le l \le \omega_j$, by (3.19), and then to form $\hat{d}_{ji}(n)$ by (3.20) as our approximation to the (appropriately normalized) eigenvector a_{jp_j}, and this is valid whether $p_j = 0$ or $p_j > 0$.

3.2.2. Approximation of the vectors $a_{ji}, 0 \le i \le p_j - 1$. Although the vector a_{jp_j} (up to a multiplicative constant) can be determined from $\hat{d}_{jp_j}(n)$ in a rather painless manner, the determination of the remaining a_{ji} from the $\hat{d}_{jl}(n)$ becomes somewhat involved. The reason for this is that the vectors $\hat{d}_{ji}$, apart from the scalar multiplicative factor $z_j^{-n-\nu-1}$, are linear combinations of the d_{jl} hence of the $a_{jl}, i \le l \le p_j$, with coefficients that vary as functions of n, as can be seen from (3.17) and (3.13), and as has been mentioned before. This means that the $\hat{d}_{ji}$ do *not* have a fixed direction with varying n.

Let us now rewrite (3.17) in the form

$$T(n) \begin{bmatrix} d_{j0} \\ d_{j1} \\ \vdots \\ d_{jp_j} \end{bmatrix} = z_j^{n+\nu+1} \begin{bmatrix} \hat{d}_{j0} \\ \hat{d}_{j1} \\ \vdots \\ \hat{d}_{jp_j} \end{bmatrix}, \tag{3.24}$$

where $T(n)$ is the upper triangular matrix

$$T(n) = \begin{bmatrix} \tau_{00} & \tau_{01} & \cdots & \tau_{0p_j} \\ & \tau_{11} & \cdots & \tau_{1p_j} \\ & & \ddots & \vdots \\ & & & \tau_{p_jp_j} \end{bmatrix}, \quad \tau_{il} = \binom{-n-\nu-1}{l-i} z_j^{-l+i} \quad \text{all } i \text{ and } l. \tag{3.25}$$

Obviously, $T(n)$ is invertible since its diagonal elements are unity. Thus,

$$\begin{bmatrix} d_{j0} \\ d_{j1} \\ \vdots \\ d_{jp_j} \end{bmatrix} = T(n)^{-1} \begin{bmatrix} \hat{d}_{j0} \\ \hat{d}_{j1} \\ \vdots \\ \hat{d}_{jp_j} \end{bmatrix} z_j^{n+\nu+1}, \tag{3.26}$$

where $T(n)^{-1}$ is also upper triangular, its diagonal elements being unity.

Now since all elements of $T(n)$ are polynomials in n, and since its determinant is unity, the elements of $T(n)^{-1}$ turn out to be polynomials in n, i.e., the matrix $T(n)^{-1}$ can grow at most polynomially as $n \to \infty$. If we denote the nonzero elements of $T(n)^{-1}$ by $\rho_{il}, i \le l \le p_j, 0 \le i \le p_j$, then we can write (3.26) in the form

$$d_{ji} = z_j^{n+\nu+1} \sum_{l=i}^{p_j} \rho_{il} \hat{d}_{jl}, \quad 0 \le i \le p_j. \tag{3.27}$$

Let us replace $\hat{d}_{jl}$ in (3.27) by $[(\hat{d}_{jl} - \hat{d}_{jl}(n)) + \hat{d}_{jl}(n)]$, and invoke (3.21). After some manipulation we obtain

$$\limsup_{n\to\infty} \left| d_{ji} - z_j^{n+\nu+1} \sum_{l=i}^{p_j} \rho_{il} \hat{d}_{jl}(n) \right|^{1/n} \le \left| \frac{\lambda_{t+1}}{\lambda_j} \right|. \tag{3.28}$$

This implies that the vector $\sum_{l=i}^{p_j} \rho_{il} \hat{d}_{jl}(n)$ aligns itself with the fixed vector d_{ji} as $n \to \infty$ practically at the rate of $|\lambda_{t+1}/\lambda_j|^n$. We leave the details of the proof of (3.28) to the reader.

We note that (3.28) shows how to construct a good approximation to d_{ji} from the $\hat{d}_{jl}(n)$ and λ_j, provided λ_j is known. Since λ_j is not known, however, the vector $\sum_{l=i}^{p_j} \rho_{il} \hat{d}_{jl}(n)$ cannot be constructed. We, therefore, propose to replace λ_j in the matrix $T(n)^{-1}$ by the known approximations $\zeta_j(n)$. Also, in this case, it can be shown that (3.28) remains valid. Again, we leave the details of the proof to the reader.

Before closing this section, we must mention that the developments of this section are meant to be theoretical in general. Although they can be used for computational purposes for small values of k, their use for large k is likely to introduce numerical instabilities in many cases. These instabilities are mainly a result of our direct use of the power iterations $u_{n+i} = A^i u_n, i = 0, 1, \ldots$. They exhibit themselves first of all through the poor computed approximations to the λ_j, which ultimately affect the computed eigenvector approximations. This problem can be remedied by observing that the approximations $F_{n,k}(z)$ that we developed and applied to the matrix eigenvalue problem are very tightly connected with Krylov subspace methods for some of which there exist computationally stable implementations. In particular, the SMPE and

STEA procedures are related to the method of Arnoldi and the method of Lanczos, respectively, as we show in detail in the next two sections.

4. General projection methods and the methods of Arnoldi and Lanczos for the matrix eigenproblem.

4.1. General projection methods. Let $\{v_1, \ldots, v_k\}$ and $\{w_1, \ldots, w_k\}$ be two linearly independent sets of vectors in $\mathbf{C}^N$, and define the $N \times k$ matrices V and W by

$$V = [v_1|v_2|\cdots|v_k] \text{ and } W = [w_1|w_2|\cdots|w_k]. \tag{4.1}$$

In addition, let us agree to denote the subspaces span $\{v_1, \ldots, v_k\}$ and span $\{w_1, \ldots, w_k\}$ by V and W, respectively. For simplicity, let us also take (x, y) to be the standard Euclidean inner product x^*y.

In projection methods one looks for an approximate eigenvalue-eigenvector pair (λ, x) with $x \in V$ that satisfies the condition

$$(y, Ax - \lambda x) = 0 \quad \text{for all } y \in W, \tag{4.2}$$

which can also be written in the equivalent form

$$W^*(A - \lambda I)V\xi = 0 \quad \text{for some } \xi \in \mathbf{C}^k. \tag{4.3}$$

Here we have used the fact that $x \in V$ implies that $x = V\xi$ for some $\xi \in \mathbf{C}^k$. Of course, (4.3) holds if and only if λ is an eigenvalue of the matrix pencil (W^*AV, W^*V), i.e., it satisfies the characteristic equation

$$\det(W^*AV - \lambda W^*V) = 0. \tag{4.4}$$

In general, (4.4) has k solutions for λ, which are known as *Ritz values* in the literature. Given that λ' is a Ritz value, the corresponding eigenvector ξ' is a solution of the homogeneous system in (4.3). The eigenvector approximation corresponding to λ' is now $x' = V\xi'$, and is known as a *Ritz vector*.

The different projection methods are characterized by the subspaces V and W that they employ. (Note that V and W are also called, respectively, the right and left subspaces.)

4.2. The method of Arnoldi. In this method V and W are Krylov subspaces given by

$$V = V_{k-1} = \text{ span}\{u_0, Au_0, \ldots, A^{k-1}u_0\} \quad \text{and} \quad W = W_{k-1} = V_{k-1}, \tag{4.5}$$

for some arbitrary vector u_0.

Arnoldi has given a very successful implementation of this method. In this implementation the vectors $A^iu_0, i = 0, 1, \ldots$, are orthogonalized by a very special Gram–Schmidt process as follows:

Step 0. Let $v_1 = u_0/\|u_0\|$.

Step 1. For $j = 1, \ldots, k-1$, do

(4.6) Determine the scalar $h_{j+1,j} > 0$ and the vector v_{j+1}, such that $h_{j+1,j}v_{j+1} = Av_j - \sum_{i=1}^{j} h_{ij}v_i, h_{ij} = (v_i, Av_j), 1 \le i \le j$, and $\|v_{j+1}\| = 1$.

