SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 1, pp. 1–34, January 1996

THE SET OF 2-BY-3 MATRIX PENCILS — KRONECKER STRUCTURES AND THEIR TRANSITIONS UNDER PERTURBATIONS*

ERIK ELMROTH† AND BO KÅGSTRÖM†

Abstract. The set (or family) of 2-by-3 matrix pencils $A - \lambda B$ comprises 18 structurally different Kronecker structures (canonical forms). The algebraic and geometric characteristics of the generic and the 17 nongeneric cases are examined in full detail. The complete closure hierarchy of the orbits of all different Kronecker structures is derived and presented in a closure graph that shows how the structures relate to each other in the 12-dimensional space spanned by the set of 2-by-3 pencils. Necessary conditions on perturbations for transiting from the orbit of one Kronecker structure to another in the closure hierarchy are presented in a labeled closure graph. The node and arc labels shows geometric characteristics of an orbit's Kronecker structure and the change of geometric characteristics when transiting to an adjacent node, respectively. Computable normwise bounds for the smallest perturbations $(\delta A, \delta B)$ of a generic 2-by-3 pencil $A-\lambda B$ such that $(A+\delta A)-\lambda(B+\delta B)$ has a specific nongeneric Kronecker structure are presented. First, explicit expressions for the perturbations that transfer $A - \lambda B$ to a specified nongeneric form are derived. In this context tractable and intractable perturbations are defined. Second, a modified GUPTRI that computes a specified Kronecker structure of a generic pencil is used. Perturbations devised to impose a certain nongeneric structure are computed in a way that guarantees one will find a Kronecker canonical form (KCF) on the closure of the orbit of the intended KCF. Both approaches are illustrated by computational experiments. Moreover, a study of the behaviour of the nongeneric structures under random perturbations in finite precision arithmetic (using the GUPTRI software) show for which sizes of perturbations the structures are invariant and also that structure transitions occur in accordance with the closure hierarchy. Finally, some of the results are extended to the general m-by-$(m+1)$ case.

Key words. matrix pencils (2-by-3), Kronecker canonical form, generalized Schur decomposition, orbit, codimension, Kronecker structure hierarchy, closest nongeneric structure, controllability

AMS subject classifications. 65F15, 15A21, 15A22

1. Introduction.

Singular matrix pencils $A - \lambda B$, where A and B are m-by-n matrices with real or complex entries, appear in several applications. Examples include problems in control theory relating to a linear system $E\dot{x}(t) = Fx(t) + Gu(t)$, where E and F are p-by-p matrices, and G is p-by-k. Solvability issues of a singular system (i.e., $\det(E) = 0$), such as the existence of a solution, consistent initial values, and its explicit solution can be revealed from the Kronecker structure of $A - \lambda B \equiv F - \lambda E$ (e.g., see [9], [20]). The problems of finding the controllable subspace, uncontrollable modes or an upper bound on the distance to uncontrollability for a controllable system $E\dot{x}(t) = Fx(t) + Gu(t)$ can all be formulated and solved in terms of certain reducing subspaces of the matrix pencil $A - \lambda B \equiv [G \quad F] - \lambda[0 \quad E]$ (e.g., see [15], [17], [18], [6]).

In most applications it is enough to transfer $A - \lambda B$ to a *generalized Schur form* (e.g., to GUPTRI form [7], [8])

$$P^H(A - \lambda B)Q = \begin{bmatrix} A_r - \lambda B_r & * & * \\ 0 & A_{\mathrm{reg}} - \lambda B_{\mathrm{reg}} & * \\ 0 & 0 & A_l - \lambda B_l \end{bmatrix}, \tag{1.1}$$

* Received by the editors January 3, 1994; accepted for publication (in revised form) by P. van Dooren February 11, 1995.

† Department of Computing Science, Umeå University, S-901 87 Umeå, Sweden (elmroth@cs.umu.se and bokg@cs.umu.se).

where P (m-by-m) and Q (n-by-n) are unitary and $*$ denotes arbitrary conforming submatrices. Here the square upper triangular block $A_{\mathrm{reg}} - \lambda B_{\mathrm{reg}}$ is regular and has the same regular structure as $A - \lambda B$ (i.e., contains all generalized eigenvalues (finite and infinite) of $A - \lambda B$). The rectangular blocks $A_r - \lambda B_r$ and $A_l - \lambda B_l$ contain the singular structure (right and left minimal indices) of the pencil and are block upper triangular. The *singular blocks of right* (column) and *left* (row) *indices of grade* j are

$$L_j \equiv \begin{bmatrix} -\lambda & 1 & & \\ & \cdot & \cdot & \\ & & -\lambda & 1 \end{bmatrix} \quad \text{and} \quad L_j^T \equiv \begin{bmatrix} -\lambda & & \\ 1 & \cdot & \\ & \cdot & -\lambda \\ & & 1 \end{bmatrix}, \tag{1.2}$$

of size j-by-$(j+1)$ and $(j+1)$-by-j, respectively. $A_r - \lambda B_r$ has only right minimal indices in its Kronecker canonical form (KCF), indeed the same L_j blocks as $A - \lambda B$. Similarly, $A_l - \lambda B_l$ has only left minimal indices in its KCF, the same L_j^T blocks as $A - \lambda B$. If $A - \lambda B$ is singular at least one of $A_r - \lambda B_r$ and $A_l - \lambda B_l$ will be present in (1.1). The explicit structure of the diagonal blocks in staircase form can be found in [8]. If $A - \lambda B$ is regular, $A_r - \lambda B_r$ and $A_l - \lambda B_l$ are not present in (1.1) and the GUPTRI form reduces to the upper triangular block $A_{\mathrm{reg}} - \lambda B_{\mathrm{reg}}$. Staircase forms that reveal the Jordan structure of the zero and infinite eigenvalues are contained in $A_{\mathrm{reg}} - \lambda B_{\mathrm{reg}}$.

Given $A - \lambda B$ in GUPTRI form we also know different pairs of reducing subspaces [18], [7]. Suppose the eigenvalues on the diagonal of $A_{\mathrm{reg}} - \lambda B_{\mathrm{reg}}$ are ordered so that the first k, say, are in Λ_1 (a subset of the spectrum of $A_{\mathrm{reg}} - \lambda B_{\mathrm{reg}}$) and the remainder are outside Λ_1. Let $A_r - \lambda B_r$ be m_r-by-n_r. Then the left and right reducing subspaces associated with Λ_1 are spanned by the leading $m_r + k$ columns of P and the leading $n_r + k$ columns of Q, respectively. When Λ_1 is empty, the corresponding reducing subspaces are called *minimal*, and when Λ_1 contains the whole spectrum the reducing subspaces are called *maximal*.

If $A - \lambda B$ is m-by-n, where $m \neq n$, then for almost all A and B it will have the same KCF, depending only on m and n (the *generic case*). The generic Kronecker structure for $A - \lambda B$ with $d = n - m > 0$ is

$$\mathrm{diag}(L_\alpha, \ldots, L_\alpha, L_{\alpha+1}, \ldots, L_{\alpha+1}), \tag{1.3}$$

where $\alpha = \lfloor m/d \rfloor$, the total number of blocks is d, and the number of $L_{\alpha+1}$ blocks is $m \bmod d$ (which is 0 when d divides m) [16], [3]. The same statement holds for $d = m - n > 0$ if we replace $L_\alpha, L_{\alpha+1}$ in (1.3) by $L_\alpha^T, L_{\alpha+1}^T$. Square pencils are generically regular, i.e., $\det(A - \lambda B) = 0$ if and only if λ is an eigenvalue. The generic singular pencils of size n-by-n have the Kronecker structures [19]:

$$\mathrm{diag}(L_j, L_{n-j-1}^T), \quad j = 0, \ldots, n-1. \tag{1.4}$$

In summary, generic rectangular pencils have only trivial reducing subspaces and no generalized eigenvalues at all. Generic square singular pencils have the same minimal and maximal reducing subspaces. Only if $A - \lambda B$ satisfies a special condition (lies in a particular manifold) does it have nontrivial reducing subspaces and generalized eigenvalues (the *nongeneric case*). Moreover, only if it is perturbed so as to move continuously within that manifold do its reducing subspaces and generalized eigenvalues also move continuously and satisfy interesting error bounds [5], [7]. These requirements are natural in many control and systems theoretic problems such as computing controllable subspaces and uncontrollable modes.

Several authors have proposed (staircase-type) algorithms for computing a generalized Schur form (e.g., see [1], [4], [11]–[14], [16], [20]). They are numerically stable in the sense that they compute the exact Kronecker structure (generalized Schur form or something similar) of a nearby pencil $A' - \lambda B'$. Let $\|\cdot\|_E$ denote the Euclidean (Frobenius) matrix norm. Then $\delta \equiv \|(A - A', B - B')\|_E$ is an upper bound on the distance to the closest $(A + \delta A, B + \delta B)$ with the KCF of (A', B'). Recently, articles about robust software with error bounds for computing the GUPTRI form of a singular $A - \lambda B$ have been published [7], [8]. Some computational experiments that use this software will be discussed later.

The existing algorithms do not guarantee that the computed generalized Schur form is the "most" nongeneric Kronecker structure within distance δ. However, if δ is of size $O(\|(A, B)\|_E \epsilon)$, where ϵ is the relative machine precision, we know that (A, B) is close to a matrix with the Kronecker structure that the algorithm reports. It would be desirable to have algorithms that could solve the following "nearness" problems:

- Compute the closest nongeneric pencil of a generic $A - \lambda B$.
- Compute the closest matrix pencil with a specified Kronecker structure.
- Compute the most nongeneric pencil within a given distance δ.

If the closest structure is not unique we are mainly interested in the most nongeneric KCF. From the perturbation theory for singular pencils [5] we know that all these problems are ill-posed in the sense that the generalized eigenvalues and reducing subspaces for a nongeneric $A - \lambda B$ can change discontinuously as a function of A and B. Therefore, to be able to solve these problems we need to regularize them by restricting the allowable perturbations as mentioned above. In this contribution we make a comprehensive study of the set of 2-by-3 pencils in order to get a greater understanding of (i) these "nearness" problems and how to solve them, and (ii) existing algorithms/software for computing the Kronecker structure of a singular pencil. The full implications of this "case study" to general m-by-n pencils are topics for further research.

In the following we give a summary of our contribution and the organization of the rest of the paper. Section 2 is devoted to algebraic and geometric characteristics of the set of 2-by-3 pencils. In §2.1 we disclose the structurally different Kronecker structures and show how all the nongeneric structures can be generated by a staircase-type algorithm, starting from the generic canonical form. Some algebraic and geometric characteristics of the 18 different Kronecker structures are summarized in three tables. Section 2.2 introduces the concepts of orbits of matrix pencils and their (co)dimensions. The codimensions of the orbits of the 2-by-3 matrix pencils, which depend only on their Kronecker structures [3], are displayed in Table 2.3. They vary between zero (the generic case) and 12 (= $2mn$) for the zero pencil (the most nongeneric case). Indeed, all 2-by-3 pencils "live" in a 12-dimensional space spanned by the set of all generic pencils. In §2.3 we derive a graph describing the closure hierarchy of the orbits of all 18 different Kronecker structures for the set of 2-by-3 pencils. The closure graph is presented in Fig. 2.1. By labeling the nodes in the closure graph with their geometric characteristics and the arcs with the change in geometric characteristics for transiting to an adjacent node, we get a labeled graph showing necessary conditions on perturbations for transiting from one Kronecker structure to another. The labeled closure graph is presented in Fig. 2.2 in §2.4.

Section 3 is devoted to an experimental study of how the nongeneric Kronecker structures behave under random perturbations in finite precision arithmetic, using the GUPTRI software [7], [8]. Assuming a fixed relative accuracy of the input data,

structure invariances and transitions of each nongeneric case are studied as a function of the size of the perturbations added. The results summarized in Table 3.1 are discussed in terms of tolerance parameters used in GUPTRI for determining the Kronecker structure. For large enough perturbations all nongeneric pencils turn generic (as expected). Some nongeneric cases transit between several nongeneric structures before turning generic. These transitions always go from higher to lower codimensions, along the arcs in the closure graph.

In §4 we present computable normwise bounds for the smallest perturbations $(\delta A, \delta B)$ of a generic 2-by-3 pencil $A - \lambda B$ such that $(A + \delta A) - \lambda(B + \delta B)$ has a specific nongeneric Kronecker structure. Two approaches to impose a nongeneric structure are considered. First, explicit expressions for the perturbations that transfer $A - \lambda B$ to a specified nongeneric form are derived in §4.1. In this context tractable and intractable perturbations are defined. We compute a perturbation $(\delta A, \delta B)$ such that $(A+\delta A)-\lambda(B+\delta B)$ is guaranteed to be in the closure of the manifold (orbit) of a certain KCF. If the KCF found is the intended KCF, then the perturbation is said to be tractable. If the KCF found is even more nongeneric then the perturbation is intractable. An intractable perturbation finds any other structure within the closure of the manifold, i.e., a structure that can be found by traveling along the arcs from the intended KCF in the closure graph. A summary of these perturbations is presented in a perturbation graph (Fig. 4.1), where the path to each KCF's node shows the tractable perturbation required to find that KCF starting from the generic KCF (an L_2 block). After illustrating intractable perturbations we derive some results regarding the closest nongeneric Kronecker structure of a generic 2-by-3 (and 1-by-2) pencil. In the second approach, we use a modified GUPTRI for computing a specified Kronecker structure of a generic pencil (§4.2). Computational experiments on random 2-by-3 pencils for the two approaches are presented in §4.3. It is the intractable perturbations, which impose the most nongeneric structure (with highest codimension) for a given size of the perturbations (e.g., the relative accuracy of the data), that are requested in applications (e.g., computing the uncontrollable subspace). Finally, in §5 we comment on the general case and extend our results for the closest nongeneric pencil to a generic m-by-$(m+1)$ pencil.

2. Algebraic and geometric characteristics of the set of 2-by-3 matrix pencils. In this section we disclose the structurally different Kronecker structures and show how all the nongeneric structures can be generated by a staircase-type algorithm, starting from the generic canonical form. Moreover, we discuss the codimensions of associated orbits and derive a closure graph, showing the Kronecker structure hierarchy of the set of 2-by-3 pencils.

2.1. Structurally different Kronecker structures. The *generic* case corresponds to A and B of size 2-by-3 both having full row rank and nonintersecting column nullspaces. This implies that $A - \lambda B$ is strictly equivalent to an L_2 block:

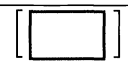
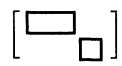
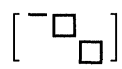
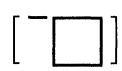
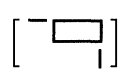
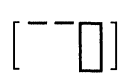
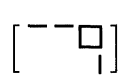
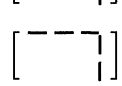
$$P^{-1}(A - \lambda B)Q = L_2 \equiv \begin{bmatrix} -\lambda & 1 & 0 \\ 0 & -\lambda & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{2.1}$$

By inspection, we see that the A- and B-parts of L_2 have row rank 2 and nonintersecting 1-dimensional column nullspaces. The generic canonical form L_2 can be obtained by deleting the last row of $J_3(0) - \lambda I_3$, a 3-by-3 Jordan block corresponding to the zero eigenvalue. $J_3(0)$ is the generic canonical form of a 3-by-3 matrix with zero as a triple eigenvalue and the associated nongeneric Jordan structures are $J_2(0) \oplus J_1(0)$

and $J_1(0) \oplus J_1(0) \oplus J_1(0)$ (i.e., a 3-by-3 zero matrix). Notice that a generic 3-by-3 matrix is diagonalizable with unspecified nonzero eigenvalues (i.e., all Jordan blocks of size 1-by-1).

In the following we disclose the structurally different nongeneric singular cases of size 2×3. By structurally different we mean that all cases have different Kronecker structures (canonical forms). There exist 17 different nongeneric singular cases. The simplest way to construct all nongeneric canonical forms of size 2×3 is to generate all possible combinations of L_1, L_0, J_2, J_1, R_1, N_1, N_2, L_0^T, and L_1^T blocks as in Table 2.1. Algorithms for computing the Kronecker structure of a singular pencil reveal the right (or left) singular structure and the Jordan structure of the zero (or infinite) eigenvalue simultaneously. Therefore, we only distinguish the zero and infinite Jordan structures and put a nonzero and finite eigenvalue in R_1, a regular 1-by-1 block with an unspecified eigenvalue. We will use R_2 to denote a 2-by-2 block with nonzero finite eigenvalues, i.e., R_2 is used to denote any of the three structures $J_1(\alpha) \oplus J_1(\beta)$, $J_1(\alpha) \oplus J_1(\alpha)$, and $J_2(\alpha)$, where α, $\beta \neq \{0, \infty\}$. Notice that if $R_2 = J_2(\alpha)$ then $A - \alpha B$ and B has $J_2(0)$ in its KCF. It is only for the case $L_0 \oplus R_2$ that we can have a $J_2(\alpha)$ block. If we treat these three cases separately we get 19 nongeneric cases, but for our purposes it is sufficient to define R_2 as above.

TABLE 2.1
2×3 pencils built from different Kronecker and Jordan blocks.

Number of cases	Block structure	KCF
1		L_2
3		$L_1 \oplus \{J_1, R_1, N_1\}$
5		$L_0 \oplus \{J_1, R_1, N_1\} \oplus \{J_1, N_1\}$
3		$L_0 \oplus \{J_2, R_2, N_2\}$
1		$L_0 \oplus L_1 \oplus L_0^T$
1		$2L_0 \oplus L_1^T$
3		$2L_0 \oplus \{J_1, R_1, N_1\} \oplus L_0^T$
1		$3L_0 \oplus 2L_0^T$

In order to get more insight into the nongeneric structures we would like to show how all the nongeneric structures can be generated by a staircase-type algorithm. By dropping the row rank of the A-part and/or B-part of L_2 (2.1) and imposing different sizes of their "common column or row nullspace(s)" (see Table 2.3) we are able to generate all 17 nongeneric cases starting from the generic canonical form (in the following denoted $A - \lambda B$). Algorithmically, we keep the rank of, for example, B constant and vary the row rank of A while imposing possible sizes of their "common

nullspace(s)." A decrease of the row rank is done by deleting a nonzero element (= 1) in the first or second row of A and/or B and the dimension of the common column nullspace is imposed by permutations of the nonzero elements. After decreasing the row rank of B by one we repeat the procedure until the row rank of B equals zero. By doing so we can generate 12 structurally different nongeneric pencils of size 2×3. These correspond to cases 2–13 in Table 2.2, where we display a case number i, the matrix pair (A_i, B_i), $\mathrm{r}(A_i), \mathrm{r}(B_i)$, the row ranks of A_i and B_i, respectively, $\mathrm{n}(A_i, B_i)$, the dimension of the common column nullspace of A_i and B_i. Finally, in the last column we display the generalized Schur forms (GUPTRI forms) which correspond to the Kronecker block structures displayed in Table 2.1.

TABLE 2.2

Summary of the 18 structurally different 2×3 pencils, numbered and presented in the order in which they are derived in §2.

i	A_i	B_i	$\mathrm{r}(A_i)$	$\mathrm{r}(B_i)$	$\mathrm{n}(A_i, B_i)$	GUPTRI form
1	$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$	2	2	0	$\begin{bmatrix} -\lambda & 1 & 0 \\ 0 & -\lambda & 1 \end{bmatrix}$
2	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$	1	2	0	$\begin{bmatrix} 1 & -\lambda & 0 \\ 0 & 0 & -\lambda \end{bmatrix}$
3	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$	0	2	1	$\begin{bmatrix} 0 & -\lambda & 0 \\ 0 & 0 & -\lambda \end{bmatrix}$
4	$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	1	2	1	$\begin{bmatrix} 0 & -\lambda & 1 \\ 0 & 0 & -\lambda \end{bmatrix}$
5	$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	2	2	1	$\begin{bmatrix} 0 & 1-\lambda & 0 \\ 0 & 0 & 1-\lambda \end{bmatrix}$
6	$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	2	1	0	$\begin{bmatrix} -\lambda & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
7	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	1	1	1	$\begin{bmatrix} 0 & -\lambda & 0 \\ 0 & 0 & 1 \end{bmatrix}$
8	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$	0	1	2	$\begin{bmatrix} 0 & 0 & -\lambda \\ 0 & 0 & 0 \end{bmatrix}$
9	$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$	1	1	2	$\begin{bmatrix} 0 & 0 & 1-\lambda \\ 0 & 0 & 0 \end{bmatrix}$
10	$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$	2	1	1	$\begin{bmatrix} 0 & 1 & -\lambda \\ 0 & 0 & 1 \end{bmatrix}$
11	$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	2	0	1	$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
12	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	1	0	2	$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$
13	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	0	0	3	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$
$1'$	$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	2	2	0	$\begin{bmatrix} -\lambda & 1 & 0 \\ 0 & 0 & 1-\lambda \end{bmatrix}$
$10'$	$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	2	1	1	$\begin{bmatrix} 0 & 1-\lambda & 0 \\ 0 & 0 & 1 \end{bmatrix}$
$4'$	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	1	2	1	$\begin{bmatrix} 0 & -\lambda & 0 \\ 0 & 0 & 1-\lambda \end{bmatrix}$
$7'$	$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	1	1	1	$\begin{bmatrix} 0 & -\lambda & 1 \\ 0 & 0 & 0 \end{bmatrix}$
$9'$	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$	1	1	2	$\begin{bmatrix} 0 & 0 & -\lambda \\ 0 & 0 & 1 \end{bmatrix}$

Case 1 in Table 2.2 corresponds to the generic structure. Cases 2–5 are obtained by keeping $\mathrm{r}(B_i) = 2$ and varying $\mathrm{r}(A_i)(2, 1, 0)$ and $\mathrm{n}(A_i, B_i)(0, 1)$. In cases 6–10 we

keep $\mathrm{r}(B_i) = 1$ and vary $\mathrm{r}(A_i)$ (as before) and $\mathrm{n}(A_i, B_i)(0, 1, 2)$. Finally, in cases 11–13 $\mathrm{r}(B_i) = 0$, $\mathrm{r}(A_i)$ and $\mathrm{n}(A_i, B_i)$ are varied ((0, 1, 2) and (1, 2, 3), respectively). In cases 8, 9, 12, and 13, the matrix pairs have a common row nullspace as well, corresponding to L_0^T blocks in their KCF. The number of L_0^T blocks equals the dimension of the common row nullspace (1 for cases 8, 9, and 12 and 2 for case 13). Notice that $\mathrm{n}(A_i, B_i) = 2$ for three of these four cases and $\mathrm{n}(A_i, B_i) = 3$ for case 13. However, $\mathrm{n}(A_i, B_i) = 2$ is neither a necessary nor a sufficient condition for a 2-by-3 matrix pair to have a common row nullspace (see cases 7' and 9' below). If we exchange the roles of A and B in the derivation of the nongeneric forms 2–13 they will appear in a different order with the N_k blocks and $J_k(0)$ blocks exchanged.

We have five more cases to retrieve, denoted 1', 10', 4', 7', and 9' in Table 2.2. Case x' denotes a case that has the same row ranks and column nullities as case x, and is obtained from case x by permuting rows or columns.

Case 1'. By swapping columns 2 and 3 in B_1 we still have a matrix pair with $\mathrm{r}(A_i) = \mathrm{r}(B_i) = 2$ and $\mathrm{n}(A_i, B_i) = 0$. We denote this pencil case 1'. As can be seen in Table 2.2, GUPTRI delivers the KCF $L_1 \oplus R_1$ for $A_{1'} - \lambda B_{1'}$. After the first step of deflation in GUPTRI (which identifies that $A_i, i = 1, 1'$ has a 1-dimensional column nullspace ($n(A_i) = 1$) and that $\mathrm{n}(A_i, B_i) = 0, i = 1, 1'$) we are left with the pencils:

$$A_1^{(1)} - \lambda B_1^{(1)} = \begin{bmatrix} 0 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad A_{1'}^{(1)} - \lambda B_{1'}^{(1)} = \begin{bmatrix} 0 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 0 & 1 \end{bmatrix}. \tag{2.2}$$

The difference is that $\mathrm{n}(A_1^{(1)}, B_1^{(1)}) = 0$ while $\mathrm{n}(A_{1'}^{(1)}, B_{1'}^{(1)}) = 1$. Is there any algebraic explanation? We find the answer in the classical characterization of a singular pencil with a right (column) index [9].

Let the matrix $R[A, B, i]$ of size $(i + 2)m \times (i + 1)n$ be defined by

$$R[A, B, i] = \begin{bmatrix} A & 0 & \cdots & 0 \\ B & A & \ddots & \vdots \\ 0 & \ddots & \ddots & 0 \\ \vdots & \ddots & B & A \\ 0 & \cdots & 0 & B \end{bmatrix}, \tag{2.3}$$

where A and B are $m \times n$ matrices. When it is clear from context we use the abbreviated notation $R[i]$ for $R[A, B, i]$. With the notation above we can state the following theorem.

THEOREM 2.1 (see [9]). *The following statements are equivalent.*

- *$A - \lambda B$ is singular with a right (column) minimal index of lowest degree $k \geq 0$, i.e., $A - \lambda B$ has no right minimal indices of degree $< k$.*
- *$A - \lambda B$ is equivalent to the pencil*

$$\begin{bmatrix} L_k & 0 \\ 0 & A' - \lambda B' \end{bmatrix}, \tag{2.4}$$

where L_k is a $k \times (k+1)$ Kronecker block. $A' - \lambda B'$ may have indices of higher degree.

- *$R[i]$ has full column rank $\mathrm{r}(R[i]) = (i+1)n$ for $i = 0, 1, \ldots, k-1$, while $\mathrm{r}(R[k]) < (k + 1)n$, or equivalently, the column nullity $\mathrm{n}(R[i]) = 0$ for $i = 0, 1, \ldots, k - 1$ and $\mathrm{n}(R[k]) > 0$.*

By applying Theorem 2.1 to cases 1 and 1' we see that $\mathrm{n}(R[1]) = 0, \mathrm{n}(R[2]) = 1$ for case 1 while $\mathrm{n}(R[1]) = 1, \mathrm{n}(R[2]) = 2$ for case 1', which justify that case 1 has

an L_2 block as its KCF and case 1' has an L_1 block in its KCF. After the second deflation of case 1', GUPTRI is left with the pencil $[1] - \lambda[1]$ which corresponds to R_1, a regular block of size 1×1.

Case 10'. By swapping columns 2 and 3 of B_{10} we still get a matrix pair with $\mathrm{r}(A_i) = 2, \mathrm{r}(B_i) = 1$ and $\mathrm{n}(A_i, B_i) \equiv \mathrm{n}(R[0]) = 1$. We denote this pencil case 10'. This swapping does not change the singular structure. However, the N_2 block in case 10 is now split into two regular 1×1 blocks N_1 and R_1, i.e., one infinite eigenvalue is turned nonzero.

To get the remaining three cases we will swap rows 1 and 2 in A_i for $i = 4, 7$, and 9.

Case 4'. If we swap rows 1 and 2 in A_4 we still get a matrix pair with $\mathrm{r}(A_i) = 1, \mathrm{r}(B_i) = 2$ and $\mathrm{n}(A_i, B_i) = 1$. We denote this pencil case 4'. The only difference is that the $J_2(0)$ block in case 4 is now split into two regular 1×1 blocks $J_1(0)$ and R_1, i.e., one zero eigenvalue is turned nonzero.

A dual form of Theorem 2.1 can be stated for a left (row) minimal index of lowest degree $k \geq 0$. Then L_k^T takes the place of L_k and $L[A, B, i]$ of size $(i+1)m \times (i+2)n$ replaces $R[A, B, i]$, where

$$L[A, B, i] = \begin{bmatrix} A & B & 0 & \cdots & 0 \\ 0 & A & B & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & A & B \end{bmatrix}, \tag{2.5}$$

and we are considering row ranks (or row nullities) of $L[A, B, i]$. (When it is clear from context we also here use the abbreviated notation $L[i]$ for $L[A, B, i]$.) We use this dual form to characterize the last two cases. Notice that $\mathrm{n}(R[A, B, 0])$ is equivalent to the dimension of the common column nullspace for A and B and that $\mathrm{n}(L[A, B, 0])$ is equivalent to the dimension of the common row nullspace for the two matrices.

Case 7'. By swapping rows 1 and 2 in A_7 we still get a matrix pair with $\mathrm{r}(A_i) = 1, \mathrm{r}(B_i) = 1$ and $\mathrm{n}(A_i, B_i) = 1$. We denote this pencil case 7'. However, this swap imposes a common row nullspace of $A_{7'}$ and $B_{7'}$ as well, and will therefore change the singular structure completely. The regular part $(J_1(0) \oplus N_1)$ disappears and is replaced by $L_1 \oplus L_0^T$, i.e., the generic singular structure of a 2-by-2 pencil [19]. $\mathrm{n}(A_i, B_i) \equiv \mathrm{n}(R[0]) = 1$ for $i = 7$ and 7'. For case 7, $\mathrm{n}(R[1]) = 2, \mathrm{n}(L[0]) = 0$ while $\mathrm{n}(R[1]) = 3, \mathrm{n}(L[0]) = 1$ for case 7'.

Case 9'. By swapping rows 1 and 2 in A_9 we still get a matrix pair with $\mathrm{r}(A_i) = 1, \mathrm{r}(B_i) = 1$ and $\mathrm{n}(A_i, B_i) = 2$. We denote this pencil case 9'. However, $A_{9'}$ and $B_{9'}$ do not have a common row nullspace. Also here the regular part disappears and $R_1 \oplus L_0^T$ turns into L_1^T, i.e., a generic 2-by-1 pencil. $\mathrm{n}(L[0]) = 1$ for case 9, while $\mathrm{n}(L[0]) = 0, \mathrm{n}(L[0]) = 1$ for case 9'.

In Table 2.3 we display ranks of A_i, B_i and nullities of $R[k]$ and $L[k]$ for some values of k together with our structurally different singular structures of the set of 2-by-3 pencils. The ordering of the cases is explained in §2.2.

2.2. Orbits and their codimensions. Each of the 18 singular canonical forms (A_i, B_i) in Table 2.3 defines a manifold of *strictly equivalent* pencils in $2mn(= 12)$-dimensional space:

$$\mathrm{orbit}(A_i - \lambda B_i) = \{P_i^{-1}(A_i - \lambda B_i)Q_i : \det(P_i)\det(Q_i) \neq 0\}.$$

TABLE 2.3
Geometric characteristics of the 18 *structurally different* 2×3 *pencils.*

Case	r(A_i)	r(B_i)	n(A_i, B_i)	n(R[1])	n(R[2])	n(L[0])	n(L[1])	KCF	Cod($A_i - \lambda B_i$)
1	2	2	0	0	1	0	0	L_2	0
1'	2	2	0	1	2	0	0	$L_1 \oplus R_1$	1
2	1	2	0	1	2	0	0	$L_1 \oplus J_1$	2
6	2	1	0	1	2	0	0	$L_1 \oplus N_1$	2
5	2	2	1	2	3	0	0	$L_0 \oplus R_2$	2
4'	1	2	1	2	3	0	0	$L_0 \oplus J_1 \oplus R_1$	3
10'	2	1	1	2	3	0	0	$L_0 \oplus N_1 \oplus R_1$	3
4	1	2	1	2	3	0	0	$L_0 \oplus J_2$	4
10	2	1	1	2	3	0	0	$L_0 \oplus N_2$	4
7	1	1	1	2	3	0	0	$L_0 \oplus J_1 \oplus N_1$	4
7'	1	1	1	3	5	1	2	$L_0 \oplus L_1 \oplus L_0^T$	5
3	0	2	1	2	3	0	0	$L_0 \oplus 2J_1$	6
11	2	0	1	2	3	0	0	$L_0 \oplus 2N_1$	6
9'	1	1	2	4	6	0	1	$2L_0 \oplus L_1^T$	6
9	1	1	2	4	6	1	2	$2L_0 \oplus R_1 \oplus L_0^T$	7
8	0	1	2	4	6	1	2	$2L_0 \oplus J_1 \oplus L_0^T$	8
12	1	0	2	4	6	1	2	$2L_0 \oplus N_1 \oplus L_0^T$	8
13	0	0	3	6	9	2	4	$3L_0 \oplus 2L_0^T$	12

The dimension of orbit$(A - \lambda B)$ is equal to the dimension of the tangent space, tan$(A - \lambda B)$, to the orbit of $A - \lambda B$. The tangent space is defined as

$$f(X, Y) = X(A - \lambda B) - (A - \lambda B)Y, \tag{2.6}$$

where X is an $m \times m$ matrix and Y is an $n \times n$ matrix [3]. Since (2.6) maps a space of dimension $m^2 + n^2$ linearly to a space of dimension $2mn$, the dimension of the tangent space is $m^2 + n^2 - d$, where d is the number of (linearly) independent solutions of $f(X, Y) = 0$.

The codimension is the dimension of the space complementary to the tangent space, i.e.,

$$\operatorname{cod}(A - \lambda B) = 2mn - \dim(\tan(A - \lambda B)) = d - (m - n)^2.$$

The codimensions of the orbits depend only on their Kronecker structures. Demmel and Edelman [3] show that the codimension of the orbit of an $m \times n$ pencil $A - \lambda B$ can be computed as the sum of separate codimensions:

$$\operatorname{cod}(A - \lambda B) = c_{\text{Jor}} + c_{\text{Right}} + c_{\text{Left}} + c_{\text{Jor,Sing}} + c_{\text{Sing}},$$

where the different components are defined as follows.

The codimension of the Jordan structure is

$$c_{\text{Jor}} = \sum_{\lambda \neq 0, \infty} (q_1(\lambda) + 3q_2(\lambda) + 5q_3(\lambda) + \cdots - 1) + \sum_{\lambda = 0, \infty} (q_1(\lambda) + 3q_2(\lambda) + 5q_3(\lambda) + \cdots),$$

where the summation is over all eigenvalues and $q_1(\lambda) \geq q_2(\lambda) \geq q_3(\lambda) \ldots$, denote the sizes of the Jordan blocks corresponding to the eigenvalue λ. The first part of c_{Jor} corresponds to unspecified eigenvalues different from zero and infinity, which explains the term -1 in the codimension count.

The codimensions of the right and left singular blocks are

$$c_{\text{Right}} = \sum_{j>k} (j - k - 1) \quad \text{and} \quad c_{\text{Left}} = \sum_{j>k} (j - k - 1),$$

respectively, where the summation for c_{Right} is over all pairs of blocks L_j and L_k, for which $j > k$, and the summation for c_{Left} is over all pairs of blocks L_j^T and L_k^T, for which $j > k$.

The codimension due to interaction between the Jordan structure and the singular blocks is

$$c_{\text{Jor,Sing}} = (\text{size of complete regular part}) \cdot (\text{number of singular blocks}).$$

The codimension due to interaction between right and left singular blocks is

$$c_{\text{Sing}} = \sum_{j,k} (j + k + 2),$$

where the summation is over all pairs of blocks L_j and L_k^T.

The codimensions of our 18 different canonical forms are displayed in the last column of Table 2.3. We have ordered the cases by increasing codimension. In general, we see that by making A and B more rank deficient and increasing their "common nullspace(s)" ($\text{n}(R[k])$ and $\text{n}(L[k])$ for $k \geq 0$) we generate nongeneric pencils with higher codimension. The generic pencil has codimension 0 while the matrix pair $(A, B) = (0_{2\times 3}, 0_{2\times 3})$ has codimension 12 ($= 2mn$), i.e., defines a "point" in 12-dimensional space.

2.3. The closure graph for different Kronecker structures. Since $\text{orbit}(L_2)$ spans the complete 12-dimensional space, it is obvious that all other structures are in the closure of the orbit of L_2, and it is just as obvious that $3L_0 \oplus 2L_0^T$ (the zero pencil) is in the closure of the orbit of any other KCF. Since all other closure relations are not that obvious, we derive a complete closure graph for the set of 2-by-3 matrix pencils.

Throughout the paper we display graphs such that orbits (nodes) with the same codimension are displayed on the same horizontal level.

THEOREM 2.2. *For the set of 2-by-3 pencils, the directed graph in Fig. 2.1 shows all closure relations as follows. One* KCF *is in the closure of the orbit of another* KCF *if and only if there exists a path to its node from the node of the* KCF *defining the closure (downwards in the graph).*

Proof. First we prove that each arc in the graph corresponds to a closure relation, and then we prove that these are all arcs that can exist. We prove that one KCF is in the closure of the orbit of another KCF by showing that the one in the closure is just a special case of the one defining the closure. We show proofs for each arc starting from the zero pencil. Since the proof is rather space demanding, we here limit ourselves to proving one of the arcs and refer the reader to Appendix A for the complete proof.

Starting at the zero pencil, the first arc with nontrivial proof corresponds to the fact that $2L_0 \oplus J_1 \oplus L_0^T$ is in the closure of $\text{orbit}(2L_0 \oplus R_1 \oplus L_0^T)$. This follows from the fact that $2L_0 \oplus J_1 \oplus L_0^T$ is the special case $\alpha = 0$ of

$$\begin{bmatrix} 0 & 0 & \alpha \\ 0 & 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & 0 & \beta \\ 0 & 0 & 0 \end{bmatrix},$$

which is equivalent to $2L_0 \oplus R_1 \oplus L_0^T$ for all other α (assuming that β is nonzero).

The proofs for all other arcs are done similarly. For some of them, an equivalence transformation is needed for transformation to KCF. □

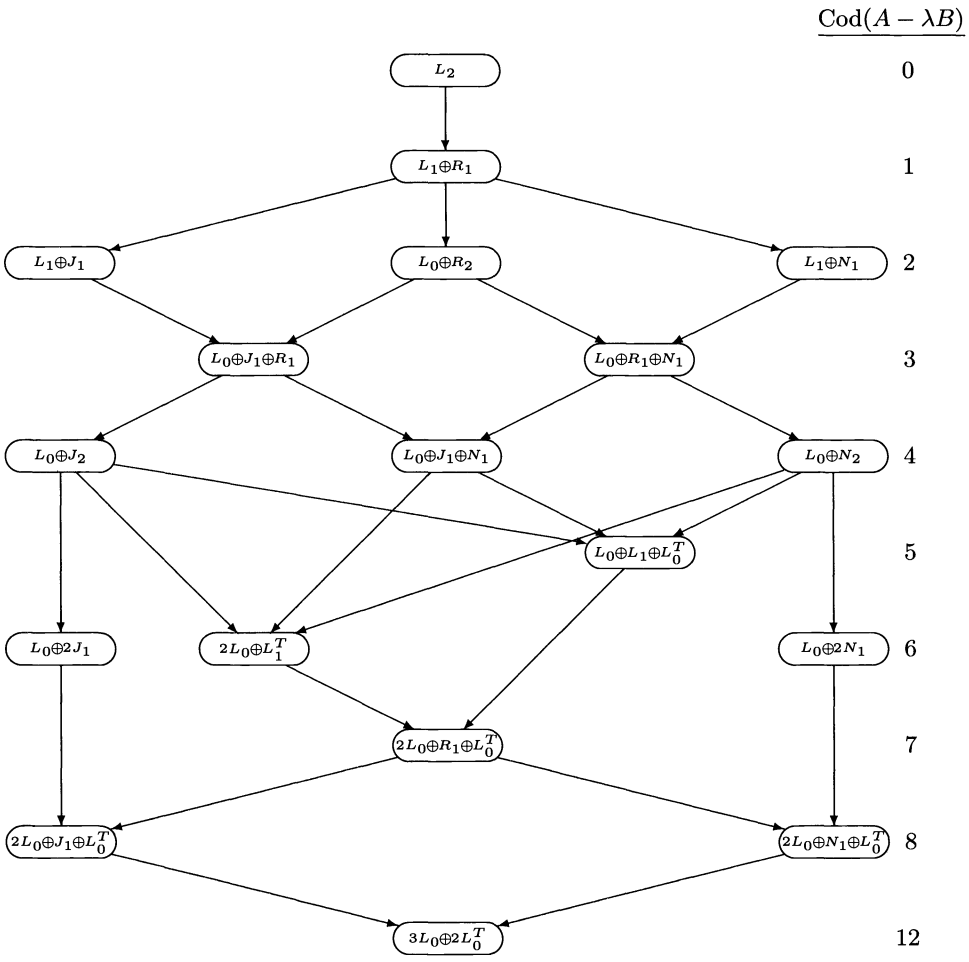


FIG. 2.1. *A graph displaying the closure hierarchy of the orbits of all* 18 *different* KCF *for the set of* 2*-by-*3 *matrix pencils.*

2.4. Labeled closure graph showing necessary conditions on perturbations for transiting from one structure to another. One way to interpret a relation in the closure hierarchy is that a KCF that is in the closure of the orbit of another KCF "lives" in the space defined by that orbit. That is, if we consider the closure of the orbit of a nongeneric KCF with certain rank-defects in Table 2.3, then to be in that closure a KCF must preserve or increase these defects. For example, since $L_1 \oplus J_1$ has $\text{rank}(A) = 1$, no KCF with $\text{rank}(A) > 1$ can be in its closure. A necessary condition for a KCF to be in the closure of orbit$(L_1 \oplus J_1)$ is that the geometric characteristics $\text{r}(A) \leq 1, \text{r}(B) \leq 2, \text{n}(A, B) \geq 0, \text{n}(R[1]) \geq 1, \text{n}(R[2]) \geq 2, \text{n}([L[0]) \geq 0$ and $\text{n}([L[1]) \geq 0$ are satisfied (see Table 2.3). Moreover, the change in geometric characteristics from, for example, $L_1 \oplus J_1$, whose orbit spans a 10-dimensional space (codimension is 2), to $L_0 \oplus J_1 \oplus R_1$, whose orbit spans a 9-dimensional space (codimension is 3), is nothing but a 1-dimensional restriction of the 10-dimensional space. We also note that $L_0 \oplus J_1 \oplus R_1$ is in the closure of orbit$(L_0 \oplus R_2)$, which also spans

a 10-dimensional space. Indeed, $L_0 \oplus J_1 \oplus R_1$ spans a 9-dimensional space in the intersection of the two 10-dimensional spaces spanned by the closures of orbit$(L_1 \oplus J_1)$ and orbit$(L_0 \oplus R_2)$.

When looking for perturbations corresponding to the arcs in the graph, a necessary condition for these perturbations is to fulfill the change in geometric characteristics. Indeed, by combining the geometric characteristics in Table 2.3 and the closure graph we get necessary conditions on perturbations $(\delta A, \delta B)$ for transiting from one structure to another.

We introduce the following labels. Let

$$[\mathrm{n}_r(A), \mathrm{n}_r(B), \mathrm{n}(A,B), \mathrm{n}(R[1]), \mathrm{n}(R[2]), \mathrm{n}([L[0]), \mathrm{n}([L[1])]$$

label the geometric characteristics for one node in the graph, where $n_r(A)$ and $n_r(B)$ denote the dimension of the row nullspace in A and B, respectively, and all other characteristics are as in Table 2.3. Moreover, we label the change in geometric characteristics for transiting from one structure to an adjacent node by

$$\langle \mathrm{n}_r(A), \mathrm{n}_r(B), \mathrm{n}(A,B), \mathrm{n}(R[1]), \mathrm{n}(R[2]), \mathrm{n}([L[0]), \mathrm{n}([L[1])\rangle.$$

In Fig. 2.2 a labeled closure graph is presented, with the geometric characteristics shown for each KCF and the change in geometric characteristics shown for each arc.

When transiting from one KCF to another, the geometric characteristics of the source node and the geometric characteristics on the arc are added to give the characteristics of the destination KCF. Since a KCF in the closure of another's orbit cannot have a smaller dimensional nullspace for any of the matrices of the labels, the values on the arcs must all be nonnegative.

Notice that the arc from $L_0 \oplus J_1 \oplus R_1$ to $L_0 \oplus J_2$ and the arc from $L_0 \oplus R_1 \oplus N_1$ to $L_0 \oplus N_2$ both have no change in the geometric characteristics. For these transitions the nonzero finite eigenvalue is turned to a zero eigenvalue and to an infinite eigenvalue, respectively. This does not affect any of the nullspaces displayed in the labels.

To transit several levels in the closure graph we just add the labels of changes in geometric characteristics for the arcs that are traveled during the transition. Each label of changes in geometric characteristics defines necessary conditions on the perturbations $(\delta A, \delta B)$ to perform the transit. Later, we will derive perturbations required to transit from L_2 to any of the nongeneric structures. In our derivation, however, for most cases we transit directly to the intended structure. There are only a few cases that require compound perturbations that transit via another KCF.

3. Structure invariances and transitions of nongeneric pencils under perturbations. Since computing the Kronecker structure of a singular pencil is a potentially ill-posed problem [5], it is interesting to see how the nongeneric cases behave under perturbations in finite precision arithmetic. We add (uniformly distributed) random perturbations of different sizes $\epsilon_n (= 10^{-10}, 10^{-9}, \ldots, 10^{-1})$ to all A_i and B_i, corresponding to the generic and 17 nongeneric cases, and compute their generalized Schur forms using GUPTRI [7], [8], assuming a fixed relative accuracy $\epsilon_u (= 10^{-8})$ of the input data. We repeat this procedure 100 times and study the structure invariances and transitions of each nongeneric case as a function of the size of the perturbations added.

GUPTRI has two input parameters EPSU (ϵ_u above) and GAP which are used to make rank decisions in order to determine the Kronecker structure of an input pencil $A - \lambda B$. Inside GUPTRI the absolute tolerances EPSUA $= \|A\|_E \cdot$ EPSU

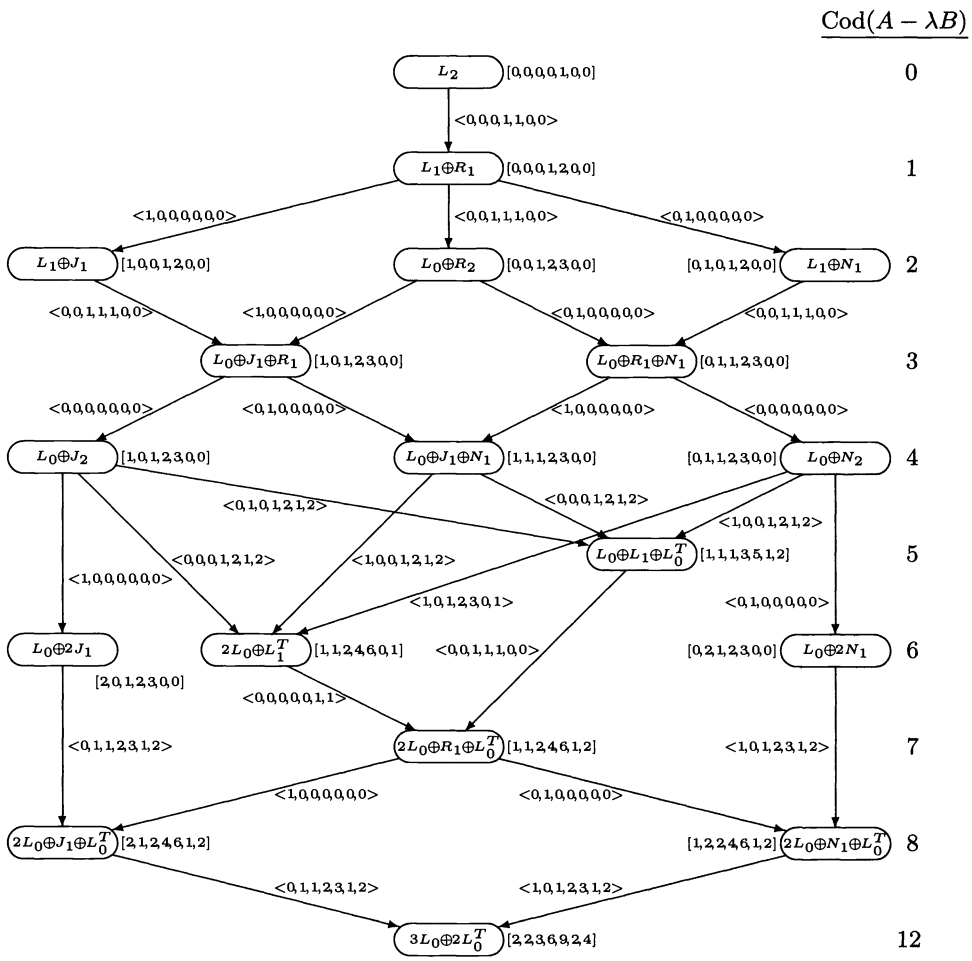


FIG. 2.2. *The labeled closure graph for all* 18 *different* KCF *for the set of* 2*-by-*3 *matrix pencils.*

and EPSUB = $\|B\|_E \cdot$ EPSU are used in all rank decisions, where the matrices A and B, respectively, are involved. Suppose the singular values of A are computed in increasing order, i.e., $0 \leq \sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_k \leq \sigma_{k+1} \leq \ldots$; then all singular values $\sigma_k <$ EPSUA are interpreted as zeros. The rank decision is made more robust in practice: if $\sigma_k <$ EPSUA but $\sigma_{k+1} \geq$ EPSUA, GUPTRI insists on a gap between the two singular values such that $\sigma_{k+1}/\sigma_k \geq$ GAP. If $\sigma_{k+1}/\sigma_k <$ GAP, σ_{k+1} is also treated as zero. This process is repeated until an appreciable gap between the zero and nonzero singular values is obtained. In all of our tests we have used EPSU = 10^{-8} and GAP = 1000.0. All computations (in §§3 and 4) are performed on a SUN SPARC station in double precision complex arithmetic with unit roundoff = $O(10^{-17})$.

In Table 3.1 we display the computed Kronecker structures of the 17 perturbed nongeneric pencils for 100 random perturbations for each ϵ_n. For each case all structure invariances and transitions are shown from left to right. The symbol $\stackrel{10^{-x}}{\longrightarrow}$ indicates that the Kronecker structure is invariant under perturbations smaller than $\epsilon_n = 10^{-x}$, and that the structure changes (at least for some of the 100 tests) for

TABLE 3.1

Computed Kronecker structures and transitions of 100 *perturbed nongeneric* 2×3 *pencils. The size* ϵ_n *of each perturbation is shown above the corresponding arrow.*

$$1'\colon L_1\oplus R_1 \xrightarrow{10^{-4}} \left\{\begin{matrix} L_2 & (81)\\ L_1\oplus R_1 & (19)\end{matrix}\right\} \xrightarrow{10^{-3}} \left\{\begin{matrix} L_2 & (98)\\ L_1\oplus R_1 & (2)\end{matrix}\right\} \xrightarrow{10^{-2}} L_2$$

$$2\colon L_1\oplus J_1 \xrightarrow{10^{-5}} \left\{\begin{matrix} L_2 & (18)\\ L_1\oplus J_1 & (82)\end{matrix}\right\} \xrightarrow{10^{-4}} \left\{\begin{matrix} L_2 & (98)\\ L_1\oplus J_1 & (2)\end{matrix}\right\} \xrightarrow{10^{-3}} L_2$$

$$6\colon L_1\oplus N_1 \xrightarrow{10^{-4}} \left\{\begin{matrix} L_2 & (92)\\ L_1\oplus R_1 & (6)\\ L_1\oplus N_1 & (2)\end{matrix}\right\} \xrightarrow{10^{-3}} \left\{\begin{matrix} L_2 & (99)\\ L_1\oplus R_1 & (1)\end{matrix}\right\} \xrightarrow{10^{-2}} L_2$$

$$5\colon L_0\oplus R_2 \xrightarrow{10^{-4}} \left\{\begin{matrix} L_2 & (63)\\ L_1\oplus R_1 & (28)\\ L_0\oplus R_2 & (9)\end{matrix}\right\} \xrightarrow{10^{-3}} \left\{\begin{matrix} L_2 & (95)\\ L_1\oplus R_1 & (4)\\ L_0\oplus R_2 & (1)\end{matrix}\right\} \xrightarrow{10^{-2}} \left\{\begin{matrix} L_2 & (99)\\ L_1\oplus R_1 & (1)\end{matrix}\right\} \xrightarrow{10^{-1}} L_2$$

$$4'\colon L_0\oplus J_1\oplus R_1 \xrightarrow{10^{-5}} \left\{\begin{matrix} L_2 & (1)\\ L_1\oplus R_1 & (17)\\ L_0\oplus J_1\oplus R_1 & (82)\end{matrix}\right\} \xrightarrow{10^{-4}} \left\{\begin{matrix} L_2 & (82)\\ L_1\oplus R_1 & (16)\\ L_1\oplus J_1 & (1)\\ L_0\oplus J_1\oplus R_1 & (1)\end{matrix}\right\} \xrightarrow{10^{-3}} L_2$$

$$10'\colon L_0\oplus N_1\oplus R_1 \xrightarrow{10^{-5}} \left\{\begin{matrix} L_2 & (90)\\ L_0\oplus N_1\oplus R_1 & (10)\end{matrix}\right\} \xrightarrow{10^{-4}} \left\{\begin{matrix} L_2 & (98)\\ L_0\oplus R_2 & (2)\end{matrix}\right\} \xrightarrow{10^{-3}} L_2$$

$$4\colon L_0\oplus J_2 \xrightarrow{10^{-5}} \left\{\begin{matrix} L_2 & (22)\\ L_0\oplus J_1\oplus R_1 & (17)\\ L_0\oplus J_2 & (61)\end{matrix}\right\} \xrightarrow{10^{-4}} L_2$$

$$10\colon L_0\oplus N_2 \xrightarrow{10^{-5}} \left\{\begin{matrix} L_2 & (10)\\ L_0\oplus N_1\oplus R_1 & (15)\\ L_0\oplus N_2 & (75)\end{matrix}\right\} \xrightarrow{10^{-4}} \left\{\begin{matrix} L_2 & (99)\\ L_0\oplus N_2 & (1)\end{matrix}\right\} \xrightarrow{10^{-3}} L_2$$

$$7\colon L_0\oplus J_1\oplus N_1 \xrightarrow{10^{-5}} \left\{\begin{matrix} L_2 & (5)\\ L_1\oplus N_1 & (13)\\ L_0\oplus J_1\oplus N_1 & (82)\end{matrix}\right\} \xrightarrow{10^{-4}} \left\{\begin{matrix} L_2 & (92)\\ L_1\oplus R_1 & (6)\\ L_1\oplus J_1 & (2)\end{matrix}\right\} \xrightarrow{10^{-3}} \left\{\begin{matrix} L_2 & (97)\\ L_1\oplus R_1 & (3)\end{matrix}\right\} \xrightarrow{10^{-2}} L_2$$

$$7'\colon L_0\oplus L_1\oplus L_0^T \xrightarrow{10^{-5}} \left\{\begin{matrix} L_2 & (3)\\ L_1\oplus R_1 & (7)\\ L_1\oplus N_1 & (12)\\ L_0\oplus L_1\oplus L_0^T & (78)\end{matrix}\right\} \xrightarrow{10^{-4}} \left\{\begin{matrix} L_2 & (93)\\ L_1\oplus R_1 & (7)\end{matrix}\right\} \xrightarrow{10^{-3}} \left\{\begin{matrix} L_2 & (98)\\ L_1\oplus R_1 & (2)\end{matrix}\right\} \xrightarrow{10^{-2}} L_2$$

$$3\colon L_0\oplus 2J_1 \xrightarrow{10^{-10}} L_2$$

$$11\colon L_0\oplus 2N_1 \xrightarrow{10^{-10}} L_2$$

$$9'\colon 2L_0\oplus L_1^T \xrightarrow{10^{-5}} \left\{\begin{matrix} L_0\oplus R_2 & (3)\\ L_0\oplus J_1\oplus R_1 & (34)\\ L_0\oplus N_2 & (15)\\ 2L_0\oplus L_1^T & (48)\end{matrix}\right\} \xrightarrow{10^{-4}} \left\{\begin{matrix} L_2 & (86)\\ L_1\oplus R_1 & (8)\\ L_0\oplus R_2 & (4)\\ L_1\oplus J_1 & (2)\end{matrix}\right\} \xrightarrow{10^{-3}} L_2$$

$$9\colon 2L_0\oplus R_1\oplus L_0^T \xrightarrow{10^{-5}} \left\{\begin{matrix} L_0\oplus R_2 & (2)\\ L_0\oplus J_1\oplus R_1 & (33)\\ L_0\oplus N_1\oplus R_1 & (16)\\ 2L_0\oplus R_1\oplus L_0^T & (49)\end{matrix}\right\} \xrightarrow{10^{-4}} \left\{\begin{matrix} L_2 & (80)\\ L_1\oplus R_1 & (17)\\ L_0\oplus R_2 & (1)\\ L_1\oplus J_1 & (1)\\ L_0\oplus J_1\oplus R_1 & (1)\end{matrix}\right\} \xrightarrow{10^{-3}} \left\{\begin{matrix} L_2 & (96)\\ L_1\oplus R_1 & (4)\end{matrix}\right\} \xrightarrow{10^{-2}} L_2$$

$$8\colon 2L_0\oplus J_1\oplus L_0^T \xrightarrow{10^{-10}} L_1\oplus N_1 \xrightarrow{10^{-6}} \left\{\begin{matrix} L_2 & (2)\\ L_1\oplus N_1 & (98)\end{matrix}\right\} \xrightarrow{10^{-5}} \left\{\begin{matrix} L_2 & (24)\\ L_1\oplus R_1 & (2)\\ L_1\oplus N_1 & (54)\end{matrix}\right\} \xrightarrow{10^{-4}} \left\{\begin{matrix} L_2 & (86)\\ L_1\oplus R_1 & (14)\end{matrix}\right\} \xrightarrow{10^{-3}} L_2$$

$$12\colon 2L_0\oplus N_1\oplus L_0^T \xrightarrow{10^{-10}} L_1\oplus J_1 \xrightarrow{10^{-5}} \left\{\begin{matrix} L_2 & (18)\\ L_1\oplus J_1 & (82)\end{matrix}\right\} \xrightarrow{10^{-4}} \left\{\begin{matrix} L_2 & (98)\\ L_1\oplus J_1 & (2)\end{matrix}\right\} \xrightarrow{10^{-3}} L_2$$

$$13\colon 3L_0\oplus 2L_0^T \xrightarrow{10^{-10}} L_2$$

perturbations of size 10^{-x}. For a size of the perturbations that has not given the same structure for all 100 tests, all KCFs found are placed within curly brackets with a number within parentheses after each KCF showing the number of that particular KCF that has been found. As before, the cases are displayed in increasing codimension order and the transit KCF forms within curly brackets are ordered similarly.

From Table 3.1 we see that for large enough perturbations all nongeneric structures turn generic (as expected). GUPTRI finds the same nongeneric structure as long as $\epsilon_n < \text{tol} \equiv \min(\text{EPSUA}, \text{EPSUB}) \cdot \text{GAP}$. This behaviour is in agreement with the perturbation theory for singular pencils [5], [7]. Only if $A - \lambda B$ lies in a particular manifold does it have a nongeneric Kronecker structure with nontrivial reducing subspaces and possibly eigenvalues. Moreover, only if it is perturbed so as to move continuously within that manifold does its original Kronecker structure remain. Actually, by choosing a tol > 0, we have thickened the manifolds so that they are no longer a set of measure zero.

All transitions from the initial case to the final generic case are clearly from cases with higher codimension to cases with lower codimension. With a closer look we can also see that all the transitions are performed upwards (or backwards) along the arcs in the closure graph (Fig. 2.1). This means that the perturbations cure the rank deficiencies in the nongeneric pencil without contributing any new singularities. GUPTRI increases the rank in A and B and decreases the size of their "common nullspace(s)," i.e., the "inverse" operations compared to what we did in §2.1. In other words, when a pencil $A - \lambda B$ with a given nongeneric KCF is perturbed, by $\delta A - \lambda \delta B$ then $A - \lambda B$ is in the closure of orbit$((A + \delta A) - \lambda(B + \delta B))$.

Even if we see that all of the cases transit via some other nongeneric structures before all 100 tests turn generic, we can also see that if for each case and each size of the perturbation we only consider the KCF that has been found in most tests, then it is only for cases 8 and 12 that a transit KCF is found. Notice that all tests for cases 8 and 12 find the same other nongeneric KCF for the smallest perturbation. In other words, when the perturbation is big enough to change the KCF for most tests of a case, then the generic KCF is the most likely to be found, except for cases 8 and 12.

How can we explain the behaviour in cases 8 and 12? For these two cases one matrix is the zero matrix. This means that tol $\equiv \min(\text{EPSUA}, \text{EPSUB}) \cdot \text{GAP} = 0$, implying that $\epsilon_n >$ tol already for the smallest perturbation, which in turn explains why case transitions occur already for the smallest perturbation. Since either EPSUA or EPSUB is zero, all singular values in the perturbed zero matrix will be interpreted as nonzero, explaining why A or B is interpreted as a full rank matrix already for the smallest perturbations. Also notice the "jumps" these transitions correspond to in the closure graph. The argumentation here also explains why the zero pencil turns generic for the smallest perturbation.

We end this section by briefly discussing how the case invariances and transitions are affected by the choice of the fixed relative accuracy of the input data (EPSU). If we choose EPSU $= \epsilon_n$ then GUPTRI will retrieve the nongeneric structure we started from for each ϵ_n considered. Notice that the distance from the input pencil to the computed Kronecker structure will normally be of size $O(\text{EPSU} \cdot \|(A, B)\|_E)$ [8]. Increasing EPSU means that the case invariances will remain longer before any case transition takes place. Decreasing EPSU will impose the generic structure sooner. For example, with EPSU equal to the relative machine precision and $\epsilon_n >$ tol, GUPTRI will always extract the generic structure. This corresponds to the fact that in infinite precision arithmetic any nongeneric $A - \lambda B$ can be made generic with arbitrary small

perturbations. Moreover, travelling upwards in the closure hierarchy can always be effected with arbitrary small perturbations, while travelling downwards may require much larger perturbations.

4. Imposing nongeneric structures by perturbing a generic pencil. In this section we study computable normwise bounds for the smallest perturbations $(\delta A, \delta B)$ of a generic 2-by-3 pencil $A - \lambda B$ such that $(A + \delta A) - \lambda(B + \delta B)$ has a specific nongeneric Kronecker structure chosen from the 17 nongeneric cases discussed earlier. Our goal is to find the closest nongeneric pencil and the closest pencil with a specified nongeneric Kronecker structure of a 2-by-3 generic pencil. We consider two approaches to impose a nongeneric structure. First we derive explicit expressions for the perturbations that transfer $A - \lambda B$ to a specified nongeneric form. Second, we have modified GUPTRI to be able to compute a specified Kronecker structure.

4.1. Explicit perturbations to impose nongeneric structures. In §2 we saw that by making A and B more rank deficient and increasing their "common nullspace(s)" we can generate nongeneric pencils with higher codimension. Here we elaborate on this fact and derive explicit expressions for the perturbations required to turn an arbitrary generic pencil into each of the 17 nongeneric cases. The norms of these explicit expressions (measured as $\|(\delta A, \delta B)\|_E$) are upper bounds for the smallest perturbations required. Indeed, for 11 of the structures, the norms are the exact sizes of the smallest perturbations required.

We need the following notation. The size of the smallest perturbations $(\delta A, \delta B)$ such that $R[A + \delta A, B + \delta B, i]$ (2.3) of size $(i + 2)m \times (i + 1)n$ has a k-dimensional column nullspace is defined as

$$d_k(R[A, B, i]) = \min_{(\delta A, \delta B)} \{\|(\delta A, \delta B)\|_E : \mathrm{n}(R[A + \delta A, B + \delta B, i]) = k\}, \tag{4.1}$$

where δA and δB vary over all m-by-n matrices with complex (or real) entries. Similarly, we define $d_k(L[A, B, i])$ as the size of the smallest perturbations that impose a k-dimensional row nullspace on $L[i]$ (2.5). When it is clear from context we use the abbreviated notation $d_k(R[i])$ and $d_k(L[i])$. Also, let $d_k(A)$ denote the size of the smallest perturbations such that $\mathrm{rank}(A + \delta A) = \min(m, n) - k$.

In general, to find $d_k(R[i])$ (or $d_k(L[i])$) is a type of a structured singular value problem. For $i \geq 1$ it is an open problem to find explicit expressions for $d_k(R[i])$ and $d_k(L[i])$. The following theorem summarizes some of their properties for the case $m = 2, n = 3$ and $k = 1$.

THEOREM 4.1. *For a generic 2-by-3 pencil (A, B) the following inequalities hold:*

$$0 \equiv d_1(R[2]) < d_1(R[1]) \leq d_1(R[0]), \tag{4.2}$$

$$d_1(R[1]) \leq d_1(A), \quad d_1(R[1]) \leq d_1(B), \tag{4.3}$$

$$d_1(R[1]) < d_1(L[1]) \leq d_1(L[0]), \tag{4.4}$$

$$d_1(A) < d_2(A), \quad d_1(B) < d_2(B), \quad d_1(R[0]) < d_2(R[0]). \tag{4.5}$$

Proof. From Theorem 2.1 it follows that $d(R[2]) = 0$ for all 2-by-3 pencils (generic or nongeneric). Decreasing the rank of the 4-by-3 $R[A, B, 0]$ by one gives that $R[A + \delta A, B + \delta B, 0]$ has only two linearly independent columns. The same perturbations make the 6-by-6 matrix $R[A + \delta A, B + \delta B, 1]$ rank deficient (a rank drop from six to four), showing that (4.2) holds. Similarly, decreasing the rank of A (or B) by one means that $A + \delta A$ (or $B + \delta B$) only has one linearly independent row. For the same perturbations $R[A + \delta A, B + \delta B, 1]$ is rank deficient with only one of the first two (or last two) rows linearly independent, resulting in the inequalities (4.3).

$L[1]$ is row rank deficient if and only if there exists at least one L_0^T or L_1^T block in the KCF. Since all KCFs with at least one L_0^T block or one L_1^T block have both A and B rank deficient (see Table 2.3), there will always exist a strictly smaller perturbation of size $d_1(A)$ that only lowers the rank in A. (The same is of course true for B.) Now applying inequality (4.3) proves the first part of (4.4). The last part follows from arguments similar to the proof of $d_1(R[1]) \leq d_1(R[0])$ above. The inequalities (4.5) follow from the definition of $d_k(\cdot)$. □

Theorem 4.1 will be used to identify the closest nongeneric Kronecker structure of a generic 2-by-3 pencil. Notice that in general we cannot say anything about the relationship between $d_1(R[0])$ and $d_1(A)$ or $d_1(B)$ (see explicit expressions below). By varying α and β in

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & \alpha \end{bmatrix}, \quad B = \begin{bmatrix} \beta & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix},$$

(i.e., a generic $A - \lambda B$ for nonzero α and β) we show that any of them can be the smallest quantity (see Table 4.1).

TABLE 4.1
The quantities $d_1(A)$, $d_1(B)$, and $d_1(R[0])$ for three examples.

Parameters	$d_1(A)$	$d_1(B)$	$d_1(R[0])$
$\alpha = \beta = 1$	1.000	1.000	**0.765**
$\alpha = 0.1,\ \beta = 1$	**0.100**	1.000	0.451
$\alpha = 1,\ \beta = 0.1$	1.000	**0.100**	0.451

The following explicit expressions, derived from the Eckart–Young and Mirsky theorem for finding the closest matrix of a given rank (e.g., see [10]), appear in our explicit bounds discussed next:

$$d_1(A) = \sigma_{\min}(A), \quad d_1(B) = \sigma_{\min}(B),$$

$$d_2(A) = \|A\|_E, \quad d_2(B) = \|B\|_E,$$

$$d_1(R[A, B, 0]) = \sigma_{\min}(R[0]), \quad d_1(L[A, B, 0]) = \sigma_{\min}(L[0]),$$

$$d_2(R[A, B, 0]) = (\sigma_{\min-1}^2(R[0]) + \sigma_{\min}^2(R[0]))^{1/2}.$$

Here, $\sigma_{\min}(X)$ and $\sigma_{\min-1}(X)$ (with $\sigma_{\min}(X) \leq \sigma_{\min-1}(X)$) denote the two smallest nonzero singular values of (a full rank) matrix X.

4.1.1. Tractable perturbations. In order to make the problem more tractable we (first) put restrictions on allowable perturbations. We can compute a perturbation $\delta A - \lambda \delta B$ such that $(A + \delta A) - \lambda(B + \delta B)$ is guaranteed to fall on the closure of the manifold (orbit) of a certain KCF. (Necessary conditions on the required perturbations are given in the labeled closure graph in Fig. 2.2.) If the KCF found is the intended KCF, then the perturbation is said to be tractable. If the KCF found is even more nongeneric (i.e., its orbit has higher codimension but belongs to the closure of the intended manifold), then the perturbation is defined intractable. In other words, a tractable perturbation finds the generic KCF (i.e., the least nongeneric KCF) in the closure of the manifold of the intended KCF. An intractable perturbation finds any other structure in the closure of the same manifold, i.e., any structure that can be found by traveling along the arcs (downwards) from the intended KCF in the closure graph in Fig. 2.1.

When computing perturbations such that $(A + \delta A) - \lambda(B + \delta B)$ is given a nongeneric KCF, we compute δA and δB such that one or more of the geometric characteristics presented in Table 2.3 for $(A+\delta A)-\lambda(B+\delta B)$ differ from the characteristics of the generic (A, B). In other words, we put restrictions on the size of the perturbed pencil's nullspaces so that at least one of them is larger than for the generic case. The space given by this restriction may contain several nongeneric matrix pencils. For example, if we restrict the set of pencils to those that have a rank deficiency in the A-part, this space contains all pencils that fulfill the condition $\operatorname{rank}(A) < 2$. However, if we compute a perturbation such that $\operatorname{rank}(A + \delta A) < 2$, the perturbed pencil will most likely be the generic (least nongeneric) KCF with a rank-deficient A-part, i.e., $L_1 \oplus J_1$. This corresponds to the KCF with rank-deficient A-part whose orbit has the smallest codimension and the corresponding perturbation $(\delta A, \delta B)$ is tractable. The perturbation is intractable if $(A + \delta A) - \lambda(B + \delta B)$ has any KCF (with $\operatorname{rank}(A) < 2$) that is more nongeneric than $L_1 \oplus J_1$. The set of possible structures are the ones that are in the closure of orbit$(L_1 \oplus J_1)$.

Eleven of the 17 nongeneric structures (2, 6, 5, 7, 7', 3, 11, 9', 8, 12, and 13) are imposed by (minimal) tractable perturbations that effectuate one of the following rank-decreasing operations:

- Rank drop in A and/or B by one or two.
- Rank drop in $R[A, B, 0]$ by one or two, i.e., imposing a common one- or two-dimensional column nullspace.
- Rank drop in $L[A, B, 0]$ by one, i.e., imposing a common row nullspace.

In Table 4.2 the size of the perturbations required to impose each of the eleven singular structures are displayed. When both $d_i(A)$ and $d_j(B)$ are involved, the size of the total perturbations is $(d_i^2(A) + d_j^2(B))^{1/2}$. The singular cases are reported in increasing codimension order (see Table 2.3). Since all these perturbations are made as the smallest possible to impose the required ranks on $A, B, R[0]$ or $L[0]$, these bounds are attained for each nongeneric form, i.e., the strongest possible, which is equivalent to the bounds in Table 4.2 being lower bounds. That these perturbations really give the forms shown in the table follows from the fact that we here are only considering tractable perturbations and these are the least nongeneric forms that have the imposed rank-deficiencies (see Table 2.3). For example, by imposing a 1-dimensional rank drop in A we have restricted the 12-dimensional space to a space that contains a subset of all nongeneric pencils. Since the perturbation is supposed to be tractable, the KCF found is the least nongeneric in that space, i.e., $L_1 \oplus J_1$.

The rank-decreasing operations performed in Table 4.2 "affect the codimension(s)"

TABLE 4.2
Minimal perturbations of a generic pencil to impose 11 *of the* 17 *nongeneric structures.*

Case	KCF	Cod(·)	$d_1(A)$	$d_1(B)$	$d_1(R[0])$	$d_1(L[0])$	$d_2(A)$	$d_2(B)$	$d_2(R[0])$
2	$L_1 \oplus J_1$	2	×						
6	$L_1 \oplus N_1$	2		×					
5	$L_0 \oplus R_2$	2			×				
7	$L_0 \oplus J_1 \oplus N_1$	4	×	×					
7′	$L_0 \oplus L_1 \oplus L_0^T$	5				×			
3	$L_0 \oplus 2J_1$	6					×		
11	$L_0 \oplus 2N_1$	6						×	
9′	$2L_0 \oplus L_1^T$	6							×
8	$2L_0 \oplus J_1 \oplus L_0^T$	8		×			×		
12	$2L_0 \oplus N_1 \oplus L_0^T$	8	×					×	
13	$3L_0 \oplus 2L_0^T$	12					×	×	

in the following way: a rank drop by one in A, B or $R[0]$ increases the codimension by two, a rank drop by one in $L[0]$ increases the codimension by five, and a rank drop by two in A, B or $R[0]$ increases the codimension by six.

Two of the remaining six nongeneric forms (4' and 10') are imposed by transiting via a nongeneric form as shown in Table 4.3. For example, to derive perturbations of the generic $A - \lambda B$ that turn $(A + \delta A, B + \delta B)$ nongeneric with KCF $L_0 \oplus J_1 \oplus R_1$ we have $(\delta A, \delta B) = (\delta A_1, \delta B_1) + (\delta A_2, \delta B_2)$, where $(\delta A_1, \delta B_1)$ is the smallest perturbation that lowers the rank of A (i.e., $\|(\delta A_1, \delta B_1)\|_E = d_1(A)$, $\delta B_1 = 0_{2\times 3}$) and $(\delta A_2, \delta B_2)$ is the smallest perturbation that imposes a common column nullspace on $(A + \delta A_1, B + \delta B_1)$ (i.e., $\|(\delta A_2, \delta B_2)\|_E = d_1(R[A + \delta A_1, B + \delta B_1, 0])$). In Table 4.3 we show how these forms are constructed. The size of the compound (total) perturbations $(\delta A, \delta B)$ for the two cases are obtained by adding the perturbations in Tables 4.2 and 4.3. $\tilde{A} = A + \delta A_1$ and $\tilde{B} = B + \delta B_1$ in Table 4.3 represent the "transit" nongeneric pencil. A rank drop by one in $R[\tilde{A}, \tilde{B}, 0]$ in Table 4.3 increases the codimension by one.

TABLE 4.3
Compound perturbations: Nongeneric structures imposed by transiting via a nongeneric form.

Case	KCF	Cod(·)	Transit KCF	$d_1(R[\tilde{A}, \tilde{B}, 0])$
4′	$L_0 \oplus J_1 \oplus R_1$	3	$L_1 \oplus J_1$	×
10′	$L_0 \oplus N_1 \oplus R_1$	3	$L_1 \oplus N_1$	×

The last four nongeneric structures (1', 4, 10 and 9) require perturbations to parts of the GUPTRI form of a transiting pencil $\tilde{A} - \lambda\tilde{B}$:

$$P^H(\tilde{A} - \lambda\tilde{B})Q = \tilde{S} - \lambda\tilde{T} \equiv \begin{bmatrix} \tilde{s}_{11} & \tilde{s}_{12} & \tilde{s}_{13} \\ 0 & \tilde{s}_{22} & \tilde{s}_{23} \end{bmatrix} - \lambda \begin{bmatrix} \tilde{t}_{11} & \tilde{t}_{12} & \tilde{t}_{13} \\ 0 & \tilde{t}_{22} & \tilde{t}_{23} \end{bmatrix}, \tag{4.6}$$

where some $\tilde{s}_{ij}, \tilde{t}_{ij}$ may be zero. The size of the perturbations $(\delta\tilde{S}, \delta\tilde{T})$ imposed on $\tilde{S}$ and/or $\tilde{T}$ are displayed in Table 4.4. Case 1', which transits via the GUPTRI form of L_2, is retrieved by imposing a common column nullspace of the A- and B-parts of the deflated 1-by-2 pencil $[\tilde{s}_{22} \quad \tilde{s}_{23}] - \lambda[\tilde{t}_{22} \quad \tilde{t}_{23}]$. For cases 4 and 10 we retrieve the requested structures by setting elements $\tilde{s}_{12} = 0$ and $\tilde{t}_{12} = 0$, respectively, in the GUPTRI forms of $\tilde{A} - \lambda\tilde{B}$ (4.6). For case 4 we impose a zero multiple eigenvalue in $\tilde{A} - \lambda\tilde{B}$. Similarly, a multiple eigenvalue is imposed at infinity for case 10. In other

TABLE 4.4

Compound perturbations: Nongeneric structures imposed by perturbing the GUPTRI *form (denoted Transit form) of the generic or some nongeneric pencils.*

Case	KCF	Cod(·)	Transit form	$d_1\left(\begin{bmatrix} \tilde{s}_{22} & \tilde{s}_{23} \\ \tilde{t}_{22} & \tilde{t}_{23} \end{bmatrix}\right)$	$d_1\left(\begin{bmatrix} \tilde{s}_{12} & \tilde{s}_{13} \\ \tilde{t}_{12} & \tilde{t}_{13} \end{bmatrix}\right)$	$\tilde{s}_{12}$	$\tilde{t}_{12}$
$1'$	$L_1 \oplus R_1$	1	L_2	×			
4	$L_0 \oplus J_2$	4	$L_0 \oplus J_1 \oplus R_1$			×	
10	$L_0 \oplus N_2$	4	$L_0 \oplus N_1 \oplus R_1$				×
9	$2L_0 \oplus R_1 \oplus L_0^T$	7	$L_0 \oplus L_1 \oplus L_0^T$		×		

words, $J_1 \oplus R_1$ and $N_1 \oplus R_1$ in $\tilde{A} - \lambda\tilde{B}$ are turned J_2 and N_2, respectively. Case 9 is obtained by giving the A- and B-parts of the L_1 block in $\tilde{A} - \lambda\tilde{B}$ a common column nullspace, which turns L_1 into $L_0 \oplus R_1$. Since P and Q in (4.6) are unitary the perturbations imposed on $\tilde{A}$ and $\tilde{B}$ are of the same size as $\delta\tilde{S}$ and $\delta\tilde{T}$. The size of the compound (total) perturbations $(\delta A, \delta B)$ for the four cases is obtained by adding the appropriate perturbations in Tables 4.2–4.4. The perturbations explicitly imposed for the four cases in Table 4.4 increase the codimensions by one, except for case 9 where the rank drop by one increases the codimension by two.

The compound perturbations discussed above are all supposedly tractable, but are not necessarily optimal. A summary of the explicit perturbations in Tables 4.2–4.4 is displayed in a perturbation graph in Fig. 4.1, where the nodes are placed at the same positions as in the closure graph (Fig. 2.1). The paths to a node indicate different ways to generate the tractable perturbation required to find the KCF of the node, starting from a generic $A - \lambda B$. Notice that some arcs are marked with a bullet and the corresponding paths from a generic pencil to a destination KCF generate perturbations that are not necessarily optimal (compound perturbations from Tables 4.3 and 4.4). All other paths correspond to optimal perturbations from Table 4.2. We clarify the notation in Fig. 4.1 with two examples. Let $(\delta A_1, \delta B_1)$ denote the optimal perturbation of size $d_1(A)$ that for a generic $A - \lambda B$ gives $\tilde{A} - \lambda\tilde{B} = (A + \delta A_1) - \lambda(B + \delta B_1)$ the Kronecker structure $L_1 \oplus J_1$. Similarly, let $(\delta\tilde{A}_2, \delta\tilde{B}_2)$ denote the optimal perturbation of size $d_1(R[\tilde{A}, \tilde{B}, 0])$ that moves $\tilde{A} - \lambda\tilde{B}$ to a pencil with Kronecker structure $L_0 \oplus J_1 \oplus R_1$. Then $(\delta A_1 + \delta\tilde{A}_2, \delta B_1 + \delta\tilde{B}_2)$ is not necessarily the optimal perturbation for moving a generic pencil to orbit$(L_0 \oplus J_1 \oplus R_1)$. Therefore the arc to orbit$(L_0 \oplus J_1 \oplus R_1)$ is marked with a bullet. On the other hand, adding the perturbations going from orbit(L_2) to orbit$(L_0 \oplus 2J_1)$ via orbit$(L_1 \oplus J_1)$ gives us the optimal perturbation, which is already shown in Table 4.2.

In order to relate our explicit perturbations to the (labeled) closure graph we consider 2-dimensional rank drops in Table 4.2 as results of two 1-dimensional rank drops. In practice, these 2-dimensional rank drops are computed directly. Some of the perturbations in Table 4.2 do not give a unique path in the graph, since the generic $A - \lambda B$ in some cases is perturbed in A and B simultaneously. For these cases all alternative paths are shown in the graph, e.g., there are three different paths to $2L_0 \oplus J_1 \oplus L_0^T$ and all of them correspond to the same perturbation (in infinite arithmetic) of size $(d_2^2(A) + d_1^2(B))^{1/2}$. From the construction of the explicit perturbations it follows that each arc in the perturbation graph connects a KCF with another KCF within its orbit's closure. Therefore, for each arc in the perturbation graph there exists a corresponding path in the closure graph. It is of course possible to find other paths in the (labeled) closure graph that give tractable perturbations.

The sizes of the perturbations are shown on the corresponding arcs in the graph,

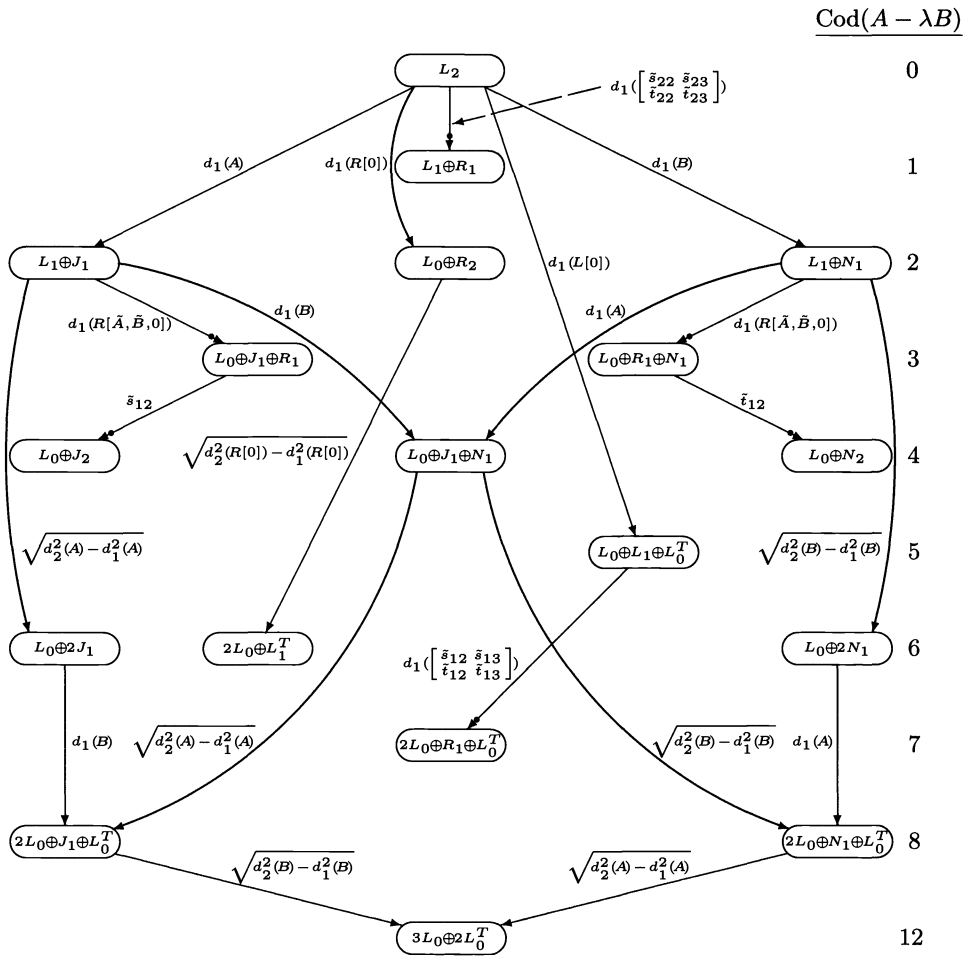


FIG. 4.1. *A graph displaying the tractable perturbations in Tables* 4.2–4.4 *of a generic* 2*-by-*3 *pencil.*

with notation as before. The reason for perturbation sizes such as $\sqrt{d_2^2(A) - d_1^2(A)}$ is that the total perturbation needed for this 2-dimensional rank drop in A is $d_2(A)$ (as shown in Table 4.2), but it is shown here as a further perturbation of a case where a perturbation of size $d_1(A)$ already has imposed a 1-dimensional rank drop in A.