Thus the $N \times k$ matrix $V = [v_1|v_2|\cdots|v_k]$ is unitary in the sense that V^*V is the $k \times k$ identity matrix. As a result, $W^*V = V^*V = I$, and the generalized eigenvalue problem of (4.3) now becomes

$$H\xi = \lambda\xi, \tag{4.7}$$

where H is the $k \times k$ upper Hessenberg matrix

$$H = \begin{bmatrix} h_{11} & h_{12} & \cdots & & h_{1k} \\ h_{21} & h_{22} & \cdots & & h_{2k} \\ & h_{32} & \cdots & & h_{3k} \\ & & \ddots & & \vdots \\ & & & h_{k,k-1} & h_{kk} \end{bmatrix}, \tag{4.8}$$

i.e., the Ritz values are the eigenvalues of H.

4.3. The method of Lanczos. In this method V and W are the Krylov subspaces

$$\begin{aligned} V &= V_{k-1} = \operatorname{span}\{u_0, Au_0, \ldots, A^{k-1}u_0\} \quad \text{and} \\ W &= W_{k-1} = \operatorname{span}\{q, A^*q, \ldots, (A^*)^{k-1}q\}, \end{aligned} \tag{4.9}$$

for some arbitrary vectors u_0 and q.

The algorithm given by Lanczos generates one set of vectors $\{v_1, \ldots, v_k\}$ from the $A^iu_0, i = 0, 1, \ldots, k-1$, and another set of vectors $\{w_1, \ldots, w_k\}$ from the $(A^*)^iq, i = 0, 1, \ldots, k-1$, that satisfy the biorthogonality condition

$$(w_i, v_j) = \delta_{ij}, \tag{4.10}$$

as long as the process does not break down. This is achieved by the following algorithm.

Step 0. Set $v_1 = \sigma u_0$ and $w_1 = \tau q$ such that $(w_1, v_1) = 1$.

Step 1. For $j = 1, \ldots, k-1$, do

(4.11) (a) Compute $\hat{v}_{j+1}$ and $\hat{w}_{j+1}$ by
$\hat{v}_{j+1} = Av_j - \alpha_j v_j - \beta_j v_{j-1}$ and $\hat{w}_{j+1} = A^*w_j - \bar{\alpha}_j w_j - \bar{\delta}_j w_{j-1}$, with
$\alpha_j = (w_j, Av_j)$. (When $j = 1$ take $\beta_1 v_0 = \bar{\delta}_1 w_0 = 0$.)
(b) Choose δ_{j+1} and β_{j+1} such that
$\delta_{j+1}\beta_{j+1} = (\hat{w}_{j+1}, \hat{v}_{j+1})$, and set
$v_{j+1} = \hat{v}_{j+1}/\delta_{j+1}$ and $w_{j+1} = \hat{w}_{j+1}/\bar{\beta}_{j+1}$.

By (4.10) the matrices V and W satisfy $W^*V = I$. As a result, the generalized eigenvalue problem of (4.3) becomes

$$H\xi = \lambda\xi, \tag{4.12}$$

where H is the $k \times k$ tridiagonal matrix

$$H = \begin{bmatrix} \alpha_1 & \beta_2 & & & & \\ \delta_2 & \alpha_2 & \beta_3 & & & \\ & \delta_3 & \alpha_3 & \beta_4 & & \\ & & \ddots & \ddots & \ddots & \\ & & & & & \beta_k \\ & & & & \delta_k & \alpha_k \end{bmatrix}, \tag{4.13}$$

and the Ritz values are the eigenvalues of H.

4.4. The case of Hermitian A. The subspaces V in (4.5) and (4.9) are identical. When A is Hermitian, i.e., $A^* = A$, and $q = u_0$, the subspaces W in (4.5) and (4.9) become identical too. Thus the methods of Arnoldi and Lanczos become equivalent for the case under consideration. Furthermore, it can be shown that the elements h_{ij} of the matrix H in the method of Arnoldi satisfy $\overline{h_{i,i+1}} = h_{i+1,i}$ so that $h_{i,i+1} = h_{i+1,i} > 0$ for $i = 1, 2, \ldots, k-1$, while $h_{ij} = 0$ for $j \geq i+2$. The diagonal elements h_{ii} are all real. That is to say, in the absence of roundoff, the matrix H is real symmetric tridiagonal. If we pick $q = u_0$ and choose $\delta_j = \beta_j = \sqrt{(\hat{v}_j, \hat{v}_j)}$ in the method of Lanczos, then the matrix H in (4.13) turns out to be real symmetric and is exactly the same as the one produced by the method of Arnoldi.

The properties of the Ritz values and Ritz vectors of the Lanczos method, as applied to Hermitian matrices, have been analyzed by Kaniel [K], Paige [Pai], and Saad [Sa1]. The paper [Sa2] gives results for non-Hermitian matrices.

5. Equivalence of rational approximation procedures and Krylov subspace methods. We now go back to the rational approximation procedures SMPE, SMMPE, and STEA. In particular, we concentrate on the poles and residues of the rational functions $F_{n,k}(z)$.

5.1. Poles of $F_{n,k}(z)$ vs. Ritz values. From the determinant representations of $F_{n,k}(z)$ that are given in Theorem 2.2 of [Si6], it follows that the denominator $Q_{n,k}(z)$ of $F_{n,k}(z)$ is a constant multiple of the determinant

$$D(\lambda) = \begin{vmatrix} 1 & \lambda & \ldots & \lambda^k \\ u_{00} & u_{01} & \ldots & u_{0k} \\ u_{10} & u_{11} & \ldots & u_{1k} \\ \vdots & \vdots & & \vdots \\ u_{k-1,0} & u_{k-1,1} & \ldots & u_{k-1,k} \end{vmatrix}, \tag{5.1}$$

where $\lambda = z^{-1}$ and u_{ij} are as defined in (1.5). This implies that the zeros of the polynomial $D(\lambda)$ are the reciprocals of the zeros of $Q_{n,k}(z)$, or, equivalently, the reciprocals of the poles of $F_{n,k}(z)$. In addition, they are the roots of a generalized eigenvalue problem as we show next.

THEOREM 5.1. *Whatever the u_{ij}, the zeros of the polynomial $D(\lambda)$ in* (5.1) *are the eigenvalues of the matrix pencil (X, T), where*

(5.2)

$$X = \begin{bmatrix} u_{01} & u_{02} & \cdots & u_{0k} \\ u_{11} & u_{12} & \cdots & u_{1k} \\ \vdots & \vdots & & \vdots \\ u_{k-1,1} & u_{k-1,2} & \cdots & u_{k-1,k} \end{bmatrix} \quad \textit{and} \quad T = \begin{bmatrix} u_{00} & u_{01} & \cdots & u_{0,k-1} \\ u_{10} & u_{11} & \cdots & u_{1,k-1} \\ \vdots & \vdots & & \vdots \\ u_{k-1,0} & u_{k-1,1} & \cdots & u_{k-1,k-1} \end{bmatrix},$$

i.e., they satisfy the equation

$$\det(X - \lambda T) = 0. \tag{5.3}$$

Proof. Multiply the $(j-1)$st column of $D(\lambda)$ by λ and subtract from the jth column for $j = k+1, k, \ldots, 2$, in this order. This results in

$$D(\lambda) = \left|\begin{array}{c|c} 1 & 0\cdots 0 \\ \hline u_{00} & \\ u_{10} & X - \lambda T \\ \vdots & \\ u_{k-1,0} & \end{array}\right| = \det(X - \lambda T), \tag{5.4}$$

thus proving the claim. □

In the remainder of this section we take $(x, y) = x^*y$.

When u_{ij} are as in (1.5), Theorem 5.1 takes on the following interesting form.