For each case in Table 4.2 is shown in Fig. 4.1 as a compound perturbation, even though it can be computed directly, the size of the total perturbation is the square root of the sum of the squares of the sizes of the components of the perturbation. For example, the case $2L_0 \oplus J_1 \oplus L_0^T$ is found by a compound perturbation $(\delta A, \delta B) = (\delta A_1, \delta B_1) + (\delta A_2, \delta B_2) + (\delta A_3, \delta B_3)$, where $\|(\delta A_1, \delta B_1)\|_E = d_1(A)$, $\|(\delta A_2, \delta B_2)\|_E = \sqrt{d_2^2(A) - d_1^2(A)}$, and $\|(\delta A_3, \delta B_3)\|_E = d_1(B)$. The size of the total perturbation is $\|(\delta A, \delta B)\|_E = (d_1^2(A) + (d_2^2(A) - d_1^2(A)) + d_1^2(B))^{1/2} = (d_2^2(A) + d_1^2(B))^{1/2}$. Notably, since the perturbation $d_1(A) = \sigma_{\min}(A)$ and $d_2(A) = \|A\|_E = (\sigma_{\min-1}^2(A) + \sigma_{\min}^2(A))^{1/2}$, the size $\sqrt{d_2^2(A) - d_1^2(A)}$ is equal to $\sigma_{\min-1}(A)$.

For each compound perturbation in Tables 4.3 and 4.4, the size of the total

perturbation is found by adding the components of the perturbation and then computing the norm of the resulting perturbation. However, an upper bound on the size of the compound perturbation can be achieved by adding the sizes of the components of the perturbation. For example, $L_0 \oplus J_1 \oplus R_1$ is found by the compound perturbation $(\delta A, \delta B) = (\delta A_1, \delta B_1) + (\delta A_2, \delta B_2)$, where $\|(\delta A_1, \delta B_1)\|_E = d_1(A)$ and $\|(\delta A_2, \delta B_2)\|_E = d_1(R[\tilde{A}, \tilde{B}, 0])$, and an upper bound on $\|(\delta A, \delta B)\|_E$ is $d_1(A) + d_1(R[\tilde{A}, \tilde{B}, 0])$.

4.1.2. Intractable perturbations and the closest nongeneric structure. The following example shows a situation where the perturbations incidentally create extra nongeneric characteristics that raise the codimension of the perturbed pencil further than devised.

$$A = \begin{bmatrix} 0 & \epsilon_1 & 0 \\ 0 & 0 & \epsilon_2 \end{bmatrix}, \quad B = \begin{bmatrix} \epsilon_3 & 0 & 0 \\ 0 & \epsilon_4 & 0 \end{bmatrix}, \quad \epsilon_2 = \min_i \epsilon_i > 0. \tag{4.7}$$

Suppose we are looking for the minimal perturbations that impose the structure $L_1 \oplus J_1$ (case 2). They are of size $d_1(A)$ with

$$\delta A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\epsilon_2 \end{bmatrix}, \quad \delta B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Incidentally, δA and δB also lower the rank of $R[0]$. (For this example, δA and δB are the minimal perturbations that cause the rank drop, i.e., $d_1(A) = d_1(R[0])$, and the minima are attained for the same perturbations.) This fact implies that the perturbations aimed to impose the nongeneric structure $L_1 \oplus J_1$ (with codimension two) result in a perturbed pencil with two zero eigenvalues corresponding to the structure $L_0 \oplus J_2$ with codimension four (case 4). One possible remedy is to further perturb the undesired nongeneric pencil. To obtain $L_1 \oplus J_1$ we add, for example, the perturbations

$$\delta A' = \begin{bmatrix} \delta & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \delta B' = \begin{bmatrix} 0 & 0 & \delta \\ 0 & 0 & 0 \end{bmatrix},$$

to $(A + \delta A, B + \delta B)$, where $\delta > 0$ is an arbitrary small number. These perturbations remove the common column nullspace ($\delta B'$) and the multiple eigenvalue at zero ($\delta A'$), making the compound perturbations tractable. If we start to look for the smallest perturbations of (A, B) that impose a common column nullspace that would normally generate the structure $L_0 \oplus R_2$ (case 5), we also get intractable perturbations and (in this case) the same structure $L_0 \oplus J_2$. We can also see from the closure graph in Fig. 2.1 that $L_0 \oplus J_2$ is in the closure of each of the two orbits defined by $L_1 \oplus J_1$ and $L_0 \oplus R_2$.

Now we turn to the problem of finding the closest nongeneric Kronecker structure of a generic 2-by-3 pencil. Assume all inequalities relating to $d_1(R[1])$ in Theorem 4.1 are strict. Then the corresponding $R[A + \delta A, B + \delta B, 1]$ is rank deficient, and for all perturbations of size $\leq d_1(R[1])$, the (perturbed) matrices $A + \delta A, B + \delta B, R[A + \delta A, B + \delta B, 0]$ and $L[A + \delta A, B + \delta B, 0]$ must be of full rank, which correspond to the case $L_1 \oplus R_1$. Since all other nongeneric cases require rank-deficiency in at least one of the matrices $A + \delta A, B + \delta B, R[A + \delta A, B + \delta B, 0]$, or $L[A + \delta A, B + \delta B, 0]$ (see necessary conditions in the labeled closure graph in Fig. 2.2 or Table 2.3), we can formulate the following corollary.

COROLLARY 4.2. *If the inequalities* (4.2) *and* (4.3) *in Theorem* 4.1 *are strict,* $L_1 \oplus R_1$ *with codimension one* (*case* 1') *is the closest* (*unique*) *nongeneric structure on distance* $d_1(R[1])$.

The presumptions of Corollary 4.2 are sufficient (but not necessary) to identify tractable perturbations that lower the rank of $R[1]$. If equality holds in any of the inequalities of Theorem 4.1 (for the same perturbations $(\delta A, \delta B)$), we are faced with intractable perturbations which will result in nongeneric structures with higher codimensions. We collect the different cases in the following corollary, where we list the closest Kronecker structure and the corresponding equality conditions. Notice that strict inequalities are assumed otherwise.

COROLLARY 4.3. *Assume strict inequalities hold in Theorem* 4.1 *when nothing else is stated. Then, if*

1. $d_1(R[1]) = d_1(R[0])$, $L_0 \oplus R_2$ (*case* 5) *is the closest nongeneric form*;
2. $d_1(R[1]) = d_1(A)$, $L_1 \oplus J_1$ (*case* 2) *is the closest nongeneric form*;
3. $d_1(R[1]) = d_1(B)$, $L_1 \oplus N_1$ (*case* 6) *is the closest nongeneric form.*

All forms in Corollary 4.3 have codimension two. Notice that if there exist some perturbations on distance $d_1(R[1])$ that do not lower the rank of $R[0], A$, and B, respectively, then $L_1 \oplus R_1$ is also at the same distance as $L_0 \oplus R_2$, $L_1 \oplus J_1$, and $L_1 \oplus N_1$ for the three cases considered.

Assume that we can have equality in different combinations of the inequalities of Theorem 4.1. As before, we collect the possible cases in a corollary.

COROLLARY 4.4. *Assume two inequalities in Theorem* 4.1 *are satisfied with equality for the same perturbations* $(\delta A, \delta B)$. *Then, if*

1. $d_1(R[1]) = d_1(R[0]) = d_1(A)$, $L_0 \oplus J_1 \oplus R_1$ (*case* 4' *with codimension* 3) *or* $L_0 \oplus J_2$ (*case* 4 *with codimension* 4) *is the closest nongeneric structure*;
2. $d_1(R[1]) = d_1(R[0]) = d_1(B)$, $L_0 \oplus R_1 \oplus N_1$ (*case* 10' *with codimension* 3) *or* $L_0 \oplus N_2$ (*case* 10 *with codimension* 4) *is the closest nongeneric Kronecker structure.*

Notice that cases 4 and 10 have higher codimensions than cases 4' and 10,' respectively, but have the same algebraic characteristics in terms of the rank of $R[k]$ and $L[k]$ matrices as is seen in Table 2.3. The reason is that the 2-by-2 regular parts of cases 4 and 10 have one Jordan block with both eigenvalues specified, which increases the codimension by one compared to cases 4' and 10' (both with one eigenvalue unspecified).

The remark following Corollary 4.3 regarding a nonunique closest Kronecker structure can also be extended to apply to Corollary 4.4.

In applications (e.g., computing the uncontrollable subspace) we are interested in finding the most nongeneric structure (with highest codimension) for a given size of the perturbations. Is it possible to find intractable perturbations that result in a closest 2-by-3 nongeneric structure with codimension > 4? The answer is no, since all other cases require a rank drop of at least two in A, B or $R[0]$ or a simultaneous rank drop in A and B. There always exist strictly smaller perturbations that drop the rank by one (see (4.5)). Similar arguments also exclude $L_0 \oplus J_1 \oplus N_1$ with codimension 4 from being the closest nongeneric pencil.

4.1.3. Closest nongeneric structures to a generic 1-by-2 pencil. Since we do not know any explicit expression for $d_1(R[1])$, it is hard to construct examples that illustrate different situations described in §4.1.2. By considering 1-by-2 pencils we overcome this problem. A generic 1-by-2 pencil has the Kronecker structure $L_1 = [-\lambda \quad 1] = [0 \quad 1] - \lambda[1 \quad 0] \equiv A - \lambda B$. The nongeneric structures of size 1-by-2 are

$L_0 \oplus R_1, L_0 \oplus J_1, L_0 \oplus N_1$, and $2L_0 \oplus L_0^T$ with codimensions 1, 2, 2, and 4, respectively.

Which form(s) can be the closest nongeneric structure of a generic 1-by-2 pencil?

- $L_0 \oplus R_1$ if there exist perturbations of size $d_1(R[0])$ that do not simultaneously decrease the rank of A or B. This is, e.g., fulfilled if $d_1(R[0]) < \min(d_1(A), d_1(B))$.
- $L_0 \oplus J_1$ if $d_1(R[0]) = d_1(A)$.
- $L_0 \oplus N_1$ if $d_1(R[0]) = d_1(B)$.

Moreover, $2L_0 \oplus L_0^T$ can never be the closest nongeneric structure. The size of the minimal perturbations that turn A and B to zero matrices is $(d_1^2(A) + d_1^2(B))^{1/2}$.

The following example illustrates a case where $d_1(R[0]) = d_1(A) = d_1(B)$ and there exist perturbations of size $d_1(R[0])$ that do not simultaneously decrease the rank of A or B. Consequently, $L_0 \oplus R_1, L_0 \oplus J_1$, and $L_0 \oplus N_1$ are all the closest nongeneric Kronecker structure.

Let $A = [1 \quad 1]$ and $B = [-1 \quad 1]$. Then $R[0]$ has the singular value decomposition

$$R[0] \equiv \begin{bmatrix} A \\ B \end{bmatrix} = U\Sigma V^T \equiv \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{2} \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}.$$

$d_1(R[0])(= \sqrt{2})$ is attained for the (minimal) perturbations

$$\begin{bmatrix} \delta A \\ \delta B \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 1 & 0 \end{bmatrix},$$

while $A + \delta A$ and $B + \delta B$ remain full rank matrices, resulting in $L_0 \oplus R_1$ as the closest nongeneric structure. The perturbations

$$\begin{bmatrix} \delta A_1 \\ \delta B_1 \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} \delta A_2 \\ \delta B_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & -1 \end{bmatrix}$$

of the same minimal size make $R[0], A$ and $R[0], B$ drop rank, respectively. These perturbations generate the nongeneric structures $L_0 \oplus J_1$ and $L_0 \oplus N_1$, respectively.

4.2. Using GUPTRI to impose nongeneric structures. We have modified GUPTRI so that, for an $m \times n$ generic pencil $A - \lambda B$ as input, it is possible to impose a generalized Schur form with a specified Kronecker structure. (The modified GUPTRI also work for imposing a Kronecker structure of higher codimension on any nongeneric pencil.) Given the block indices that define the specified Kronecker structure (n_i's and r_i's of the RZ-staircase and LI-staircase forms [8]), GUPTRI imposes the necessary rank deflations in order to compute the specified (nongeneric) structure. The perturbations induced by these rank deflations are usually tractable. If the perturbations imposed by GUPTRI are intractable, GUPTRI computes the corresponding nongeneric structure of higher codimension. The resulting generalized Schur decomposition can be expressed in finite arithmetic as

$$P^H((A + \delta A) - \lambda(B + \delta B))Q = \begin{bmatrix} A_r - \lambda B_r & * & * \\ 0 & A_{\text{reg}} - \lambda B_{\text{reg}} & * \\ 0 & 0 & A_l - \lambda B_l \end{bmatrix}, \tag{4.8}$$

where $*$ denotes arbitrary conforming submatrices. Let δ_σ^2 denote the sum of the squares of all deleted singular values (imposed as zeros) during the reduction to GUPTRI form. Then δ_σ is an accurate estimate of $\|(\delta A, \delta B)\|_E$ in (4.8). One interpretation

is that GUPTRI computes an exact generalized Schur decomposition (with the specified Kronecker structure) for a pencil $A' - \lambda B'$ within distance δ_σ from the input pencil $A - \lambda B$. Moreover, δ_σ is an upper bound on the distance from $A - \lambda B$ to the nearest pencil with the Kronecker structure specified as input to GUPTRI.

Furthermore, this gives us a method for computing an upper bound on the distance from a generic m-by-n pencil to the closest nongeneric pencil.

- Compute the structure indices (n_i's and r_i's of the RZ-staircase and LI-staircase forms [8]) for all q structurally different nongeneric GUPTRI forms of size $m \times n$. This is a finite integer matching problem.
- Use the modified version of GUPTRI to impose the q nongeneric structures:

$$A_i - \lambda B_i = P_i^H((A + \delta A_i) - \lambda(B + \delta B_i))Q_i, \quad i = 1, \ldots, q. \tag{4.9}$$

- Compute the matrix pairs corresponding to the q nongeneric structures:

$$\hat{A}_i = P_i A_i Q_i^H, \quad \hat{B}_i = P_i B_i Q_i^H, \qquad i = 1, \ldots, q. \tag{4.10}$$

- Compute

$$\delta = \min_{1 \le i \le q} \delta_i, \quad \delta_i = \|(A - \hat{A}_i, B - \hat{B}_i)\|_E. \tag{4.11}$$

Now δ is an upper bound on the closest nongeneric pencil to $A - \lambda B$ and the δ_i's are upper bounds on the closest nongeneric pencils with the Kronecker structure of $A_i - \lambda B_i$ in (4.9).

The method described above is quite expensive already for moderate m and n (see §5) but is perfectly parallel. In a distributed memory environment it is possible to distribute the block indices for the different Kronecker structures evenly over the p ($\le q$) processors. Each processor also holds A and B and computes its local δ using the method above. Finally, a global minimum operation over all p processors gives us δ in (4.11).

4.3. Computational experiments on random 2-by-3 pencils. We have performed computational experiments on 100 random 2-by-3 pencils $A - \lambda B$. The elements of A and B are chosen uniformly distributed in $(0, 1)$. For each random pencil we impose the 17 nongeneric structures using the two approaches discussed in §§4.1 and 4.2.

Table 4.5 displays the mean values of perturbations required to impose each of the 17 nongeneric forms for 100 random examples. We measure the perturbations for each example and nongeneric form as $\|(A - \tilde{A}, B - \tilde{B})\|_E$, where $\tilde{A} - \lambda\tilde{B}$ denotes a nongeneric pencil. The matrices A and B are normalized such that $\|A\|_E = \|B\|_E$ and $\|(A, B)\|_E = 1$.

Columns 2 and 3 of Table 4.5 show the δ_i's in (4.11) computed by modified GUPTRI for the pencils $A - \lambda B$ and $B - \mu A$, respectively. Column 4 shows the explicit perturbations of Tables 4.2–4.4. The explicit perturbations that are proved to be the smallest possible are marked with the superscript *.

In Table 4.6 we display the smallest perturbations (measured as above) required to impose nongeneric forms of each possible codimension for the same 100 random 2-by-3 examples. For example, we have three nongeneric structures with codimension 2, so the smallest perturbations in this case are determined from 300 random examples. The singular structures (cases) that give the smallest perturbations are shown in columns directly following columns 2, 4, and 6 of Table 4.6.

TABLE 4.5
Mean values of perturbations (measured as $\|(A-\tilde{A}, B-\tilde{B})\|_F$*) required to impose each of the 17 nongeneric forms for 100 random* $A-\lambda B$ *of size 2-by-3.*

Case	$A-\lambda B$	$B-\mu A$	Explicit	Cod$(A-\lambda B)$	Comment
1	0.000	0.000	0.000	0	
1′	0.160	0.154	0.127	1	
2	**0.181**	0.394	0.181*	2	
6	0.378	**0.190**	0.190*	2	
5	0.235	0.227	0.140*	2	
4′	0.218	0.268	0.211	3	
10′	0.287	0.227	0.220	3	
4	*0.456*	0.533	0.461	4	
10	0.538	*0.481*	0.524	4	
7	0.437	0.434	0.281*	4	
7′	0.589	0.602	0.326*	5	
3	**0.707**	**0.707**	0.707*	6	$A = 0_{2\times 3}$
11	**0.707**	**0.707**	0.707*	6	$B = 0_{2\times 3}$
9′	0.399	0.399	0.353*	6	
9	0.466	0.460	0.390	7	
8	**0.737**	**0.737**	0.737*	8	$A = 0_{2\times 3}$
12	**0.736**	**0.736**	0.736*	8	$B = 0_{2\times 3}$
13	**1.000**	**1.000**	1.000*	12	$A = B = 0_{2\times 3}$

TABLE 4.6
Minimum perturbations (measured as $\|(A-\tilde{A}, B-\tilde{B})\|_F$*) required to impose nongeneric forms of each possible codimension for 100 random* $A-\lambda B$ *of size 2-by-3.*

Cod$(A-\lambda B)$	$A-\lambda B$	Case	$B-\mu A$	Case	Explicit	Case
0	0.000	1	0.000	1	0.000	1
1	$2\cdot 10^{-4}$	1′	$3\cdot 10^{-4}$	1′	$1\cdot 10^{-4}$	1′
2	0.011	2	0.010	5	0.009	5
3	**0.036**	4′	0.037	4′	0.036	4′
4	0.111	4	**0.106**	10	0.106	7
5	0.192	7′	**0.119**	7′	0.119	7′
6	0.163	9′	0.163	9′	0.153	9′
7	0.233	9	0.224	9	0.184	9
8	**0.707**	12	**0.707**	12	0.707	12
12	**1.000**	13	**1.000**	13	1.000	13

Numbers in bold font in Tables 4.5 and 4.6 indicate that the size of the perturbations (distances) computed by modified GUPTRI are the same as for the explicit perturbations, which for these cases are also shown to be the minimal perturbations. Numbers marked in italic font in Table 4.5 indicate that modified GUPTRI computed smaller upper bounds than corresponding bounds for the explicit perturbations.

All explicit perturbations of the 100 2-by-3 random pencils turned out to be tractable. The results show that the smallest distance from $A-\lambda B$ to a nongeneric structure with fixed codimension k increases with increasing k, in accordance with the Kronecker structure hierarchy in Fig. 2.1. Case 1' with KCF $L_1 \oplus R_1$ is the closest nongeneric pencil. Our explicit bound for case 1' is not proved to be the smallest possible.

5. Some comments on the general case. The complexity and the intricacies of the problems considered are well exposed in §§2–4. In the following we discuss some extensions to general m-by-n pencils. The number of different KCFs grows rapidly with increasing m and n. Some cases are displayed in Table 5.1.

We have been able to generate 20098 structurally different KCFs for $m = 10, n =$

TABLE 5.1
Number of structurally different Kronecker forms of size m-by-n ($m \le n$).

m	n: 1	2	3	4	5	6	7	8	9	10
1	4	5	5	5	5	5	5	5	5	5
2		14	18	19	19	19	19	19	19	19
3			41	54	58	59	59	59	59	59
4				110	145	159	163	164	164	164
5					271	358	397	411	415	416

20. Notice that for a given m the number of different structures is fixed for $n \ge 2m$. For $m > n$ the number of KCFs are the same as for the transposed pencil. As an example we show all structurally different 3-by-4 Kronecker forms in Table 5.2, where as before we let R_2 denote a 2-by-2 regular block with any nonzero finite eigenvalues (see §2.1) and, similarly, we let R_3 denote a regular 3-by-3 block.

TABLE 5.2
All 54 structurally different 3-by-4 pencils.

KCF		
L_3	$L_0 \oplus R_2 \oplus N_1$	$2L_0 \oplus L_2^T$
$L_2 \oplus N_1$	$L_0 \oplus R_3$	$2L_0 \oplus N_2 \oplus L_0^T$
$L_2 \oplus R_1$	$L_0 \oplus J_3$	$2L_0 \oplus N_1 \oplus L_1^T$
$L_2 \oplus J_1$	$L_0 \oplus J_2 \oplus N_1$	$2L_0 \oplus 2N_1 \oplus L_0^T$
$L_1 \oplus N_2$	$L_0 \oplus J_2 \oplus R_1$	$2L_0 \oplus R_1 \oplus L_1^T$
$L_1 \oplus 2N_1$	$L_0 \oplus J_1 \oplus N_2$	$2L_0 \oplus R_1 \oplus N_1 \oplus L_0^T$
$L_1 \oplus R_1 \oplus N_1$	$L_0 \oplus J_1 \oplus 2N_1$	$2L_0 \oplus R_2 \oplus L_0^T$
$L_1 \oplus R_2$	$L_0 \oplus J_1 \oplus R_1 \oplus N_1$	$2L_0 \oplus J_2 \oplus L_0^T$
$L_1 \oplus J_2$	$L_0 \oplus J_1 \oplus R_2$	$2L_0 \oplus J_1 \oplus L_1^T$
$L_1 \oplus J_1 \oplus N_1$	$L_0 \oplus J_1 \oplus J_2$	$2L_0 \oplus J_1 \oplus N_1 \oplus L_0^T$
$L_1 \oplus J_1 \oplus R_1$	$L_0 \oplus 2J_1 \oplus N_1$	$2L_0 \oplus J_1 \oplus R_1 \oplus L_0^T$
$L_1 \oplus 2J_1$	$L_0 \oplus 2J_1 \oplus R_1$	$2L_0 \oplus 2J_1 \oplus L_0^T$
$2L_1 \oplus L_0^T$	$L_0 \oplus 3J_1$	$2L_0 \oplus L_1 \oplus 2L_0^T$
$L_0 \oplus N_3$	$L_0 \oplus L_2 \oplus L_0^T$	$3L_0 \oplus L_0^T \oplus L_1^T$
$L_0 \oplus N_1 \oplus N_2$	$L_0 \oplus L_1 \oplus L_1^T$	$3L_0 \oplus N_1 \oplus 2L_0^T$
$L_0 \oplus 3N_1$	$L_0 \oplus L_1 \oplus N_1 \oplus L_0^T$	$3L_0 \oplus R_1 \oplus 2L_0^T$
$L_0 \oplus R_1 \oplus N_2$	$L_0 \oplus L_1 \oplus R_1 \oplus L_0^T$	$3L_0 \oplus J_1 \oplus 2L_0^T$
$L_0 \oplus R_1 \oplus 2N_1$	$L_0 \oplus L_1 \oplus J_1 \oplus L_0^T$	$4L_0 \oplus 3L_0^T$

It is possible to extend Theorem 4.1 to general m-by-$(m+1)$ pencils.

THEOREM 5.1. *For a generic m-by-$(m+1)$ pencil (A, B) the following inequalities hold:*

$$0 \equiv d_1(R[m]) < d_1(R[m-1]) \le \cdots \le d_1(R[0]), \tag{5.1}$$

$$d_1(R[m-1]) \le d_1(A), \quad d_1(R[m-1]) \le d_1(B), \tag{5.2}$$

$$d_1(R[m-1]) \le d_1(L[m-1]) \le \cdots \le d_1(L[0]), \tag{5.3}$$

$$\left.\begin{aligned} d_k(A) &< d_{k+1}(A) \\ d_k(B) &< d_{k+1}(B) \\ d_k(R[0]) &< d_{k+1}(R[0]) \end{aligned}\right\} \quad k = 1, \ldots, m-1. \tag{5.4}$$

Proof. From Theorem 2.1 it follows that $d_1(R[m]) = 0$ for all m-by-$(m+1)$ pencils (generic or nongeneric). A perturbation that lowers the column rank in $R[k-1]$ will always lower the rank in $R[k]$, since a dependence between columns in $R[k-1]$ will make the corresponding columns in

$$R[k] = \begin{bmatrix} A & 0 \\ B & R[k-1] \\ 0 & \end{bmatrix}$$

linearly dependent, proving (5.1). A perturbation that reduces the rank in A (or B) will cause a linear dependence among the m first (or last) rows of

$$R[m-1] = \begin{bmatrix} A & 0 \\ & \cdots \\ 0 & B \end{bmatrix}.$$

Since $R[m-1]$ is square $(m^2+m) \times (m^2+m)$, the row rank-deficiency is equivalent to $R[m-1]$ being column rank deficient, which proves (5.2). The relations between $d_1(L[k]), k = 0, \ldots, m-1$ in (5.3) can be similarly proved as the corresponding relations between the $R[k]$-matrices in (5.1). For the first inequality in (5.3) we recall the fact that a row rank-deficient $L[m-1]$ is equivalent to at least one L_k^T block ($k = 0, \ldots,$ or $m-1$) in the KCF. To match the dimensions of the pencil, the KCF must contain at least one L_i block ($i = 0, \ldots,$ or $m-2$) which is equivalent to $R[i]$ being column rank deficient. Hence row rank-deficient $L[m-1]$ is equivalent to $R[i]$ being column rank deficient for some $i = 0, \ldots, m-2$. Now, the first inequality of (5.3) is obtained by applying (5.1) to the relation between $R[i]$ and $R[m-1]$. As in Theorem 4.1, the inequalities (5.4) follow from the definition of $d_k(\cdot)$. □

We can see that the closest nongeneric structure to a generic m-by-$(m+1)$ pencil is on distance $d_1(R[m-1])$. Notably, when all inequalities relating to $d_1(R[m-1])$ in Theorem 5.1 are strict, (5.1) excludes any L_k blocks for $k < m-1$ in the KCF of any pencil on distance $d_1(R[m-1])$ from the generic case. Similarly, (5.2) excludes any J_i or N_i blocks, and (5.3) the existence of L_k^T blocks. Altogether, this extends Corollary 4.2 to m-by-$(m+1)$ pencils.

COROLLARY 5.2. *If all inequalities relating to* $d_1(R[m-1])$ *in Theorem* 5.1 *are strict, the closest nongeneric structure to a generic* m*-by-*$(m+1)$ *pencil is* $L_{m-1} \oplus R_1$ (*with codimension* 1) *on distance* $d_1(R[m-1])$.

Corollary 5.2 can be used to characterize the distance to uncontrollability for a single input single output linear system $E\dot{x}(t) = Fx(t) + Gu(t)$, where E and F are p-by-p matrices, G is p-by-1, and E is assumed to be nonsingular. The linear system is completely controllable (i.e., the dimension of the controllable subspace equals p) if and only if $A - \lambda B \equiv [G|F - \lambda E]$ is generic. Under the assumptions in Corollary 5.2, the closest uncontrollable system is on distance $d_1(R[p-1])$, corresponding to the nongeneric structure $L_{p-1} \oplus R_1$ (with the eigenvalue of R_1 finite and nonzero but otherwise unspecified).

Since B has full row rank $A - \lambda B \equiv [G|F - \lambda E]$ can have neither infinite eigenvalues nor L_j^T blocks in its KCF. Therefore, it can only have finite eigenvalues and L_j blocks in its KCF (and GUPTRI form) and the number of L_j blocks is equal to the number of columns of G. For $p = 2$ the possible uncontrollable systems correspond to cases 1', 2, 5, 4', 4 and 3 of Table 2.3.

Generalizations of Corollaries 4.3 and 4.4 to m-by-$(m+1)$ pencils are straightforward, but there are several more cases to distinguish. The formulations and technicalities are omitted here.

Some results for general matrix pencils relating to problems studied here are presented in [2]. Eigenvalue perturbation bounds are used to develop computational bounds on the distance from a given pencil to one with a qualitatively different Kronecker structure.

Appendix A. Proof of Theorem 2.2.

Proof. First we prove that each arc in the graph corresponds to a closure relation, and then we prove that these are all arcs that can exist. We prove that one KCF is in the closure of the orbit of another KCF by showing that the one in the closure is just a special case of the one defining the closure. We show proofs for each arc starting from the zero pencil.

Before looking at each arc we note that there is a symmetry regarding row ranks and column nullities between the Kronecker structures with J_i and N_i blocks replaced (see Table 2.3). From this we see that some of the proofs below that are shown for J_i blocks can be similarly done for the corresponding case with N_i blocks. Typically we must work with specific elements in A instead of B or vice versa. For these cases we will just mention this similarity without repeating the computations.

In the following, α, β, γ, δ, and ϵ are supposed to be nonzero elements when nothing else is stated.

- $3L_0 \oplus 2L_0^T$ is in the closure of orbit$(2L_0 \oplus J_1 \oplus L_0^T)$, since $3L_0 \oplus 2L_0^T$ is the special case $\alpha = 0$ of

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & 0 & \alpha \\ 0 & 0 & 0 \end{bmatrix},$$

  which is equivalent to $2L_0 \oplus J_1 \oplus L_0^T$ for all nonzero α.
- $3L_0 \oplus 2L_0^T$ is in the closure of orbit$(2L_0 \oplus N_1 \oplus L_0^T)$ follows from similar arguments based on the symmetry between J_i and N_i blocks.
- $2L_0 \oplus J_1 \oplus L_0^T$ is in the closure of orbit$(2L_0 \oplus R_1 \oplus L_0^T)$, since $2L_0 \oplus J_1 \oplus L_0^T$ is the special case $\alpha = 0$ of

$$\begin{bmatrix} 0 & 0 & \alpha \\ 0 & 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & 0 & \beta \\ 0 & 0 & 0 \end{bmatrix},$$

  which is equivalent to $2L_0 \oplus R_1 \oplus L_0^T$ for all nonzero α.
- $2L_0 \oplus N_1 \oplus L_0^T$ is in the closure of orbit$(2L_0 \oplus R_1 \oplus L_0^T)$ follows from similar arguments.
- $2L_0 \oplus J_1 \oplus L_0^T$ is in the closure of orbit$(L_0 \oplus 2J_1)$, since $2L_0 \oplus J_1 \oplus L_0^T$ is the special case $\alpha = 0$ of

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & 0 & \beta \\ 0 & \alpha & 0 \end{bmatrix},$$

  which multiplied by a permutation matrix can be shown to be equivalent to

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & \alpha & 0 \\ 0 & 0 & \beta \end{bmatrix},$$

  and this pencil is equivalent to $L_0 \oplus 2J_1$ for all nonzero α.

- $2L_0 \oplus N_1 \oplus L_0^T$ is in the closure of orbit$(L_0 \oplus 2N_1)$ follows from similar arguments.
- $2L_0 \oplus R_1 \oplus L_0^T$ is in the closure of orbit$(2L_0 \oplus L_1^T)$, since $2L_0 \oplus R_1 \oplus L_0^T$ is the special case $\beta = 0$ of

$$\begin{bmatrix} 0 & 0 & \alpha \\ 0 & 0 & \beta \end{bmatrix} - \lambda \begin{bmatrix} 0 & 0 & \gamma \\ 0 & 0 & 0 \end{bmatrix},$$

which for nonzero β is shown to be equivalent to $2L_0 \oplus L_1^T$ by the following equivalence transformation

$$\begin{bmatrix} \frac{1}{\gamma} & \frac{-\alpha}{\beta\gamma} \\ 0 & \frac{1}{\beta} \end{bmatrix} \left(\begin{bmatrix} 0 & 0 & \alpha \\ 0 & 0 & \beta \end{bmatrix} - \lambda \begin{bmatrix} 0 & 0 & \gamma \\ 0 & 0 & 0 \end{bmatrix} \right) \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

- $2L_0 \oplus R_1 \oplus L_0^T$ is in the closure of orbit$(L_0 \oplus L_1 \oplus L_0^T)$, since $2L_0 \oplus R_1 \oplus L_0^T$ is the special case $\beta = 0$ of

$$\begin{bmatrix} 0 & 0 & \alpha \\ 0 & 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & \beta & \gamma \\ 0 & 0 & 0 \end{bmatrix},$$

which for nonzero β is shown to be equivalent to $L_0 \oplus L_1 \oplus L_0^T$ by the following equivalence transformation

$$\begin{bmatrix} \frac{1}{\alpha} & 0 \\ 0 & 1 \end{bmatrix} \left(\begin{bmatrix} 0 & 0 & \alpha \\ 0 & 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & \beta & \gamma \\ 0 & 0 & 0 \end{bmatrix} \right) \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{\alpha}{\beta} & \frac{-\gamma}{\beta} \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

- $2L_0 \oplus L_1^T$ is in the closure of orbit$(L_0 \oplus J_1 \oplus N_1)$, since $2L_0 \oplus L_1^T$ is the special case $\gamma = 0$ of

$$\begin{bmatrix} 0 & 0 & \alpha \\ 0 & 0 & \beta \end{bmatrix} - \lambda \begin{bmatrix} 0 & \gamma & \delta \\ 0 & 0 & 0 \end{bmatrix}. \tag{A.1}$$

This is shown by the following equivalence transformation:

$$\begin{bmatrix} \frac{1}{\delta} & \frac{-\alpha}{\beta\delta} \\ 0 & \frac{1}{\beta} \end{bmatrix} \left(\begin{bmatrix} 0 & 0 & \alpha \\ 0 & 0 & \beta \end{bmatrix} - \lambda \begin{bmatrix} 0 & 0 & \delta \\ 0 & 0 & 0 \end{bmatrix} \right) \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix},$$

which is identical to $2L_0 \oplus L_1^T$. That the pencil (A.1) is equivalent to $L_0 \oplus J_1 \oplus N_1$ for all nonzero γ follows from the equivalence transformation:

$$\begin{bmatrix} 1 & \frac{-\alpha}{\beta} \\ 0 & \frac{1}{\beta} \end{bmatrix} \left(\begin{bmatrix} 0 & 0 & \alpha \\ 0 & 0 & \beta \end{bmatrix} - \lambda \begin{bmatrix} 0 & \gamma & \delta \\ 0 & 0 & 0 \end{bmatrix} \right) \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\gamma} & \frac{-\delta}{\gamma} \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{A.2}$$

- $2L_0 \oplus L_1^T$ is in the closure of orbit$(L_0 \oplus J_2)$, since $2L_0 \oplus L_1^T$ is a permutation of

$$\begin{bmatrix} 0 & 0 & \alpha \\ 0 & 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \beta \end{bmatrix},$$

which is the special case $\gamma = 0$ of

$$\begin{bmatrix} 0 & 0 & \alpha \\ 0 & 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & \gamma & 0 \\ 0 & 0 & \beta \end{bmatrix},$$

and this pencil is equivalent to $L_0 \oplus J_2$ for all nonzero γ.

- $2L_0 \oplus L_1^T$ is in the closure of orbit$(L_0 \oplus N_2)$ follows from similar arguments.
- $L_0 \oplus L_1 \oplus L_0^T$ is in the closure of orbit$(L_0 \oplus J_2)$, since $2L_0 \oplus L_1^T$ is the special case $\beta = 0$ of

$$\begin{bmatrix} 0 & 0 & \alpha \\ 0 & 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & \gamma & 0 \\ 0 & 0 & \beta \end{bmatrix},$$

which is equivalent to $L_0 \oplus J_2$ for all nonzero β.

- $L_0 \oplus L_1 \oplus L_0^T$ is in the closure of orbit$(L_0 \oplus N_2)$ follows from similar arguments.
- $L_0 \oplus L_1 \oplus L_0^T$ is in the closure of orbit$(L_0 \oplus J_1 \oplus N_1)$, since $L_0 \oplus L_1 \oplus L_0^T$ is the special case $\beta = 0$ of

$$\begin{bmatrix} 0 & 0 & \alpha \\ 0 & 0 & \beta \end{bmatrix} - \lambda \begin{bmatrix} 0 & \gamma & \delta \\ 0 & 0 & 0 \end{bmatrix}. \tag{A.3}$$

This follows from the equivalence transformation

$$\begin{bmatrix} \frac{1}{\alpha} & 0 \\ 0 & 1 \end{bmatrix} \left(\begin{bmatrix} 0 & 0 & \alpha \\ 0 & 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & \gamma & \delta \\ 0 & 0 & 0 \end{bmatrix} \right) \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{\alpha}{\gamma} & -\frac{\delta}{\gamma} \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

That (A.3) is equivalent to $L_0 \oplus J_1 \oplus N_1$ for nonzero β is shown in (A.2).

- $L_0 \oplus 2J_1$ is in the closure of orbit$(L_0 \oplus J_2)$, since $L_0 \oplus 2J_1$ is the special case $\alpha = 0$ of

$$\begin{bmatrix} 0 & 0 & \alpha \\ 0 & 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & \beta & 0 \\ 0 & 0 & \gamma \end{bmatrix},$$

which is equivalent to $L_0 \oplus J_2$ for all nonzero α.

- $L_0 \oplus 2N_1$ is in the closure of orbit$(L_0 \oplus N_2)$ follows from similar arguments.
- $L_0 \oplus J_2$ is in the closure of orbit$(L_0 \oplus J_1 \oplus R_1)$, since $L_0 \oplus J_2$ is the special case $\beta = 0$ of

$$\begin{bmatrix} 0 & 0 & \alpha \\ 0 & 0 & \beta \end{bmatrix} - \lambda \begin{bmatrix} 0 & \gamma & 0 \\ 0 & 0 & \delta \end{bmatrix},$$

which for nonzero β is shown to be equivalent to $L_0 \oplus J_1 \oplus R_1$ (with eigenvalue β/δ) by the following equivalence transformation

$$\begin{bmatrix} 1 & -\frac{\alpha}{\beta} \\ 0 & 1 \end{bmatrix} \left(\begin{bmatrix} 0 & 0 & \alpha \\ 0 & 0 & \beta \end{bmatrix} - \lambda \begin{bmatrix} 0 & \gamma & 0 \\ 0 & 0 & \delta \end{bmatrix} \right) \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\gamma} & \frac{\alpha}{\beta\gamma} \\ 0 & 0 & \frac{1}{\delta} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \frac{\beta}{\delta} \end{bmatrix} - \lambda \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

- $L_0 \oplus N_2$ is in the closure of orbit$(L_0 \oplus N_1 \oplus R_1)$ follows from similar arguments.
- $L_0 \oplus J_1 \oplus N_1$ is in the closure of orbit$(L_0 \oplus J_1 \oplus R_1)$, since $L_0 \oplus J_1 \oplus N_1$ is the special case $\gamma = 0$ of

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \alpha \end{bmatrix} - \lambda \begin{bmatrix} 0 & \beta & 0 \\ 0 & 0 & \gamma \end{bmatrix},$$

which is equivalent to $L_0 \oplus J_1 \oplus R_1$ for all nonzero γ.

- $L_0 \oplus J_1 \oplus N_1$ is in the closure of orbit$(L_0 \oplus N_1 \oplus R_1)$ follows from similar arguments.
- $L_0 \oplus J_1 \oplus R_1$ is in the closure of orbit$(L_1 \oplus J_1)$, since $L_0 \oplus J_1 \oplus R_1$ is equivalent to $L_0 \oplus R_1 \oplus J_1$, which is the special case $\alpha = 0$ of

$$\begin{bmatrix} \alpha & \beta & 0 \\ 0 & 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & \gamma & 0 \\ 0 & 0 & \delta \end{bmatrix},$$

which for nonzero α is shown to be equivalent to $L_1 \oplus J_1$ by the following equivalence transformation

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \left(\begin{bmatrix} \alpha & \beta & 0 \\ 0 & 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 0 & \gamma & 0 \\ 0 & 0 & \delta \end{bmatrix} \right) \begin{bmatrix} -\frac{\beta}{\alpha\gamma} & \frac{1}{\alpha} & 0 \\ \frac{1}{\gamma} & 0 & 0 \\ 0 & 0 & \frac{1}{\delta} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

- $L_0 \oplus N_1 \oplus R_1$ is in the closure of orbit$(L_1 \oplus N_1)$ follows from similar arguments.
- $L_0 \oplus J_1 \oplus R_1$ is in the closure of orbit$(L_0 \oplus R_2)$, since $L_0 \oplus J_1 \oplus R_1$ is the special case $\alpha = 0$ of

$$\begin{bmatrix} 0 & \alpha & 0 \\ 0 & 0 & \beta \end{bmatrix} - \lambda \begin{bmatrix} 0 & \gamma & 0 \\ 0 & 0 & \delta \end{bmatrix},$$

which is equivalent to $L_0 \oplus R_2$ for all nonzero α.
- $L_0 \oplus J_1 \oplus R_1$ is in the closure of orbit$(L_0 \oplus R_2)$ follows from similar arguments.
- $L_1 \oplus J_1$ is in the closure of orbit$(L_1 \oplus R_1)$, since $L_1 \oplus J_1$ is the special case $\beta = 0$ of

$$\begin{bmatrix} \alpha & 0 & 0 \\ 0 & 0 & \beta \end{bmatrix} - \lambda \begin{bmatrix} 0 & \gamma & 0 \\ 0 & 0 & \delta \end{bmatrix},$$

which is equivalent to $L_1 \oplus R_1$ for all nonzero β.
- $L_1 \oplus N_1$ is in the closure of orbit$(L_1 \oplus R_1)$ follows from similar arguments.
- $L_0 \oplus R_2$ is in the closure of orbit$(L_1 \oplus R_1)$, since $L_0 \oplus R_2$ is the special case $\alpha = 0$ of

$$\begin{bmatrix} \alpha & \beta & 0 \\ 0 & 0 & \gamma \end{bmatrix} - \lambda \begin{bmatrix} 0 & \delta & 0 \\ 0 & 0 & \epsilon \end{bmatrix},$$

which for nonzero α is shown to be equivalent to $L_1 \oplus R_1$ (with eigenvalue γ/ϵ) by the following equivalence transformation

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \left(\begin{bmatrix} \alpha & \beta & 0 \\ 0 & 0 & \gamma \end{bmatrix} - \lambda \begin{bmatrix} 0 & \delta & 0 \\ 0 & 0 & \epsilon \end{bmatrix} \right) \begin{bmatrix} \frac{1}{\alpha} & -\frac{\beta}{\alpha\delta} & 0 \\ 0 & \frac{1}{\delta} & 0 \\ 0 & 0 & \frac{1}{\epsilon} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & \frac{\gamma}{\epsilon} \end{bmatrix} - \lambda \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

- $L_1 \oplus R_1$ is in the closure of orbit(L_2), since L_2 spans the complete 12-dimensional space.

Now we have shown that all arcs in the graph are valid. It remains to show that there are no arcs missing. This can be done by examining the KCFs that cannot be in the closure of each other.

First we remark that one necessary condition for a KCF to be in the closure of the orbit of another is that it must have higher codimension than the one defining the closure.

Since $L_0 \oplus N_1 \oplus R_1$, $L_0 \oplus N_2$, and $L_0 \oplus 2N_1$ all require that A has full rank $(= 2)$, none of them can be in the closure of orbit$(L_1 \oplus J_1)$, since that KCF requires A to have rank $= 1$. (Of course this also implies that none of these three KCFs can be in the closure of the orbit of $L_0 \oplus J_1 \oplus R_1$, $L_0 \oplus 2J_1$, or any other KCF that is in the closure of orbit$(L_1 \oplus J_1)$.)

From the symmetry for J_i and N_i blocks, we see that neither $L_0 \oplus J_1 \oplus R_1$, nor $L_0 \oplus J_2$, nor $L_0 \oplus 2J_1$ can be in the closure of orbit$(L_1 \oplus N_1)$, since they require B to have full rank and $L_1 \oplus N_1$ has rank$(B) = 1$.

Since $2L_0 \oplus J_1 \oplus L_0^T$ and $2L_0 \oplus R_1 \oplus L_0^T$ have a B of rank 1, neither of them can be in the closure of orbit$(L_0 \oplus 2N_1)$ since that KCF requires a 2-dimensional rank deficiency in B. By similar arguments for the rank of A we see that $2L_0 \oplus N_1 \oplus L_0^T$ and $2L_0 \oplus R_1 \oplus L_0^T$ cannot be in the closure of orbit$(L_0 \oplus 2J_1)$. Since we have investigated all presumptive KCFs the proof is complete. □

Acknowledgments. We are grateful to Alan Edelman and the referees for constructive comments, which have improved both the content and the organization of the paper.

REFERENCES

[1] T. Beelen and P. van Dooren, *An improved algorithm for the computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 105 (1988), pp. 9–65.

[2] D. L. Boley, *Estimating the sensitivity of the algebraic structure of pencils with simple eigenvalue estimates*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 632–643.

[3] J. Demmel and A. Edelman, *The dimension of matrices (matrix pencils) with given Jordan (Kronecker) canonical forms*, Report LBL-31839, Mathematics Department, Lawrence Berkeley Laboratories, University of California, Berkeley, 1992, Linear Algebra Appl., to appear.

[4] J. Demmel and B. Kågström, *Stably computing the Kronecker structure and reducing subspaces of singular pencils $A - \lambda B$ for uncertain data*, in Large Scale Eigenvalue Problems, J. Cullum and R. A. Willoughby, eds., North-Holland, Amsterdam, 1986, pp. 283–323. Mathematics Studies Series Vol. 127, Proceedings of the IBM Institute Workshop on Large Scale Eigenvalue Problems, July 8-12, 1985, Oberlech, Austria.

[5] ———, *Computing stable eigendecompositions of matrix pencils*, Linear Algebra Appl., 88/89 (1987), pp. 139–186.

[6] ———, *Accurate solutions of ill-posed problems in control theory*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 126–145.

[7] ———, *The generalized Schur decomposition of an arbitrary pencil $A - \lambda B$: Robust software with error bounds and applications. Part* I*: Theory and algorithms*, ACM Trans. Math. Software, 19 (1993), pp. 160–174.

[8] ———, *The generalized Schur decomposition of an arbitrary pencil $A - \lambda B$: Robust software with error bounds and applications. Part* II*: Software and applications*, ACM Trans. Math. Software, 19 (1993), pp. 175–201.

[9] F. Gantmacher, *The Theory of Matrices*, Vols. I and II (transl.), Chelsea, New York, 1959.

[10] G. Golub and C. Van Loan, *Matrix Computations*, Second Edition. Johns Hopkins University Press, Baltimore, MD, 1989.

[11] B. Kågström, *The generalized singular value decomposition and the general $A - \lambda B$ problem*, BIT, 24 (1984), pp. 568–583.

[12] ———, RGSVD—*an algorithm for computing the Kronecker canonical form and reducing subspaces of singular matrix pencils $A - \lambda B$*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 185–211.

[13] V. B. Khazanov and V. Kublanovskaya, *Spectral problems for matrix pencils. Methods and algorithms.* I., Sov. J. Numer. Anal. Math. Modelling, 3 (1988), pp. 337–371.

[14] V. KUBLANOVSKAYA, *AB-algorithm and its modifications for the spectral problem of linear pencils of matrices*, Numer. Math., 43 (1984), pp. 329–342.

[15] C. PAIGE, *Properties of numerical algorithms related to computing controllability*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 130–138.

[16] P. VAN DOOREN, *The computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 27 (1979), pp. 103–141.

[17] ———, *The generalized eigenstructure problem in linear system theory*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 111–129.

[18] ———, *Reducing subspaces: Definitions, properties and algorithms*, in Matrix Pencils, B. Kågström and A. Ruhe, eds., Springer-Verlag, Berlin, 1983, pp. 58–73. Lecture Notes in Mathematics, Vol. 973, Proceedings, Pite Havsbad, 1982.

[19] W. WATERHOUSE, *The codimension of singular matrix pairs*, Linear Algebra Appl., 57 (1984), pp. 227–245.

[20] J. H. WILKINSON, *Linear differential equations and Kronecker's canonical form*, in Recent Advances in Numerical Analysis, C. de Boor and G. Golub, eds., Academic Press, 1978, pp. 231–265.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 1, pp. 35–46, January 1996

ON THE STABILITY OF CHOLESKY FACTORIZATION FOR SYMMETRIC QUASIDEFINITE SYSTEMS*

PHILIP E. GILL†, MICHAEL A. SAUNDERS‡, AND JOSEPH R. SHINNERL†

Abstract. Sparse linear equations $Kd = r$ are considered, where K is a specially structured symmetric *indefinite* matrix that arises in numerical optimization and elsewhere. Under certain conditions, K is *quasidefinite*. The Cholesky factorization $PKP^T = LDL^T$ is then known to exist for any permutation P, even though D is indefinite.

Quasidefinite matrices have been used successfully by Vanderbei within barrier methods for linear and quadratic programming. An advantage is that for a sequence of K's, P may be chosen once and for all to optimize the sparsity of L, as in the positive-definite case.

A preliminary stability analysis is developed here. It is observed that a quasidefinite matrix is closely related to an unsymmetric positive-definite matrix, for which an LDM^T factorization exists. Using the Golub and Van Loan analysis of the latter, conditions are derived under which Cholesky factorization is stable for quasidefinite systems. Some numerical results confirm the predictions.

Key words. indefinite systems, symmetric quasidefinite (sqd) systems, unsymmetric positive-definite systems, backward stability, condition number, barrier methods, linear programming

AMS subject classifications. 49D37, 65F05, 65K05, 90C30

1. Introduction.

We define a matrix K to be *symmetric quasidefinite* (sqd) if there exists a permutation matrix Π that reorders K to the form

$$\Pi K \Pi^T = \begin{pmatrix} H & A^T \\ A & -G \end{pmatrix}, \tag{1.1}$$

where $H \in \mathbb{R}^{n\times n}$ and $G \in \mathbb{R}^{m\times m}$ are symmetric and positive definite. Such a K is indefinite and nonsingular. Vanderbei [Van91], [Van94] has shown that sqd matrices are *strongly factorizable*; i.e., for *every* permutation P there exist a diagonal D and a unit lower-triangular L such that

$$PKP^T = LDL^T. \tag{1.2}$$

We refer to (1.2) as a Cholesky factorization, while emphasizing that K is indefinite and D has both positive and negative diagonals. The usual stability analysis therefore does not apply, and the factorization may be unstable.

An example sqd matrix is

$$K = \begin{pmatrix} 1 & 1 \\ 1 & -\epsilon \end{pmatrix} = \begin{pmatrix} 1 & \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & \\ & -(1+\epsilon) \end{pmatrix} \begin{pmatrix} 1 & 1 \\ & 1 \end{pmatrix}. \tag{1.3}$$

The Cholesky factors exist for all values of ϵ, and can be computed accurately in finite precision for any ϵ. The symmetrically permuted system

$$PKP^T = \begin{pmatrix} -\epsilon & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & \\ -\frac{1}{\epsilon} & 1 \end{pmatrix} \begin{pmatrix} -\epsilon & \\ & 1+\frac{1}{\epsilon} \end{pmatrix} \begin{pmatrix} 1 & -\frac{1}{\epsilon} \\ & 1 \end{pmatrix} \tag{1.4}$$

* Received by the editors July 13, 1993; accepted for publication (in revised form) by S. Hammarling January 17, 1995. This research was supported by Department of Energy contract DE-FG03-92ER25117, National Science Foundation grants DMI-9204208 and DMI-9204547, and Office of Naval Research grant N00014-90-J-1242.

† Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0012 (pgill@ucsd.edu and jshinnerl@ucsd.edu).

‡ Department of Operations Research, Stanford University, Stanford, CA 94305-4022 (mike@sol-michael.stanford.edu).

has Cholesky factors for any nonzero ϵ, but as noted in [Van91], the factorization becomes unstable in finite-precision arithmetic as $|\epsilon| \to 0$.

Strong factorizability is particularly attractive when K is large and sparse and a direct factorization method is used to solve the linear system of equations

$$Kd = r. \tag{1.5}$$

As with positive-definite systems, we choose P in (1.2) to reduce fill-in during the Cholesky factorization. If several similar systems are to be solved, we would like to use the same "ideal" P for each system, as long as the associated factorizations are stable. In this paper we examine conditions under which Cholesky factorization may be used reliably on an sqd matrix K. For the example in (1.3)–(1.4), the analysis predicts (of course) that $|\epsilon|$ should not be too small. For certain systems arising in constrained optimization, it predicts that Cholesky factorization should be stable until the iterates are in a small neighborhood of the solution.

1.1. Notation. When discussing permutations P, we speak of "sparsity interchanges" and "stability interchanges" to indicate the usual criteria for choosing P. The spectral condition number is $\kappa_2(K) \equiv \|K\|_2\|K^{-1}\|_2$. The following symbols are used for matrices:

- A, G, and H are the block components of an arbitrary sqd matrix K.
- B is an arbitrary square nonsingular matrix whose triangular factorization $B = LU = LDM^T$ exists in exact arithmetic without row or column interchanges. (D is diagonal, L and M are unit lower triangular, and $DM^T \equiv U$. Although such a factorization does not exist for all nonsingular B, when it does exist, the factors are unique.)
- C is a square matrix that is unsymmetric but positive definite.
- T and S are the symmetric and skew-symmetric parts of C: $T = (C+C^T)/2$, $S = (C - C^T)/2$, and $C = T + S$.
- LDL^T denotes Cholesky factors of a symmetric matrix: L unit triangular, D diagonal and possibly *indefinite*.
- LBL^T denotes factors of a symmetric indefinite matrix: L unit triangular, B block-diagonal with blocks of order 1 or 2.

2. Connection with the unsymmetric positive-definite case. We seek conditions on sqd matrices K that allow stable computation of the Cholesky factorization $PKP^T = LDL^T$ for every permutation P. There is no loss of generality in assuming $\Pi = I$ in (1.1). With this convention, observe that

$$K = \begin{pmatrix} H & A^T \\ A & -G \end{pmatrix} \equiv \bar{K}\bar{I}, \tag{2.1}$$

where

$$\bar{K} \equiv \begin{pmatrix} H & -A^T \\ A & G \end{pmatrix} \quad \text{and} \quad \bar{I} \equiv \begin{pmatrix} I_n & \\ & -I_m \end{pmatrix}, \tag{2.2}$$

where I_n and I_m denote the identity matrices of order n and m. The matrix $\bar{K}$ is *unsymmetric positive definite*; i.e., $x^T\bar{K}x > 0$ for all nonzero x. The main idea of this paper is that (2.1) can be used to characterize the stability of algorithms for symmetric quasidefinite matrices in terms of the stability of Gaussian elimination for unsymmetric positive-definite matrices.

With $\bar{I}$ as above, let $\tilde{I} \equiv P\bar{I}P^T$ for some permutation P. The matrix $\tilde{I}$ is diagonal with diagonal entries 1 and -1; thus, in any product of the form $\tilde{A} = A\tilde{I}$, $\tilde{A}$ is equal to A with some of its columns scaled by -1. Now for every permutation P,

$$PKP^T = P\bar{K}\bar{I}P^T = P\bar{K}P^T(P\bar{I}P^T) = (P\bar{K}P^T)\tilde{I}. \tag{2.3}$$

It follows that

$$PKP^T = LDL^T \quad \text{if and only if} \quad P\bar{K}P^T = L\tilde{D}M^T, \tag{2.4}$$

where $\tilde{D} \equiv D\tilde{I}$ and $M \equiv \tilde{I}L\tilde{I}$. The matrices D and $\tilde{D}$ are diagonal, and L and M are unit lower triangular as required. Relations (2.1) and (2.4) can be construed as an alternative proof of Vanderbei's theorem on the strong factorizability of symmetric quasidefinite matrices. For if K is sqd, then $\bar{K}$ and hence $P\bar{K}P^T$ are positive definite; therefore the LU factorization of $P\bar{K}P^T$ exists (cf. [GV89, p. 140]); hence, by (2.4), the LDL^T factorization of PKP^T exists as well.

Since only column signs are involved, it is trivial to show that (2.4) holds in finite precision. If $\hat{L}$ and $\hat{D}$ are the computed factors of PKP^T, then $\hat{L}$, $\hat{D}\tilde{I}$, and $\hat{M}^T = \tilde{I}\hat{L}^T\tilde{I}$ are the computed LU factors of $P\bar{K}P^T$. Hence any conditions that ensure stability for the factorization $P\bar{K}P^T = LDM^T$ will also ensure stability for $PKP^T = LDL^T$. In particular, it is safe to factor the quasidefinite matrix PKP^T *without stability interchanges* if and only if it is safe to factor the unsymmetric positive-definite matrix $P\bar{K}P^T$ without stability interchanges.

3. When stability interchanges are unnecessary. Throughout this section, we assume that C is an unsymmetric positive-definite matrix. Let T and S be the symmetric and skew-symmetric parts of C. Then it is safe to factor C without stability interchanges if

(i) S is not too large compared to T; and
(ii) T is not too ill-conditioned.

This follows from results of Golub and Van Loan [GV79], [GV89], which we summarize next.

3.1. Theorems of Golub and Van Loan. Let $C = LDM^T$. In the backward error analysis of Gaussian elimination, it is shown that the computed solution $\hat{x}$ to the system $Cx = r$ is the exact solution of the perturbed system $(C + \Delta C)\hat{x} = r$, where the size of ΔC is bounded by an expression involving the sizes of the computed factors of C; say, $\hat{L}$, $\hat{D}$, and $\hat{M}^T$ (cf. (3.1) below). Algorithms that produce $\hat{L}$, $\hat{D}$, and $\hat{M}^T$ of sufficiently bounded size are therefore considered stable.

For general C, row or column interchanges are necessary to ensure the existence of the factors, and to prevent them from having large elements. For positive-definite C, however, the following theorems can be used with Assumption 3.1 to obtain a satisfactory bound on the sizes of the computed factors without stability interchanges.

(When applied to vectors or matrices, the symbols $|\cdot|$ and $\leq$ are to be interpreted componentwise. The symbol $\mathbf{u}$ denotes the unit round-off, and all floating-point calculations are assumed to conform to the "standard model" described in [GV89, pp. 61–62].)

ASSUMPTION 3.1 (see [GV89, p. 141]). *For some scalar γ of moderate size,*

$$\| |\hat{L}|\,|\hat{D}|\,|\hat{M}^T| \|_F \leq \gamma \| |L|\,|D|\,|M^T| \|_F.$$

THEOREM 3.1 (see [GV79, p. 88]). *Let $C \in \mathbb{R}^{n\times n}$ be positive definite and set $T = (C + C^T)/2$ and $S = (C - C^T)/2$. If $C = LDM^T$, then*

$$\| |L|\,|D|\,|M^T| \|_F \leq n \left(\|T\|_2 + \|ST^{-1}S\|_2 \right).$$

THEOREM 3.2 (see [GV89, p. 136, Eqn. (4.1.3)]). *Let $B \in \mathbb{R}^{n\times n}$ be a matrix whose LDM^T factorization exists, and let $\hat{L}$, $\hat{D}$, and $\hat{M}$ be the computed factors. Let $\hat{x}$ denote the computed solution to the system $Bx = b$, obtained by the usual methods of forward and backward substitution* (cf. [GV89, p. 97, Algorithm 3.2.3]). *Then $(B + \Delta B)\hat{x} = b$, with*

$$|\Delta B| \leq n\mathbf{u} \left(3|B| + 5|\hat{L}|\,|\hat{D}|\,|\hat{M}^T| \right) + \mathcal{O}(\mathbf{u}^2). \tag{3.1}$$

From Assumption 3.1 and these theorems, it follows that the computed solution $\hat{x}$ to the positive-definite system $Cx = r$ satisfies $(C + \Delta C)\hat{x} = r$, with

$$\|\Delta C\|_2 \leq \|\Delta C\|_F \leq \mathbf{u} \left(3n\|C\|_F + 5\gamma n^2 \left(\|T\|_2 + \|ST^{-1}S\|_2 \right) \right) + \mathcal{O}(\mathbf{u}^2). \tag{3.2}$$

Since $\|T\|_2 \leq \|C\|_2$, we have

$$\|T\|_2 + \|ST^{-1}S\|_2 \leq \left(1 + \frac{\|ST^{-1}S\|_2}{\|C\|_2} \right) \|C\|_2. \tag{3.3}$$

RESULT 3.1 (see [GV79, p. 92] and [GV89, p. 141]). *If C is positive definite, the factorization $C = LDM^T$ is stable if $\omega(C)$ is not too large, where*

$$\omega(C) \equiv \frac{\|ST^{-1}S\|_2}{\|C\|_2}. \tag{3.4}$$

3.2. An alternative indicator. Because it may not always be clear how the structure of the matrix $ST^{-1}S$ depends on the structure of the original matrix C, we observe that $\omega(C) \leq \theta(C)$, where $\theta(C)$ is defined next.

RESULT 3.2. *If C is positive definite, the factorization $C = LDM^T$ is stable if $\theta(C)$ is not too large, where*

$$\theta(C) \equiv \left(\frac{\|S\|_2}{\|T\|_2} \right)^2 \kappa_2(T). \tag{3.5}$$

When $\|S\|$ is not much larger than $\|T\|$, and T is not too ill-conditioned, $\theta(C)$ may provide an adequate guarantee of numerical stability. The straightforward dependence of $\theta(C)$ on T and S makes it easier to estimate than $\omega(C)$.

In certain contexts, however, $\theta(C)$ may be arbitrarily larger than $\omega(C)$. For example, suppose C has the form of $\bar{K}$ in (2.2), with $H = \beta I$, $G = (1/\beta)I$, and $A = \beta^2 I$. It is easily shown that as $\beta \to \infty$, $\theta(C) = \mathcal{O}(\beta)\omega(\mathcal{C})$. Thus, a large value of $\theta(C)$ should not be automatically interpreted to mean that stability interchanges are necessary.

4. Application to quasidefinite matrices. For our purposes, the role of C is played by $P\bar{K}P^T$ in §2. Since it is easily shown that ω and θ are invariant under symmetric permutations of their arguments, we assume $C = \bar{K}$ in (2.2). In this case,

$$T = \frac{1}{2}(\bar{K} + \bar{K}^T) = \begin{pmatrix} H & \\ & G \end{pmatrix} \quad \text{and} \quad S = \frac{1}{2}(\bar{K} - \bar{K}^T) = \begin{pmatrix} & -A^T \\ A & \end{pmatrix},$$

so that

$$ST^{-1}S = \begin{pmatrix} -A^TG^{-1}A & \\ & -AH^{-1}A^T \end{pmatrix}.$$

From the Golub and Van Loan analysis, stability of the factorization can be guaranteed if $\omega(\bar{K}) \equiv \|ST^{-1}S\|_2/\|\bar{K}\|_2$ is not too large. In terms of K rather than $\bar{K}$, we therefore have the following result for sqd matrices of the form (1.1).

RESULT 4.1. *If K is sqd, the factorization $PKP^T = LDL^T$ is stable for every permutation P if $\omega(K)$ is not too large, where*

$$\omega(K) \equiv \frac{\max\{\|A^TG^{-1}A\|_2, \|AH^{-1}A^T\|_2\}}{\|K\|_2}. \tag{4.1}$$

As in (3.4)–(3.5), we have $\omega(K) \le \theta(K)$, where the latter is readily computed in terms of A, H, H^{-1}, G, and G^{-1}.

RESULT 4.2. *If K is sqd, the factorization $PKP^T = LDL^T$ is stable for every permutation P if $\theta(K)$ is not too large, where*

$$\theta(K) \equiv \left(\frac{\|A\|_2}{\max\{\|G\|_2, \|H\|_2\}}\right)^2 \max\{\kappa_2(G), \kappa_2(H)\}. \tag{4.2}$$

For example, suppose $\|H\|_2 \ge \|G\|_2$ and $\|G^{-1}\|_2 \ge \|H^{-1}\|_2$. Then

$$\theta(K) \le \frac{\|A\|_2^2}{\|H\|_2}\|G^{-1}\|_2.$$

In general,

(i) $\|A\|_2$ must not be too large compared to $\|H\|_2$ and $\|G\|_2$; and
(ii) diag(H, G) must not be too ill-conditioned.

5. The condition number of a quasidefinite system. To assess the accuracy of computed solutions to $Kd = r$ with K sqd as in (2.1), we must consider both the backward stability of the factorization $PKP^T = LDL^T$ and the forward sensitivity of d to perturbations in K. That is, given that our computed solution $\hat{d}$ satisfies the perturbed system

$$(K + \Delta K)\,\hat{d} = r, \tag{5.1}$$

how close is $\hat{d}$ to d, the true solution? The usual sensitivity bound takes the form

$$\frac{\|d - \hat{d}\|}{\|d\|} \le \frac{\alpha}{1 - \alpha}, \quad \text{where} \quad \alpha = \frac{\|\Delta K\|}{\|K\|}\kappa_2(K). \tag{5.2}$$

For general K, the relative perturbation $\|\Delta K\|/\|K\|$ cannot be suitably bounded without the use of stability interchanges. When K is sqd, however, (3.2) and (3.3) give a bound on this perturbation that is essentially proportional to $1+\omega(C)$, with $C = \bar{K}$ (2.2).

Combining the results of §§3–4, we obtain the following in terms of K rather than $\bar{K}$. (Let $\hat{L}$ and $\hat{D}$ be the computed factors of K, and note that $\kappa_2(\bar{K}) = \kappa_2(K) = \kappa_2(PKP^T)$.)

ASSUMPTION 5.1. *For some scalar γ of moderate size,*

$$\| |\hat{L}|\,|\hat{D}|\,|\hat{L}^T| \|_F \le \gamma \| |L|\,|D|\,|L^T| \|_F.$$

THEOREM 5.1. *If K is symmetric quasidefinite as in* (2.1), *and if $\hat{d}$ is the computed solution of $Kd = r$,*

$$\frac{\|d - \hat{d}\|}{\|d\|} \le \mathbf{u}\gamma n_K c_K \phi(K), \tag{5.3}$$

where n_K is the dimension of K, c_K depends linearly on n_K, $\omega(K)$ is defined in (4.1), *and*

$$\phi(K) \equiv (1+\omega(K))\,\kappa_2(K). \tag{5.4}$$

A similar result holds with $\omega(K)$ replaced by $\theta(K)$ in (4.2). For the example in (1.3)–(1.4), the condition number is $\phi(K) \approx 1/|\epsilon|$, as we might expect.

Under Assumption 5.1, then, arbitrary symmetric permutations of $Kd = r$ (such as those reducing fill-in) can be solved stably without further permutations as long as $\phi(K)$ is not too large. We therefore interpret $\phi(K)$ to be the condition number of Cholesky factorization without interchanges, applied to an sqd system. In algorithms where sequences of sqd systems are solved, techniques that either reduce $\phi(K)$ or delay its increase will, by postponing the need for stability permutations and hence allowing the unhampered use of sparsity permutations, decrease the total computation time for solving $Kd = r$.

Note that the reduction of $\phi(K)$ is sufficient, but not necessary, for ensuring the accurate solution of $Kd = r$ without interchanging rows and columns for stability. Indeed, Golub and Van Loan [GV79] exhibit a family of unsymmetric positive-definite systems C for which $\omega(C)$ increases without bound but whose computed solutions remain accurate without the use of stability interchanges. Their example suggests that in special cases it may be possible to refine the above results to obtain a sharper bound.

6. An application in numerical optimization. The standard linear programming (LP) problem is

$$\begin{array}{ll} \underset{x}{\text{minimize}} & c^T x \\ \text{subject to} & Ax = b, \quad l \le x \le u, \end{array} \tag{6.1}$$

where $A \in \mathbb{R}^{m\times n}$ $(m \le n)$. Barrier methods for computing primal and dual solutions (x, π) generate a series of sparse symmetric systems; for example, see [LMS92]. Most authors reduce these to the positive-definite form $AH^{-1}A^T\Delta\pi = v$, for which Cholesky factorization is often efficient, as long as A contains no dense columns. We discuss some alternatives.

6.1. Regularized LP. In [GMPS91], [GMPS94], we treat the *regularized* LP problem

$$\begin{array}{ll} \underset{x,\,p}{\text{minimize}} & c^Tx + \frac{1}{2}\|\gamma x\|^2 + \frac{1}{2}\|p\|^2 \\ \text{subject to} & Ax + \delta p = b, \quad l \le x \le u, \end{array} \tag{6.2}$$

where γ and δ are small scalar parameters, typically 10^{-5}. When the optimal (x, π) is not unique, choosing a positive γ and δ (respectively) aids convergence to a solution with minimum $\|x\|$ and $\|\pi\|$. If the constraints $Ax = b$, $l \le x \le u$ have no solution, a positive δ also permits convergence to a meaningful point.

The systems to be solved are

$$K \begin{pmatrix} \Delta x \\ -\Delta\pi \end{pmatrix} = \begin{pmatrix} w \\ r \end{pmatrix}, \qquad K \equiv \begin{pmatrix} H_0 + \gamma^2 I & A^T \\ A & -\delta^2 I \end{pmatrix}, \tag{6.3}$$

where H_0 is diagonal with $(H_0)_{jj} \ge 0$. Choosing $\gamma > 0$ and $\delta > 0$ ensures that K is sqd (though barely!). This was not the original motivation, but in view of Vanderbei's work it raises the question: under what conditions is $PKP^T = LDL^T$ stable for any permutation P (with D diagonal but indefinite)?

In the notation of §4, we have $H = H_0 + \gamma^2 I$ and $G = \delta^2 I$. It is safe to assume that $\|A\| \approx 1$ after the LP problem is suitably scaled. As iterations proceed, some elements of H_0 become large and cause $\|K\|$ and $\kappa_2(K)$ to appear large. We eliminate this artificial ill-conditioning by symmetrically scaling the large diagonals of K down to 1. System (6.3) is then equivalent to an sqd system $Kd = r$ in which

$$\|K\| \approx 1, \quad \|A\| \approx 1, \quad \|H\| = 1, \quad \|H^{-1}\| \approx \gamma^{-2}, \quad \|G\| = \delta^2, \quad \|G^{-1}\| = \delta^{-2},$$

with the 2-norm used throughout. The scaling does not alter $AH^{-1}A^T$. Result 4.1 then gives

$$\begin{aligned} \omega(K) &\approx \max\{\delta^{-2}\|A^TA\|, \|AH^{-1}A^T\|\} \\ &\le \|A\|^2 \max\{\delta^{-2}, \|H^{-1}\|\} \\ &\approx \max\{\delta^{-2}, \gamma^{-2}\}. \end{aligned}$$

Recalling Theorem 5.1, we have the following.

RESULT 6.1. *Using $PKP^T = LDL^T$, the effective condition number for solving the sqd system* (6.3) *with small γ and δ is $\phi(K) = \max\{\delta^{-2}, \gamma^{-2}\}\kappa_2(K)$.*

On a typical LP problem, the barrier algorithm generates 20 to 30 K's that are increasingly ill-conditioned (even after the large diagonals are scaled to 1). With reasonable values of γ and δ, we can expect $PKP^T = LDL^T$ to be stable until the iterates are close to an optimal solution.

6.2. Numerical experiments. To confirm this prediction, we applied our barrier code PDQ1 [GMPS91] to some of the more difficult problems in the Netlib collection [Gay85]. Table 6.1 defines some terms and Table 6.2 lists the problem statistics. We requested 6 digits of accuracy in x and π on a DEC Alpha 3000/400 workstation with about 16 digits of precision. For regularization we set $\gamma = \delta$ in the range 10^{-3} to 10^{-5}. (Larger values perturb the problem noticeably, while smaller values leave little room for the LDL^T factorization to be stable.)

In PDQ1 Version 1.0, the indefinite solver MA27 [DR82], [DR83] is used to factorize either K itself, or certain reduced matrices K_B (obtained by pivoting on diagonals of

TABLE 6.1
Definitions associated with the barrier code PDQ1 for solving linear programs.

K	Full KKT system as in (6.3)
K_B	Reduced KKT system after pivoting on part of H
$nz(K)$	Number of nonzeros in K
PDQ1	Code for solving sparse LP and QP problems [GMPS91], [GMPS94]
MA27	Code for solving sparse symmetric $Kd = r$ [DR82], [DR83]
LDL^T	Sparse factors of permuted K or K_B with D diagonal
LBL^T	Sparse factors of permuted K or K_B with B block-diagonal
Htol	PDQ1's stability tolerance for pivots on H (default = 10^{-6})
factol	MA27's stability tolerance "u" (default = 0.01)
ndense	Nonzeros in a "dense" column of A (default = 10)
residual	$\|r - K\hat{d}\|/\|r\|$, where $\hat{d}$ is the computed d
restol	Tolerance for invoking iterative refinement (default = 10^{-5})

TABLE 6.2
LP test problems: Approximate dimensions of the constraint matrix A, the full KKT matrix K, and a typical reduced KKT matrix K_B.

	m	n	$nz(A)$	Size of K	Typical K_B
grow22	450	950	6000	1400	900
25*fv47*	800	1900	11000	2700	1100
pilotja	900	2000	15000	2900	1300

H that are larger than *Htol* and have fewer than *ndense* entries in the corresponding column of A).

The Analyze phase of MA27 typically predicts very sparse LDL^T factors, but to retain stability on indefinite systems, the Factor phase forms LBL^T factors if necessary. These factors grow increasingly dense as the iterations proceed (more so than the combined Analyze/Factor approach used by Fourer and Mehrotra [FM93]).

Stability is measured by testing residuals after the factors of K are used to solve $Kd = r$. If *residual* > *restol*, one step of iterative refinement is performed to correct $\hat{d}$. (The effects of refinement with an unstable factorization are analyzed in [ADD89].) If *residual* still exceeds *restol*, the factors are considered unreliable and *factol* is increased in stages towards 1. In the experiments cited here, once the LDL^T factors were abandoned, the remaining LBL^T solves were performed reliably with *factol* = 0.01.

6.3. Factorizing K. We first caused the full K to be used every iteration (*Htol* = 10^{20}). With the default stability tolerance (*factol* = 0.01), MA27 computed LBL^T factors at all iterations except the first few. Iterative refinement was seldom needed, but the factors were two to four times as dense as Analyze predicted. On problem *grow22*, $nz(LBL^T)$ increased steadily from 20000 to 80000 over 18 iterations, giving a relatively long runtime.

With *factol* = 0.001 (a little more dangerous), the LBL^T solves were again reliable, and the factors somewhat more sparse. The values of γ and δ had little effect on the sparsity of the factors.

We then allowed MA27 to compute LDL^T factors as long as possible (*factol* = 10^{-20}). Table 6.3 shows the number of iterations for which the Cholesky solves were reliable, for various values of γ and δ. Times are in cpu seconds. With the larger regularizations, most Cholesky factorizations were stable and efficient. On problem *grow22*, $nz(LDL^T)$ was 20000. With regularizations 10^{-5}, 10^{-4}, 10^{-3}, refinement was first requested at iterations 15, 16, 17, and first failed at iterations 16, 17, 17.

TABLE 6.3
Performance of PDQ1 with various regularizations (γ, δ), factorizing full KKT systems. The column labeled LDL^T shows how many iterations were performed reliably with (indefinite) Cholesky factors of K. The remaining iterations used LBL^T factors, which become increasingly dense.

	γ, δ	*factol*	Analyze	LDL^T	LBL^T	time
grow22	10^{-5}	0.01	1	0	18	13.4
	10^{-5}	10^{-20}	1	15	3	6.7
	10^{-4}	10^{-20}	1	16	2	5.6
	10^{-3}	10^{-20}	1	16	2	5.5
25fv47	10^{-5}	0.01	1	0	23	23.0
	10^{-5}	10^{-20}	1	5	18	24.2
	10^{-4}	10^{-20}	1	17	6	20.4
	10^{-3}	10^{-20}	1	21	2	16.8
pilotja	10^{-5}	0.01	1	0	27	37.3
	10^{-5}	10^{-20}	1	5	22	38.6
	10^{-4}	10^{-20}	1	18	9	33.5
	10^{-3}	10^{-20}	1	23	3	26.6

For the last two or three iterations, $nz(LBL^T)$ jumped to 80000.

In general, iterative refinement saved several Cholesky factorizations before a switch was made to LBL^T. The larger the regularization, the later the need for refinement (and the later the switch to LBL^T). The best performance was obtained with the largest regularization, 10^{-3}.

Some sensitivity was noted regarding the test for refinement. Earlier experience with PDQ1 on the first 70 Netlib problems suggested using *restol* $= 10^{-4}$, but the present experiments with Cholesky factors revealed an occasional increase in total iterations, indicating some unnoticed instability. With *restol* $= 10^{-5}$, the results here err on the side of "fewer iterations at the expense of earlier refinement, and hence possible earlier switch to LBL^T factors." Perhaps the tests in [ADD89] would increase the number of iterations for which Cholesky factors could be safely used.

6.4. Reduced KKT systems. We next followed the original PDQ1 strategy of pivoting on most of the diagonals of H (*Htol* $= 10^{-6}$, *ndense* $= 10$). Partitioning $H = \text{diag}(H_N, H_B)$, $A = (N \;\; B)$ and pivoting on H_N gives a reduced matrix of the form

$$K_B = \begin{pmatrix} H_B & B^T \\ B & -NH_N^{-1}N^T - \delta^2 I \end{pmatrix}. \tag{6.4}$$

The aim is to help the Factor phase of MA27, since K_B is smaller and "less indefinite" than K. A penalty is that a new Analyze is needed whenever the makeup of K_B changes.

Note that Result 6.1 still applies, since we still have a Cholesky factorization of the full K, permuted by a different P. Table 6.4 therefore shows qualitatively similar results. The best performance was obtained with $\gamma = \delta = 10^{-3}$ as before, because Analyze was needed only once, and most iterations survived with LDL^T factors.

6.5. Fully reduced systems. Table 6.5 gives results when K was fully reduced to $-(AH^{-1}A^T + \delta^2 I)$ via *Htol* $= 10^{-20}$, *factol* $= 0.0$, *ndense* $= 100$. We write this matrix as $AH^{-1}A^T$ for short. It is the one used in most barrier implementations, such as OB1 [LMS92]. A single Analyze is sufficient for the Cholesky factorizations.

TABLE 6.4

Performance of PDQ1 with various regularizations, factorizing reduced KKT systems K_B. Htol is 10^{-20} initially, but is increased to 10^{-6} after Cholesky factors become unstable. A new Analyze is then needed each time the size of K_B changes. Best results are obtained with maximum regularization ($\gamma = \delta = 10^{-3}$) because the size of K_B depends only on ndense; a single Analyze suffices.

	γ, δ	*factol*	Analyze	LDL^T	LBL^T	time
grow22	10^{-5}	0.01	11	0	18	19.6
	10^{-5}	10^{-20}	5	14	4	11.2
	10^{-4}	10^{-20}	4	15	3	8.6
	10^{-3}	10^{-20}	1	18	1	5.7
25*fv*47	10^{-5}	0.01	14	0	23	18.9
	10^{-5}	10^{-20}	4	20	3	15.3
	10^{-4}	10^{-20}	3	21	2	14.2
	10^{-3}	10^{-20}	1	21	2	12.5
pilotja	10^{-5}	0.01	15	0	27	38.0
	10^{-5}	10^{-20}	15	5	22	38.7
	10^{-4}	10^{-20}	3	25	2	25.7
	10^{-3}	10^{-20}	1	25	1	20.9

TABLE 6.5

Performance of PDQ1, factorizing $AH^{-1}A^T$. This is often the most effective method, but $AH^{-1}A^T$ must be formed efficiently. Not applicable if A contains dense columns.

	γ, δ	Analyze	LDL^T	time
grow22	10^{-3}	1	17	5.5
25*fv*47	10^{-3}	1	23	12.4
pilotja	10^{-3}	1	27	29.6

Regularization is essential, given the way "free variables" are handled. (If x_j has infinite bounds, $(H_0)_{jj} = 0$. Problem *pilotja* has 88 free variables.) We used $\gamma = \delta = 10^{-3}$ to match the best results in the other tables.

Somewhat surprisingly, $AH^{-1}A^T$ was not a clear winner. Since A had no dense columns in these examples, the Cholesky factors of $AH^{-1}A^T$ were more sparse than the LDL^T or LBL^T factors in Tables 6.3 and 6.4, yet the factorization times were slightly greater. A possible explanation is that the off-diagonals of $AH^{-1}A^T$ are formed as a long list of entries from the sparse rank-one matrices $(1/H_{jj})a_j a_j^T$, which MA27 must accumulate before commencing the factorization. (The same accumulation is used for partially reduced KKT systems, but to a lesser degree.)

6.6. Use of MA47. We have recently implemented PDQ1 Version 2.0, in which MA27 is replaced by the new indefinite solver MA47 [DGR91], [DR94]. Following [FM93], we have also experimented with looser pivot tolerances in both codes to improve the sparsity of the numerical factors. In particular, we have initialized *factol* at 10^{-8} (increasing it by a factor of 10 whenever refinement fails), and we have run a larger set of test problems.

With MA27, we do obtain significantly improved performance, though iterative refinement and tolerance increases are frequently needed as before. In some cases, *factol* reaches 0.01 or even 0.1.

With MA47, we have found unexpectedly that refinement is *almost never needed*. Reduced KKT systems again give the best performance ($Htol = 10^{-8}$), and milder regularization seems adequate ($\gamma = \delta = 10^{-4}$). The first 53 Netlib problems solved to 8 digits of accuracy with a total of only three refinements, two of which caused *Htol*

and *factol* to be raised to 10^{-7}. With tolerances of this nature, most factorizations are simply LDL^T with the Analyze ordering. Any LDL^T or LBL^T factorizations with revised orderings are almost equally sparse. The ability to *do* the reordering provides stability at negligible cost.

It appears that two features are contributing to MA47's performance: new stability tests [DGR91], and the default strategy of amalgamating tree nodes to reduce indirect addressing. (By themselves, MA27 *with* amalgamation and MA47 *without* amalgamation were not equally successful.) We hope to give fuller results elsewhere.