THEOREM 5.2. *Define the $N \times k$ matrices V and W by*

$$V = [u_n | u_{n+1} | \cdots | u_{n+k-1}] \tag{5.5}$$

and

$$\begin{aligned} &W = V \text{ for SMPE}, \\ &W = [q_1 | q_2 | \cdots | q_k] \text{ for SMMPE}, \\ &W = [q | A^*q | \cdots | (A^*)^{k-1}q] \text{ for STEA}. \end{aligned} \tag{5.6}$$

Then, with u_{ij} as defined by (1.5), *the zeros of $D(\lambda)$ are the eigenvalues of the matrix pencil (W^*AV, W^*V), i.e., they satisfy*

$$\det(W^*AV - \lambda W^*V) = 0. \tag{5.7}$$

Consequently, the reciprocals of the poles of the rational approximations $F_{n,k}(z)$ obtained from the SMPE *or* SMMPE *or* STEA *procedures are the Ritz values of the Krylov subspace methods whose right and left subspaces are column spaces of V and W, respectively.*

Proof. Since Theorem 5.1 applies, all we need to show is that $X = W^*AV$ and $T = W^*V$ there. That $T = W^*V$ follows from (1.5), (5.2), (5.5), and (5.6). From (1.5), (5.2), and (5.6), we similarly have $X = W^*[u_{n+1} | \cdots | u_{n+k}]$. Now using the fact that $u_{j+1} = Au_j, j \geq 0$, we also have $[u_{n+1} | \cdots | u_{n+k}] = AV$. Consequently, $X = W^*AV$. Again, from $u_{j+1} = Au_j, j \geq 0$, we realize, in addition, that the right subspace for all three methods is none other than the Krylov subspace span $\{u_n, Au_n, \ldots, A^{k-1}u_n\}$. This completes the proof. □

5.2. Residues of $F_{n,k}(z)$ vs. Ritz vectors. Turning Theorem 5.2 around, what we have is that the Ritz values obtained by applying the Krylov subspace methods whose left and right subspaces are column spaces of V and W, respectively, are, in fact, the reciprocals of the poles of the corresponding rational approximations $F_{n,k}(z)$ to the meromorphic function $F(z) = \sum_{i=0}^{\infty} u_i z^i$. An immediate question that arises is, of course, whether there is any connection between the Ritz *vectors* and the $F_{n,k}(z)$. The answer, which is in the affirmative, is provided in Theorem 5.3 below.

THEOREM 5.3. *Let $\hat{\lambda}$ be a Ritz value of the Krylov subspace methods whose right and left subspaces are column spaces of, respectively, V and W in Theorem* 5.2. *Denote the corresponding Ritz vector by $\hat{x}$. Let $\nu = -1$ in the corresponding rational approximation $F_{n,k}(z)$, cf.* (1.2). *Provided $\hat{\lambda}$ is simple, $\hat{x}$ is a constant multiple of the residue of $F_{n,k}(z)$ at the pole $\hat{z} = 1/\hat{\lambda}$.*

Proof. Let us first determine the residue of $F_{n,k}(z)$ at the pole $\hat{z} = 1/\hat{\lambda}$. With $\nu = -1$

$$\text{Res } F_{n,k}(z)|_{z=\hat{z}} = \frac{P_{n,k}(\hat{z})}{Q'_{n,k}(\hat{z})} = \frac{\sum_{r=0}^{k} c_r \hat{z}^{k-r} F_{n+r-1}(\hat{z})}{Q'_{n,k}(\hat{z})}, \tag{5.8}$$

since $Q'_{n,k}(\hat{z}) \neq 0$ that follows from the assumption that $\hat{\lambda}$ is simple, which implies that $\hat{z}$ is a simple pole. By $F_{n+s}(z) = F_{n-1}(z) + \sum_{m=n}^{n+s} u_m z^m$ and $\sum_{r=0}^{k} c_r \hat{z}^{k-r} = 0$, we can rewrite (5.8) in the form

$$\text{Res } F_{n,k}(z)|_{z=\hat{z}} = \frac{1}{Q'_{n,k}(\hat{z})} \sum_{r=1}^{k} c_r \hat{z}^{k-r} \sum_{m=n}^{n+r-1} u_m \hat{z}^m = \frac{\hat{z}^{n+k-1}}{Q'_{n,k}(\hat{z})} \sum_{m=0}^{k-1} \eta_m u_{n+m}, \tag{5.9}$$

where

$$\eta_m = \sum_{r=m+1}^{k} c_r \hat{\lambda}^{r-m-1}, \qquad m = 0, 1, \ldots, k-1. \tag{5.10}$$

Let us now denote $\eta = (\eta_0, \eta_1, \ldots, \eta_{k-1})^T$. Then (5.9) implies that Res $F_{n,k}(z)|_{z=\hat{z}}$ is a scalar multiple of $V\eta$. Recall that the Ritz vector corresponding to $\hat{\lambda}$ is $V\hat{\xi}$, where $\hat{\xi} \in C^k$ and satisfies $W^*(A - \hat{\lambda}I)V\hat{\xi} = 0$, which, on account of Theorem 5.2, is the same as $(X - \hat{\lambda}T)\hat{\xi} = 0$. Thus in order to show that Res $F_{n,k}(z)|_{z=\hat{z}}$ is a constant multiple of the Ritz vector corresponding to the Ritz value $\hat{\lambda}$, it is sufficient to show that

$$(X - \hat{\lambda}T)\eta = 0. \tag{5.11}$$

From (5.2), the $(i+1)$st component of the k-dimensional vector $\tau = (X - \hat{\lambda}T)\eta, i = 0, 1, \ldots, k-1$, is

$$\tau_i = \sum_{m=0}^{k-1} (u_{i,m+1} - \hat{\lambda} u_{im}) \eta_m, \tag{5.12}$$

which, by (5.10), becomes

$$\tau_i = \sum_{m=0}^{k-1} (u_{i,m+1} - \hat{\lambda} u_{im}) \sum_{r=m+1}^{k} c_r \hat{\lambda}^{r-m-1}. \tag{5.13}$$

Expanding and rearranging this summation, we obtain

$$\tau_i = -u_{i0} \left(\sum_{r=1}^{k} c_r \hat{\lambda}^r \right) + \sum_{m=1}^{k} u_{im} c_m. \tag{5.14}$$

Recalling that $\sum_{r=0}^{k} c_r \hat{\lambda}^r = 0$, we can rewrite (5.14) as

$$\tau_i = \sum_{m=0}^{k} u_{im} c_m. \tag{5.15}$$

Finally, from the assumption that $c_k = 1$ and from the fact that $c_0, c_1, \ldots, c_{k-1}$ satisfy the linear equations in (1.4), we conclude that

$$\tau_i = 0, \quad i = 0, 1, \ldots, k-1. \tag{5.16}$$

This completes the proof. □

5.3. Summary of $F_{n,k}(z)$ vs. Krylov subspace methods. We now combine the results of Theorems 5.2 and 5.3 to state the following equivalence theorem, which forms the main result of this section, and one of the main results of this work.

THEOREM 5.4. *Let $F_{n,k}(z)$ be the rational approximation obtained by applying the* SMPE *or* SMMPE *or* STEA *procedure to the vector-valued power series $\sum_{m=0}^{\infty} u_m z^m$, where $u_m = A^m u_0, m = 0, 1, \ldots$, are power iterations. Denote the reciprocals of the poles of $F_{n,k}(z)$ by $\lambda'_1, \ldots, \lambda'_k$. Setting $\nu = -1$ in the numerator of $F_{n,k}(z)$, denote the corresponding residues of $F_{n,k}(z)$ by $x'_1, \ldots, x'_k$. Next, denote by $\lambda''_1, \ldots, \lambda''_k$ and $x''_1, \ldots, x''_k$, respectively, the Ritz values and corresponding Ritz vectors produced by the Krylov subspace methods whose right subspace is* $\operatorname{span}\{u_n, Au_n, \ldots, A^{k-1}u_n\}$ *and left subspaces are the column spaces of the matrices W in* (5.6). *Then*

$$\lambda'_j = \lambda''_j, \qquad j = 1, \ldots, k, \tag{5.17}$$

and

$$x'_j \propto x''_j, \text{ provided } \lambda'_j \text{ is simple.} \tag{5.18}$$

More can be said about the SMPE and STEA procedures versus the methods of Arnoldi and Lanczos, and this is done in Corollary 5.5 below.

COROLLARY 5.5. *With $F_{n,k}(z), \lambda'_j, x'_j, j = 1, \ldots, k$, as in Theorem* 5.4, *let λ''_j, x''_j, $j = 1, \ldots, k$, be the Ritz values and Ritz vectors produced by applying the k-step Arnoldi or Lanczos methods to the matrix A, starting with the vector $u_n = A^n u_0$. (That is to say, replace the initial vector u_0 in Step* 0 *of* (4.6) *or* (4.11) *by the nth power iteration u_n.) In addition, let q be the same vector for the* STEA *procedure and the Lanczos method. Then the* SMPE *and* STEA *procedures are equivalent to the methods of Arnoldi and Lanczos, respectively, precisely in the sense of* (5.17) *and* (5.18).