7. Conclusions. Diverse techniques have been combined here to obtain some new theoretical and practical results. In the context of barrier methods for linear programming, full KKT matrices K are known to have advantages over $AH^{-1}A^T$ in the presence of dense columns and free variables. In [GMPS91] we attempted to improve the performance of MA27's LBL^T factorizations on severely indefinite systems, but with limited success. Regularization was included there for "numerical analysis" reasons, ensuring uniqueness and boundedness of solutions.

Around the same time, Vanderbei introduced quasidefinite systems and exploited the efficiency of LDL^T factors on KKT-like matrices. Recognizing that regularized KKT systems are quasidefinite, and that a closely related system is positive definite, we were led to the results of Golub and Van Loan on LU factorization without interchanges. From these, we established an effective condition number $\phi(K)$ (5.4) for Cholesky factorization of sqd systems. Result 6.1 justifies LDL^T factorization of sqd matrices K for the special case of barrier methods for linear programming.

Note that our analysis does *not* explain the remarkable success that Vanderbei has had with his LDL^T factors of sqd systems. In particular, Vanderbei does not resort to regularization. Instead, some innovative problem formulation and partitioning gives a multilevel ordering scheme in which certain diagonal pivots are deferred (notably zeros). An sqd principal submatrix is chosen and factored as LDL^T. The Schur complement then has an sqd principal submatrix, and so on. We hope that a direct analysis will eventuate.

Meanwhile, the numerical results obtained here suggest the following approach to systems $Kd = r$ of the form (6.3): Choose the regularizing parameters γ, δ reasonably large (e.g., 10^{-3} or 10^{-4}) and pivot on all entries of H for which the column of A is not too dense. A single Analyze will then suffice, and LDL^T factorization should be efficient and reliable until a good estimate of the solution is reached.

For higher accuracy, we must not forget that implementations based on $AH^{-1}A^T$ are surprisingly reliable and efficient on most real-world problems [Lus94]. Otherwise, Vanderbei's indefinite Cholesky approach is an answer to dense columns and free variables, as are the LBL^T factors in [FM93], [GMPS91], with MA47 now providing a very welcome boost.

Acknowledgments. We are grateful to Nicholas Higham and a second referee for many helpful suggestions.

REFERENCES

[ADD89] M. Arioli, J. W. Demmel, and I. S. Duff, *Solving sparse linear systems with sparse backward error*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 165–190.

[DGR91] I. S. Duff, N. I. M. Gould, J. K. Reid, J. A. Scott, and K. Turner, *The factorization of sparse symmetric indefinite matrices*, IMA J. Numer. Anal., 11 (1991), pp. 181–204.

[DR82] I. S. DUFF AND J. K. REID, MA27: *A set of Fortran subroutines for solving sparse symmetric sets of linear equations*, Report R-10533, Computer Science and Systems Division, AERE Harwell, Oxford, England, 1982.

[DR83] ———, *The multifrontal solution of indefinite sparse symmetric linear equations*, ACM Trans. Math. Software, 9 (1983), pp. 302–325.

[DR94] ———, MA47: *A Fortran code for direct solution of indefinite sparse symmetric linear systems*, 1994, manuscript.

[FM93] R. FOURER AND S. MEHROTRA, *Solving symmetric indefinite systems in an interior-point method for linear programming*, Math. Prog., 62 (1993), pp. 15–39.

[Gay85] D. M. GAY, *Electronic mail distribution of linear programming test problems*, Math. Programming Society COAL Newsletter, December 1985.

[GMPS91] P. E. GILL, W. MURRAY, D. B. PONCELEÓN, AND M. A. SAUNDERS, *Solving reduced* KKT *systems in barrier methods for linear and quadratic programming*, Report SOL 91-7, Department of Operations Research, Stanford University, Stanford, CA, 1991.

[GMPS94] ———, *Solving reduced* KKT *systems in barrier methods for linear programming*, in Numerical Analysis 1993, G. A. Watson and D. Griffiths, eds., Pitman Research Notes in Mathematics 303, Longmans Press, 1994, pp. 89–104.

[GV79] G. H. GOLUB AND C. F. VAN LOAN, *Unsymmetric positive definite linear systems*, Linear Algebra Appl., 28 (1979), pp. 85–98.

[GV89] ———, *Matrix Computations*, second edition, The Johns Hopkins University Press, Baltimore, 1989.

[Lus94] I. J. LUSTIG, *Comments on the performance of the* CPLEX *barrier algorithm*, private communication, 1994.

[LMS92] I. J. LUSTIG, R. E. MARSTEN, AND D. F. SHANNO, *On implementing Mehrotra's predictor-corrector interior point method for linear programming*, SIAM J. Optim., 2 (1992), pp. 435–449.

[Van91] R. J. VANDERBEI, *Symmetric quasi-definite matrices*, Report SOR 91-10, Department of Civil Engineering and Operations Research, Princeton University, Princeton, NJ, 1991.

[Van94] ———, *Symmetric quasi-definite matrices*, SIAM J. Optim., 5 (1995), pp. 100–113.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 1, pp. 47–68, January 1996

003

PRECONDITIONING REDUCED MATRICES*

STEPHEN G. NASH† AND ARIELA SOFER†

Abstract. We study preconditioning strategies for linear systems with positive-definite matrices of the form Z^TGZ, where Z is rectangular and G is symmetric but not necessarily positive definite. The preconditioning strategies are designed to be used in the context of a conjugate-gradient iteration, and are suitable within algorithms for constrained optimization problems. The techniques have other uses, however, and are applied here to a class of problems in the calculus of variations. Numerical tests are also included.

Key words. preconditioning, conjugate-gradient method, reduced Hessian, nonlinear programming

AMS subject classifications. 65F10, 90C06, 90C30

1. Introduction.

We are interested in solving linear systems of the form

$$Z^TGZp = d \tag{1.1}$$

via the preconditioned conjugate-gradient method. The matrix Z^TGZ is assumed to be positive definite, although G need not be. Our primary concern is the choice of a preconditioner for this system.

We intend to apply the techniques within algorithms for solving large nonlinear optimization problems:

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && g(x) = 0, \quad h(x) \leq 0, \end{aligned}$$

where $f(x)$, $g(x)$, and $h(x)$ are nonlinear functions ($f(x)$ is scalar valued, $g(x)$ and $h(x)$ are vector valued). Many algorithms for this problem solve a sequence of linear systems of the form (1.1). This is true for active set methods [6] as well as stabilized penalty methods [13]. It is also true for sequential quadratic programming algorithms [6]. The matrix $G = G(x)$ may represent the Hessian of f at the point x, or it may represent the Hessian of the corresponding Lagrangian function. The matrix Z is a basis for the tangent subspace defined by the active constraints at x. When the number of variables is large, it is often appropriate to apply an iterative method to (1.1), such as a truncated-Newton method [2].

As the solution to the optimization problem is approached, the optimality conditions guarantee that Z^TGZ will be positive semidefinite. In nondegenerate cases, Z^TGZ will be positive definite in a neighborhood of the solution. It is always the case that Z^TGZ will be symmetric. For these reasons, the conjugate-gradient method is normally used to solve (1.1), with safeguards in case Z^TGZ is not positive definite [2], [10].

---

*Received by the editors March 10, 1993; accepted for publication (in revised form) by L. Kaufman January 20, 1995. This work was supported by National Science Foundation grant DDM–9104670.

†Operations Research and Engineering Department, George Mason University, Fairfax, VA 22030 (snash@gmu.edu, asofer@gmu.edu).

As a simple special case, consider a quadratic programming problem of the form

$$\begin{aligned} &\text{minimize}_x \quad f(x) = \tfrac{1}{2}x^TGx - c^Tx \\ &\text{subject to} \quad Ax = b, \end{aligned}$$

where A is an $m \times n$ matrix with $m < n$, and G is positive definite on the null space of A. The first-order optimality conditions for the problem are

$$\begin{pmatrix} G & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} c \\ b \end{pmatrix}. \tag{1.2}$$

(The vector λ is the vector of Lagrange multipliers for the constraints.) Then $x = x_0 + Zp$ where p is the solution to

$$Z^TGZp = Z^T(c - Gx_0),$$

x_0 is any solution to the constraints $Ax = b$, and Z is a matrix whose columns form a basis for the null space of A (i.e., $AZ = 0$). Auxiliary calculations are needed to compute λ.

In this simple case the solution is obtained by solving a single system of the form (1.1). In more complicated cases a sequence of systems of this form must be solved, with perhaps both G and Z changing from one system to the next. The null-space matrix Z can change not only in its entries, but also in its size as the dimension of the relevant null-space changes.

Some optimization algorithms work directly with the linear system (1.2). This system is symmetric but not positive definite, and so the traditional conjugate-gradient method cannot be applied. It is also a larger system than (1.1), having $n+m$ variables instead of $n - m$. Preconditioning strategies for (1.2) are discussed in [5], [14]. We concentrate on the solution of (1.1).

In exact arithmetic, the number of iterations required by the conjugate-gradient method is bounded by the number of distinct eigenvalues of Z^TGZ. In addition, from iteration to iteration the method displays a linear rate of convergence with rate constant $(\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)$, where κ is the condition number of Z^TGZ [7].

A preconditioner is a (symmetric) positive-definite matrix K such that $K \approx Z^TGZ$. System (1.1) is equivalent to the *preconditioned* system

$$K^{-1}Z^TGZp = K^{-1}d.$$

The goal is to choose K so that the preconditioned matrix $K^{-1}Z^TGZ$ will have a smaller number of distinct eigenvalues, or a smaller condition number, or both. For practical purposes, it must be possible to solve linear systems of the form $Ky = v$ for arbitrary v.

If an approximation $K \approx Z^TGZ$ is provided then the situation is straightforward, since the approximation could be used directly within the preconditioned conjugate-gradient method. In the optimization setting, however, we think it more likely that only an approximation $M \approx G$ would be available. Such a matrix M might be generated by the optimization algorithm itself, using for example the techniques in [11]. Or it might be suggested by the optimization problem. For example, if the objective function $f(x)$ were derived from a differential equation, an approximation to the corresponding differential operator G might be available, such as a fast Poisson

solver. It seems less likely that an approximation to the reduced matrix Z^TGZ would be provided, particularly in cases where Z changes frequently.

In some cases, explicitly forming Z^TGZ or Z^TMZ will be undesirable. In many large problems the matrices G and A will be sparse or have special structure, and the special structure of A will often carry over to Z. Forming the reduced matrices can destroy this structure. For example, if $Ax = b$ corresponds to the simple constraint

$$\sum x_i = 1,$$

and an orthogonal null-space matrix Z is used, then Z^TGZ will be dense even if G is sparse.

For these reasons we are interested in constructing $K \approx Z^TGZ$ given Z and a (symmetric) positive-definite matrix $M \approx G$. (Z is used within the optimization algorithm and so is available.) We describe a variety of approximations K that could be used, even though some of them are likely to be too computationally expensive for routine use. We also discuss the computation of Z and its influence on the conditioning of the reduced matrices. Finally, we apply the ideas to a class of problems arising in the calculus of variations. Numerical tests are presented to illustrate the techniques.

The simplest of the preconditioners that we derive is

$$K^{-1} = W^TM^{-1}W \approx (Z^TGZ)^{-1},$$

where W^T is a left inverse for Z (i.e., $W^TZ = I$). In many cases, W can be obtained as a by-product of the calculations used to obtain Z. The computational effort required to apply the preconditioner is that required for the "unreduced" preconditioner M, plus that for W and W^T. The numerical tests in §8 indicate that the formula can be an effective preconditioner for reduced systems. It is this preconditioner that we would recommend for most applications. Hence, within the conjugate-gradient method, applying the preconditioner to a vector v would mean calculating $K^{-1}v = W^TM^{-1}Wv$.

The more elaborate preconditioners that we derive require considerably more computational effort to use, but they can further accelerate the convergence of the conjugate-gradient method. Whether these preconditioners would be appropriate would depend on the relative computational costs of applying the preconditioner (forming $K^{-1}v$) and computing a matrix-vector product (forming Z^TGZv). In cases where the latter is expensive (see below), the more elaborate preconditioners might be worthwhile.

In truncated-Newton software, it is common to compute a Hessian-vector product via finite differencing. In this case $G = \nabla^2 f(x)$ for some nonlinear function $f(x)$, and

$$Z^TGZv = Z^T\nabla^2 f(x)Zv \approx Z^T\frac{\nabla f(x + hZv) - \nabla f(x)}{h}$$

for some "small" value of h. Thus the cost of the matrix-vector product is proportional to the cost of evaluating the gradient $\nabla f(x)$. If evaluating the gradient is expensive, then the matrix-vector product will be expensive, and could easily become the dominant part of the conjugate-gradient iteration. In such cases we envision the more elaborate preconditioners being used.

Here is an outline of the paper. Basic topics are in §2. A family of preconditioners is derived in §3. They are based on an infinite series, whose convergence is the topic

of §4. Section 5 extends the results to the case where Z is a projection matrix that is not of full rank. Section 6 focuses on a specific choice of Z commonly used in large-scale optimization. Section 7 shows how the techniques can be applied to a class of problems in the calculus of variations. Numerical tests are in §8 and conclusions in §9.

2. Basics. For simplicity we consider the simple case (1.2), treating it, if appropriate, as a single instance of a sequence of linear systems of the same structure. Hence A is an $m \times n$ matrix with $m < n$ corresponding to the (perhaps linearized) constraints. (If $m = n$ then the solution is determined entirely by the constraints.) For convenience we assume that $\text{rank}(A) = m$. As before, Z is a matrix whose columns form a basis for the null space of A, so that $AZ = 0$ and $\text{rank}(Z) = n - m$.

There are two traditional ways of forming Z. The first uses an orthogonal factorization of A^T:

$$A^T = QR \equiv \begin{pmatrix} Y & Z \end{pmatrix} \begin{pmatrix} R_1 \\ 0 \end{pmatrix},$$

where R_1 is an $m \times m$ upper triangular matrix. Then $Z^TZ = I$. This technique is often used on small problems where dense-matrix methods are appropriate.

The second technique (called *variable reduction*) identifies a subset of the variables of size m, called a basis. If the first m variables were to be used we would write

$$A = \begin{pmatrix} B & N \end{pmatrix},$$

where B is nonsingular. The corresponding matrix Z is given by

$$Z = \begin{pmatrix} -B^{-1}N \\ I \end{pmatrix}.$$

This only requires a factorization of the submatrix B, and is better suited to large sparse problems. It is easily checked that $AZ = 0$, but $Z^TZ \neq I$ unless $N = 0$.

In our formulas we will require a left-inverse for Z, i.e., a matrix W^T satisfying $W^TZ = I$. If Z is available, a left-inverse for Z is usually available at little or no additional cost. For example, if Z is formed via an orthogonal factorization of A^T, we can choose $W^T = Z^T$. If Z is computed via the variable reduction method, we can choose $W^T = (0, I)$.

2.1. Some lemmas. Not a great deal can be said in general about the relationship between G and Z^TGZ, so only limited conclusions can be drawn about the quality of the preconditioners we derive. However, the following lemmas provide some information. In the discussion that follows, $\|\cdot\| = \|\cdot\|_2$.

The first lemma discusses the case where G and M share an eigenvalue-vector pair. Note that $M^{-1}G$ then has an eigenvalue equal to one. The lemma shows that spectral information for the original matrix can be used in constructing a preconditioner for the reduced matrix.

LEMMA 1. *Suppose that* $Gv = \lambda v$ *and* $Mv = \lambda v$, *where* $v \neq 0$. *If* $v \in \text{range}(Z)$ *then*

$$(Z^TGZ)w = \lambda w \quad \textit{and} \quad (Z^TMZ)w = \lambda w,$$

where $Zw = v$.

The next result gives a bound on the norm of a reduced matrix.

LEMMA 2. *Let* H *be an* $n \times n$ *matrix. Then* $\|Z^THZ\| \leq \|H\| \cdot \|Z\|^2$.

The lemma may be used to bound the norm of the difference between the reduced matrix and its approximation:

$$\|(Z^TMZ) - (Z^TGZ)\| = \|Z^T(M-G)Z\| \leq \|M-G\| \cdot \|Z\|^2.$$

In particular, if Z is an orthogonal matrix the bound becomes

$$\|(Z^TMZ) - (Z^TGZ)\| \leq \|M-G\|.$$

When Z is an orthogonal matrix, we also have the following interlacing property (see [7]).

LEMMA 3. *Let Z be an $n \times l$ orthogonal matrix and G an $n \times n$ symmetric matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$. Let $\omega_1 \geq \cdots \geq \omega_l$ be the eigenvalues of Z^TGZ. Then*

$$\lambda_i \geq \omega_i \geq \lambda_{i+n-l}, \quad i = 1, \ldots, l.$$

An immediate conclusion is that orthogonal reduction will not increase the condition number of a positive-definite matrix. This is stated in the next lemma.

LEMMA 4. *Let Z be an orthogonal matrix, and let G be a positive definite matrix. Then* $\operatorname{cond}(Z^TGZ) \leq \operatorname{cond}(G)$.

Unfortunately, the bounds provided in the above lemmas can be tight. For example, consider Lemma 4 in the case where G is a diagonal matrix and Z corresponds to a set of bound constraints. Then Z^TGZ is just a principal submatrix of G. If that submatrix includes the extreme eigenvalues of G then $\operatorname{cond}(G) = \operatorname{cond}(Z^TGZ)$. Therefore, it is not possible to make the results more precise. Any more specific results would require precise information about the particular submatrix that had been chosen. Similar pessimistic examples can be found for the other lemmas.

3. Formulas for preconditioners. From a theoretical standpoint, it would be ideal to precondition Z^TGZ using Z^TGZ itself. If G is nonsingular, then from the identity (see [3, p. 87])

$$Z(Z^TGZ)^{-1}Z^T = G^{-1} - G^{-1}A^T(AG^{-1}A^T)^{-1}AG^{-1}$$

it follows that the inverse of Z^TGZ satisfies the identity

$$(Z^TGZ)^{-1} = W^TG^{-1}W - W^TG^{-1}A^T(AG^{-1}A^T)^{-1}AG^{-1}W,$$

where W^T is a left-inverse for Z.

In cases where it is inconvenient to use G explicitly (or if G is singular), and a positive-definite preconditioner $M \approx G$ is given, it is natural to consider using Z^TMZ as a preconditioner for Z^TGZ. Linear systems of the form $Z^TMZy = v$ can be solved using the formula

$$(Z^TMZ)^{-1} = W^TM^{-1}W - W^TM^{-1}A^T(AM^{-1}A^T)^{-1}AM^{-1}W. \tag{3.1}$$

We assume that linear systems of the form $My = v$ can be easily solved. A preconditioner for the conjugate-gradient method must be positive definite. This will be true for the preconditioners that we derive because M is positive definite.

The preconditioners we consider are based on approximations to the right-hand side of (3.1). The accuracy of the approximation can be varied, but the more accurate the approximation, the more computations are required to implement it.

In certain cases it is possible to use (3.1) in its original form. For example, if m is small then $AM^{-1}A^T$ can be formed explicitly and then either factored or inverted. (In the case where $n-m$ is small and G is available, Z^TGZ can be formed and there is no need to use an iterative method to solve (1.1).) Even if the component parts of the right-hand side of (3.1) can be formed, it requires two applications of M^{-1} as well as an application of $(AM^{-1}A^T)^{-1}$ to use it:

$$(Z^TMZ)^{-1}v = W^TM^{-1}(I - A^T(AM^{-1}A^T)^{-1}AM^{-1})Wv,$$

making it more than twice as expensive to precondition Z^TGZ as to precondition G.

If neither m nor $n-m$ is small, then some approximation to (3.1) must be used. The simplest available is

$$(Z^TMZ)^{-1} \approx W^TM^{-1}W. \tag{3.2}$$

This approximation requires only one application of M^{-1} and so is economical. Since

$$(Z^TMZ)^{-1} - W^TM^{-1}W = -W^TM^{-1}A^T(AM^{-1}A^T)^{-1}AM^{-1}W,$$

there is no guarantee that $\|(Z^TMZ)^{-1} - (W^TM^{-1}W)\|$ is small. However, if either m or $n-m$ is small then the difference is of low rank and (3.2) may be satisfactory. (If the difference between the matrix and the preconditioner is of rank l then the preconditioned conjugate-gradient method will converge in at most $l+1$ iterations with exact arithmetic.)

The preconditioned matrix is

$$(W^TM^{-1}W)(Z^TMZ) = W^T(M^{-1}WZ^TM)Z.$$

The eigenvalues of the inner matrix $(M^{-1}WZ^TM)$ are all either zero or one, since the corresponding eigenvectors are the columns of $M^{-1}W$ and $M^{-1}A^T$:

$$\begin{aligned}(M^{-1}WZ^TM)(M^{-1}W) &= M^{-1}W,\\ (M^{-1}WZ^TM)(M^{-1}A^T) &= 0.\end{aligned}$$

This suggests that the preconditioned matrix may be well-suited for the conjugate-gradient method, since the convergence of this method depends on the number of distinct eigenvalues of the matrix.

More sophisticated (and more expensive) preconditioners can be obtained by approximating the inverse of the matrix (Z^TMZ). They are based on the power series expansion

$$\begin{aligned}X^{-1} &= [Y - (Y - X)]^{-1}\\ &= Y^{-1}[I + (I - XY^{-1}) + (I - XY^{-1})^2 + \cdots]\end{aligned} \tag{3.3}$$

where $Y \approx X$. The series converges if the eigenvalues of $(I - XY^{-1})$ are less than one in absolute value.

If we let

$$X = Z^TMZ \quad \text{and} \quad Y^{-1} = W^TM^{-1}W$$

in (3.3), then using the first k terms of the series gives the approximation

$$(Z^TMZ)^{-1} \approx (W^TM^{-1}W)\sum_{j=0}^{k}(I - T)^j, \tag{3.4}$$

where $T = (Z^TMZ)(W^TM^{-1}W)$. (We assume here that the series on the right converges as $k \to \infty$; see §4.) Taking for example $k = 0$, we obtain the "inverse preconditioner" (3.2). (Formula (3.4) provides the *inverse* of the preconditioner; all our formulas will be of this nature, and the preconditioners themselves will not be specified.) Taking $k = 1$ we obtain the inverse preconditioner

$$(Z^TMZ)^{-1} \approx (W^TM^{-1}W)\,(2I - (Z^TMZ)(W^TM^{-1}W))\,.$$

Since the inverse preconditioner need not be formed explicitly, it is available at no expense (once Z and W are available). The expense lies in applying it to a vector:

$$(Z^TMZ)^{-1}v \approx (W^TM^{-1}W)\,(2I - (W^TM^{-1}W)(Z^TMZ))\,v.$$

We can derive an alternative expression for the inverse preconditioner that is more efficient for computation. Using the relationship

$$\sum_{j=0}^{k}(I-T)^j = \sum_{j=0}^{k}(-1)^j\tbinom{k+1}{j+1}T^j, \tag{3.5}$$

we obtain

$$(Z^TMZ)^{-1} \approx (W^TM^{-1}W)\sum_{j=0}^{k}(-1)^j\tbinom{k+1}{j+1}T^j, \tag{3.6}$$

where as before, $T = (Z^TMZ)(W^TM^{-1}W)$. This is the form of the inverse preconditioner used in our computations. However in the following sections we continue to use the representation given in (3.4), since it is more convenient for mathematical manipulation.

3.1. An alternative formula. It might appear that the power series expansion could be used in another way to obtain additional inverse preconditioners. If $X = AM^{-1}A^T$ and R is a matrix such that $RA^T = I$ then a "plausible" choice for an approximate inverse of X is

$$Y^{-1} = RMR^T.$$

Applying the power series to approximate $(AM^{-1}A^T)^{-1}$ on the right-hand side of (3.1) we obtain

$$(Z^TMZ)^{-1} \approx W^TM^{-1}W - W^TM^{-1}A^TY^{-1}\left(\sum_{j=0}^{k-1}(I-U)^j\right)AM^{-1}W, \tag{3.7}$$

where $U = XY^{-1} = (AMA^T)(RM^{-1}R^T)$. As an illustration, if only one term in the power series is taken, we obtain the inverse preconditioner

$$(Z^TMZ)^{-1} \approx W^TM^{-1}[I - A^TY^{-1}AM^{-1}]W.$$

A left-inverse for A^T is usually available from the computation of Z at little or no additional cost. For example if Z is obtained from an orthogonal factorization of A^T then the left-inverse matrix $R = (AA^T)^{-1}A$ can be computed easily from the factors of the orthogonalization.

The following lemma shows that the inverse preconditioners (3.4) and (3.7) are mathematically identical, and so no new preconditioners are obtained—just an alternative formula. In some applications one formula may be easier to apply than the other. (See also §4.)

The lemma assumes that $A^TR + WZ^T = I$. This assumption ensures that R and W^T are chosen in a "consistent" fashion. It is satisfied when Z is computed via an orthogonal factorization of A, $R = (AA^T)^{-1}A$, and $W = Z$. It is also satisfied when Z is computed via the variable reduction method, with

$$Z = \begin{pmatrix} -B^{-1}N \\ I \end{pmatrix}, \quad R = \begin{pmatrix} B^{-T} & 0 \end{pmatrix}, \quad W = \begin{pmatrix} 0 \\ I \end{pmatrix}.$$

(In general, given a full-rank matrix Z that is a null-space matrix for A, and a matrix R that is a left-inverse for A^T, there exists a unique left-inverse matrix W^T for Z such that $A^TR + WZ^T = I$.)

LEMMA 5. *Let $RA^T = I$ $W^TZ = I$, and suppose that $A^TR + WZ^T = I$. Let $Y^{-1} = RMR^T$, $U = (AM^{-1}A^T)(RMR^T)$ and $T = (Z^TMZ)(W^TM^{-1}W)$. Then for $k \geq 1$,*

$$W^TM^{-1}W - W^TM^{-1}A^TY^{-1}\left(\sum_{j=0}^{k-1}(I-U)^j\right)AM^{-1}W = (W^TM^{-1}W)\sum_{j=0}^{k}(I-T)^j.$$

Proof. It is sufficient to prove that

$$-W^TM^{-1}A^TY^{-1}\left((I-U)^{k-1}\right)AM^{-1}W = (W^TM^{-1}W)(I-T)^k. \tag{3.8}$$

The proof is by induction. For $k = 1$ we have

$$\begin{aligned}
-W^TM^{-1}A^TY^{-1}&((I-U)^{k-1})AM^{-1}W \\
&= -W^TM^{-1}A^TY^{-1}AM^{-1}W \\
&= -W^TM^{-1}A^TRMR^TAM^{-1}W \\
&= -W^TM^{-1}(I-WZ^T)M(I-ZW^T)M^{-1}W \\
&= -W^TM^{-1}(W - WZ^TW - MZW^TM^{-1}W + WZ^TMZW^TM^{-1}W) \\
&= -W^TM^{-1}\left(-MZ(W^TM^{-1}W) + W(Z^TMZ)(W^TM^{-1}W)\right) \\
&= -W^TM^{-1}\left(-MZ + W(Z^TMZ)\right)(W^TM^{-1}W) \\
&= \left(I - (W^TM^{-1}W)(Z^TMZ)\right)(W^TM^{-1}W) \\
&= (W^TM^{-1}W)(I-T).
\end{aligned}$$

Assume now that (3.8) is true for $k-1$. We shall prove that it is also true for k. Note from the proof for $k = 1$ that

$$A^TY^{-1}AM^{-1}W = \left(-MZ + W(Z^TMZ)\right)(W^TM^{-1}W).$$

We conclude from this that

$$\begin{aligned}
UAM^{-1}W &= AM^{-1}A^TY^{-1}AM^{-1}W \\
&= AM^{-1}\left(-MZ + W(Z^TMZ)\right)(W^TM^{-1}W) \\
&= AM^{-1}W(Z^TMZ)(W^TM^{-1}W) \\
&= AM^{-1}WT.
\end{aligned}$$

Using this relation and the induction hypothesis we obtain

$$\begin{aligned}
&-W^TM^{-1}A^TY^{-1}\left((I-U)^{k-1}\right)AM^{-1}W \\
&\qquad = -W^TM^{-1}A^TY^{-1}\left((I-U)^{k-2}\right)(AM^{-1}W - UAM^{-1}W) \\
&\qquad = -W^TM^{-1}A^TY^{-1}\left((I-U)^{k-2}\right)(AM^{-1}W(I-T)) \\
&\qquad = (W^TM^{-1}W)(I-T)^{k-1}(I-T) \\
&\qquad = (W^TM^{-1}W)(I-T)^k. \qquad \square
\end{aligned}$$

There are several remaining questions, such as what to do if the series does not converge, how many terms to use, and which of the two equivalent formulas to use. Convergence is discussed in the next section, the choice of the number of terms to use is mentioned in §8 when we discuss computational results, and the particular formula to use would probably just be a matter of which was more convenient, perhaps depending on the relative magnitudes of m and $n-m$.

4. Guaranteeing convergence. In developing the polynomial preconditioners, we tacitly assumed that the corresponding infinite series converged. We now explore the conditions for convergence of the series and discuss strategies to handle the case where the conditions are not met.

Consider the power series (3.3). It is convergent if the eigenvalues of $(I-XY^{-1})$ are less than 1 in absolute value, or equivalently if the eigenvalues of XY^{-1} lie strictly between 0 and 2. In our application, where X and Y^{-1} are positive definite, the eigenvalues of XY^{-1} will be positive, since XY^{-1} is similar to a positive-definite matrix:

$$Y^{-\frac{1}{2}}(XY^{-1})Y^{\frac{1}{2}} = Y^{-\frac{1}{2}}(X)Y^{-\frac{1}{2}}.$$

Let $\lambda_1(\cdot)$ and $\lambda_n(\cdot)$ denote the largest and smallest eigenvalues of a matrix, respectively. Then for a positive scalar α such that $\alpha < (2/\lambda_1(XY^{-1}))$, the eigenvalues of αXY^{-1} will lie strictly between 0 and 2. In turn, the eigenvalues of $(I-\alpha XY^{-1})$ will be smaller than 1 in absolute value. We can now use the approximation

$$\begin{aligned}
X^{-1} &= \alpha(\alpha X)^{-1} = \alpha[Y-(Y-\alpha X)]^{-1} \\
&= \alpha Y^{-1}[I+(I-\alpha XY^{-1})+(I-\alpha XY^{-1})^2+\cdots].
\end{aligned}$$

The expression on the right is a convergent sequence. Notice that it differs from (3.3) only in that Y^{-1} is replaced by αY^{-1}. The rate of convergence of the sequence depends on the particular choice of the scaling parameter α; ideally the eigenvalues of αXY^{-1} should be close to 1. A possible choice is the scaling parameter that minimizes the largest deviation of the eigenvalues of the scaled matrix from 1. This gives $\alpha = 2/(\lambda_1(XY^{-1}) + \lambda_n(XY^{-1}))$.

Choosing $\alpha < 2/\lambda_1(XY^{-1})$ guarantees that the resulting preconditioner will be positive definite. To see this, denote by η_i the eigenvalues of $I-\alpha XY^{-1}$; the selection of α guarantees that $-1 < \eta_i < 1$. Then the eigenvalues of $\sum_{j=0}^{k}(I-XY^{-1})^j$ will be $\bar{\eta}_i = \sum_{j=0}^{k}\eta_i^j > 0$.

We now obtain the explicit form for the inverse preconditioner that includes the scaling parameter. Consider first the family of inverse preconditioners represented by (3.4), and denote as before

$$T = (W^TM^{-1}W)(Z^TMZ).$$

Let α be a scalar such that $\alpha < 2/\lambda_1(T)$. Applying the above results we obtain

$$(Z^TMZ)^{-1} \approx \alpha(W^TM^{-1}W)\sum_{j=0}^{k}(I-\alpha T)^j.$$

Alternatively, if we choose to use an inverse preconditioner of the form (3.7), we obtain

$$(Z^TMZ)^{-1} \approx W^TM^{-1}W - \alpha W^TM^{-1}A^TY^{-1}\left(\sum_{j=0}^{k-1}(I-\alpha U)^j\right)AM^{-1}W,$$

where $U = (AM^{-1}A^T)(RMR^T)$ and $\alpha < 2/(\lambda_1(U)$.

In practice, the largest eigenvalues of U and T are not available. Our tests (§8) indicate that in many cases it is sufficient to use only the first term in the power series. There is then no need to compute the scaling parameter. In other cases we must estimate the eigenvalues. (For the purpose of estimation it may be easier to use bounds on the norms of the matrices.) The following lemma shows that if R and W satisfy $A^TR + WZ^T = I$, then the eigenvalues of U and T are closely related.

LEMMA 6. *Suppose that $A^TR + WZ^T = I$. If $m \le n - m$ then any nonunit eigenvalue of U is also an eigenvalue of T, and all other eigenvalues of T are equal to 1. If $n - m \le m$ then any nonunit eigenvalue of T is also an eigenvalue of U, and all other eigenvalues of U are equal to 1.*

Proof. Let $E = AM^{-1}W$ and $F = Z^TMR^T$. Recalling that $AR^T = I$ we obtain

$$U = AM^{-1}A^TRMR^T = AM^{-1}(I - WZ^T)MR^T = I - EF.$$

Similarly, recalling that $Z^TW = I$ we obtain

$$T = Z^TMZW^TM^{-1}W = Z^TM(I - R^TA)M^{-1}W = I - FE.$$

Suppose that $m \le n - m$. We first show that any nonzero eigenvalue of EF is an eigenvalue for FE. Let λ and v be an eigenvalue/vector pair for EF with $\lambda \neq 0$. Then $EFv = \lambda v$, and since $\lambda \neq 0$, $Fv \neq 0$. Now $FE(Fv) = F\lambda v = \lambda(Fv)$. Hence λ is an eigenvalue for FE with associated eigenvector Fv.

Next we show that any eigenvalue of FE that is not an eigenvalue of EF must be zero. Suppose that $FEx = \mu x$ for some nonzero vector x. Then

$$EF(Ex) = E(FEx) = \mu(Ex).$$

The above relation can occur in one of three situations: (i) $Ex \neq 0$ and $\mu \neq 0$; in this case μ is a nonzero eigenvalue of EF. (ii) $Ex \neq 0$ but $\mu = 0$. (iii) $Ex = 0$; this implies that $\mu = 0$. The proof for the case $n - m \le n$ is similar. □

The lemma indicates that in estimating the scaling parameter one can use either U or T, whichever is more convenient. Denote the resulting inverse preconditioners by

$$K_1^{-1}(\alpha,k) = W^TM^{-1}W - \alpha W^TM^{-1}A^TY^{-1}\left(\sum_{j=0}^{k-1}(I-\alpha U)^j\right)AM^{-1}W,$$

and

$$K_2^{-1}(\alpha,k) = \alpha(W^TM^{-1}W)\sum_{j=0}^{k}(I-\alpha T)^j.$$

The following result shows the relationship between K_1 and K_2. When $\alpha = 1$ they are equal (as pointed out in §3). Otherwise, assuming they converge, their difference goes to zero as k increases.

LEMMA 7. *Let $RA^T = I$ and $W^TZ = I$, and suppose that $A^TR + WZ^T = I$. Then for $k \geq 0$*

$$K_1^{-1}(\alpha, k) = K_2^{-1}(\alpha, k) + (1 - \alpha)(W^TM^{-1}W)(I - \alpha T)^k.$$

Proof. Here we just sketch the proof. First, using arguments similar to those in Lemma 5, we can show that for $j \geq 0$,

$$-W^TM^{-1}A^TY^{-1}\left((I - \alpha U)^j\right)AM^{-1}W = (W^TM^{-1}W)(I - T)(I - \alpha T)^j.$$

Next, the relationship $\alpha(I - T) = (I - \alpha T) - (1 - \alpha)I$ gives

$$\begin{aligned}
&W^TM^{-1}W - \alpha W^TM^{-1}A^TY^{-1}\left(\sum_{j=0}^{k-1}(I - \alpha U)^j\right)AM^{-1}W \\
&\quad = W^TM^{-1}W + \alpha(W^TM^{-1}W)\sum_{j=0}^{k-1}(I - T)(I - \alpha T)^j \\
&\quad = W^TM^{-1}W + (W^TM^{-1}W)\sum_{j=0}^{k-1}(I - \alpha T)^{j+1} \\
&\qquad\qquad - (1 - \alpha)(W^TM^{-1}W)\sum_{j=0}^{k-1}(I - \alpha T)^j \\
&\quad = (W^TM^{-1}W)\sum_{j=0}^{k}(I - \alpha T)^{j+1} - (1 - \alpha)(W^TM^{-1}W)\sum_{j=0}^{k-1}(I - \alpha T)^j \\
&\quad = \alpha\sum_{j=0}^{k}(W^TM^{-1}W)(I - \alpha T)^j + (1 - \alpha)(W^TM^{-1}W)(I - \alpha T)^k. \qquad \square
\end{aligned}$$

5. Using an orthogonal projection matrix. In the previous sections we assumed that the matrix Z that generates the null space of A has full column rank. In this section we focus on the case where the null-space matrix is an orthogonal projection $P = I - A^T(AA^T)^{-1}A$, where A is an $m \times n$ matrix of full row rank. (To avoid confusion we use the notation P rather than Z.) P is an $n \times n$ matrix of rank $n - m$. We are concerned with the solution of the system

$$PGPp = d. \tag{5.1}$$

Why use an orthogonal projection? First, if A is sparse it is possible to apply the projection in a way that utilizes the sparsity. Second, applying an orthogonal projection does not increase the norm of a matrix; that is, $\|PGP\| \leq \|G\|$.

In the typical systems that arise in optimization, the vector d can be written as $d = Pg$ for some vector g. If G is nonsingular, then (5.1) is consistent and a solution is

$$p = (PGP)^+Pg,$$

where $(PGP)^+$ is the Moore–Penrose generalized inverse of PGP.

It is possible to solve (5.1) using the linear conjugate-gradient method. If G is positive definite on the null space of A, then in exact arithmetic the conjugate-gradient method will terminate in at most $n - m$ iterations with the solution vector [8].

We shall extend the ideas of §3 to obtain a preconditioner for PGP. As before, we assume that we have a positive-definite approximation $M \approx G$. The "inverse" preconditioner will be of the form $(PMP)^+$. An explicit expression for this matrix is provided in the following lemma.

LEMMA 8.

$$\begin{aligned}(PMP)^+ &= M^{-1} - M^{-1}A^T(AM^{-1}A^T)^{-1}AM^{-1} \\ &= P(M^{-1} - M^{-1}A^T(AM^{-1}A^T)^{-1}AM^{-1})P.\end{aligned}$$

Proof. It is easy to verify that

$$(PMP)(M^{-1} - M^{-1}A^T(AM^{-1}A^T)^{-1}AM^{-1}) = P.$$

The lemma follows immediately. □

Our preconditioners are based on the ideas of §3. Let $R = (AA^T)^{-1}A$ be a left-inverse for A^T. We use a power series expansion for $(AM^{-1}A^T)^{-1}$ with RMR^T as the approximate inverse. Denoting $U = (AM^{-1}A^T)(RMR^T)$, we obtain

$$(PMP)^+ \approx PM^{-1}P - PM^{-1}A^T(RMR^T)\left(\sum_{j=0}^{k-1}(I-U)^j\right)AM^{-1}P.$$

Now let $T = (PMP)(PM^{-1}P) = PMPM^{-1}P$. The following result is analogous to Lemma 5.

LEMMA 9.

$$PM^{-1}P - PM^{-1}A^TY^{-1}\left(\sum_{j=0}^{k-1}(I-U)^j\right)AM^{-1}P = (PM^{-1}P)\sum_{j=0}^{k}(I-T)^j.$$

Proof. It is similar to the proof of Lemma 5. □

The expression on the right is another form for our inverse preconditioner. A formula analogous to (3.5) could also be derived. Techniques similar to those described in §4 can be used to guarantee convergence of the appropriate infinite series.

5.1. Application to linear programming. There is possibly another application of these results. In recent years several successful interior-point algorithms have been developed for solving linear programs. Interior point methods typically require few iterations, but each iteration requires the solution of a system of the form

$$AMA^Tp = d \tag{5.2}$$

to provide a search direction. The matrix M is diagonal and changes from one iteration to another (while A remains constant). The principal cost lies in (repeatedly) computing the Cholesky factorization of AMA^T.

Here we propose an approach for approximately solving (5.2) that requires only one Cholesky factorization. We start by using the approximation

$$M^{-1} - M^{-1}A^T(AM^{-1}A^T)^{-1}AM^{-1} \approx (PM^{-1}P)\sum_{j=0}^{k}(I-T)^j.$$

Premultiplying by RM (where $R = (AA^T)^{-1}A$), postmultiplying by (MR^T), and reordering gives

$$(AM^{-1}A^T)^{-1} \approx RMR^T - RMPM^{-1}P\left(\sum_{j=0}^{k}(I-T)^j\right)MR^T.$$

The expression on the right involves only P, R, the diagonal matrix M, and its inverse. The main work needed is to factor AA^T (just once). The solution of (5.2) requires $3k+4$ applications of $(AA^T)^{-1}$, $2k+4$ applications of M or M^{-1}, and one application each of A and A^T per conjugate-gradient iteration.

6. Using variable reduction. If $Z^TZ = I$, Lemma 2 shows that Z^TGZ cannot be more ill-conditioned than G. However if Z is obtained from variable reduction, as would be more likely in sparse problems, this need not be true. We would like to have a better understanding of how the choice of Z affects the conditioning of the problem, and if the conditioning can be monitored or controlled as the optimization problem is being solved.

The ideas in this section are motivated by the following lemma.

LEMMA 10. *If Z_0 is an orthogonal null-space matrix, and Z is a null-space matrix obtained from some other technique such as variable reduction, then*

$$\operatorname{cond}(Z^TGZ) \le \operatorname{cond}(Z_0^TGZ_0)\cdot\operatorname{cond}(Z)^2.$$

Proof. Since Z_0 and Z are both null-space matrices, $Z = Z_0T$ for some nonsingular matrix T. The largest eigenvalue of Z^TGZ satisfies

$$\begin{aligned}
\lambda_{\max}(Z^TGZ) &= \max_{v\neq 0}\frac{v^TZ^TGZv}{v^Tv}\\
&= \max_{v\neq 0}\frac{v^TT^TZ_0^TGZ_0Tv}{v^Tv}\\
&= \max_{v\neq 0}\frac{(Tv)^T(Z_0^TGZ_0)(Tv)}{(Tv)^T(Tv)}\,\frac{v^TT^TTv}{v^Tv}\\
&= \max_{v\neq 0}\frac{(Tv)^T(Z_0^TGZ_0)(Tv)}{(Tv)^T(Tv)}\,\frac{v^T(Z_0T)^T(Z_0T)v}{v^Tv}\\
&\le \max_{w\neq 0}\frac{w^T(Z_0^TGZ_0)w}{w^Tw}\,\max_{v\neq 0}\frac{v^T(Z^TZ)v}{v^Tv}\\
&= \lambda_{\max}(Z_0^TGZ_0)\lambda_{\max}(Z^TZ).
\end{aligned}$$

An analogous result is true for the smallest eigenvalue. The lemma follows immediately. □

The lemma shows that it is desirable to keep $\operatorname{cond}(Z)$ small. If an orthogonal factorization is used, then $\operatorname{cond}(Z) = 1$. More generally, $\operatorname{cond}(Z) = \sigma_{\max}(Z)/\sigma_{\min}(Z)$, the ratio of the largest and smallest singular values of Z. When variable reduction is used, the smallest singular value can be determined, as the following lemma indicates.

LEMMA 11. *Let Z be a null-space matrix obtained from variable reduction applied to an $m \times n$ matrix $A = (\,B \quad N\,)$, so that*

$$Z = \begin{pmatrix} -B^{-1}N \\ I \end{pmatrix}.$$

Then

$$\sigma_{\min}(Z) \geq \sqrt{1 + \frac{\sigma_{\min}(N)}{\sigma_{\max}(B)}} \geq 1.$$

If the dimension of the null space of N is k then the k smallest singular values of Z are equal to one. In particular, if $n > 2m$ then $\sigma_{\min}(Z) = 1$.

A simple consequence of the lemma is that

$$\text{cond}(Z) = \frac{\sigma_{\max}(Z)}{\sigma_{\min}(Z)} \leq \sigma_{\max}(Z) = \|Z\|.$$

Hence

$$\begin{aligned}
\text{cond}(Z)^2 &\leq \max_{\|u\|=1} \|Zu\|^2 \\
&= \max_{\|u\|=1} \left\| \begin{pmatrix} -B^{-1}Nu \\ u \end{pmatrix} \right\|^2 \\
&= \max_{\|u\|=1} \| - B^{-1}Nu\|^2 + \|u\|^2 \\
&= 1 + \max_{\|u\|=1} \| - B^{-1}Nu\|^2 \\
&= 1 + \|B^{-1}N\|^2.
\end{aligned}$$

Thus, keeping $\text{cond}(Z)$ small is closely related to keeping $\|B^{-1}N\|$ small. If $n > 2m$ they are the same.

There is considerable choice in the selection of B and N. Any subset of m variables that results in a nonsingular basis matrix B is acceptable. The procedure of selecting a basis, called a "crash" in the context of linear programming, can be based on various criteria. For example, columns of A can be selected to try to produce a matrix B that is sparse or nearly triangular. We would like to control $\|B^{-1}N\|$ as columns are added incrementally to the basis, but the matrix is only defined when B is invertible. However, we have the bound

$$\|B^{-1}N\| \leq \|B^{-1}\| \cdot \|N\| = \frac{\sigma_{\max}(N)}{\sigma_{\min}(B)}$$

and this bound is defined even if B is only partially formed.

Suppose that the columns of A are ordered according to some auxiliary criterion, such as sparsity, and that the columns will be considered for membership in B based on this ordering. We would like to estimate $\sigma_{\max}(N)/\sigma_{\min}(B)$ each time a column is added to the (trial) basis and then reject columns that cause the bound to be large. If N has many columns this would be expensive, so this is not likely to be useful as a general technique. In some circumstances, for example if all the constraints in the optimization problem were linear and so the basis would be used many times, it might be worth the effort.

In many cases we would expect that $\sigma_{\max}(N)$ would not vary greatly as B changed. This would be true if the columns of A were all scaled to have norm 1, or if the norms of particular columns were not pathologically large or small. For this reason it should be sufficient in many circumstances merely to monitor $\sigma_{\min}(B)$ as B is formed.

Traditional condition number estimators, such as the Linpack estimator, completely factor the matrix B before using the factorization to estimate the condition

number. Since the goal is to examine B as each new column is added, these techniques are not appropriate.

Incremental condition number estimators have been proposed [1], but they only apply to triangular matrices. They would be appropriate if a QR factorization of B were computed, but LU factorizations are more commonly used. To overcome this difficulty we propose exploiting the following upper bound:

$$\sigma_{\min}(B) = \sigma_{\min}(LU) \geq \sigma_{\min}(L)\sigma_{\min}(U).$$

This can be a considerable overestimate, but it is certainly true that if $\sigma_{\min}(L)$ and $\sigma_{\min}(U)$ are reasonably sized then so is $\sigma_{\min}(B)$.

The algorithm used to factor B will generally ensure that L is well conditioned. (Some software packages monitor U instead of L, in which case the comments below should be adjusted appropriately.) The diagonal entries in L will be equal to one, and the subdiagonal entries will be bounded (by one in the dense case, by a somewhat larger number in the sparse case). Thus, it would not be unreasonable to estimate simply $\sigma_{\min}(U)$. The incremental condition number estimator [1] can monitor this value as U is formed one column (or row) at a time.

To summarize, ill conditioning can be controlled, even when variable reduction is used to form Z. It requires, however, that the conditioning of the component matrices be monitored as they are formed. The trade-offs between cost and security are clear.

7. Calculus of variations. A classical problem in the calculus of variations is

$$\text{minimize}_{x(t)} \int_a^b J(t, x(t), x'(t))dt, \qquad x(a) = x_a,\ x(b) = x_b, \tag{7.1}$$

where $x(t)$ is some smooth real-valued function. The problem can be converted to a finite-dimensional problem by, for example, discretizing $x(t)$ and approximating it using a cubic spline. The Hessian of the finite-dimensional problem then has the form Z^TGZ and it is possible to apply the preconditioning ideas of the previous sections, even though this is an "unconstrained" problem and the matrix Z does not correspond to an explicit set of constraints. We examine this idea here. (The use of a cubic spline was suggested in [4], although the specific approach used here is different.)

The finite-dimensional analogs of (7.1) can be difficult to solve [12], with the Hessian having many small eigenvalues as the solution is approached. However, the components Z and G of the Hessian are easy to compute and G is easy to invert. The "null-space matrix" Z corresponds to the formulas for the cubic spline, and is independent of the formulas in (7.1). The matrix G is block diagonal, with 4×4 diagonal blocks if cubic splines are used. The formulas for the partial derivatives of $J(t, x, x')$ must be specified, but since J is only a function of three variables, this is not difficult.

To derive the formulas for G and Z, we first discretize the problem using

$$a = t_0 < t_1 < \cdots < t_n < t_{n+1} = b.$$

For simplicity we use equally spaced points: $t_i = a + ih$, where $h = (b - a)/(n + 1)$, although an adaptive mesh could also be used. The variables in the finite-dimensional problem will be $x_i \equiv x(t_i)$.

The function $x(t)$ will be approximated by a cubic spline $s(t)$ that interpolates the values $\{ x_i \}$ at the points $\{ t_i \}$. We use the representation described in [9]. On

the ith subinterval $[x_i, x_{i+1}]$,

$$s(t) = x_i b_1(t) + d_i b_2(t) + x_{i+1} b_3(t) + d_{i+1} b_4(t) \equiv \sum_{k=1}^{4} \alpha_{i,k} b_k(t),$$

where

$$\begin{aligned} b_1(t) &= \frac{2}{h^3}(t - t_{i+1})^2(t - t_i + h/2), \\ b_2(t) &= \frac{1}{h^2}(t - t_i)(t - t_{i+1})^2, \\ b_3(t) &= -\frac{2}{h^3}(t - t_i)^2(t - t_{i+1} - h/2), \\ b_4(t) &= \frac{1}{h^2}(t - t_i)^2(t - t_{i+1}), \end{aligned}$$

and d_i and d_{i+1} are estimates of $x'(t_i)$ and $x'(t_{i+1})$, respectively. It is straightforward to verify that $s(t_j) = x_j$ and $s'(t_j) = d_j$ for $j = i, i+1$. The coefficients $\{d_i\}$ are uniquely determined by the requirements that the function $s''(t)$ be continuous, together with two auxiliary conditions (usually involving the derivatives of $s(t)$ at the endpoints of the interval $[a, b]$) [9]. If we define $x = (x_0, \ldots, x_{n+1})^T$ and $d = (d_0, \ldots, d_{n+1})^T$ and use the auxiliary conditions

$$d_0 = \frac{x_1 - x_0}{h} \quad \text{and} \quad d_{n+1} = \frac{x_{n+1} - x_n}{h},$$

then d can be determined by solving a linear system of the form

$$Td = Sx,$$

where T and S are tridiagonal matrices. With our choice of auxiliary conditions,

$$S = \begin{pmatrix} -\frac{1}{h} & \frac{1}{h} & & & & \\ -3 & 0 & 3 & & & \\ & -3 & 0 & 3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -3 & 0 & 3 \\ & & & & -\frac{1}{h} & \frac{1}{h} \end{pmatrix}$$

and

$$T = \begin{pmatrix} 1 & 0 & & & & \\ h & 4h & h & & & \\ & h & 4h & h & & \\ & & \ddots & \ddots & \ddots & \\ & & & h & 4h & h \\ & & & & 0 & 1 \end{pmatrix}.$$

Only the nonzero entries of these matrices need be stored.

We are now in a position to define the matrices Z and G. We begin with Z. The first step is to write

$$\begin{pmatrix} x \\ d \end{pmatrix} = \begin{pmatrix} I \\ T^{-1}S \end{pmatrix} x \equiv Z_1 x.$$

(Of course, T^{-1} would not be formed explicitly since it is a dense matrix; Gaussian elimination would be used to perform the necessary calculations.)

The next step is to determine $\{\alpha_{i,k}\}$ from x and d. Let

$$\alpha = (\alpha_{0,1}, \alpha_{0,2}, \alpha_{0,3}, \alpha_{0,4}, \alpha_{1,1}, \ldots)^T$$

be the vector of spline coefficients. Then α is just a re-ordering of the variables $\{x_i\}$ and $\{d_i\}$ so that

$$\alpha = Z_2 \begin{pmatrix} x \\ d \end{pmatrix},$$

where Z_2 is a matrix with 0/1 entries. For example, if $n = 1$ then

$$\begin{pmatrix} \alpha_{0,1} \\ \alpha_{0,2} \\ \alpha_{0,3} \\ \alpha_{0,4} \\ \alpha_{1,1} \\ \alpha_{1,2} \\ \alpha_{1,3} \\ \alpha_{1,4} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ d_0 \\ d_1 \\ d_2 \end{pmatrix}.$$

This re-ordering of the variables is used so that the matrix G will be block diagonal. The matrix Z_2 would not be formed explicitly, but would instead be represented by a set of rules for obtaining $\{\alpha_{i,j}\}$ from $\{x_i\}$ and $\{d_i\}$. Then $Z = Z_2 Z_1$.

In this application the values x_0 and x_{n+1} are specified, so they are not really variables. An additional matrix Z_0 of the form

$$Z_0 = \begin{pmatrix} 0 & \cdots & 0 \\ & I & \\ 0 & \cdots & 0 \end{pmatrix}$$

could be used to remove them from the problem, giving $Z = Z_2 Z_1 Z_0$. In the numerical tests in §8, Z_0 was not used.

To determine the matrix G we must first estimate the integral in (7.1). We apply a quadrature rule separately on each subinterval $[t_i, t_{i+1}]$. If we use a four-point Gaussian quadrature rule [9] then

$$\int_{t_i}^{t_{i+1}} J(t, x(t), x'(t))dt \approx \sum_{j=1}^{4} w_j J(t_i + \theta_j, s(t_i + \theta_j), s'(t_i + \theta_j)),$$

where $\{w_j\}$ are the weights and $\{\theta_j\}$ are the abscissas for the quadrature rule:

$$\theta \approx h \begin{pmatrix} .0694318442030 \\ .3300094782075 \\ .6699905217925 \\ .9305681557970 \end{pmatrix} \quad \text{and} \quad w \approx h \begin{pmatrix} .1739274225685 \\ .3260725774315 \\ .3260725774315 \\ .1739274225685 \end{pmatrix}.$$

The weights and abscissas are the same for all subintervals.

The matrix G is block-diagonal, with one block for each subinterval. Within each block the (k,l) entry is the second partial derivative of the quadrature formula with respect to $\alpha_{i,k}$ and $\alpha_{i,l}$:

$$\sum_{j=1}^{4} w_j \left[\frac{\partial^2 J}{\partial s^2} b_k(t_i+\theta_j) b_l(t_i+\theta_j) + \frac{\partial^2 J}{\partial s \partial s'} b'_k(t_i+\theta_j) b_l(t_i+\theta_j) \right.$$
$$\left. + \frac{\partial^2 J}{\partial s \partial s'} b_k(t_i+\theta_j) b'_l(t_i+\theta_j) + \frac{\partial^2 J}{\partial s'^2} b'_k(t_i+\theta_j) b'_l(t_i+\theta_j) \right].$$

The formulas for $b_k(t)$ and $b'_k(t)$ depend only on those for the spline $s(t)$. The partial derivatives of J with respect to s and s' must be derived, but this is the only calculation that depends on J. Hence, once the partial derivatives of J with respect to s and s' have been specified, the derivatives of the discretized calculus of variations problem can be computed in a straightforward, general-purpose manner that is independent of the particular problem being solved. Finally, the Hessian of the discretized calculus of variations problem with respect to the variables $\{ x_i \}$ is given by $Z^T G Z$.

A left-inverse W^T is easy to obtain. It can be chosen as a column permutation of the matrix $(I, 0)$. For example it is possible to choose $W_{1,1} = 1$ and then $W_{i,4i-5} = 1$ for $i > 1$, and all other entries $W_{i,j} = 0$.

It is also possible to determine the left-inverse $W^T = (Z^T Z)^{-1} Z^T$ in a computationally efficient manner, even though $(Z^T Z)$ is a dense matrix. Since $Z = Z_2 Z_1$, $Z^T Z = Z_1^T Z_2^T Z_2 Z_1$. It is straightforward to show that

$$Z_2^T Z_2 = \begin{pmatrix} D & 0 \\ 0 & D \end{pmatrix},$$

where D is an $(n+2)\times(n+2)$ diagonal matrix with diagonal entries $D_{1,1} = D_{n+2,n+2} = 1$ and $D_{i,i} = 2$ for $2 < i < n+1$. Then

$$Z^T Z = Z_1^T Z_2^T Z_2 Z_1 = D + S^T T^{-T} D T^{-1} S$$

in terms of the tridiagonal matrices T and S defined earlier. The last formula is the Schur complement of $-TD^{-1}T^T$ in the block 2×2 matrix

$$H \equiv \begin{pmatrix} -TD^{-1}T^T & S \\ S^T & D \end{pmatrix}.$$

If we define $\bar{Z}^T = (0\ I)$, then

$$(Z^T Z)^{-1} = (D + S^T T^{-T} D T^{-1} S)^{-1} = \bar{Z}^T H^{-1} \bar{Z}.$$

The matrix H is a permutation of a 7-diagonal matrix, which can be factored inexpensively. With this approach, the left-inverse $W^T = (Z^T Z)^{-1} Z^T$ can be formed.

8. Computational experiments. We tested the inverse preconditioners (3.6) for various values of k on three examples. The calculations were done using MATLAB on a Sun SPARCstation computer, with machine precision $\approx 2.2 \times 10^{-16}$.

In the first example, G was a diagonal matrix with $G_{i,i} = 2\gamma^{i-1}$, where γ was chosen so that $G_{n,n} = 10^7$. The $m \times n$ constraint matrix A was of the form $U \Sigma V^T$, where Σ was a diagonal $m \times n$ matrix with diagonal entries $\Sigma_{i,i} = i$, and U and V were

random square orthogonal matrices. The random number generator was initialized using the MATLAB command `randn('seed',0)` before each run, and random numbers were generated using the `randn` function. (This complicated method of generating A allowed us to control its singular values.)

The second example is based on the calculus of variations problem of §7. The integral was evaluated over the interval $[a, b] = [0, 1]$. The variables were given the values $x_i = 1 - t_i^2$. The kernel was chosen as the convex function

$$J(t, x, x') = x^4 + (x')^4,$$

for which G is positive definite.

For the first problem, two choices of Z were used: an orthogonal Z with $W = Z(Z^TZ)^{-1} = Z$, and a Z based on variable reduction (the first m variables forming the basis) together with the "special purpose" $W^T = (0, I)$. For the second problem Z is given, but we tested two choices of W: $W = Z(Z^TZ)^{-1}$ and the "simple" W mentioned in §7 (i.e., W^T is a permutation of the matrix $(I, 0)$).

The third problem uses matrices of the form

$$G = \begin{pmatrix} G_1 & I \\ I & G_2 \end{pmatrix} \quad \text{and} \quad Z = \begin{pmatrix} I \\ 0 \end{pmatrix},$$

where all the blocks are the same size. The reduced matrix is $Z^TGZ = G_1$. The matrix G_1 was chosen as in the first problem (note that it is a matrix of order $n/2$). The diagonal matrix G_2 was chosen so that the eigenvalues of the preconditioned matrix were the fifth roots of the eigenvalues of G_1. For this problem, both choices of Z are the same.

We were interested in the behavior of the preconditioners under "ideal" circumstances. We therefore used $M = G$ in all problems. To ensure convergence of the series defining the inverse preconditioners, we chose the scaling factor so that α^{-1} was the largest eigenvalue of T (see (3.4)).

The first problem was tested with $n = 100$ and $m = 20$ so that the reduced matrix was 80×80. The second problem was tested with $n = 60$ so that the reduced matrix was 62×62 and G was 244×244. The third problem was tested with $n = 200$ and $m = 100$ so that the reduced matrix was 100×100. In all cases the effect of the preconditioner was assessed in two ways: (i) by the condition number of the preconditioned matrix, and (ii) by the number of conjugate-gradient iterations required to solve a linear system with right-hand side $(1, \ldots, 1)^T$. The conjugate-gradient iterations were terminated when the norm of the residual was less than 10^{-8}. The results are listed in Tables 1 and 2.

The tables indicate that the preconditioning strategies can greatly reduce the number of iterations required to solve a linear system using the conjugate-gradient method. The effect on the condition number is less pronounced. The choice of the "generic" left-inverse $W^T = (Z^TZ)^{-1}Z^T$ worked well in all these cases (as in many others that we tried). The other "special-purpose" choices of W (i.e., those constructed from permutations of $(0, I)$) were less predictable. In the case of variable reduction we found them to be effective (in fact, better than other choices). However, for the calculus of variations problems we were unable to find a special-purpose choice of W that worked well.

For Problems 1 and 2 (and for most examples that we tried), the most dramatic improvement comes with the simplest of the inverse preconditioners, $W^TM^{-1}W$. This

TABLE 1
Effect of preconditioning using $W^T = (Z^TZ)^{-1}Z^T$.

Preconditioner	Problem 1		Problem 2		Problem 3	
	cond	iter	cond	iter	cond	iter
None	1.0×10^6	756	1.3×10^5	164	2.2×10^3	240
$k = 0$	6.8×10^4	34	5.4×10^4	39	4.6×10^2	149
$k = 1$	3.4×10^4	31	2.7×10^4	39	2.3×10^2	125
$k = 2$	2.3×10^4	29	1.8×10^4	38	1.5×10^2	111
$k = 3$	1.7×10^4	27	1.4×10^4	40	1.2×10^2	99
$k = 4$	1.4×10^4	25	1.1×10^4	39	9.3×10^1	88

TABLE 2
Effect of preconditioning using special-purpose W.

Preconditioner	Problem 1		Problem 2	
	cond	iter	cond	iter
None	1.5×10^5	529	1.3×10^5	164
$k = 0$	4.8×10^1	13	5.4×10^7	395
$k = 1$	2.4×10^1	13	2.7×10^7	396
$k = 2$	1.6×10^1	13	1.8×10^7	379
$k = 3$	1.2×10^1	13	1.3×10^7	372
$k = 4$	9.9×10^0	13	1.1×10^7	374

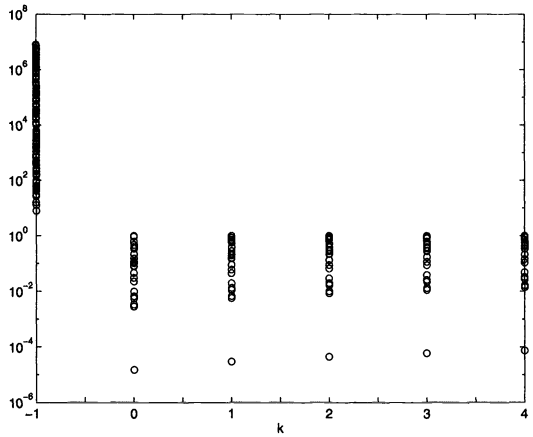


FIG. 1. *Eigenvalues of the preconditioned matrix for Problem* 1.

is reassuring, since it has lower costs than the others. It also does not require that α be selected to ensure convergence of the series used in the derivation. Adding more terms in the series leads to further improvements, but whether these improvements are cost-effective would depend on the specifics of the particular application. For Problem 3, the later terms lead to considerable reductions in the number of iterations required, indicating the potential of these techniques.

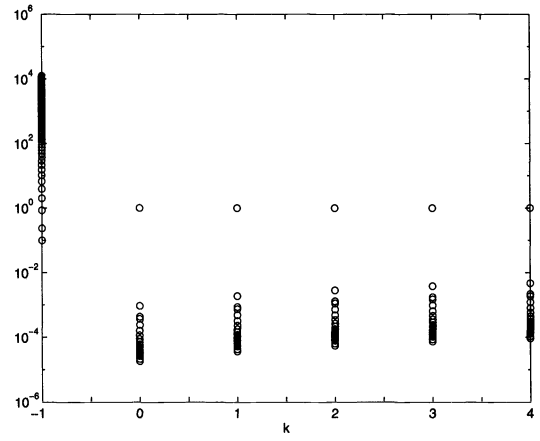


FIG. 2. *Eigenvalues of the preconditioned matrix for Problem* 2.

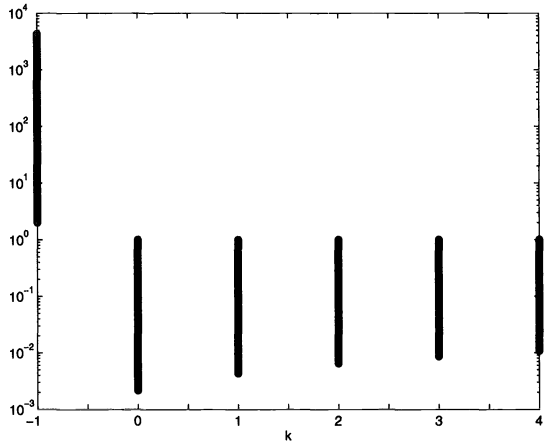


FIG. 3. *Eigenvalues of the preconditioned matrix for Problem* 3.

As a final indication of the behavior of these preconditioners, Figs. 1–3 show the eigenvalues of the preconditioned matrix, corresponding to the results in Table 1. In each figure, column -1 shows the eigenvalues of the original matrix, and columns 0–4 show the eigenvalues of the preconditioned matrix for $k = 0, \ldots, 4$. Note that the range of the eigenvalues has been compressed (hence the reduced condition number) and in some cases there is increased clustering of eigenvalues, a feature that the conjugate-gradient method can exploit.

9. Conclusions. We have described a set of preconditioners for positive-definite matrices of the form Z^TGZ, using information about the individual matrices Z and G to construct approximations to $(Z^TGZ)^{-1}$. Matrices of this type arise in constrained optimization problems, in particular in interior-point methods for linear programming. The techniques can also be applied to a class of problems in the calculus of variations. Although some of the preconditioning formulas may be too expensive for routine use,

numerical tests suggest that the simplest of the formulas (3.2) can be an effective preconditioner for general use. The more elaborate formulas would be appropriate in cases where a product Gv is computationally expensive. Because (3.2) and the other formulas only require information about Z and G *separately* (and do not require that an approximation to Z^TGZ be provided), the preconditioners can be used within a wide variety of optimization algorithms. This is significant, since the structure of G often comes from the optimization model, whereas the form of Z is determined by the optimization software.

Acknowledgments. We wish to thank an anonymous referee for the careful reading of our paper and for the insightful comments that led to many improvements.

REFERENCES

[1] C.H. BISCHOF, *Incremental condition estimation,* SIAM J. Matrix Anal. Appl., 11 (1990), pp. 312–322.

[2] R.S. DEMBO AND T. STEIHAUG, *Truncated-Newton algorithms for large-scale unconstrained optimization,* Math. Prog., 26 (1983), pp. 190–212.

[3] R. FLETCHER, *Practical Methods of Optimization, Volume* 2 : *Constrained Optimization,* Wiley, New York, 1981.

[4] P.E. GILL AND W. MURRAY, *The numerical solution of a problem in the calculus of variations,* in Recent Mathematical Developments in Control, D.J. Bell, ed., Academic Press, London, 1979, pp. 97–122.

[5] P.E. GILL, W. MURRAY, D.B. PONCELEÓN, AND M.A. SAUNDERS, *Preconditioners for indefinite systems arising in optimization,* SIAM J. Matrix Anal. Appl., 13 (1992), pp. 292–311.

[6] P.E. GILL, W. MURRAY, AND M.H. WRIGHT, *Practical Optimization,* Academic Press, New York, 1981.

[7] G.H. GOLUB AND C. VAN LOAN, *Matrix Computations,* 2nd ed., The Johns Hopkins University Press, Baltimore, 1989.

[8] M.R. HESTENES, *Conjugate-Direction Methods in Optimization,* Springer-Verlag, New York, 1980.

[9] D.K. KAHANER, C. MOLER, AND S.G. NASH, *Numerical Methods and Software,* Prentice-Hall, Englewood Cliffs, NJ, 1989.

[10] S.G. NASH, *Newton-like minimization via the Lanczos method,* SIAM J. Numer. Anal., 21 (1984), pp. 770–788.

[11] ———, *Preconditioning of truncated-Newton methods,* SIAM J. Sci. Statist. Comput., 6 (1985), pp. 599–616.

[12] S.G. NASH AND J. NOCEDAL, *A numerical study of the limited memory* BFGS *method and the truncated-Newton method for large scale optimization,* SIAM J. Optim., 1 (1991), pp. 358–372.

[13] S.G. NASH AND A. SOFER, *A barrier method for large-scale constrained optimization,* ORSA J. Comput., 5 (1993), pp. 40–53.

[14] T. RUSTEN AND R. WINTHER, *A preconditioned iterative method for saddlepoint problems,* SIAM J. Matrix Anal. Appl., 13 (1992), pp. 887–904.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 1, pp. 69–82, January 1996

004

RESIDUAL BOUNDS ON APPROXIMATE SOLUTIONS FOR THE UNITARY EIGENPROBLEM*

JI-GUANG SUN†

Abstract. Let A be an $n \times n$ unitary matrix, and let the columns of an $n \times l\,(l < n)$ matrix $\tilde{X}_1$ form an orthonormal basis for an approximate eigenspace $\tilde{\mathcal{X}}_1$ of A. Then there are two problems: How near is $\tilde{\mathcal{X}}_1$ to an eigenspace of A? How can we make use of the $l \times l$ matrix $\tilde{X}_1^H A \tilde{X}_1$ to get l approximate eigenvalues of A? This paper gives solutions to these problems. In particular, this paper reveals such a fact: One can use the eigenvalues of the unitary polar factor of $\tilde{X}_1^H A \tilde{X}_1$ (or the eigenvalues of the matrix $\tilde{X}_1^H A \tilde{X}_1$) as l approximate eigenvalues of A, and the precision of the eigenvalues of the unitary polar factor of $\tilde{X}_1^H A \tilde{X}_1$ (or the eigenvalues of $\tilde{X}_1^H A \tilde{X}_1$) as l approximate eigenvalues of A is higher than that of $\tilde{\mathcal{X}}_1$ as an approximate eigenspace of A.

Key words. unitary matrix, eigenvalues, eigenvectors, eigenspaces, perturbation bounds, residual bounds, Rayleigh quotient matrix, backward perturbation analysis

AMS subject classifications. 15A18, 15A42, 65F15

1. Introduction.

In recent years, computational methods and perturbation theory for the unitary eigenproblem have been developed (e.g., see [1], [3], [4], [8], [9]). The purpose of this paper is to derive several a posteriori error bounds for the eigenproblem.

Let A be an $n \times n$ unitary matrix. Assume that the columns of an $n \times l\,(l < n)$ matrix $\tilde{X}_1$ form an orthonormal basis for an approximate eigenspace $\tilde{\mathcal{X}}_1$ of A. For instance, $\tilde{X}_1$ may come from a numerical algorithm for approximating eigenspaces. Then there are two problems: How near is $\tilde{\mathcal{X}}_1$ to an eigenspace of A? How can we make use of the $l \times l$ matrix $\tilde{X}_1^H A \tilde{X}_1$ to get l approximate eigenvalues of A? This paper gives solutions to these problems.

In §2 we cite and prove some lemmas. In §§3–4 we derive a posteriori error bounds from approximate eigenspaces. All the a posteriori error bounds are residual bounds. In §5 we give a numerical example to illustrate our main results.

It is well known that if A is an $n \times n$ Hermitian matrix, and if the columns of an $n \times l$ matrix $\tilde{X}_1$ form an orthonormal basis for an approximate eigenspace $\tilde{\mathcal{X}}_1$ of A, then one can use the eigenvalues of the Rayleigh quotient matrix $\tilde{X}_1^H A \tilde{X}_1$ as l approximate eigenvalues of A, and the precision of the eigenvalues of $\tilde{X}_1^H A \tilde{X}_1$ as l approximate eigenvalues of A is higher than that of $\tilde{\mathcal{X}}_1$ as an approximate eigenspace of A [17, pp. 254–257], [18]. This paper shows that the unitary eigenproblem has the same property. Moreover, this paper reveals such a fact: For a unitary matrix A, one can also use the eigenvalues of the unitary polar factor P_1 of $\tilde{X}_1^H A \tilde{X}_1$ as l approximate eigenvalues of A, and the precision of the eigenvalues of P_1 as l approximate eigenvalues of A is higher than that of $\tilde{\mathcal{X}}_1$ as an approximate eigenspace of A.

Throughout this paper we use the following notation and definitions. The symbol $\mathcal{C}^{m \times n}$ denotes the set of complex $m \times n$ matrices, and $\mathcal{C}^n = \mathcal{C}^{n \times 1}$. A^T stands for the transpose of a matrix A, A^H for the conjugate transpose of A, and $A^\dagger$ for the Moore–Penrose inverse of A. I is the identity matrix, $I^{(n)}$ is the identity matrix of

* Received by the editors June 15, 1994; accepted for publication (in revised form) by A. Bunse-Gerstner January 23, 1995. This work 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).

order n, and 0 is the null matrix. $\mathcal{R}(A)$ denotes the column space of A. $\lambda(A)$ denotes the set of all eigenvalues of A. $\sigma_j(A)$, $j = 1,\dots,n$, denote the singular values of $A \in \mathcal{C}^{n\times n}$ arranged in decreasing order $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_n(A)$. $\sigma_{\min}(A)$ is the smallest singular value of A. $H \geq 0$ denotes that the matrix H is Hermitian and positive semidefinite. For a Hermitian positive semidefinite matrix H, the matrix $H^{1/2}$ denotes the unique Hermitian positive semidefinite square root of H. $\emptyset$ is the empty set. $\|\ \|_2$ denotes the spectral norm, and $\|\ \|_{\mathrm{F}}$ the Frobenius norm.

Let $A \in \mathcal{C}^{n\times n}$ be unitary, and let $\mathcal{X}_1$ be a subspace of $\mathcal{C}^n$. The subspace $\mathcal{X}_1$ is an eigenspace of A if $A\mathcal{X}_1 \subseteq \mathcal{X}_1$.

Let the columns of $X_1 \in \mathcal{C}^{n\times l}$ form an orthonormal basis for a subspace $\mathcal{X}_1$. Then it is easy to prove that $\mathcal{X}_1$ is an eigenspace of a unitary matrix A if and only if there is a unitary $Z_1 \in \mathcal{C}^{l\times l}$ such that $AX_1 = X_1Z_1$.

Let $\mathcal{X}_1 = \mathcal{R}(X_1)$ and $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1)$, in which $X_1, \tilde{X}_1 \in \mathcal{C}^{n\times l}$ with $X_1^HX_1 = \tilde{X}_1^H\tilde{X}_1 = I$. Define

$$\Theta(X_1,\tilde{X}_1) = \arccos(X_1^H\tilde{X}_1\tilde{X}_1^HX_1)^{1/2} \geq 0. \tag{1.1}$$

Then it is known that for every unitarily invariant norm $\|\ \|$, $\|\sin\Theta(X_1,\tilde{X}_1)\|$ is a generalized chordal distance between the subspaces $\mathcal{X}_1$ and $\tilde{\mathcal{X}}_1$ (e.g., see [17, p. 94]). Moreover, we have

$$\begin{aligned}\|\sin\Theta(X_1,\tilde{X}_1)\| &= \|\tan\Theta(X_1,\tilde{X}_1)\left(1+\tan^2\Theta(X_1,\tilde{X}_1)\right)^{-1/2}\| \\ &\leq \frac{\|\tan\Theta(X_1,\tilde{X}_1)\|}{[1+\sigma_{\min}^2(\tan\Theta(X_1,\tilde{X}_1))]^{1/2}} \leq \|\tan\Theta(X_1,\tilde{X}_1)\|.\end{aligned} \tag{1.2}$$

Let $B \in \mathcal{C}^{l\times l}$ and $C \in \mathcal{C}^{m\times m}$. $\mathrm{sep}_p(B,C)$, the separation of B and C, is defined by [15]

$$\mathrm{sep}_p(B,C) = \inf_{\substack{P\in\mathcal{C}^{m\times l}\\ \|P\|_p=1}} \|PB - CP\|_p, \tag{1.3}$$

where $p = 2, F$.

2. Lemmas. The following lemmas will be used in the next section.

LEMMA 2.1. *Let $B, C \in \mathcal{C}^{n\times l}$ be given. Define*

$$\mathcal{A} = \{A \in \mathcal{C}^{n\times n} \ :\ A^HA = I,\ AB = C\}. \tag{2.1}$$

Then $\mathcal{A} \neq \emptyset$ if and only if B, C satisfy

$$B^HB = C^HC, \tag{2.2}$$

and in the case of $\mathcal{A} \neq \emptyset$, any $A \in \mathcal{A}$ can be expressed by

$$A = CB^\dagger + FG^H, \tag{2.3}$$

where $F, G \in \mathcal{C}^{n\times(n-r)}$ with $r = \mathrm{rank}(B) = \mathrm{rank}(C)$ satisfy

$$F^HF = G^HG = I \quad \text{and} \quad G^HB = F^HC = 0. \tag{2.4}$$

Proof. It is evident that if $\mathcal{A} \neq \emptyset$ then B, C satisfy (2.2).

Now we assume that B, C satisfy (2.2). Let

$$B = U \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix} V^H = U_1 \Sigma_1 V_1^H \tag{2.5}$$

be the singular value decomposition of B, where $\Sigma_1 = \operatorname{diag}(\sigma_j)$ with $\sigma_1 \geq \cdots \geq \sigma_r > 0$, $U = (U_1, U_2) \in \mathcal{C}^{n \times n}$ and $V = (V_1, V_2) \in \mathcal{C}^{l \times l}$ are unitary with $U_1 \in \mathcal{C}^{n \times r}, V_1 \in \mathcal{C}^{l \times r}$. Substituting (2.5) into (2.2) gives $CV_2 = 0$ and $CV_1 = Q_1 \Sigma_1$, in which $Q_1 \in \mathcal{C}^{n \times r}$ with $Q_1^H Q_1 = I$. Thus, we have

$$C = Q \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix} V^H, \tag{2.6}$$

where $Q = (Q_1, Q_2) \in \mathcal{C}^{n \times n}$ is unitary. The relations (2.5) and (2.6) imply that the unitary matrix $A = QU^H \in \mathcal{A}$. Therefore, $\mathcal{A} \neq \emptyset$.

Let A be an arbitrary matrix in $\mathcal{A}$, and let

$$E = A - CB^\dagger, \quad P = A^H Q, \tag{2.7}$$

where Q is the unitary matrix in (2.6). Then from $AB = C$ and (2.5)–(2.7)

$$P^H U = \begin{pmatrix} I^{(r)} & 0 \\ 0 & W_2 \end{pmatrix} \quad \text{with unitary } W_2 \in \mathcal{C}^{(n-r) \times (n-r)}.$$

Combining it with (2.7) we get

$$E = QP^H UU^H - Q \begin{pmatrix} I^{(r)} & 0 \\ 0 & 0 \end{pmatrix} U^H = Q_2 W_2 U_2^H.$$

Let $F = Q_2 W_2$ and $G = U_2$. Then A is expressed by (2.3), where F, G satisfy (2.4).

Conversely, if a matrix A can be expressed by (2.3) and (2.4), then we have

$$AB = CB^\dagger B = CC^\dagger C = C,$$

and

$$\begin{aligned} A^H A &= {B^\dagger}^H C^H C B^\dagger + G F^H F G^H = {B^\dagger}^H B^H B B^\dagger + G G^H \\ &= BB^\dagger + GG^H = U_1 U_1^H + U_2 U_2^H = I. \end{aligned}$$

Therefore, $A \in \mathcal{A}$. Note that the equality $GG^H = U_2 U_2^H$ is due to the fact that from (2.5) and $G^H B = 0$ the matrix G can be expressed by $G = U_2 Z$ with a unitary $Z \in \mathcal{C}^{(n-l) \times (n-l)}$. The proof is completed.

LEMMA 2.2. *Given unitary matrices $A, X = (X_1, X_2) \in \mathcal{C}^{n \times n}$ with $X_1 \in \mathcal{C}^{n \times l}$. Then*

$$\min_{\substack{Z_1 \in \mathcal{C}^{l \times l} \\ Z_1^H Z_1 = I}} \|X_1 Z_1 - A X_1\|_F = \|X_1 P_1 - A X_1\|_F, \tag{2.8}$$

where P_1 is the unitary polar factor of $X_1^H A X_1$.

Proof. See [7, p. 582] or [12, pp. 431–432] (see also [6] and [10] for the polar decomposition and polar factors).

LEMMA 2.3 [2]. *Let $B, C \in \mathcal{C}^{n\times n}$ be unitary matrices, and*

$$\Gamma = \operatorname{diag}(\gamma_i), \quad 0 \le \gamma_1 \le \cdots \le \gamma_n.$$

Then for every unitarily invariant norm $\|\ \|$

$$\|B\Gamma - \Gamma C\| \ge \gamma_1 \|B - C\|. \tag{2.9}$$

Lemma 2.3 has been proved in [2] by two steps: The first step is to apply von Neumann's Maximum Principle [21, Theorem 1] (see also [17, Chap. II, Lemma 3.4]) and Fan's Dominance Theorem [5, Theorem 4] (see also [17, Chap. II, Theorem 3.17 and Corollary 3.18]) to prove that the inequality (2.9) is true for Hermitian matrices B and C. The second step is to construct Hermitian matrices $\hat{B}, \hat{C}$ and a diagonal matrix $\hat{\Gamma}$ by

$$\hat{B} = \begin{pmatrix} 0 & B \\ B^H & 0 \end{pmatrix}, \quad \hat{C} = \begin{pmatrix} 0 & C \\ C^H & 0 \end{pmatrix}, \quad \hat{\Gamma} = \begin{pmatrix} \Gamma & 0 \\ 0 & \Gamma \end{pmatrix},$$

and to derive (2.9) from

$$\|\hat{B}\hat{\Gamma} - \hat{\Gamma}\hat{C}\| \ge \gamma_1 \|\hat{B} - \hat{C}\|.$$

LEMMA 2.4. *Let $B \in \mathcal{C}^{m\times n}$ $(m \ge n)$ have the polar decomposition $B = PH$, where $P \in \mathcal{C}^{m\times n}$ satisfies $P^H P = I$, and $H \in \mathcal{C}^{n\times n}$ is Hermitian positive semidefinite. Then*

$$\|B - P\| \le \frac{\|B^H B - I\|}{1 + \sigma_{\min}(B)},$$

for every unitarily invariant norm $\|\ \|$.

Proof. From

$$P(B^H B - I) = P(H^2 - I) = P(H - I)(H + I) = (B - P)(H + I),$$

we see that for every unitarily invariant norm $\|\ \|$

$$\begin{aligned} \|B^H B - I\| &= \|P(B^H B - I)\| \ge \|(H + I)^{-1}\|_2^{-1} \|B - P\| \\ &= [1 + \sigma_{\min}(B)]\|B - P\|. \end{aligned}$$

□

3. Residual bounds (I). In this section we first use the method of backward perturbation analysis (see [19]; see also [13], [14], [17], [22]) to derive an a posteriori error bound for an approximate eigenspace of a unitary matrix. The key step is to derive an explicit expression for the optimal backward perturbation from the approximate eigenspace (see Theorem 3.1). Conventional perturbation theory can then be used to get an a posteriori error bound for the approximate eigenspace (see Theorem 3.2).