Now that we have shown the equivalence of the methods of Arnoldi and Lanczos with the generalized power methods based on the SMPE and STEA approximation procedures, we realize that those results we proved in §3 for the latter and which pertain to the nondefective as well as defective eigenvalues of A are, in fact, new results for the former. That is to say, if we apply the methods of Arnoldi or Lanczos of order k to the matrix A starting with the nth power iteration $u_n = A^n u_0$ for large n, then the Ritz values are approximations to the k largest distinct eigenvalues of A counted according to the multiplicities that appear in (2.2). Similarly, the Ritz vectors can be used for constructing the approximations to the corresponding invariant subspaces. These points will be considered in greater detail in the next section.

Judging from Theorems 3.1 and 3.2, we conclude that applying Krylov subspace methods beginning with $u_n = A^n u_0, n > 0$, rather than with u_0, may be advantageous, especially when the eigenvalues that are largest in modulus *and* the corresponding eigenvectors and invariant subspaces are needed. Specifically, a given level of accuracy may be achieved for smaller values of k as n is increased. We recall that k is also the number of vectors $v_1, v_2, \ldots$, in (4.1) that need to be saved. Thus we see that the strategy in which Krylov subspace methods are applied to u_n with n sufficiently large

may result in substantial savings in storage. In addition, smaller k means savings in the computational overhead caused by the arithmetic operations that lead to the matrices V and W, and, subsequently, to the Ritz vectors. (For a detailed discussion of this point we refer the reader to §7 Example 7.2.) All this was observed to be the case in various examples done by the author.

5.4. Optimality properties of the Arnoldi method. In §1 we mentioned that the coefficients of $c_i^{(n,k)}$ of the denominator polynomial $Q_{n,k}(z)$ of $F_{n,k}(z)$ for the SMPE procedure are the solution to the optimization problem given in (1.6). If we now pick the vectors u_m as the power iterations $u_m = A^m u_0, m = 0, 1, \ldots$, then (1.6) reads

$$\min_{c_0,c_1,\ldots,c_{k-1}} \left\| \left(\sum_{j=0}^{k-1} c_j A^j + A^k \right) u_n \right\|. \tag{5.19}$$

Exploiting the fact that the method of Arnoldi is equivalent to the generalized power method based on the SMPE approximation procedure, we can state the following optimality properties for the Arnoldi method as applied to a *general* matrix A.

THEOREM 5.6. *Let $\lambda_j', x_j', j = 1, 2, \ldots, k$, be the Ritz values and appropriately normalized Ritz vectors, respectively, produced by applying the k-step Arnoldi method to the matrix A starting with the power iteration $u_n = A^n u_0$. Let $\mathcal{P}_k$ denote the set of monic polynomials of degree exactly k, while π_k denotes the set of polynomials of degree at most k. Then for $k < k_0$, cf.* (2.4),

$$\left\| \left[\prod_{i=1}^{k} (A - \lambda_i' I) \right] u_n \right\| = \min_{f \in \mathcal{P}_k} \|f(A) u_n\| \equiv \varepsilon_{n,k}, \tag{5.20}$$

$$x_j' = \left[\prod_{\substack{i=1 \\ i \neq j}}^{k} (A - \lambda_i' I) \right] u_n, \tag{5.21}$$

$$(A - \lambda_j' I) x_j' = \left(\sum_{i=0}^{k-1} c_i^{(n,k)} A^i + A^k \right) u_n = \sum_{i=0}^{k-1} c_i^{(n,k)} u_{n+i} + u_{n+k}, \tag{5.22}$$

$$\begin{aligned} \|(A - \lambda_j' I) x_j'\| &= \min_{\lambda \in \mathbf{C}, g \in \mathcal{P}_{k-1}} \|(A - \lambda I) g(A) u_n\|, \\ &= \min_{\lambda \in \mathbf{C}} \|(A - \lambda I) x_j'\|, \\ &= \min_{g \in \mathcal{P}_{k-1}} \|(A - \lambda_j' I) g(A) u_n\|, \\ &= \varepsilon_{n,k} \text{ independently of } j, \end{aligned} \tag{5.23}$$

and

$$((A - \lambda_j' I) x_j', g(A) u_n) = 0 \quad \text{all } g \in \pi_{k-1}. \tag{5.24}$$

For $k = k_0$, we have $A x_j' = \lambda_j' x_j'$.

Proof. We start by noting that (5.24) is nothing but a restatement of the requirement that $Ax'_j - \lambda'_j x'_j$ be orthogonal to the left subspace of the Arnoldi method, which is also its right subspace $V = \{g(A)u_n : g \in \pi_{k-1}\}$.

Since the Ritz values $\lambda'_j, j = 1, \dots, k$, are the zeros of the monic polynomial $\hat{Q}_{n,k}(\lambda) = \sum_{i=0}^{k-1} c_i^{(n,k)} \lambda^i + \lambda^k$, we can write

$$\hat{Q}_{n,k}(\lambda) = \prod_{i=1}^{k} (\lambda - \lambda'_i). \tag{5.25}$$

Thus

$$\hat{Q}_{n,k}(A) = \sum_{i=0}^{k-1} c_i^{(n,k)} A^i + A^k = \prod_{i=1}^{k} (A - \lambda'_i I). \tag{5.26}$$

Combining (5.26) with (5.19), we obtain (5.20).

Provided x'_j is as given by (5.21), the proofs of (5.22) and (5.23) are immediate.

To prove the validity of (5.21) it is sufficient to show that $x'_j \in V$ and that $(A - \lambda'_j I)x'_j$ is orthogonal to all the vectors in V. That $x'_j \in V$ is obvious from (5.21) itself. The fact that $c_i^{(n,k)}, i = 0, 1, \dots, k-1$, are the solution of the optimization problem in (5.19) implies that the vector $\hat{Q}_{n,k}(A)u_n$ is orthogonal to every vector in V. But $\hat{Q}_{n,k}(A)u_n = (A - \lambda'_j I)x'_j$, as can be seen from (5.26). This completes the proof. □

Note that the proofs of (5.20) and (5.21) for Hermitian matrices can also be found in [Par2, Chap. 12, pp. 239–240].

A few historical notes on the methods of Arnoldi and Lanczos are now in order.

Following the work of Arnoldi the equivalent form in (5.19) was suggested in a paper by Erdelyi [E], in the book by Wilkinson [W, pp. 583–584], and in the papers by Manteuffel [M] and Sidi and Bridger [SiB]. The equivalence of the different approaches does not seem to have been noticed, however. For instance, [W] discusses both approaches without any attempt to explore the connection between them. With the exception of [SiB], these works all consider the case $n = 0$. The case $n > 0$ and the limit as $n \to \infty$ are considered in [SiB] and [Si3].

In his discussion of the power iterations in [H, Chap. 7], Householder gives determinantal representations of certain polynomials whose zeros are approximations to the largest eigenvalues of the matrix being considered. One of these representations, namely, the one given in (16) in [H, p. 186], coincides with the determinant $D(\lambda)$ in (5.1) of the present work pertaining to the STEA approximation procedure with $n \geq 0$. It is shown there that the zeros of $D(\lambda)$ tend to the k largest eigenvalues of the matrix A as $n \to \infty$, but a theorem as detailed as our Theorem 3.1 is not given. It is also mentioned in the same place that, apart from a constant multiplicative factor, the polynomials $D(\lambda)$ with $n = 0$ are precisely the so-called Lanczos polynomials given in (10) of [H, p. 23] that are simply $\det(\lambda I - H)$ with H as given in (4.13). As we pointed out in this section, up to a constant multiplicative factor, $D(\lambda)$ with $n > 0$ is itself the Lanczos polynomial $\det(\lambda I - H)$ when the Lanczos method is being applied with u_0 replaced by $u_n = A^n u_0$. It is not clear to the author whether this connection between $D(\lambda)$ with $n > 0$ and the Lanczos method has been observed before or not.

6. Stable numerical implementations. In this section we concentrate on the implementation of the generalized power methods based on the SMPE and the STEA

approximation procedures as these are related to the methods of Arnoldi and Lanczos, respectively, and as good implementations for the latter are known. For example, the implementations in (4.6) and (4.11) are usually quite stable.

6.1. General computational considerations. The theoretical results of §3 all involve the limiting procedure $n \to \infty$. When $|\lambda_1|$ is larger (smaller) than 1, we may have difficulties in implementing the procedures above due to possible overflow (underflow) in the computation of the vectors u_m for large m. This situation can be remedied easily as will be shown below.

We first observe that the denominator polynomial $Q_{n,k}(z)$ of the vector-valued rational approximation $F_{n,k}(z)$ remains unchanged when the vectors $u_n, u_{n+1}, u_{n+2}, \dots$, are all multiplied by the same scalar, say α, and so do its zeros. Consequently, the vectors $\hat{d}_{ji}(n)$ defined in Theorem 3.2 remain the same up to the multiplicative factor α. That is to say, as far as the matrix eigenvalue problem is concerned, multiplication of the vectors $u_n, u_{n+1}, \dots$, by the scalar α leaves the eigenvalue approximations unchanged and multiplies the eigenvector approximations by α.

For the purpose of numerical implementation we propose to pick $\alpha = 1/\|u_n\|$, and we achieve this by the following simple algorithm that is also used in the classical power method.

(6.1)
Step 0. Pick u_0 arbitrarily such that $\|u_0\| = 1$.
Step 1. For $m = 1, 2, \dots, n$, do
$w_m = Au_{m-1}$
$u_m = w_m/\|w_m\|$.

6.2. Treatment of defective eigenvalues. When the eigenvalue λ_j is defective and has $\omega_j > 1$ in (2.2), then, under the conditions of Theorem 3.1, there are precisely ω_j Ritz values $\lambda_{jl}(n), 1 \le l \le \omega_j$, that tend to λ_j, each with the rate of convergence $O([n^{\bar{p}}|\lambda_{t+1}/\lambda_j|^n]^{1/\omega_j})$ as $n \to \infty$. That is to say, the Ritz values for a defective eigenvalue are not as effective as the ones for nondefective eigenvalues. However, $\hat{\lambda}_j(n)$ and $\tilde{\lambda}_j(n)$ that are defined in Theorem 3.1 do enjoy the property that they tend to λ_j with the optimal rate of convergence $O(n^{\bar{p}}|\lambda_{t+1}/\lambda_j|^n)$ as $n \to \infty$, as in the case of a nondefective eigenvalue.

As for the invariant subspaces $Y_i, i = 0, 1, \dots, p_j, p_j = \omega_j - 1$, the most basic result to use is Theorem 3.2. According to this theorem and the subsequent developments, the building blocks for the invariant subspaces are the vectors $\hat{d}_{ji,l}(n)$ that are defined by (3.19). Now the vector $\hat{d}_{ji,l}(n)$ is a constant multiple of $\operatorname{Res} F_{n,k}(z)|_{z=z_{jl}(n)}$, where $z_{jl}(n) = 1/\lambda_{jl}(n)$, which, when $\nu = -1$, is a constant multiple of the Ritz vector corresponding to $\lambda_{jl}(n)$ by Theorem 5.4. That is, once the Ritz vectors have been computed, they can be used to construct the vectors $\hat{d}_{ji,l}(n)$ which, in turn, are used in constructing the approximate invariant subspaces Y_i with optimal accuracy.

Let us now show how the vector $\hat{d}_{ji,l}(n)$ is expressed in terms of the corresponding Ritz vector. For simplicity of notation we shall write $\hat{z} = z_{jl}(n) = 1/\lambda_{jl_{(n)}}$. The Ritz vector corresponding to $\lambda_{jl}(n)$ is $\hat{x} = \sum_{i=1}^{k} \xi_i v_i$, where $v_1 = u_n$ and $(u_n, u_n) = 1$ by (6.1). We recall that for the method of Arnoldi the vectors $v_1, v_2, \dots, v_k$ are actually the ones that would be obtained by orthogonalizing the power iterations $u_n, Au_n, \dots, A^{k-1}u_n$ by the Gram–Schmidt process. For the method of Lanczos the vectors $v_1, v_2, \dots, v_k$ are obtained by biorthogonalizing $u_n, Au_n, \dots, A^{k-1}u_n$ against the vectors $q, A^*q, \dots, (A^*)^{k-1}q$. In both cases we have

$$AV = VH + R, \tag{6.2}$$

where H is the upper Hessenberg matrix of (4.8) for the Arnoldi method or the tridiagonal matrix of (4.13) for the Lanczos method, and thus it is upper Hessenberg in both cases. The matrix R has all of its first $k-1$ columns equal to zero, and its kth column is $h_{k+1,k}v_{k+1}$.

From the way the vectors $v_1, v_2, \ldots, v_k$ are constructed it is easy to see that

$$V = [u_n | Au_n | \cdots | A^{k-1}u_n]B, \tag{6.3}$$

where B is the upper triangular matrix

$$B = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1k} \\ & \beta_{22} & \cdots & \beta_{2k} \\ & & \ddots & \vdots \\ & & & \beta_{kk} \end{bmatrix}, \tag{6.4}$$

whose entries β_{ij} are required. Substituting (6.3) in (6.2), we have

$$[Au_n | A^2u_n | \cdots | A^ku_n]B = [u_n | Au_n | \cdots | A^{k-1}u_n]BH + R. \tag{6.5}$$

By equating the jth columns of both sides of (6.5) for $j < k$, we obtain

$$\sum_{i=1}^{j}(A^iu_n)\beta_{ij} = \sum_{i=0}^{j}(A^iu_n)(BH)_{i+1,j} \tag{6.6}$$

as the matrices B and BH are upper triangular and upper Hessenberg, respectively. From the linear independence of the vectors $A^iu_n, i = 0, 1, \ldots, k-1$, (6.6) reduces to

$$\beta_{ij} = (BH)_{i+1,j}, \qquad 0 \le i \le j;\ \beta_{0j} \equiv 0 \text{ all } j \ge 1. \tag{6.7}$$

Now $\beta_{11} = 1$ since $v_1 = u_n$. These equations can be solved in the order $i = 0, 1, \ldots, j, j = 1, 2, \ldots, k-1$, which amounts to computing the 1st, 2nd, $\ldots$, kth columns of the matrix B, in this order. This can be accomplished as $h_{j+1,j} > 0$ for all j. Thus by letting $i = 0$ in (6.7), we obtain $\sum_{r=1}^{j+1}\beta_{1r}h_{rj} = 0$, which we solve for $\beta_{1,j+1}$. Next, letting $i = 1$, we obtain $\beta_{1j} = \sum_{r=1}^{j+1}\beta_{2r}h_{rj}$, which we solve for $\beta_{2,j+1}$. By letting $i = 2, 3, \ldots, j$, we obtain $\beta_{i+1,j+1}, i = 2, 3, \ldots, j$, in this order.

Suppose that the Ritz vector $\hat{x}$ has been computed in the form $\hat{x} = \sum_{i=1}^{k}\xi_iv_i$ and that the ξ_i have been saved. Then, recalling also that $u_{n+i} = A^iu_n, i = 0, 1, \ldots, k-1$,

$$\hat{x} = \sum_{i=0}^{k-1}\sigma_iu_{n+i}, \tag{6.8}$$

and the coefficient of u_n is given by

$$\sigma_0 = \sum_{j=1}^{k}\beta_{1j}\xi_j. \tag{6.9}$$

Similarly, from (3.19), the coefficient of u_n in $\hat{d}_{ji,l}(n)$ (setting $\nu = -1$ there) is given by

$$\sigma_0' = (\hat{z} - \zeta_j(n))^i\frac{\sum_{r=1}^{k}c_r^{(n,k)}\hat{z}^{k-r}}{\sum_{r=0}^{k}c_r^{(n,k)}(k-r)\hat{z}^{k-r-1}} = -(\hat{z} - \zeta_j(n))^i\frac{c_0^{(n,k)}\hat{z}^k}{Q_{n,k}'(\hat{z})}. \tag{6.10}$$

Now if we denote the Ritz values by $\lambda'_1, \ldots, \lambda'_k$ and set $z'_i = 1/\lambda'_i, i = 1, \ldots, k$, then we can show that

$$\sigma'_0 = -(\hat{z} - \zeta_j(n))^i \frac{\hat{z}}{\prod_{\substack{r=1 \\ z'_r \neq \hat{z}}}^{k} (1 - z'_r/\hat{z})}, \tag{6.11}$$

so that

$$\hat{d}_{ji,l}(n) = \frac{\sigma'_0}{\sigma_0}\hat{x} = -\frac{(\hat{z} - \zeta_j(n))^i}{\prod_{\substack{r=1 \\ z'_r \neq \hat{z}}}^{k} (1 - z'_r/\hat{z})} \frac{\hat{z}}{\sum_{j=1}^{k} \beta_{1j}\xi_j}\hat{x}, \tag{6.12}$$

which is the desired result.