THEOREM 3.1. *Let $A, \tilde{X} = (\tilde{X}_1, \tilde{X}_2) \in \mathcal{C}^{n\times n}$ be unitary with $\tilde{X}_1 \in \mathcal{C}^{n\times l}$, and $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1)$ approximate an eigenspace of A. Define*

$$\mathcal{W} = \{W \in \mathcal{C}^{n\times n} : (A + W)^H (A + W) = I,\ (A + W)\tilde{\mathcal{X}}_1 \subseteq \tilde{\mathcal{X}}_1\}, \tag{3.1}$$

$$\tilde{A}_k = \tilde{X}_k^H A \tilde{X}_k, \quad k = 1, 2, \tag{3.2}$$

and let

$$\tilde{A}_k = P_k H_k \tag{3.3}$$

be the polar decompositions of $\tilde{A}_k$, *where* P_k *are unitary, and* H_k *are Hermitian positive definite,* $k = 1, 2$. *Moreover, let*

$$R_k = \tilde{X}_k P_k - A\tilde{X}_k \tag{3.4}$$

be residuals of A *with respect to* $\tilde{X}_k$ *and* P_k, $k = 1, 2$. *Then there is a matrix* $W^{(0)} \in \mathcal{W}$ *expressed by*

$$W^{(0)} = \tilde{X}\mathrm{diag}(P_1, P_2)\tilde{X}^H - A \tag{3.5}$$

such that $\|W^{(0)}\|_F = \min_{W\in\mathcal{W}} \|W\|_F$, *and*

$$\|W^{(0)}\|_F = \sqrt{\|R_1\|_F^2 + \|R_2\|_F^2}. \tag{3.6}$$

Proof. From (3.1) we know that a matrix $W \in \mathcal{W}$ if and only if W is a solution to the equation

$$W\tilde{X}_1 = \tilde{X}_1 Z_1 - A\tilde{X}_1 \tag{3.7}$$

for an arbitrarily fixed unitary matrix $Z_1 \in \mathcal{C}^{l\times l}$. By Lemma 2.1, $\mathcal{W} \neq \emptyset$, and any solution W to (3.7) can be expressed by

$$W = \tilde{X}_1 Z_1 \tilde{X}_1^H + Q_2 \tilde{X}_2^H - A, \tag{3.8}$$

where $Q_2 \in \mathcal{C}^{n\times(n-l)}$ satisfies

$$Q_2^H Q_2 = I, \quad Q_2^H \tilde{X}_1 = 0. \tag{3.9}$$

The relation (3.9) implies that the matrix Q_2 can be expressed by

$$Q_2 = \tilde{X}_2 Z_2 \quad \text{with} \quad Z_2 \in \mathcal{C}^{(n-l)\times(n-l)}, \ Z_2^H Z_2 = I.$$

Thus, the solution W expressed by (3.8) can be rewritten as

$$\begin{aligned} W &= \tilde{X}_1 Z_1 \tilde{X}_1^H + \tilde{X}_2 Z_2 \tilde{X}_2^H - A \\ &= (\tilde{X}_1 Z_1 - A\tilde{X}_1, \tilde{X}_2 Z_2 - A\tilde{X}_2)\tilde{X}^H. \end{aligned} \tag{3.10}$$

Consequently, we have

$$\|W\|_F = \sqrt{\|\tilde{X}_1 Z_1 - A\tilde{X}_1\|_F^2 + \|\tilde{X}_2 Z_2 - A\tilde{X}_2\|_F^2}. \tag{3.11}$$

By Lemma 2.2,

$$\min_{Z_k^H Z_k = I} \|\tilde{X}_k Z_k - A\tilde{X}_k\|_F = \|\tilde{X}_k P_k - A\tilde{X}_k\|_F = \|R_k\|_F, \quad k = 1, 2.$$

Combining it with (3.10) and (3.11) we get (3.5) and (3.6). □

Applying Theorem 3.1 and Stewart's result [15, Theorem 4.11] on perturbation bounds for invariant subspaces, we can derive an a posteriori error bound for an approximate eigenspace of a unitary matrix.

THEOREM 3.2. *Let* $A, \tilde{X} = (\tilde{X}_1, \tilde{X}_2), W^{(0)}, \tilde{A}_k, P_k\,(k = 1.2)$ *be as in Theorem* 3.1, $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1)$ *be an approximate eigenspace of* A, *and let*

$$\hat{R}_1 = \tilde{X}_1\tilde{A}_1 - A\tilde{X}_1 \tag{3.12}$$

be the residual of A *with respect to* $\tilde{X}_1$ *and* $\tilde{A}_1$. *If for* $p = 2, F$

$$\delta_p \equiv \operatorname{sep}_p(P_1, P_2) - (\|\tilde{A}_1 - P_1\|_p + \|\tilde{A}_2 - P_2\|_p) > 0, \quad \frac{2\|\hat{R}_1\|_p}{\delta_p} < 1, \tag{3.13}$$

then there is an eigenspace $\mathcal{X}_1 = \mathcal{R}(X_1)$ *of* A, *where* $X_1 \in \mathcal{C}^{n\times l}$ *with* $X_1^H X_1 = I$, *such that*

$$\|\tan\Theta(X_1, \tilde{X}_1)\|_p < \frac{2\|\hat{R}_1\|_p}{\delta_p}, \tag{3.14}$$

where $\Theta(X_1, \tilde{X}_1)$ *and* sep_p *are defined by* (1.1) *and* (1.3), *respectively.*

Proof. From (3.5) we get

$$\tilde{X}^H(A + W^{(0)})\tilde{X} = \begin{pmatrix} P_1 & 0 \\ 0 & P_2 \end{pmatrix} \tag{3.15}$$

and

$$\tilde{X}^H(-W^{(0)})\tilde{X} = \begin{pmatrix} \tilde{A}_1 - P_1 & \tilde{X}_1^H A\tilde{X}_2 \\ \tilde{X}_2^H A\tilde{X}_1 & \tilde{A}_2 - P_2 \end{pmatrix}.$$

By [15, Theorem 4.11], if for $p = 2, F$

$$\delta_p \equiv \operatorname{sep}_p(P_1, P_2) - (\|\tilde{A}_1 - P_1\|_p + \|\tilde{A}_2 - P_2\|_p) > 0, \quad \frac{\|\tilde{X}_2 A\tilde{X}_1\|_p}{\delta_p} < \frac{1}{2}, \tag{3.16}$$

then there is an eigenspace $\mathcal{X}_1$ of A satisfying

$$\|\tan\Theta(X_1, \tilde{X}_1)\|_p < \frac{\|\tilde{X}_2^H A\tilde{X}_1\|_p}{\delta_p}. \tag{3.17}$$

Observe that for $p = 2, F$

$$\begin{aligned} \|\tilde{X}_2^H A\tilde{X}_1\|_p &= \|\tilde{X}^H(\tilde{X}_1\tilde{A}_1 - A\tilde{X}_1)\|_p \\ &= \|\tilde{X}_1\tilde{A}_1 - A\tilde{X}_1\|_p = \|\hat{R}_1\|_p. \end{aligned} \tag{3.18}$$

Hence, (3.16) and (3.17) can be rewritten as (3.13) and (3.14), respectively. □

The following two results (Theorem 3.3 and Corollary 3.5) will reveal such a fact: If $\mathcal{R}(\tilde{X}_1)$ is close to an eigenspace of A, then the eigenvalues $\mu_1, \ldots, \mu_l$ of the unitary polar factor P_1 of $\tilde{X}_1^H A\tilde{X}_1$ are l approximate eigenvalues of A, and the precision of the eigenvalues $\mu_1, \ldots, \mu_l$ of P_1 as l approximate eigenvalues of A is higher than that of $\mathcal{R}(\tilde{X}_1)$ as an approximate eigenspace of A.

THEOREM 3.3. *Let* $A, \tilde{X} = (\tilde{X}_1, \tilde{X}_2), \tilde{\mathcal{X}}_1, \tilde{A}_1, \hat{R}_1$ *and* $P_k, H_k\,(k = 1, 2)$ *be as in Theorems* 3.1–3.2. *Suppose that the condition* (3.13) *is satisfied, and let* $\mathcal{X}_1 = \mathcal{R}(X_1)$ *be the eigenspace of* A *satisfying* (3.14). *Moreover, let*

$$A_1 = X_1^H A X_1, \quad \lambda(A_1) = \{\lambda_j\}_{j=1}^l, \quad \lambda(P_1) = \{\mu_j\}_{j=1}^l. \tag{3.19}$$

If $\|\sin\Theta(X_1, \tilde{X}_1)\|_2 < 1$, *then there are permutations* π_f *and* π_2 *of* $\{1, \ldots, l\}$ *such that*

$$\sqrt{\sum_{j=1}^{l} |\lambda_{\pi_f(j)} - \mu_j|^2} \leq \frac{1}{\sqrt{1-\sigma_1^2}} \cdot \left(\sigma_1 + \frac{\|\hat{R}_1\|_2}{1+\alpha}\right) \cdot \|\hat{R}_1\|_F, \tag{3.20}$$

and

$$|\lambda_{\pi_2(j)} - \mu_j| \leq \frac{1}{\sqrt{1-\sigma_1^2}} \cdot \left(\sigma_1 + \frac{\|\hat{R}_1\|_2}{1+\alpha}\right) \cdot \|\hat{R}_1\|_2, \quad j = 1, \ldots, l, \tag{3.21}$$

where

$$\sigma_1 = \|\sin\Theta(X_1, \tilde{X}_1)\|_2, \quad \alpha = \sigma_{\min}(\tilde{A}_1). \tag{3.22}$$

Proof. We first use the technique described in [18] to make some simplification on the forms of the matrices $A, X_1, \tilde{X}_1$ and $\tilde{X}_2$.

By the CS decomposition [16, Theorem A.1], there are unitary matrices $Q \in \mathcal{C}^{n\times n}$, $U_1, V_1 \in \mathcal{C}^{l\times l}$, and $V_2 \in \mathcal{C}^{(n-l)\times(n-l)}$ such that

$$X_1 = QX_{10}U_1, \quad \tilde{X}_1 = Q\tilde{X}_{10}V_1, \quad \tilde{X}_2 = Q\tilde{X}_{20}V_2,$$

where

$$X_{10} = \begin{pmatrix} I^{(l)} \\ 0 \end{pmatrix}, \quad \tilde{X}_{10} = \begin{pmatrix} \Gamma \\ \Sigma \end{pmatrix}, \quad \tilde{X}_{20} = \begin{pmatrix} -\Sigma^T \\ \hat{\Gamma} \end{pmatrix}, \tag{3.23}$$

$$\Gamma = \begin{cases} \Gamma_1 & \text{if } 2l \leq n, \\ \operatorname{diag}(\Gamma_1,\ I^{(2l-n)}) & \text{if } 2l \geq n, \end{cases} \qquad \Gamma_1 = \operatorname{diag}(\gamma_j), \tag{3.24}$$

$$\Sigma = \begin{cases} \begin{pmatrix} \Sigma_1 \\ 0 \end{pmatrix} & \text{if } 2l \leq n, \\ (\Sigma_1,\ 0) & \text{if } 2l \geq n, \end{cases} \qquad \Sigma_1 = \operatorname{diag}(\sigma_j), \tag{3.25}$$

$$\hat{\Gamma} = \begin{cases} \operatorname{diag}(\Gamma_1,\ I^{(n-2l)}) & \text{if } 2l \leq n, \\ \Gamma_1 & \text{if } 2l \geq n, \end{cases} \tag{3.26}$$

and

$$0 \leq \gamma_1 \leq \gamma_2 \leq \cdots \leq 1, \quad 1 \geq \sigma_1 \geq \sigma_2 \geq \cdots \geq 0, \quad \gamma_j^2 + \sigma_j^2 = 1 \ \ \forall j. \tag{3.27}$$

Let

$$A_0 = Q^H A Q, \quad A_{10} = U_1 A_1 U_1^H, \quad \hat{R}_{10} = Q^H \hat{R}_1 V_1^H,$$

$$\tilde{A}_{k0} = V_k \tilde{A}_k V_k^H, \quad P_{k0} = V_k P_k V_k^H, \quad H_{k0} = V_k H_k V_k^H, \quad k = 1, 2.$$

Then

$$A_{10} = X_{10}{}^H A_0 X_{10}, \quad \hat{R}_{10} = \tilde{X}_{10}\tilde{A}_{10} - A_0\tilde{X}_{10},$$

and

$$\tilde{A}_{k0} = \tilde{X}_{k0}^H A_0 \tilde{X}_{k0} = P_{k0}H_{k0}, \quad k = 1, 2.$$

Furthermore, from $AX_1 = X_1A_1$ we get $A_0X_{10} = X_{10}A_{10}$, and thus

$$A_0 = \begin{pmatrix} A_{10} & 0 \\ 0 & A_{20} \end{pmatrix},$$

where A_{20} is an $(n-l) \times (n-l)$ unitary matrix. Observe that

$$\lambda(A_0) = \lambda(A), \quad \lambda(A_{10}) = \lambda(A_1), \quad \lambda(\tilde{A}_{10}) = \lambda(\tilde{A}_1),$$

$$\mathrm{sep}_p(P_{10}, P_{20}) = \mathrm{sep}_p(P_1, P_2), \quad \|\tilde{A}_{k0} - P_{k0}\|_p = \|\tilde{A}_k - P_k\|_p, \quad k = 1, 2,$$

and

$$\|\sin\Theta(X_{10}, \tilde{X}_{10})\|_p = \|\sin\Theta(X_1, \tilde{X}_1)\|_p, \quad \|\hat{R}_{10}\|_p = \|\hat{R}_1\|_p, \quad p = 2, F.$$

Hence, without loss of generality we may assume that the matrices $A, X_1, \tilde{X}_1$, and $\tilde{X}_2$ have the following reduced forms:

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}, \quad X_1 = \begin{pmatrix} I^{(l)} \\ 0 \end{pmatrix}, \quad \tilde{X}_1 = \begin{pmatrix} \Gamma \\ \Sigma \end{pmatrix}, \quad \tilde{X}_2 = \begin{pmatrix} -\Sigma^T \\ \hat{\Gamma} \end{pmatrix},$$

where A_1, A_2 are unitary, Γ, Σ, and $\hat{\Gamma}$ are expressed by (3.23)–(3.27).

By (3.12), we have

$$\hat{R}_1 = \begin{pmatrix} \Gamma\tilde{A}_1 - A_1\Gamma \\ \Sigma\tilde{A}_1 - A_2\Sigma \end{pmatrix}, \tag{3.28}$$

and

$$\tilde{X}^H\hat{R}_1 = \begin{pmatrix} 0 \\ K \end{pmatrix}, \quad K = -\tilde{X}_2^H A\tilde{X}_1. \tag{3.29}$$

Furthermore, from (3.28) and (3.29),

$$\begin{aligned} \Gamma\tilde{A}_1 - A_1\Gamma &= (I, 0)\hat{R}_1 = (I, 0)\tilde{X}\begin{pmatrix} 0 \\ K \end{pmatrix} \\ &= (I, 0)\tilde{X}_2 K = -\Sigma^T K. \end{aligned} \tag{3.30}$$

Thus, the relations (3.29) and (3.30) give

$$\|\Gamma\tilde{A}_1 - A_1\Gamma\|_p \le \|\Sigma\|_2\|K\|_p = \|\Sigma\|_2\|\hat{R}_1\|_p, \quad p = 2, F. \tag{3.31}$$

On the other hand, for $p = 2, F$

$$\begin{aligned}\|\Gamma\tilde{A}_1 - A_1\Gamma\|_p &\geq \|\Gamma P_1 - A_1\Gamma\|_p - \|\Gamma(\tilde{A}_1 - P_1)\|_p \\ &\geq \gamma_1\|P_1 - A_1\|_p - \|\tilde{A}_1 - P_1\|_p \quad \text{(by Lemma 2.3 and } \|\Gamma\|_2 \leq 1) \\ &\geq \gamma_1\|P_1 - A_1\|_p - \|\tilde{A}_1^H\tilde{A}_1 - I\|_p/[1 + \sigma_{\min}(\tilde{A}_1)] \quad \text{(by Lemma 2.4).}\end{aligned}$$

Combining it with (3.31), (3.32), and

$$\|\Sigma\|_2 = \sigma_1 = \|\sin\Theta(X_1, \tilde{X}_1)\|_2, \quad \gamma_1 = \sqrt{1 - \sigma_1^2} > 0,$$

we get

$$\|P_1 - A_1\|_p \leq \frac{1}{\sqrt{1 - \sigma_1^2}} \cdot \left(\sigma_1\|\hat{R}_1\|_p + \frac{\|\tilde{A}_1^H\tilde{A}_1 - I\|_p}{1 + \alpha}\right). \tag{3.32}$$

Moreover, by the definition (3.2) of $\tilde{A}_1$, we have

$$\begin{aligned}\|\tilde{A}_1^H\tilde{A}_1 - I\|_p &= \|\tilde{X}_1^H A^H \tilde{X}_1\tilde{X}_1^H A\tilde{X}_1 - I\|_p \\ &= \|\tilde{X}_1^H A^H(I - \tilde{X}_2\tilde{X}_2^H)A\tilde{X}_1 - I\|_p \\ &= \|\tilde{X}_1^H A^H\tilde{X}_2\tilde{X}_2^H A\tilde{X}_1\|_p \leq \|\tilde{X}_2^H A\tilde{X}_1\|_2\|\tilde{X}_2^H A\tilde{X}_1\|_p \\ &= \|\hat{R}_1\|_2\|\hat{R}_1\|_p \quad \text{(by (3.18)).}\end{aligned}$$

Substituting it into (3.32) gives

$$\|P_1 - A_1\|_p \leq \frac{1}{\sqrt{1 - \sigma_1^2}} \cdot \left(\sigma_1 + \frac{\|\hat{R}_1\|_2}{1 + \alpha}\right) \cdot \|\hat{R}_1\|_p. \tag{3.33}$$

Observe that by the Hoffman–Wielandt Theorem [11] and the Bhatia–Davis Theorem [1], there are permutations π_f and π_2 of $\{1, \ldots, l\}$ such that

$$\sqrt{\sum_{j=1}^{l} |\lambda_{\pi(j)} - \mu_j|^2} \leq \|P_1 - A_1\|_F, \tag{3.34}$$

and

$$|\lambda_{\pi_2(j)} - \mu_j| \leq \|P_1 - A_1\|_2, \quad j = 1, \ldots, l. \tag{3.35}$$

Hence, from (3.33)–(3.35) we derive (3.20)–(3.21). □

REMARK 3.4. Let $A, X_1, \tilde{X}_1, \mathcal{X}_1, \tilde{\mathcal{X}}_1$, and $\hat{R}_1$ be as in Theorem 3.3. Then by (3.28)

$$\|\hat{R}_1\|_p^2 \leq \|\Gamma\tilde{A}_1 - A_1\Gamma\|_p^2 + \|\Sigma\tilde{A}_1 - A_2\Sigma\|_p^2, \tag{3.36}$$

where

$$\begin{aligned}\|\Gamma\tilde{A}_1 - A_1\Gamma\|_p &\leq \|\Sigma\|_2\|\hat{R}_1\|_p \quad \text{(by (3.31))} \\ &= \|\sin\Theta(X_1, \tilde{X}_1)\|_2\|\hat{R}_1\|_p,\end{aligned} \tag{3.37}$$

and

$$\begin{aligned} \|\Sigma\tilde{A}_1 - A_2\Sigma\|_p &\le (\|\tilde{A}_1\|_2 + \|A_2\|_2)\|\Sigma\|_p \\ &\le 2\|\sin\Theta(X_1, \tilde{X}_1)\|_p. \end{aligned} \tag{3.38}$$

Substituting (3.37) and (3.38) into (3.36) shows that if $\|\sin\Theta(X_1, \tilde{X}_1)\|_2 < 1$, then

$$\|\hat{R}\|_p \le \frac{2}{\sqrt{1 - \|\sin\Theta(X_1, \tilde{X}_1)\|_2^2}} \cdot \|\sin\Theta(X_1, \tilde{X}_1)\|_p, \quad p = 2, F. \tag{3.39}$$

The inequality (3.39) presents a quantitative description of such a fact: The closer to $\mathcal{X}_1 = \mathcal{R}(X_1)$ the subspace $\tilde{\mathcal{X}}_1 = \mathcal{R}(\tilde{X}_1)$ is, the smaller the norms $\|\hat{R}_1\|_p$ for $p = 2, F$ are. Moreover, combining the inequality (3.39) with the estimates (3.20)–(3.21) gives a theoretical result: Under the condition (3.13) the precision of the eigenvalues $\mu_1, \ldots, \mu_l$ of the unitary polar factor P_1 of $\tilde{X}_1^H A\tilde{X}_1$ as approximations of the eigenvalues $\lambda_1, \ldots, \lambda_l$ of A is higher than that of $\mathcal{R}(\tilde{X}_1)$ as an approximation of the eigenspace $\mathcal{R}(X_1)$ of A.

From (3.20)–(3.21) and (3.39) we can prove that if $\|\sin\Theta(X_1, \tilde{X}_1)\|_2 < 0.2952$, that is, if $\mathcal{R}(\tilde{X}_1)$ is pretty close to $\mathcal{R}(X_1)$, then

$$\sqrt{\sum_{j=1}^{l} |\lambda_{\pi_f(j)} - \mu_j|^2} < \|\hat{R}_1\|_F, \quad |\lambda_{\pi_2(j)} - \mu_j| < \|\hat{R}_1\|_2, \quad j = 1, \ldots, l.$$

Furthermore, from Theorems 3.2–3.3 (see the estimates (3.14), (3.20), and (3.21)) and the relation (1.2), we get the following corollary.

COROLLARY 3.5. *Let* $A, \tilde{X} = (\tilde{X}_1, \tilde{X}_2), X_1, A_1, \tilde{A}_1, \hat{R}_1, P_k, R_k$ $(k = 1, 2)$ *and* λ_j, μ_j $(j = 1, \ldots, l)$ *be as in Theorem* 3.3. *Then under the condition* (3.13) *there are permutations* π_f *and* π_2 *of* $\{1, \ldots, l\}$ *such that*

$$\sqrt{\sum_{j=1}^{l} |\lambda_{\pi_f(j)} - \mu_j|^2} \le \frac{\delta_F}{2\sqrt{1 - \rho_F^2}} \cdot \left(1 + \frac{\delta_F}{2(1+\alpha)}\right) \cdot \left(\frac{2\|\hat{R}_1\|_F}{\delta_F}\right)^2, \tag{3.40}$$

and

$$|\lambda_{\pi_2(j)} - \mu_j| \le \frac{\delta_2}{2\sqrt{1 - \rho_2^2}} \cdot \left(1 + \frac{\delta_2}{2(1+\alpha)}\right) \cdot \left(\frac{2\|\hat{R}_1\|_2}{\delta_2}\right)^2, \quad j = 1, \ldots, l, \tag{3.41}$$

where δ_p *are defined by* (3.13), α *by* (3.22), *and* ρ_p *by*

$$\rho_p = \frac{2\|\hat{R}_1\|_p}{\delta_p}, \quad p = 2, F.$$

Comparing the a posteriori error estimates (3.40)–(3.41) with (3.14) we see again that under the condition (3.13) the precision of the eigenvalues $\mu_1, \ldots, \mu_l$ of the unitary polar factor P_1 of $\tilde{X}_1^H A\tilde{X}_1$ as approximations of the eigenvalues $\lambda_1, \ldots, \lambda_l$ of A is higher than that of $\mathcal{R}(\tilde{X}_1)$ as an approximation of the eigenspace $\mathcal{R}(X_1)$ of A.

4. Residual bounds (II). Assume that $A \in \mathcal{C}^{n\times n}$ is unitary and the columns of $\tilde{X}_1 \in \mathcal{C}^{n\times l}$ $(l < n)$ form an orthonormal basis for an approximate eigenspace of A. Let $\tilde{A}_1 = \tilde{X}_1^H A \tilde{X}_1$. We have proved in §3 that the eigenvalues of the unitary polar factor P_1 of $\tilde{A}_1$ are l approximate eigenvalues of A. However, it may well be asked: How does one relate the eigenvalues of $\tilde{A}_1$ to those of A? In this section we study this problem.

We first cite a perturbation theorem of the eigenvalues of a normal matrix [20, Theorem 1.1].

THEOREM 4.1 [20]. *Let $A \in \mathcal{C}^{n\times n}$ be normal with $\lambda(A) = \{\lambda_j\}$, and let $\tilde{A} \in \mathcal{C}^{n\times n}$ be nonnormal with $\lambda(\tilde{A}) = \{\tilde{\lambda}_j\}$. Then there is a permutation π of $\{1,\dots,n\}$ such that*

$$\sqrt{\sum_{j=1}^{n} |\tilde{\lambda}_{\pi(j)} - \lambda_j|^2} \le \sqrt{n}\|\tilde{A} - A\|_F. \tag{4.1}$$

By Theorem 4.1 we have proved the following corollary.

COROLLARY 4.2 [20]. *Let $A, \tilde{X} = (\tilde{X}_1, \tilde{X}_2), \tilde{A}_1$ and $\hat{R}_1$ be as in Theorems 3.1–3.2. If $\lambda(A) = \{\lambda_j\}_{j=1}^n$, and $\lambda(\tilde{A}_1) = \{\tilde{\lambda}_j\}_{j=1}^l$, then there are $\lambda_{1'},\dots,\lambda_{l'} \in \lambda(A)$ such that*

$$\sqrt{\sum_{j=1}^{l} |\lambda_{j'} - \tilde{\lambda}_j|^2} \le \sqrt{n}\|\hat{R}_1\|_F. \tag{4.2}$$

Combining the estimate (4.2) with the inequality (3.39) shows that if $\mathcal{R}(\tilde{X}_1)$ is an approximate eigenspace of A, then the eigenvalues $\tilde{\lambda}_1,\dots,\tilde{\lambda}_l$ of $\tilde{X}_1^H A \tilde{X}_1$ are l approximate eigenvalues of A. The following result gives a new estimate.

THEOREM 4.3. *Let $A, \tilde{X} = (\tilde{X}_1, \tilde{X}_2), X_1, \Theta(X_1, \tilde{X}_1), A_1, \tilde{A}_1, \hat{R}_1, \lambda(A)$, and $\lambda(A_1)$ be as in Theorem 3.3, and let $\lambda(\tilde{A}_1) = \{\tilde{\lambda}_j\}_{j=1}^l$. Then if $\|\sin\Theta(X_1, \tilde{X}_1)\|_2 < 1$, there is a permutation π of $\{1,\dots,l\}$ such that*

$$\sqrt{\sum_{j=1}^{l} |\lambda_{\pi(j)} - \tilde{\lambda}_j|^2} \le \frac{1}{\sqrt{1-\sigma_1^2}} \cdot \left(\sigma_1\|\hat{R}_1\|_F + \frac{\sqrt{l}\sigma_1^2}{1+\sqrt{1-\sigma_1^2}}\right), \tag{4.3}$$

where $\sigma_1 = \|\sin\Theta(X_1, \tilde{X}_1)\|_2$.

Proof. By the proof of Theorem 3.3 (see (3.31)), we have

$$\|\Gamma\tilde{A}_1 - A_1\Gamma\|_F \le \|\Sigma\|_2\|\hat{R}_1\|_F = \sigma_1\|\hat{R}_1\|_F. \tag{4.4}$$

On the other hand, (3.24) and (3.27) give

$$\begin{aligned} \|\Gamma\tilde{A}_1 - A_1\Gamma\|_F &\ge \|\Gamma(\tilde{A}_1 - A_1)\|_F - \|\Gamma A_1 - A_1\Gamma\|_F \\ &\ge \gamma_1\|\tilde{A}_1 - A_1\|_F - \sqrt{l}\max_{i,j}|\gamma_i - \gamma_j| \\ &\ge \gamma_1\|\tilde{A}_1 - A_1\|_F - \sqrt{l}(1-\gamma_1). \end{aligned} \tag{4.5}$$

By the hypothesis, $\gamma_1 > 0$. Consequently, from the inequalities (4.4) and (4.5),

$$\|\tilde{A}_1 - A_1\|_F \le \frac{1}{\gamma_1}[\sigma_1\|\hat{R}_1\|_F + \sqrt{l}(1-\gamma_1)]. \tag{4.6}$$

Observe that by Theorem 4.1, there is a permutation π of $\{1,\dots,l\}$ such that

$$\sqrt{\sum_{j=1}^{l}|\lambda_{\pi(j)}-\tilde{\lambda}_j|^2}\le\sqrt{l}\|\tilde{A}_1-A_1\|_F.$$

This together with (4.6) and

$$\gamma_1=\sqrt{1-\sigma_1^2},\quad 1-\gamma_1=1-\sqrt{1-\sigma_1^2}=\sigma_1^2/(1+\sqrt{1-\sigma_1^2}),$$

shows the estimate (4.3). □

From Theorem 3.2, Theorem 4.3, and the relation (1.2), we get the following corollary.

COROLLARY 4.4. *Let* $A,\tilde{X}=(\tilde{X}_1,\tilde{X}_2),X_1,A_1,\tilde{A}_1,\hat{R}_1$ *and* $\lambda_j,\tilde{\lambda}_j\,(j=1,\dots,l)$ *be as in Theorem* 4.3. *Then under the condition* (3.13) *there is a permutation* π *of* $\{1,\dots,l\}$ *such that*

$$\sqrt{\sum_{j=1}^{l}|\lambda_{\pi(j)}-\tilde{\lambda}_j|^2}\le\frac{1}{\sqrt{1-\rho_F^2}}\cdot\left(\frac{\delta_F}{2}+\frac{\sqrt{l}}{1+\sqrt{1-\rho_F^2}}\right)\cdot\left(\frac{2\|\hat{R}_1\|_F}{\delta_F}\right)^2,\tag{4.7}$$

where δ_F *is defined by* (3.13), *and* $\rho_F=2\|\hat{R}_1\|_F/\delta_F$.

Comparing (4.5) with (3.14) we see that under the condition (3.13) the precision of the eigenvalues $\tilde{\lambda}_1,\dots,\tilde{\lambda}_l$ of the Rayleigh quotient matrix $\tilde{X}_1^HA\tilde{X}_1$ as approximations of the eigenvalues $\lambda_1,\dots,\lambda_l$ of A is higher than that of $\mathcal{R}(\tilde{X}_1)$ as an approximation of the eigenspace $\mathcal{R}(X_1)$ of A.

5. A numerical example. We first give a brief summary of the main results of this paper. Here we only use the Frobenius norm.

Let two unitary matrices $A,\tilde{X}=(\tilde{X}_1,\tilde{X}_2)\in\mathcal{C}^{n\times n}$ be given, where $\tilde{X}_1\in\mathcal{C}^{n\times l}$, and $\tilde{\mathcal{X}}_1=\mathcal{R}(\tilde{X}_1)$ approximates an eigenspace of A. Define $\tilde{A}_k=\tilde{X}_k^HA\tilde{X}_k$, and let P_k be the unitary polar factors of $\tilde{A}_k$, $k=1,2$. Moreover, define

$$\hat{R}_1=\tilde{X}_1\tilde{A}_1-A\tilde{X}_1,\quad k=1,2.$$

Then if

$$\delta_F\equiv\operatorname{sep}_F(P_1,P_2)-(\|\tilde{A}_1-P_1\|_F+\|\tilde{A}_2-P_2\|_F)>0,\quad \rho_F\equiv\frac{2\|\hat{R}_1\|_F}{\delta_F}<1,\tag{5.1}$$

we have the following conclusions.

(1) There is an eigenspace $\mathcal{X}_1=\mathcal{R}(X_1)$ of A, where $X_1\in\mathcal{C}^{n\times l}$ with $X_1^HX_1=I$, such that

$$\|\tan\Theta(X_1,\tilde{X}_1)\|_F<\rho_F.\tag{5.2}$$

(2) Let $A_1=X_1^HAX_1$, $\lambda(A_1)=\{\lambda_j\}_{j=1}^l$, $\lambda(P_1)=\{\mu_j\}_{j=1}^l$, $\lambda(\tilde{A}_1)=\{\tilde{\lambda}_j\}_{j=1}^l$, and define

$$d(\lambda(A_1),\lambda(P_1))=\min_\pi\sqrt{\textstyle\sum_{j=1}^l|\lambda_{\pi(j)}-\mu_j|^2},$$

$$d(\lambda(A_1),\lambda(\tilde{A}_1))=\min_\pi\sqrt{\textstyle\sum_{j=1}^l|\lambda_{\pi(j)}-\tilde{\lambda}_j|^2},$$

where π ranges over all permutations of $\{1, 2, \ldots, l\}$. Then

$$d(\lambda(A_1), \lambda(P_1)) \leq \frac{\delta_F}{2\sqrt{1-\rho_F^2}} \cdot \left(1 + \frac{\delta_F}{2(1+\alpha)}\right) \cdot \rho_F^2 \equiv \mu_F, \tag{5.3}$$

and

$$d(\lambda(A_1), \lambda(\tilde{A}_1)) \leq \frac{1}{\sqrt{1-\rho_F^2}} \cdot \left(\frac{\delta_F}{2} + \frac{\sqrt{l}}{1+\sqrt{1-\rho_F^2}}\right) \cdot \rho_F^2 \equiv \lambda_F, \tag{5.4}$$

where δ_F, ρ_F are defined by (5.1), and $\alpha = \sigma_{\min}(\tilde{A}_1)$.

Observe that $\delta_F \leq 2$, and $\rho_F \ll 1$ implies $\alpha \approx 1$. Hence, if $\rho_F \ll 1$, then by (5.3)–(5.4) we have

$$\mu_F \lesssim \frac{3}{2}\rho_F^2, \quad \lambda_F \lesssim \left(1 + \frac{\sqrt{l}}{2}\right)\rho_F^2.$$

We now illustrate the main results with a simple example.

Example 5.1. Let

$$A = \mathrm{diag}(1,\ i,\ -1,\ -(1+i)/\sqrt{2},\ -i),$$

$$F_0 = \begin{pmatrix} 3+i & -2 & 5-2i & -1 & 6 \\ -5 & 6+i & -3 & 7 & i \\ 1 & 3-3i & -6 & 2+3i & 8 \\ 7 & 4i & -i & 5 & 9+i \\ -i & 2 & -5 & 7+2i & 3 \end{pmatrix},$$

$$G = I^{(5)} + \epsilon F_0, \quad \epsilon > 0,$$

and let $\tilde{X} = (\tilde{X}_1, \tilde{X}_2)$ be the unitary QR factor of G, where $\tilde{X}_1 \in \mathcal{C}^{5\times 2}$. Define $\tilde{A}_k, P_k, \hat{R}_1, \delta_F, \rho_F$ as above. Note that $X = I^{(5)} = (X_1, X_2)$, where $X_1 \in \mathcal{C}^{5\times 2}$. Some numerical results obtained by (5.1)–(5.4) and by using MATLAB are listed in Table 1, where μ_F and λ_F are defined by (5.3) and (5.4).

TABLE 1

ϵ	3.40e-02	1.00e-02	1.00e-04	1.00e-06
δ_F	9.8120e-01	1.3330e+00	1.4136e+00	1.4142e+00
ρ_F	9.9894e-01	2.4113e-01	2.3663e-03	2.3661e-05
$\|\tan\Theta(X_1, \tilde{X}_1)\|_F$	2.8688e-01	9.0839e-02	9.4302e-04	9.4340e-06
μ_F	1.3351e+01	5.3298e-02	5.3560e-06	5.3583e-10
$d(\lambda(A_1), \lambda(P_1))$	4.2986e-02	3.6174e-03	3.5507e-07	3.5499e-11
λ_F	3.9915e+01	8.2930e-02	7.9166e-06	7.9174e-10
$d(\lambda(A_1), \lambda(\tilde{A}_1))$	9.0007e-02	8.9523e-03	9.4443e-07	9.4494e-11

The numerical results listed in Table 1 show the strength of the upper bounds ρ_F, μ_F and λ_F.

Acknowledgment. I would like to thank the referees for their helpful comments and valuable suggestions.

REFERENCES

[1] R. BHATIA AND C. DAVIS, *A bound for the spectral variation of a unitary operator*, Linear Multilinear Algebra, 15 (1984), pp. 71–76.

[2] R. BHATIA, C. DAVIS, AND F. KITTANEH, *Some inequalities for commutators and an application to spectral variation*, Aequationes Math., 41 (1991), pp. 70–78.

[3] A. BUNSE-GERSTNER AND L. ELSNER, *The Schur parameter form for the solution of the unitary eigenproblem*, Linear Algebra Appl., 154-156 (1991), pp. 741–778.

[4] L. ELSNER AND C. HE, *Perturbation and interlace theorems for the unitary eigenvalue problem*, Linear Algebra Appl., 188/189 (1993), pp. 207–229.

[5] K. FAN, *Maximum properties and inequalities for the eigenvalues of completely continuous operators*, Proc. Nat. Acad. Sci. U.S.A., 37 (1951), pp. 760–766.

[6] K. FAN AND J. HOFFMAN, *Some metric inequalities in the space of matrices*, Proc. Amer. Math. Soc., 6 (1955), pp. 111–116.

[7] G. GOLUB AND C. VAN LOAN, *Matrix Computations*, 2nd edition, The Johns Hopkins University Press, Baltimore, 1989.

[8] W. B. GRAGG, *The QR algorithm for unitary Hessenberg matrices*, J. Comput. Appl. Math., 16 (1986), pp. 1–8.

[9] W. B. GRAGG AND L. REICHEL, *A divide and conquer method for unitary and orthogonal eigenproblems*, Numer. Math., 57 (1990), pp. 695–718.

[10] N. J. HIGHAM, *Computing the polar decomposition—with applications*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1160–1174.

[11] A. J. HOFFMAN AND H. W. WIELANDT, *The variation of the spectrum of a normal matrix*, Duke Math. J., 20 (1953), pp. 37–39.

[12] R. HORN AND C. JOHNSON, *Matrix Analysis*, Cambridge University Press, New York, 1985.

[13] W. M. KAHAN, B. N. PARLETT, AND E. JIANG, *Residual bounds on approximate eigensystems of nonnormal matrices*, SIAM J. Numer. Anal., 19 (1982), pp. 470–484.

[14] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice Hall, Englewood Cliffs, NJ, 1980.

[15] G. W. STEWART, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–767.

[16] ———, *On the perturbation of pseudo-inverses, projections, and linear least squares problems*, SIAM Rev., 19 (1977), pp. 634–662.

[17] G. W. STEWART AND J. -G. SUN, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[18] J. -G. SUN, *Eigenvalues of Rayleigh quotient matrices*, Numer. Math., 59 (1991), pp. 603–614.

[19] ———, *Backward perturbation analysis of certain characteristic subspaces*, Numer. Math., 65 (1993), pp. 357–382.

[20] ———, *On the variation of the spectrum of a normal matrix*, Report UMINF-94.06, ISSN-0348-0542, Department of Computing Science, Umeå University, 1994.

[21] J. VON NEUMANN, *Some matrix-inequalities and metrization of matric-space*, Tomsk Univ. Rev., 1 (1937), pp. 286–300. *Collected Works*, A. H. Taub, ed., Pergamon Press, New York, 1962, pp. 205–219.

[22] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, England, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 1, pp. 83–109, January 1996

A QL PROCEDURE FOR COMPUTING THE EIGENVALUES OF COMPLEX SYMMETRIC TRIDIAGONAL MATRICES*

JANE K. CULLUM† AND RALPH A. WILLOUGHBY‡

Abstract. We present a storage efficient procedure for computing all of the eigenvalues of a complex symmetric tridiagonal matrix. This procedure mimics the implicit QL procedure for computing all of the eigenvalues of a real symmetric tridiagonal matrix, modified by heuristics for monitoring and maintaining numerical stability. Numerical experiments demonstrate the capabilities c^c this procedure.

Key words. eigenvalues, complex symmetric, tridiagonal matrices, QL procedure

AMS subject classifications. 15A18, 15-04

1. Introduction.

A matrix is symmetric if and only if it is equal to its transpose. In this paper we are interested in matrices that are both complex and symmetric. Complex symmetric matrices arise in many different applications. See, for example, [1], [2], [17]–[19], [21]. Variants of nonsymmetric Lanczos recursions are being used to solve complex symmetric systems of equations and to compute eigenvalues of complex symmetric matrix eigenvalue problems [9], [13].

The current work was motivated by the authors' development of Lanczos procedures for computing eigenvalues of both complex symmetric and general nonsymmetric matrices [5], [6], [8], [9]. Those procedures transform the original matrix eigenvalue problem into a family of complex symmetric tridiagonal matrix eigenvalue problems. Approximations to eigenvalues of the original problem are obtained by computing eigenvalues of one or more of the complex symmetric tridiagonal problems.

The tridiagonal procedure that we present is also applicable to real nonsymmetric tridiagonal matrices such as those generated by various real versions of the nonsymmetric Lanczos eigenvalue procedures. See, for example, [26], [27]. The amount of spectral information that can be obtained by using a nonsymmetric Lanczos recursion depends both upon the spectral properties of the given problem and upon the size of the complex symmetric or real nonsymmetric Lanczos matrices that can be resolved.

Other procedures for computing eigenvalues of complex symmetric tridiagonal matrices require $O(n^2)$ storage and $O(n^3)$ arithmetic operations where n is the size of the tridiagonal matrix. Eberlein [12] proposes a norm reducing Jacobi procedure that is applicable to a general complex symmetric matrix, but the procedure does not preserve the nonzero structure of the matrix.