With this we can now go on to compute the approximations to the eigenvector a_{jp_j} and the vectors $a_{ji}, 0 \leq i < p_j - 1$, precisely as described in §§3.2.1 and 3.2.2, respectively. For example, the vector $\hat{d}_{jp_j}(n) = \sum_{l=1}^{\omega_j} \hat{d}_{jp_j,l}(n)$ is the approximation to the eigenvector a_{jp_j} the error in which is, roughly speaking, $O(|\lambda_{t+1}/\lambda_j|^n)$ as $n \to \infty$.

7. Numerical examples. In this section we demonstrate by numerical examples the validity of some of the theory and claims of the previous sections. The computations for this section were done in double precision arithmetic on an IBM-370 machine.

Example 7.1. Consider the 11×11 real symmetric matrix

$$A = 0.06 \times \begin{bmatrix} 5 & 2 & 1 & 1 & & & & & & & \\ 2 & 6 & 3 & 1 & 1 & & & & & & \\ 1 & 3 & 6 & 3 & 1 & 1 & & & & & \\ 1 & 1 & 3 & 6 & 3 & 1 & 1 & & & & \\ & 1 & 1 & 3 & 6 & 3 & 1 & 1 & & & \\ & & 1 & 1 & 3 & 6 & 3 & 1 & 1 & & \\ & & & 1 & 1 & 3 & 6 & 3 & 1 & 1 & \\ & & & & 1 & 1 & 3 & 6 & 3 & 1 & 1 \\ & & & & & 1 & 1 & 3 & 6 & 3 & 1 \\ & & & & & & 1 & 1 & 3 & 6 & 2 \\ & & & & & & & 1 & 1 & 2 & 5 \end{bmatrix}. \tag{7.1}$$

This matrix has 10 distinct positive eigenvalues, the smallest and largest being 0.0313 $\cdots$ and 0.896 $\cdots$, respectively. We applied the SMPE and SMMPE procedures to approximate its eigenvalues. With $u_0 = (1, 1, \ldots, 1)^T$, only 6 of the 10 eigenvalues appear in the spectral decomposition of u_m for all m. To five-digit accuracy these eigenvalues are $\lambda_1 = 0.89651, \lambda_2 = 0.52971, \lambda_3 = 0.26440, \lambda_4 = 0.24775, \lambda_5 = 0.19029$, and $\lambda_6 = 0.031337$.

In Tables 7.1.1 and 7.1.2 we give the errors $e_j(n) = \lambda_j - \lambda_j(n)$ in the approximations $\lambda_j(n), j = 1, 2, 3$, that were obtained by, respectively, the SMMPE and the SMPE procedures with $k = 3$. Here $\lambda_j(n)$ stands for $\lambda_{j1}(n)$, and we know that $\omega_j = 1$ for all j. Recall that for the SMPE procedure these $\lambda_j(n)$ are simply the Ritz values obtained by the Arnoldi method of order $k = 3$ as this method is being applied to u_n. They are also the Ritz values obtained by the Lanczos method of order $k = 3$ as this method is being applied to u_n with $q = u_n$ in (4.9). (The $\lambda_j(n)$ were actually obtained by solving the polynomial equation $\sum_{i=0}^{k} c_i^{(n,k)} \lambda^i = 0$ with

Table 7.1.1.
Errors in $\lambda_j(n)$ *obtained from* SMMPE *procedure with* $k = 3$ *on the matrix* A *in Example* 7.1. *The vector* u_0 *is* $(1, 1, \ldots, 1)^T$. *Here* $e_j(n) = \lambda_j - \lambda_j(n)$, $j = 1, 2, 3$.

n	$e_1(n)$	$e_2(n)$	$e_3(n)$
0	2.01D−02	1.15D−01	6.54D−02
1	5.02D−03	4.31D−02	3.78D−02
2	1.30D−03	1.82D−02	3.03D−02
3	3.38D−04	7.87D−03	2.61D−02
4	8.57D−05	3.39D−03	2.28D−02
5	2.15D−05	1.45D−03	2.00D−02
6	5.37D−06	6.21D−04	1.77D−02
7	1.35D−06	2.68D−04	1.58D−02
8	3.44D−07	1.17D−04	1.42D−02
9	8.84D−08	5.13D−05	1.29D−02
10	2.31D−08	2.29D−05	1.18D−02
11	6.10D−09	1.03D−05	1.09D−02
12	1.63D−09	4.72D−06	1.02D−02
13	4.42D−10	2.18D−06	9.54D−03
14	1.21D−10	1.01D−06	8.99D−03
15	3.32D−11	4.74D−07	8.52D−03

Table 7.1.2.
Errors in $\lambda_j(n)$ *obtained from* SMPE *procedure with* $k = 3$ *on the matrix* A *in Example* 7.1. *The vector* u_0 *is* $(1, 1, \ldots, 1)^T$. *Here* $e_j(n) = \lambda_j - \lambda_j(n)$, $j = 1, 2, 3$.

n	$e_1(n)$	$e_2(n)$	$e_3(n)$
0	7.01D−05	6.92D−03	2.26D−02
1	1.11D−06	3.64D−04	9.64D−03
2	5.43D−08	5.23D−05	6.84D−03
3	2.91D−09	8.15D−06	5.18D−03
4	1.65D−10	1.34D−06	4.11D−03
5	9.93D−12	2.36D−07	3.40D−03
6	6.41D−13	4.41D−08	2.88D−03
7	4.42D−14	8.73D−09	2.50D−03
8	3.62D−15	1.80D−09	2.18D−03
9	1.11D−15	3.81D−10	1.92D−03
10	6.25D−16	8.21D−11	1.70D−03

the $c_i^{(n,k)}$ determined from (1.6) and $c_k^{(n,k)} = 1$.) Note that the errors are all positive, and, for the SMPE procedure, this is consistent with the asymptotic result of [Si3, Thm. 2.1]. In addition, we have $e_j(n) = O(|\lambda_4/\lambda_j|^n)$ as $n \to \infty$ for SMMPE procedure (cf. (3.4)) and $e_j(n) = O(|\lambda_4/\lambda_j|^{2n})$ as $n \to \infty$ for SMPE procedure (cf. (3.11)). These can be verified numerically by computing $r_j(n) = e_j(n+1)/e_j(n)$, for which, $\lim_{n\to\infty} r_j(n) = \lambda_4/\lambda_j$ for SMMPE procedure and $\lim_{n\to\infty} r_j(n) = (\lambda_4/\lambda_j)^2$ for SMPE procedure. Indeed, $r_j(n)$ do approach their respective limits with increasing n.

The vectors $q_1, \ldots, q_k$ in the SMMPE procedure were taken to be the first k standard basis vectors for this example.

We should note, of course, that as n is increased, roundoff errors cause the vectors $u_n, u_{n+1}, \ldots$, to have contributions from *all* eigenvalues of A. With the precision we are using, at $n = 15$ the roundoff errors are still not sufficiently effective to cause this to happen in appreciable amounts.

Finally, if the above is repeated with $k = 4$, a significant improvement in the convergence rates of the $\lambda_j(n)$ is observed, as predicted by the theory of §3. This point has been explained in the third paragraph following the statement of Theorem 3.1.

Example 7.2. Consider the $m^2 \times m^2$ block tridiagonal matrix

$$A = \begin{bmatrix} B & -I & & \\ -I & B & -I & \\ & \ddots & \ddots & \ddots \\ & & & -I \\ & & -I & B \end{bmatrix}, \tag{7.2}$$

where I is the $m \times m$ identity matrix, B is the $m \times m$ real nonsymmetric tridiagonal matrix given by

$$B = \begin{bmatrix} 4 & a & & & \\ b & 4 & a & & \\ & \ddots & \ddots & \ddots & \\ & & b & 4 & a \\ & & & b & 4 \end{bmatrix}, \quad a = -1 + \frac{\gamma}{2(m+1)}, \quad b = -1 - \frac{\gamma}{2(m+1)}. \tag{7.3}$$

This matrix, with $\gamma = 1$, appears in [Sa1, Example 4.2.2], where it is treated with the help of the Arnoldi method when $m = 15$. It arises from central difference discretization of the elliptic operator $-(\partial^2/\partial x^2 + \partial^2/\partial y^2) + \gamma(\partial/\partial x)$ on the unit square with Dirichlet boundary conditions, the number of points of discretization interior to the unit square being m in each direction.

It can be shown that A is diagonalizable and that its eigenvalues are given by

$$\Lambda_{p,q}(\gamma) = 4 - 2\cos\frac{q\pi}{m+1} - 2\sqrt{1 - \left[\frac{\gamma}{2(m+1)}\right]^2}\cos\frac{p\pi}{m+1}, \quad p, q = 1, 2, \ldots, m. \tag{7.4}$$

To be able to compare our numerical results with those of [Sa1], we also applied the Arnoldi method to the matrix A with $\gamma = 1$ and $m = 15$. In this case all eigenvalues of A are real and positive. In Table 7.2.1 we give the errors in the Ritz values λ_1' and λ_2' that are approximations to the first two largest eigenvalues of A, namely, $\Lambda_{m,m}(1)$ and $\Lambda_{m-1,m}(1)$, for $k = 1, 2, \ldots$, obtained by applying the Arnoldi method of order k to a randomly generated vector u_0, as is commonly done. We also give the l_2-norms of the residuals $Ax_j' - \lambda_j' x_j', j = 1, 2$, where x_j' is the Ritz vector corresponding to λ_j', and $(x_j', x_j') = 1$. These norms are computed precisely in the way described in [Sa1, Eq. (3.14)]. In Tables 7.2.2 and 7.2.3 we do the same, except that we now apply the Arnoldi method to $u_n = A^n u_0$, with $n = 100$ and $n = 200$, respectively, u_0 being again a randomly generated vector.

Comparison of the results in Tables 7.2.1–7.2.3 shows first of all that the largest Ritz values converge much faster in k for $n = 100$ and $n = 200$ than for $n = 0$. Also the

TABLE 7.2.1.

Errors in the two largest Ritz values and l_2-norms of the residuals of corresponding Ritz vectors obtained from the Arnoldi method on the matrix A of Example 7.2 with $\gamma = 1$ and $m = 15$. The method is applied to the randomly generated vector u_0. Here $e_j^{(k)} = |\lambda_j - \lambda'_j|$ and $w_j^{(k)} = \|Ax'_j - \lambda'_j x'_j\|$, (λ'_j, x'_j) being pairs of Ritz values and Ritz vectors obtained from the Arnoldi method of order k, and $\|x'_j\| = 1$.

k	$e_1^{(k)}$	$w_1^{(k)}$	$e_2^{(k)}$	$w_2^{(k)}$
1	6.70D+00	1.89D+00		
2	2.54D+00	1.72D+00	7.44D+00	7.78D−01
3	1.19D+00	1.16D+00	4.55D+00	1.44D+00
4	6.56D−01	7.62D−01	3.03D+00	1.24D+00
5	3.88D−01	5.77D−01	2.02D+00	1.06D+00
6	2.44D−01	4.41D−01	1.40D+00	8.75D−01
7	1.62D−01	3.30D−01	1.03D+00	6.94D−01
8	1.16D−01	2.61D−01	7.82D−01	6.08D−01
9	8.34D−02	2.28D−01	5.78D−01	5.63D−01
10	5.85D−02	2.08D−01	4.13D−01	5.06D−01
11	4.10D−02	1.77D−01	3.04D−01	4.16D−01
12	2.74D−02	1.57D−01	2.24D−01	3.60D−01
13	1.86D−02	1.23D−01	1.74D−01	2.83D−01
14	1.31D−02	9.71D−02	1.40D−01	2.31D−01
15	9.18D−03	8.89D−02	1.12D−01	2.26D−01
16	6.71D−03	7.18D−02	8.98D−02	1.97D−01
17	4.37D−03	7.78D−02	6.34D−02	2.32D−01
18	2.42D−03	6.91D−02	3.77D−02	2.12D−01
19	1.22D−03	5.30D−02	2.08D−02	1.65D−01
20	5.72D−04	3.88D−02	1.11D−02	1.24D−01
21	2.28D−04	2.92D−02	5.32D−03	9.84D−02
22	1.00D−04	1.87D−02	2.74D−03	6.73D−02
23	5.02D−05	1.27D−02	1.42D−03	4.94D−02
24	2.86D−05	7.78D−03	7.01D−04	3.25D−02
25	1.85D−05	5.48D−03	3.02D−04	2.48D−02
26	1.16D−05	4.83D−03	1.53D−06	2.43D−02
27	7.95D−06	3.12D−03	1.63D−04	1.73D−02
28	6.46D−06	1.58D−03	2.15D−04	9.42D−03
29	5.60D−06	9.05D−04	2.19D−04	5.73D−03
30	4.80D−06	5.57D−04	2.06D−04	3.75D−03

cost, both storagewise and computational, of obtaining a high level accuracy is larger when $n = 0$ than when $n > 0$ and is sufficiently large. For instance, the accuracy attained for λ'_1 with $n = 0$ and $k = 30$ can be attained with $n = 100$ and $k = 5$. In the former we must store 30 vectors, whereas in the latter we need to store 5 vectors. Roughly speaking, the computational effort in the former case is the equivalent of about 232 matrix-vector products, whereas in the latter this number is 144.

We determine computational cost in the following way. First of all, if we are interested only in the eigenvalues, then the computational cost is the sum of (i) the n matrix-vector products to get to u_n along with the n normalizations for $u_0, u_1, \ldots, u_{n-1}$, cf. (6.1), and (ii) the cost of forming the matrix V_{k-1}, cf. (4.6). The cost of (i) is n matrix-vector products, n scalar products, and n scalar-vector multiplications. The cost of (ii) is $k-1$ matrix-vector products, $\frac{1}{2}k(k+1)$ scalar products, $\frac{1}{2}k(k+1)$

TABLE 7.2.2.

Errors in the two largest Ritz values and l_2-norms of the residuals of corresponding Ritz vectors obtained from the Arnoldi method on the matrix A of Example 7.2 with $\gamma = 1$ and $m = 15$. The method is applied to the vector $u_n = A^n u_0$ with $n = 100$, where u_0 is a randomly generated vector. Here $e_j^{(k)} = |\lambda_j - \lambda_j'|$ and $w_j^{(k)} = \|Ax_j' - \lambda_j' x_j'\|, (\lambda_j', x_j')$ being pairs of Ritz values and Ritz vectors obtained from the Arnoldi method of order k, and $\|x_j'\| = 1$.

k	$e_1^{(k)}$	$w_1^{(k)}$	$e_2^{(k)}$	$w_2^{(k)}$
1	1.46D−02	4.98D−02		
2	1.58D−03	1.97D−02	1.65D−02	7.12D−02
3	8.74D−06	5.19D−03	1.56D−05	2.59D−02
4	2.14D−06	6.15D−04	1.33D−04	4.26D−03
5	2.79D−06	9.87D−05	1.61D−04	8.65D−04
6	2.33D−06	3.64D−05	1.17D−04	4.66D−04
7	4.30D−07	1.62D−05	2.45D−05	3.12D−04
8	1.29D−06	3.19D−06	1.86D−05	7.40D−05
9	8.28D−06	1.02D−06	6.39D−05	2.78D−05
10	1.60D−06	7.35D−08	2.19D−05	2.06D−06
11	9.04D−09	1.64D−07	7.99D−06	5.23D−06
12	2.09D−07	5.17D−08	1.02D−05	1.85D−06
13	2.56D−09	5.06D−08	7.76D−06	2.48D−06
14	1.16D−07	5.36D−08	1.04D−05	3.96D−06
15	1.88D−08	4.49D−08	5.33D−06	1.22D−05
16	2.14D−08	1.28D−08	9.42D−06	5.25D−06
17	1.10D−08	2.01D−08	9.72D−08	3.00D−05
18	3.90D−09	5.96D−09	7.69D−06	2.06D−05
19	3.93D−09	6.15D−09	1.42D−05	6.70D−05
20	8.85D−10	2.74D−09	1.09D−05	2.05D−04

scalar-vector multiplications, and $\frac{1}{2}k(k-1)$ vector additions. If we agree to consider a scalar product as consisting of a scalar-vector multiplication and a vector addition, the total number of operations will be $n + k - 1$ matrix-vector products, $2n + k^2 + k$ scalar-vector multiplications, and $n + k^2$ vector additions. Finally, let us make the simplification that addition and multiplication have the same cost. All this, of course, is not most accurate, but gives a reasonable account of the cost. In our example, one matrix-vector product is very nearly equivalent to five scalar-vector multiplications and four vector additions.

The approximation λ_1' that corresponds to $n = 100$ and $k = 20$ in Table 7.2.2 has about the same accuracy as that given in [Sa1]. But the way the approximation of [Sa1] is obtained is much more complicated and also more expensive computationally.

Now with $\gamma = 1$, the matrix A is close to being symmetric, and one may attribute the good results shown in Tables 7.2.2 and 7.2.3 to this fact. We, therefore, applied the Arnoldi method with larger values of γ that cause A to become highly nonsymmetric. Our results and conclusions were invariably the same. Actually, when the Arnoldi method was applied with large values of γ, e.g., $\gamma = 10$, the quality of the Ritz values with $n = 0$ deteriorated, whereas the quality of those with $n = 100$ remained almost the same.

Finally, we have also applied the Arnoldi method to $M = I - \frac{1}{4}A$ with $\gamma = 0$.

TABLE 7.2.3.

Errors in the two largest Ritz values and l_2-norms of the residuals of corresponding Ritz vectors obtained from the Arnoldi method on the matrix A of Example 7.2 with $\gamma = 1$ and $m = 15$. The method is applied to the vector $u_n = A^n u_0$ with $n = 200$, where u_0 is a randomly generated vector. Here $e_j^{(k)} = |\lambda_j - \lambda'_j|$ and $w_j^{(k)} = \|Ax'_j - \lambda'_j x'_j\|, (\lambda'_j, x'_j)$ being pairs of Ritz values and Ritz vectors obtained from the Arnoldi method of order k, and $\|x'_j\| = 1$.

k	$e_1^{(k)}$	$w_1^{(k)}$	$e_2^{(k)}$	$w_2^{(k)}$
1	5.61D−02	5.76D−02		
2	8.43D−05	4.08D−02	7.03D−05	8.33D−03
3	3.57D−06	1.61D−03	7.28D−05	5.34D−04
4	6.00D−07	1.15D−04	2.49D−05	6.79D−05
5	2.10D−07	2.48D−05	5.51D−05	2.24D−05
6	1.56D−07	1.01D−06	5.53D−05	1.14D−06
7	5.25D−07	2.53D−07	4.72D−05	7.12D−07
8	5.23D−07	1.56D−08	6.31D−05	5.12D−08
9	1.03D−08	8.23D−08	5.09D−05	6.30D−06
10	1.03D−08	4.88D−09	6.12D−05	6.02D−07
11	2.86D−09	8.36D−09	2.13D−05	2.02D−05
12	3.58D−10	1.08D−09	2.96D−05	1.28D−05
13	2.56D−10	3.96D−10	3.47D−06	1.31D−05
14	1.96D−10	6.09D−10	5.71D−06	2.25D−05
15	4.51D−11	1.94D−10	1.98D−06	1.01D−05
16	1.94D−11	5.69D−11	7.79D−07	3.21D−06
17	1.99D−11	6.86D−11	1.65D−06	4.12D−06
18	1.01D−11	5.74D−11	6.09D−07	4.05D−06
19	5.32D−12	2.48D−11	8.41D−07	1.87D−06
20	4.67D−12	2.84D−11	2.44D−07	2.44D−06

This matrix is real symmetric and its spectrum is in $(-1, 1)$ and is symmetric with respect to the origin. Again the results obtained from the Arnoldi (now equivalent to symmetric Lanczos) method with $n > 0$ and large were superior to those obtained with $n = 0$.

REFERENCES

[A] W. E. ARNOLDI, *The principle of minimized iterations in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

[E] I. ERDELYI, *An iterative least-square algorithm suitable for computing partial eigensystems*, SIAM J. Numer. Anal., 2 (1965), pp. 421–436.

[FSi] W. F. FORD AND A. SIDI, *Recursive algorithms for vector extrapolation methods*, Appl. Numer. Math., 4 (1988), pp. 477–489.

[GV] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, Second Edition, Johns Hopkins University Press, Baltimore, 1989.

[GW] G. H. GOLUB AND J. H. WILKINSON, *Ill-conditioned eigensystems and the computation of the Jordan canonical form*, SIAM Rev., 18 (1976), pp. 578–619.

[H] A. S. HOUSEHOLDER, *The Theory of Matrices in Numerical Analysis*, Blaisdell, New York, 1964.

[K] S. KANIEL, *Estimates for some computational techniques in linear algebra*, Math. Comp., 20 (1966), pp. 369–378.

[L] C. LANCZOS, *An iteration method for the solution of the eigenvalue problem of linear differential*

and integral operators, J. Res. Nat. Bur. Stand., 45 (1950), pp. 255–282.

[M] T. A. MANTEUFFEL, *Adaptive procedure for estimating parameters for the nonsymmetric Tchebychev iteration*, Numer. Math., 31 (1978), pp. 183–203.

[O] A. M. OSTROWSKI, *On the convergence of the Rayleigh quotient iteration for the computation of characteristic roots and vectors*, IV *and* VI, Arch. Rat. Mech. Anal., 3 (1959), pp. 341–347 and 4 (1959/60), pp. 153–165.

[Pai] C. C. PAIGE, *The Computation of Eigenvalues and Eigenvectors of Very Large Sparse Matrices*, Ph.D. thesis, University of London, 1971.

[Par1] B. N. PARLETT, *Global convergence of the basic QR algorithm on Hessenberg matrices*, Math. Comp., 22 (1968), pp. 803–817.

[Par2] ———, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[ParPo] B. N. PARLETT AND W. G. POOLE, *A geometric theory for the QR, LU and power iterations*, SIAM J. Numer. Anal., 10 (1973), pp. 389–412.

[Sa1] Y. SAAD, *Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices*, Linear Algebra Appl., 34 (1980), pp. 269–295.

[Sa2] ———, *On the rates of convergence of the Lanczos and the block Lanczos methods*, SIAM J. Numer. Anal., 17 (1980), pp. 687–706.

[Si1] A. SIDI, *Convergence and stability properties of minimal polynomial and reduced rank extrapolation algorithms*, SIAM J. Numer. Anal., 23 (1986), pp. 197–209. Originally appeared as NASA TM-83443 (1983).

[Si2] ———, *Extrapolation vs. projection methods for linear systems of equations*, J. Comput. Appl. Math., 22 (1988), pp. 71–88.

[Si3] ———, *On extensions of the power method for normal operators*, Linear Algebra Appl., 120 (1989), pp. 207–224.

[Si4] ———, *Quantitative and constructive aspects of the generalized Koenig's and de Montessus's theorems for Padé approximants*, J. Comput. Appl. Math., 29 (1990), pp. 257–291.

[Si5] ———, *Efficient implementation of minimal polynomial and reduced rank extrapolation methods*, J. Comput. Appl. Math., 36 (1991), pp. 305–337.

[Si6] ———, *Rational approximations from power series of vector-valued meromorphic functions*, J. Approx. Theory, 76 (1994), pp. 89–111.

[SiB] A. SIDI AND J. BRIDGER, *Convergence and stability analyses for some vector extrapolation methods in the presence of defective iteration matrices*, J. Comput. Appl. Math., 22 (1988), pp. 35–61.

[SiFSm] A. SIDI, W. F. FORD, AND D. A. SMITH, *Acceleration of convergence of vector sequences*, SIAM J. Numer. Anal., 23 (1986), pp. 178–196. Originally appeared as NASA TP-2193 (1983).

[SmFSi] D. A. SMITH, W. F. FORD, AND A. SIDI, *Extrapolation methods for vector sequences*, SIAM Rev., 29 (1987), pp. 199–233. Erratum: Correction to *Extrapolation methods for vector sequences*, SIAM Rev., 30 (1988), pp. 623–624.

[W] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.