The procedure we present is a complex symmetric analog of the implicit, orthogonal QL procedure for real symmetric tridiagonal matrices. The complex symmetric tridiagonal structure is preserved as the computations proceed. It requires only $O(n)$ storage and $O(n^2)$ arithmetic operations. It is not, however, a straightforward unitary implementation of the real symmetric tridiagonal algorithm. Unitary QL factorizations do not preserve the complex symmetric tridiagonal structure. This structure

* Received by the editors June 22, 1987; accepted for publication (in revised form) by L. Kaufman January 30, 1995.

† Mathematical Sciences Department, IBM Research Division, T.J. Watson Research Center, Yorktown Heights, NY 10598, and Department of Computer Science, Institute for Advanced Computer Studies, and Institute for Systems Research, University of Maryland, College Park, MD 20742 (cullumj@watson.ibm.com). This author was supported by National Science Foundation grant GER-9450081.

‡ IBM Research Division, retired.

is preserved by introducing complex orthogonal QL factorizations where Q is complex, $Q^TQ = I$ but $Q^HQ \neq I$. See Definitions 2.3 and 2.4. The resulting complex symmetric tridiagonal algorithm consists of a sequence of complex orthogonal QL factorizations.

It is not difficult however to construct matrices that do not have complex orthogonal QL factorizations. See Example 3.1 in §3. Therefore, it is necessary to introduce heuristics to monitor the numerical stability of the complex orthogonal transformations generated at intermediate steps in the procedure and additional heuristics to modify the procedure if potential problems are indicated. As we will see in §5, these problems occur because there are nonzero complex numbers, for example, $a = 1$ and $b = \sqrt{-1}$, such that $a^2 + b^2 = 0$. For a history of the development and the use of unitary QR and QL procedures see [22], [29].

In §2 we summarize the notation and definitions required. In §3 we list relevant properties of complex symmetric tridiagonal matrices and prove several lemmas that are needed in other sections of this paper. In §4 we outline a basic QL procedure for complex symmetric tridiagonal matrices, summarize relevant results from the literature, and obtain a convergence theorem under the assumption that the eigenvalues have distinct magnitudes. In §5 we consider an implicit version of the procedure outlined in §4. We prove that any iteration of the implicit procedure can be completed if and only if the current matrix iterate has the required complex orthogonal decomposition. We also introduce heuristics to stabilize these computations. In §6 we summarize results of numerical experiments on several families of test matrices which indicate that this procedure performs well in practice. Tridiagonal Lanczos matrices generated by both a complex and a real version of the nonsymmetric Lanczos recursions are considered, as well as matrices generated randomly, using several different probability distributions.

2. Notation and definitions. The following notation and definitions are used throughout the paper.

2.1. Notation.

$A = (a_{ij})$, $1 \leq i, j \leq n$, denotes a $n \times n$ matrix

$\bar{A} = (\bar{a}_{ij})$, $A^T = (a_{ji})$, $A^H = (\bar{a}_{ji})$, denote respectively the complex conjugate, the transpose, and the complex conjugate transpose of A

$H = (h_{ij})$, $h_{ij} = 0$ for $i > j+1$ denotes an upper Hessenberg matrix

$T = (t_{ij})$, $t_{ij} = 0$ for $i \neq j-1, j, j+1$ denotes a tridiagonal matrix

$L = (l_{ij})$, $l_{ij} = 0$ for $i < j$ and $R = (r_{ij})$, $r_{ij} = 0$ for $i > j$ denote respectively, lower and upper triangular matrices

$D = \text{diag } \{d_1, \ldots, d_n\}$, denotes a $n \times n$ diagonal matrix

J denotes the $n \times n$ matrix with $J(i,j) = 1$ for $i + j = n + 1$ and $J(i,j) = 0$ otherwise

$T_{1,j}$ and $T_{n,n-j+1}$ denote respectively the $j \times j$ leading principal minors of T and of JTJ

L_j, R_j, A_j denote respectively the leading principal minors of L, R and A

$A_k(i,j)$ denotes (i,j) entry of the subscripted matrix A_k

$\lambda_j(A)$, $1 \leq j \leq n$, denote the eigenvalues of A where $|\lambda_1| \leq \cdots \leq |\lambda_n|$

$\omega(A)$ denotes the set of all eigenvalues of A

$\sigma_j(A)$, $1 \leq j \leq n$, denote the singular values of A where $\sigma_1 \geq \cdots \geq \sigma_n$
$\Sigma = \text{diag } \{\sigma_1, \ldots, \sigma_n\}$
$\|A\| = \|A\|_2 = \sigma_1(A)$, $\|x\|_2 = \sqrt{\Sigma x_j^2}$
$\kappa_2(A) = \sigma_1(A)/\sigma_n(A)$ denotes the condition number of A
$W_k = \text{sp}\{w_1, \ldots, w_k\}$ denotes the space spanned by the vectors w_j
$\mathcal{C}^n$ denotes $n-$dimensional complex space
$d(\mathcal{S},\mathcal{U})$ denotes the distance between two $k-$dimensional subspaces $\mathcal{S}$ and $\mathcal{U}$
$\mathcal{S}^\perp$ denotes the orthogonal complement in $\mathcal{C}^n$ of the subspace $\mathcal{S}$
$A \in \mathcal{C}^{n\times n}$ denotes a $n \times n$ matrix with complex entries
e_j denotes the jth coordinate vector
I_j denotes the $j \times j$ identity matrix

2.2. Definitions.

DEFINITION 2.1. *A symmetric tridiagonal T is* irreducible *if and only if each* $t_{j+1,j} \neq 0$.

DEFINITION 2.2. *A matrix $P_k \in \mathcal{C}^{n\times n}$ is a* complex symmetric rotation *if and only if*

$$P_k = \begin{pmatrix} I_{k-1} & & & \\ & -c_k & s_k & \\ & s_k & c_k & \\ & & & I_{n-k-1} \end{pmatrix}, \tag{1}$$

where c_k and s_k are complex scalars such that ${c_k}^2 + {s_k}^2 = 1$.

DEFINITION 2.3. *A matrix $Q \in \mathcal{C}^{n\times n}$ is* complex orthogonal *if and only if $Q^TQ = I$.*

DEFINITION 2.4. *A matrix A has a* complex orthogonal QL decomposition *if and only if there exists a complex orthogonal matrix Q and a lower triangular matrix L such that $A = QL$.*

DEFINITION 2.5. *A_1 is* essentially equal *to A_2 if and only if there exists a diagonal matrix D with $D^2 = I$ such that $A_1 = DA_2$ or $A_1 = A_2D$.*

DEFINITION 2.6. *A factorization $A = QL$ is* essentially unique *if and only if $A = Q_1L_1 = Q_2L_2$ implies that there exists a diagonal matrix D with $D^2 = I$ such that*

$$Q_2 = Q_1D \quad \text{and} \quad L_2 = DL_1. \tag{2}$$

DEFINITION 2.7. *A sequence of matrices, B_k, $k = 1, 2, \ldots$* converges essentially *to a matrix B if and only if for some M and all k, $\|B_k\| \leq M$, and all limit points of this sequence are of the form BD or DB where D is a diagonal matrix with $D^2 = I$.*

DEFINITION 2.8 (see [28]). *An eigenvalue algorithm is* numerically stable *if and only if for each computed eigenpair (μ, z) associated with a matrix A there exists a matrix E with small norm compared to that of A for which μ and z are an exact eigenpair of the matrix $A + E$. For each μ the corresponding $\|E\|/\|A\|$ is a* backward error estimate *for μ.*

In §6 we estimate the accuracy of each computed eigenvalue approximation by computing backward error estimates. If the size of a matrix or vector is clear from the context, subscripts specifying sizes may be dropped. Unless explicitly stated otherwise, QL decomposition will mean complex orthogonal QL decomposition.

3. Relevant properties and lemmas. We use T to denote the complex symmetric tridiagonal matrix whose eigenvalues are to be computed where

$$T \equiv \begin{pmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \beta_3 & \alpha_3 & \ddots & \\ & & \ddots & \ddots & \beta_k \\ & & & \beta_k & \alpha_k \end{pmatrix}. \tag{3}$$

In this paper we assume that T is diagonalizable. This simplifies the discussions and the numerical studies. Defective eigenvalues introduce an additional level of uncertainty and complexity which we do not address. There is, however, no a priori reason why the proposed procedure, which we denote by CMTQL1, cannot be used on defective matrices.

We will also assume that T is irreducible [15]. However, this is not a restriction on the use of the procedure since at each iteration the procedure works with the irreducible submatrices of the current iterate. Theoretically, irreducible, diagonalizable tridiagonal matrices cannot have multiple eigenvalues.

LEMMA 3.1. *If T is an irreducible tridiagonal matrix with multiple eigenvalues, then one or more of the eigenvalues of T is defective and T is not diagonalizable.*

Lemma 3.1 is an immediate consequence of the fact that for any scalar μ, the determinant of the minor of $T-\mu I$ corresponding to the $(1,n)$ element is equal to the nonzero product of the subdiagonal elements of T. In practice, however, it is easy to generate irreducible, diagonalizable complex (or real) symmetric tridiagonal matrices with numerically multiple eigenvalues.

If T is diagonalizable there exist a diagonal matrix of eigenvalues Λ and a complex orthogonal basis of right eigenvectors X such that

$$T = X\Lambda X^T. \tag{4}$$

If T has distinct eigenvalues and an ordering of the eigenvalues λ_j is specified, then X is uniquely specified. Since T is symmetric, any right eigenvector of T is also a left eigenvector of T [8]. Therefore using the definitions of condition number given, for example, in [29], we obtain the following lemma expressing the condition of each eigenvalue in terms of the right and the left eigenvectors of T.

LEMMA 3.2. *If T has distinct eigenvalues and Λ and X satisfy* (4)*, then for each λ_j and corresponding right eigenvector x_j,*

$$\mathrm{cond}(\lambda_j) = \|x_j\|^2/|x_j^T x_j|, \tag{5}$$

where $\mathrm{cond}(\lambda_j)$ *denotes the condition number of λ_j.*

Lemma 3.3 states that any irreducible, real or complex, nonsymmetric tridiagonal matrix is diagonally similar to an irreducible complex symmetric tridiagonal matrix [9]. Therefore, the eigenvalues of any irreducible nonsymmetric tridiagonal matrix can be computed by applying CMTQL1 to a corresponding complex symmetric matrix.

LEMMA 3.3 (see [9]). *Let T_a be any irreducible tridiagonal matrix. Let $\alpha_j = T_a(j,j)$, $\beta_{j+1} = T_a(j,j+1)$ and $\gamma_{j+1} = T_a(j+1,j)$. Then the eigenvalues of T_a depend only upon the α_j and the products $\beta_{j+1}\gamma_{j+1}$. Furthermore, T_a and the complex symmetric matrix T_{cs} defined by $T_{cs}(j,j) = \alpha_j$ and $T_{cs}(j+1,j) = \sqrt{\beta_{j+1}\gamma_{j+1}}$ have the same eigenvalues.*

Lemma 3.4 indicates the relationship between a complex symmetric rotation and a complex but not unitary Givens rotation [15].

LEMMA 3.4. *Any complex symmetric rotation P_k satisfies $P_k^T = P_k = P_k^{-1}$. $G_k \equiv JP_kDJ$, where D is diagonal with $d_{11} = -1$ and $d_{22} = 1$ is a complex Givens plane rotation with $G_k(k+1,k+1) = G_k(k,k) = c_k$, and $G_k(k+1,k) = -G_k(k,k+1) = s_k$.*

The following lemmas require the matrix to be nonsingular. However, the eigenvalue procedure is not limited to nonsingular matrices. Shifting is an integral part of the procedure and leads the computations through various parts of $\omega(T)$, including through any zero eigenvalue.

LEMMA 3.5 (see [15]). *Let B be symmetric (real or complex) and nonsingular. Then a nonsingular L exists such that $B = LL^T$ if and only if all of the leading principal submatrices of B are nonsingular. Similarly, a nonsingular L exists such that $B = L^TL$ if and only if all of the principal submatrices of JBJ are nonsingular.*

The statement of sufficiency in Lemma 3.5 is proved in [15] for real symmetric matrices but that proof uses only the symmetry. Necessity is readily established by observing that each L_j is nonsingular and each $B_j = L_jL_j^T$. The second part of the proof is obtained by applying a similar argument to the matrix JBJ and using the fact that for any L, the matrix JLJ is upper triangular. Lemma 3.6 exhibits the connection between complex L^TL decompositions and complex orthogonal QL decompositions.

LEMMA 3.6. *Let A be a nonsingular symmetric matrix. Then a nonsingular matrix L exists such that $A^TA = L^TL$ if and only if a complex orthogonal Q exists such that $A = QL$. Moreover, these factorizations are essentially unique.*

Proof. Let $A = QL$. Since $Q^TQ = I$, $A^TA = L^TL$. Conversely, if $A^TA = L^TL$, define $Q = AL^{-1}$. Clearly, $Q^TQ = I$ and $A = QL$. If $A = Q_1L_1 = Q_2L_2$ or equivalently $A^TA = L_1^TL_1 = L_2^TL_2$, then $Q_1^TQ_2 = L_1L_2^{-1} = L_1^{-T}L_2^T = D$, where D must be diagonal. Therefore, $D^2 = I$, and $Q_2 = Q_1D$, $L_1 = DL_2$, and $L_2 = DL_1$. □

CMTQL1 requires a QL decomposition at each iteration. Example 3.1 illustrates that not every nonsingular, irreducible, complex symmetric tridiagonal matrix has a QL decomposition. Therefore, there is no a priori guarantee that each stage of the proposed procedure will be well defined. We address this question in §§5 and 6.

Example 3.1. Let

$$T = \begin{bmatrix} 1 & b^{-1} & 0 \\ b^{-1} & -i & 1 \\ 0 & 1 & 2+i \end{bmatrix}, \quad \text{where } b = \sqrt{1+i}. \tag{6}$$

T is irreducible, complex symmetric, tridiagonal and nonsingular. The determinant of T equals $-3i/(1+i) \neq 0$. By Lemma 3.6, T has a QL factorization if and only if there exists L such that $T^2 = L^TL$. But by Lemma 3.5 this factorization exists if and only if each principal minor of JT^2J is nonsingular. However, the second principal minor of JT^2J is the singular matrix

$$\begin{bmatrix} b^{-2} & 2 \\ 2 & 4(1+i) \end{bmatrix}. \tag{7}$$

LEMMA 3.7. *For any matrix A*

$$\|A\|_2 = \|A^T\|_2. \tag{8}$$

We have the following simple relationship between the condition number of a complex orthogonal matrix and its norm.

LEMMA 3.8. *If Q is a complex orthogonal matrix then*

$$\kappa_2(Q) = \|Q\|_2^2. \tag{9}$$

The following lemma is needed in §6. It is used to obtain backward error estimates for each of the CMTQL1 eigenvalues computed for each test matrix.

LEMMA 3.9. *Let A be any matrix, μ be any scalar and x_μ be any vector. Define*

$$E_\mu^* = r_\mu x_\mu^H / \|x_\mu\|_2^2 \text{ and } E_\mu = r_\mu x_\mu^T / x_\mu^T x_\mu, \text{ where } r_\mu \equiv -Ax_\mu + \mu x_\mu. \tag{10}$$

Then

$$(A + E_\mu^*)x_\mu = \mu x_\mu, \ (A + E_\mu)x_\mu = \mu x_\mu, \text{ and} \tag{11}$$

$$\begin{aligned} \|E_\mu^*\|_2 &= \|r_\mu\|_2 / \|x_\mu\|_2, \\ \|E_\mu\|_2 &= \|r_\mu\|_2 \|x_\mu\|_2 / |x_\mu^T x_\mu|, \\ \|E_\mu\|_2 &= \|x_\mu\|_2^2 \|E_\mu^*\|_2 / |x_\mu^T x_\mu|. \end{aligned} \tag{12}$$

Proof. Since for any B, $\|B\|_2^2 = \|B^H B\|_2$,

$$|x_\mu^T x_\mu|^2 \|E_\mu\|_2^2 = \|r_\mu\|_2^2 \|\bar{x}_\mu x_\mu^T\|_2. \text{ But, } \|\bar{x}_\mu x_\mu^T\|_2^2 = \|\bar{x}_\mu x_\mu^T\|_2 \|x_\mu\|_2^2. \tag{13}$$

A similar argument yields the expression for $\|E_\mu^*\|_2$. □

Clearly, $\|E_\mu\|_2 \geq \|E_\mu^*\|_2$. We use the following lemmas in the discussion of convergence in §4 where K and Y in these lemmas reduce to $K = \Lambda$ and $Y = X^T$ in (4).

LEMMA 3.10 (see [24]). *Let H be an irreducible upper Hessenberg matrix. Let K be any lower Jordan form for H. Then there exists a nonsingular matrix Y such that $H = Y^{-1}KY$ and Y permits a triangular decomposition $Y = LR$ where L is unit lower triangular and R is upper triangular.*

LEMMA 3.11. *Let T be a diagonalizable, complex symmetric, tridiagonal matrix with $T = X\Lambda X^T$, where $X^T X = I$. Then there exist R and L such that $X^T = RL$, where R is unit upper triangular and L is lower triangular.*

Proof. From Lemma 3.10 there exist an upper triangular matrix $\bar{R}$ and a unit lower triangular matrix $\bar{L}$ such that $X^T = \bar{L}\bar{R}$. Therefore, $X = \bar{R}^T \bar{L}^T = \tilde{L}\tilde{R}$. Since X is nonsingular $\tilde{L}$ and $\tilde{R}$ are nonsingular. Therefore, $X^T = X^{-1} = \tilde{R}^{-1}\tilde{L}^{-1} = RL$. □

In §4 we outline the basic complex symmetric QL procedure and obtain a convergence theorem for this basic procedure under the assumptions that each step in the procedure is well defined and that the eigenvalues of T have distinct magnitudes.

4. Basic procedure and convergence. Basic CMTQL1.

1. Set $T_1 = T$.
2. For $k = 1, 2, \ldots$ specify a shift ϕ_k and compute a complex orthogonal factorization of $T_k - \phi_k I = Q_k L_k$.

3. Increment k and define $T_{k+1} = L_k Q_k + \phi_k I = Q_k^T T_k Q_k$.
4. Check for convergence. If not converged go to step 2.

In practice convergence occurs in stages, typically a few eigenvalues converge every few iterations. Each time one or more eigenvalues converge, the current iterate T_k is deflated to smaller irreducible problems to which the procedure is then applied.

Since each $Q_k^T = Q_k^{-1}$ the eigenvalues of each T_k equal the eigenvalues of T. Define

$$\mathcal{Q}_k \equiv Q_1 Q_2 \dots Q_k \text{ and } \mathcal{L}_k \equiv L_k L_{k-1} \dots L_1. \tag{14}$$

The following lemma is easily verified by induction [32].

LEMMA 4.1. *For each k, $\mathcal{Q}_k$ is complex orthogonal,*

$$T_{k+1} = \mathcal{Q}_k^T T \mathcal{Q}_k, \quad \text{and} \quad p_k(T) = \mathcal{Q}_k \mathcal{L}_k, \tag{15}$$

where $p_k(z) \equiv (z - \phi_1) \dots (z - \phi_k)$ is the kth degree polynomial defined by the shifts.

Watkins and Elsner [30], [32] develop a general framework for procedures for computing eigenvalues of a given matrix A that includes factorizations of the form $A_k - \phi_k I = G_k R_k$ (or $G_k L_k$) with the next iterate $A_{k+1} \equiv G_k^{-1} A_k G_k$. They call these GR procedures, and focus on the connections between such procedures and simultaneous iteration procedures [25]. They emphasize the fact that any GR procedure is a nested subspace iterations procedure. Theoretically, for each j with $j \leq n$, the subspace spanned by the first j columns of $\mathcal{G}_k \equiv G_1 \dots G_k$ is identical to the subspace

$$\begin{aligned} \mathcal{S}_j^k &\equiv \operatorname{sp}\{p_k(A)e_1, \dots, p_k(A)e_j\}, \text{ and} \\ \mathcal{T}_j^k &\equiv \operatorname{sp}\{(p_k(A))^{-H} e_{j+1}, \dots, (p_k(A))^{-H} e_n\} \end{aligned} \tag{16}$$

is the orthogonal complement of $\mathcal{S}_j^k$.

Watkins and Elsner [30], [32] obtain a convergence theorem for simultaneous iterations that demonstrates the convergence of the $\mathcal{S}_j^k$ to invariant subspaces $\mathcal{U}_j$ of A under certain conditions on the polynomials p_k. Since the distance $d(\mathcal{S}_j^k, \mathcal{U}) = d(\mathcal{T}_j^k, \mathcal{U}^\perp)$, the $\mathcal{T}_j^k$ subspaces must also converge. The arguments require an additional condition that is always satisfied when A is an irreducible Hessenberg matrix [23].

The nominal objective of any GR procedure is to determine a coordinate system in which the linear operator is block triangular (in our case block diagonal). Subspace convergence is not sufficient to guarantee the convergence of the A_k to block triangular form. Under the additional assumption that the condition numbers of the $\mathcal{G}_k$ are uniformly bounded over k, Watkins and Elsner [30], [32] demonstrate convergence to block triangular form, subject to certain assumptions on the polynomials p_k. The following theorem is a combination of theorems from [32] as they apply to the basic CMTQL1 procedure with all shifts set to zero ($p_k = z$ for all k). For details the reader is referred to [30], [32].

THEOREM 4.1 (see [32]). *Let T be a diagonalizable complex symmetric tridiagonal matrix. Suppose j is such that $|\lambda_j| < |\lambda_{j+1}|$ and set $\rho = |\lambda_j|/|\lambda_{j+1}|$. Let $\mathcal{U}_j$ and $\mathcal{W}_j$ be the invariant subspaces associated with $\{\lambda_1, \dots, \lambda_j\}$ and $\{\lambda_{j+1}, \dots, \lambda_n\}$, respectively. Let T_k be the sequence of complex symmetric tridiagonal iterates generated by* CMTQL1 *with all shifts equal zero. Let $\mathcal{S}_j^k$ be the subspaces generated by the last $m = n - j$ columns of $\mathcal{Q}_k$. If there exists a constant Θ such that $\kappa_2(\mathcal{Q}_k) \leq \Theta$ for all k, then T_k converges to block diagonal form, in the following sense. Let*

$$T_k = \begin{pmatrix} T_{11}^k & T_{12}^k \\ T_{12}^k & T_{22}^k \end{pmatrix}, \tag{17}$$

where $T_{22}^k \in \mathcal{C}^{m\times m}$. *Then*

$$\|T_{12}^k\|_2 \leq 2\sqrt{2}\Theta\|T\|_2 d(\mathcal{S}_j^k, \mathcal{W}_j), \tag{18}$$

where for some constant Ω *independent of* k,

$$d(\mathcal{S}_j^k, \mathcal{W}_j) \leq \Omega\|X\|_2{}^2\rho^k. \tag{19}$$

If we can prove that the condition numbers of the $\mathcal{Q}_k$ are uniformly bounded over k, then Theorem 4.1 provides a proof of convergence of the basic CMTQL1 procedure under the assumption that all shifts are zero and that each step of CMTQL1 is well defined. We can obtain such a result if the eigenvalues of T have distinct magnitudes. However, that proof is itself a proof of the convergence.

Before proceeding we note that the spaces $\mathcal{T}_j^k$ in (17) correspond to the Hermitian conjugate of A. Any real eigenvalues of A will be eigenvalues of A^H. If A is real, then A and A^H have the same eigenvalues. However, if A is complex, the complex eigenvalues of A need not occur in conjugate pairs.

When T is complex and symmetric, we could consider the complex orthogonal complement of the $\mathcal{S}_j^k$ in (16),

$$\mathcal{V}_j^k \equiv \text{sp}\{(p_k(T))^{-1}e_{j+1}, \ldots, (p_k(T))^{-1}e_n\}. \tag{20}$$

These subspaces are not orthogonal to the $\mathcal{S}_j^k$ so convergence of the $\mathcal{S}_j^k$ does not guarantee the convergence of the $\mathcal{V}_j^k$. However, the speed of convergence observed in the numerical tests indicates that these corresponding subspaces are probably playing a role in the rapid convergence.

THEOREM 4.2. *Let* T *be a nonsingular, irreducible, diagonalizable, complex symmetric tridiagonal matrix with no eigenvalues equal in magnitude. Let* $T = X\Lambda X^T$ *where* $|\lambda_1| < \cdots < |\lambda_n|$. *Apply* CMTQL1 *with all shifts equal to zero. Then there exists* Θ *such that*

$$\kappa_2(\mathcal{Q}_k) \leq \Theta \text{ and} \tag{21}$$

$$\mathcal{Q}_k \to X(\text{essentially}) \quad \text{and} \quad T_k \to \Lambda \text{ as } k \to \infty. \tag{22}$$

The ordering of the λ_j uniquely determines X. The proof of Theorem 4.2 uses the following lemmas. Detailed proofs of these lemmas are in [7].

LEMMA 4.2. *Let* E_k, $k = 1, 2, \ldots$ *denote a family of symmetric matrices such that* $\|E_k\| \to 0$ *as* $k \to \infty$. *Then for large* k, *the matrices* $I + E_k$ *have* $L_k^T L_k$ *decompositions where* $L_k \to I$ *(essentially) as* $k \to \infty$.

Proof. For large k each $J(I + E_k)J$ and its principal submatrices are diagonally dominant. Therefore by Lemma 3.5 the required factorizations exist. It is easy to prove that each $L_k(j,j)^2 \to 1$, from which it is then easy to prove by induction that all nondiagonal entries of the L_k must converge to 0 as $k \to \infty$. □

LEMMA 4.3. *Let* F_k, $k = 1, 2, \ldots$ *be such that* $\|F_k\| \to 0$ *as* $k \to \infty$. *Then for any* $\epsilon > 0$ *and large* k, *there exist complex orthogonal decompositions,* $I + F_k = Q_k L_k$ *such that*

$$\|Q_k\| \leq (1 + \epsilon)/(1 - \epsilon), \tag{23}$$

$Q_k \to I$ *(essentially) and* $L_k \to I$ *(essentially) as* $k \to \infty$.

COROLLARY 4.1. *Under the assumptions of Lemma* 4.3 *for any* $\epsilon > 0$ *and for large enough* k,

$$\kappa_2(Q_k) = \|Q_k\|^2 \leq [(1+\epsilon)/(1-\epsilon)]^2. \tag{24}$$

Proof. From Lemmas 4.2 and 3.6, it is straightforward to prove that $Q_k \to I$ (essentially). If we set $E_k = (F_k + F_k^T + F_k^T F_k)$, then inequality (23) follows readily from the fact that $(I+E_k)^{-1} = I - E_k(I+E_k)^{-1}$ and $\|E_k\| \to 0$. The corollary is an immediate consequence of Lemmas 4.3 and 3.8. □

The following lemma relates the preceding lemmas to the $\mathcal{Q}_k$. Using this lemma we can then prove that $\kappa_2(\mathcal{Q}_k)$ are uniformly bounded over k.

LEMMA 4.4. *Under the assumptions of Theorem* 4.2 *the matrices* $\mathcal{Q}_k$ *defined in* (14) *are of the form* $X(I+\tilde{F}_k)D_k$, *where each* D_k *is diagonal with* $D_k^2 = I$ *and* $\|\tilde{F}_k\| \to 0$ *as* $k \to \infty$.

Proof. From Lemma 3.11 there exist R and L such that R is unit upper triangular, L is lower triangular, and

$$X^T = RL \text{ and therefore } T^k = \mathcal{Q}_k\mathcal{L}_k = X(\Lambda^k R \Lambda^{-k})(\Lambda^k L). \tag{25}$$

Clearly, $(\Lambda^k L)$ is lower triangular and

$$\Lambda^k R \Lambda^{-k} = I + F_k \text{ where } F_k(i,j) = R_{ij}[\lambda_i/\lambda_j]^k \text{ for } j > i \tag{26}$$

and zero otherwise. Since the eigenvalues have distinct magnitudes $\|F_k\| \to 0$ as $k \to \infty$.

Therefore, by Lemma 4.3, for large k there exist complex orthogonal decompositions $\bar{Q}_k$ and $\bar{L}_k$ such that $\bar{Q}_k \to I$ (essentially), $\bar{L}_k \to I$ (essentially), and

$$T^k = \mathcal{Q}_k\mathcal{L}_k = (X\bar{Q}_k)(\bar{L}_k\Lambda^k L). \tag{27}$$

By Lemma 3.6 there exist diagonal matrices D_k with ${D_k}^2 = I$ such that

$$\mathcal{Q}_k = X\bar{Q}_k D_k \text{ and } \mathcal{L}_k = D_k\bar{L}_k\Lambda^k L. \tag{28}$$ □

Proof (Theorem 4.2). From (28) we have

$$\mathcal{Q}_k \to X(\text{essentially}). \tag{29}$$

Therefore the condition numbers of $\mathcal{Q}_k$ are uniformly bounded over k. Moreover,

$$T_{k+1} = D_k\bar{Q}_k^T\Lambda\bar{Q}_k D_k \to \Lambda, \tag{30}$$

where the convergence is ordinary since Λ is diagonal. □

If one or more eigenvalues of T are equal in magnitude then only the $F_k(i,j)$ in (26) outside of triangular regions corresponding to indices of eigenvalues equal in magnitude are guaranteed to converge to zero. However, if certain factorizations exist, the proof can be extended to this case with convergence to a block diagonal matrix where each block corresponds to some subset of eigenvalues that are equal in magnitude. The practical importance of an extension of the arguments to this case is not clear since any nonzero shift can be expected to separate eigenvalues that are not equal but have equal magnitudes. Shifts that are not equal to eigenvalues also yield nonsingular matrices.

5. Practical numerical procedure. Explicit subtraction and addition of shifts can lead to significant computational errors. We can, however, obtain an implicit version of CMTQL1 analogous to the implicit version of the real symmetric orthogonal QL procedure [22]. On each iteration of the implicit procedure the new matrix iterate T_{k+1} is obtained from $T_k - \phi_k I$ by the application of $n-1$ complex symmetric plane rotations.

DEFINITION 5.1. *Let T be a complex symmetric tridiagonal matrix. Define $\hat{T}_n = T$ and the complex symmetric rotation P_{n-1} whose last column is a multiple of the last column of T. Set $\hat{T}_{n-1} = P_{n-1}TP_{n-1}$. For $k = 2, \ldots n-1$ define $\hat{T}_{n-k} = P_{n-k}\hat{T}_{n-k+1}P_{n-k}$ with each P_{n-k} a complex symmetric rotation defined to zero out the $(n-k, n-k+2)$ and $(n-k+2, n-k)$ entries in $\hat{T}_{n-k+1}$. The successive two-sided application of the P_{n-k} to the $\hat{T}_{n-k+1}$ to obtain $\hat{T}_1$ is called a* (bottom up) sweep *across T.*

The following theorem which was proved in [7] states that a bottom up sweep can be completed if and only if the starting matrix T has a complex orthogonal factorization QL, and that the matrix resulting from the sweep equals Q^TTQ. This equivalence provides a mechanism, namely, the successful completion of a sweep, to guarantee that the implicit CMTQL1 procedure is performing the required QL factorizations and recompositions. Watkins and Elsner [31] extend this theorem to the GR class of procedures, which includes a complex orthogonal QR variant of CMTQL1.

Shifts are not explicitly considered in Theorem 5.1. Shifts are, however, used in CMTQL1. A shift ϕ_k is incorporated into the sweep implicitly as follows. The P_{n-1} rotation is defined using the last column of $T_k - \phi_k I$. This rotation is applied implicitly to $T_k - \phi_k I$ using the fact that the update formulas corresponding to each of the complex symmetric rotations generated within the sweep involve only differences of the diagonal entries of the matrix being transformed. The actual computations are done on T_k so that the resulting matrix is $Q^TT_kQ = LQ + \phi_k I$. Such a sweep will be called an implicit sweep.

THEOREM 5.1. *Let T be a nonsingular, irreducible, complex symmetric tridiagonal matrix. T has a complex orthogonal factorization $T = QL$ if and only if a full bottom up sweep is defined for T. Moreover, upon completion of the sweep T has been transformed into $T^+ \equiv Q^TTQ$.*

Proof. Assume $T = QL$ exists, then L is nonsingular and $T^2 = L^TL$. A sweep can be completed if and only if each of the complex rotations P_{n-k} is well defined. P_{n-k} is well defined if and only if the denominator of c_{n-k} is nonzero. By Lemma 3.5 each submatrix ${T^2}_{n,j}$ is nonsingular. An induction and continuation argument demonstrates that the denominator of c_{n-k} is a simple multiple of the determinant of ${T^2}_{n,n-k+1}$. Therefore, each rotation is well defined and the sweep can be completed.

Conversely, if a full sweep is defined, set $\tilde{T}_n = T$,

$$Q \equiv P_{n-1} \ldots P_1, \text{ and for each } k \text{ define } \tilde{T}_{n-k} \equiv P_{n-k}\tilde{T}_{n-k+1}. \tag{31}$$

Then

$$Q^TQ = I \text{ and } \tilde{T}_1 = Q^TT. \tag{32}$$

A straightforward induction argument demonstrates that the $n-k+1$ to n columns of $\tilde{T}_{n-k}$ are lower triangular. Therefore, $\tilde{T}_1 = L$ is lower triangular, and $T = QL$ is a complex orthogonal decomposition of T. Furthermore, $T^+ = \tilde{T}_1Q = Q^TTQ$. ∎

5.1. Numerical stability. The successful completion of a sweep is equivalent to the successful computation of a QL decomposition of the current matrix iterate $T_k - \phi_k I$, coupled with the formation of the next iterate $T_{k+1} = Q^T T Q$. If for some sequence of shifts we can construct a successful sweep for each shift, this is an implementation of the CMTQL1 procedure. The convergence behavior will however be correlated to the particular set of shifts which are used.

In practice we must deal with additional considerations. Theorem 5.1 does not give any indication of the numerical condition of the complex symmetric rotations used within each sweep or of the overall transformation matrices $\mathcal{Q}_k$. By construction, at least theoretically, each matrix iterate has the same eigenvalues as T. However, the eigenspaces of the iterates differ from those of T by the matrix $\mathcal{Q}_k$. For numerical stability we need to control the condition of these matrices.

Therefore, we introduce a heuristic that both monitors and controls local numerical stability. We consider the inner plane rotations but drop the subscripts. Within each sweep in CMTQL1 the computation of each c and s in each P is implemented by computing

$$1/\sqrt{1+w^2} \quad \text{and} \quad w/\sqrt{1+w^2}, \tag{33}$$

where $w = a+bi$ is a complex number. Each w is defined as the ratio of the two relevant entries from the current inner iteration matrix iterate ordered so that $\|w\| \leq 1$. Therefore, $a^2 \leq 1$ and $b^2 \leq 1$. For specific details see [7].

Each c and s are well defined numerically if and only if $1 + w^2$ is not too close to 0 or equivalently w is not too close to $i = \sqrt{-1}$. Difficulties can occur if a is small and b is close to 1. Therefore at each step in each sweep and for some prespecified q the procedure checks for satisfaction of the following inequality:

$$(1 + a^2 - b^2) \leq 10^{-q} b^2 \text{ or equivalently } 1 + a^2 \leq (1 + 10^{-q}) b^2. \tag{34}$$

If (34) is satisfied, cancellation has occurred in the computation of $1 + w^2$ and the current sweep is restarted using a randomly generated complex shift scaled by an estimate of the norm of T. Exceptional shifts have been used by other authors. See, for example, [10].

5.2. Choice of shifts. From (20) we expect the procedure to behave as though it is simultaneously a method using $p_k(T)$ and $(p_k(T))^{-1}$, where p_k is the kth degree polynomial defined by the shifts $\phi_j, 1 \leq j \leq k$. Therefore, we expect the rate of convergence to improve if a shift ϕ_k is close to an eigenvalue of T. For almost all of the iterations the shift is set equal to the Wilkinson shift, that eigenvalue of the uppermost nontrivial 2×2 submatrix of the current iterate T_k which is closest to the $(1,1)$ element of that submatrix.

The starting shift, however, is chosen to be a randomly generated complex number. There are two reasons for this choice. First, we force the computations into complex arithmetic. If for example, T is a symmetrized real, nonsymmetric tridiagonal matrix, then its diagonal entries are real and its nonzero off diagonal entries are either real or purely imaginary. Therefore, at least initially w in (33) could be purely imaginary which increases the probability of difficulties with the denominator $1 + w^2$. Second, to avoid pathological cases such as the following example, where of course B

is not irreducible, not nonsingular, and is in fact defective.

$$B = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{35}$$

5.3. Practical implementation.

CMTQL1, practical implementation.

1. Set $T_1 = T$.
2. Check if the current matrix T_k is reducible. If so replace T_k by a smaller irreducible matrix or matrices and apply the procedure to each irreducible submatrix.
3. For $k = 1$ generate a random complex shift ϕ_1. Otherwise, set ϕ_k equal to the eigenvalue of the uppermost 2×2 submatrix of T_k which is closest to the first diagonal entry of T_k.
4. Initiate a bottom up implicit sweep on $T_k - \phi_k I$. If inequality (34) is satisfied at any stage in the sweep, terminate that sweep, generate a random complex shift η scaled by an estimate of the $\|T\|_2$ and restart the sweep setting $\phi_k = \eta$. Repeat as necessary until a complete sweep is achieved.
5. Increment k and define T_{k+1} as the tridiagonal matrix obtained from the last successful sweep.
6. Check for convergence. If not converged, set $k = k + 1$ and go to step 2.

Specifically, in step 2 we use the following test.

$$|T(j+1,j)| \leq \sqrt{2}\epsilon_{\text{mach}} \left(|T(j,j)| + |T(j+1,j+1)| \right),$$

where ϵ_{mach} is the computer machine epsilon. In §6 we examine the performance of CMTQL1 on three different families of test problems.

6. Numerical tests. The practical value of any numerical procedure can only be determined by applying it to relevant test problems and tracking its behavior on such problems. The eigenvalues of a diagonalizable, complex symmetric tridiagonal matrix may be anywhere in the complex plane, and matrices of corresponding eigenvectors may be arbitrarily ill conditioned. Small perturbations in such a matrix may cause large perturbations in the computed eigenvalues.

In our experiments we computed both backward error estimates, using Lemma 3.9, and two-sided Rayleigh quotient differences. For some of the smaller problems, $n = 225$, we also compared the eigenvalues computed by CMTQL1 to those computed using the subroutine COMQR from [28]. In each case these two sets of eigenvalues agreed to 10 or more digits. We note however that if the two sets of eigenvalues had not agreed, we could not have concluded that either procedure was incorrect unless it was known a priori that the eigenvalues of the test matrices were well conditioned. In these tests we can only hope to demonstrate that the eigenvalues obtained from CMTQL1 are eigenvalues of matrices which are small perturbations of the original matrix whose eigenvalues are to be computed.

For each test matrix T and each corresponding CMTQL1 eigenvalue μ, we applied one step of inverse iteration to compute an approximate eigenvector x_μ. This vector was then used to compute the corresponding backward error estimates $\|E_\mu^*\|_2$

defined in Lemma 3.9. Estimates of the condition numbers, $\|x_\mu\|_2^2/|x_\mu^T x_\mu|$, and the corresponding unnormalized estimates $\|E_\mu\|_2$ were also computed. In the figures we normalized these values, plotting $\log_{10}(\|E_\mu^*\|_2/\|T\|_2)$ versus eigenvalue number with the eigenvalues ordered by magnitude.

Each x_μ was also used to compute the following two-sided Rayleigh quotient differences.

$$\epsilon_\mu^{rq} \equiv \mu - \left(x_\mu^T T x_\mu / x_\mu^T x_\mu\right). \tag{36}$$

Since T is complex symmetric, two-sided Rayleigh quotients reduce to one-sided quotients. Since the magnitudes of the real and of the imaginary parts of μ can differ radically, in the comparisons the real and the imaginary parts of each ϵ_μ^{rq} were normalized and plotted separately, using, respectively, $+$ and $\diamond$ symbols.

A primary interest in this procedure is its use in nonsymmetric Lanczos procedures for computing eigenvalues of large nonsymmetric matrices where (when there is no look-ahead) a general nonsymmetric problem is reduced to a family of complex symmetric tridiagonal problems. We therefore considered test matrices generated by applying the nonsymmetric Lanczos procedures to matrix 425 in file 13 (425boe13), to matrix 479 in file 14 (479boe14), and to the sherman4 and sherman5 matrices in the Boeing–Harwell collection [11]. These Lanczos matrices have the property that as the size is increased, significant numbers of eigenvalues of these matrices are not only nearly equal in magnitude but also nearly equal (numerically).

We used COMQR in [28] to compute the eigenvalues of three of these matrices. Both 425boe13 ($n = 425$) and 479boe14 ($n = 479$) have many complex conjugate pairs of eigenvalues and many real eigenvalues. The eigenvalues of 425boe13 surround the origin like a cloud. Eigenvalues range in magnitude from a real eigenvalue .12 to complex eigenvalues $.148 \pm 5.28i$. The gaps between nearest-neighbor eigenvalues range from .03 to .75 and there is a double real eigenvalue -1.0. The eigenvalues of 479boe14 form a swath parallel to the real axis with large outliers close to both the real and the imaginary axes. Eigenvalues range in magnitude from a real eigenvalue .00017 to complex eigenvalues $.009 \pm 1700$. The gaps between eigenvalues range from .00043 at .00017 to 1580 at $.009 \pm 1700.i$ of size 1700. The eigenvalues of sherman4 ($n = 1104$) are positive real. 1.0 is an eigenvalue of multiplicity 558. The other eigenvalues range in size from .0307 to 66.5, and the gaps range in size from .0016 to 2.22. We did not compute the eigenvalues of sherman5 ($n = 3312$), but Lanczos computations indicate that the eigenvalues are real or nearly real with many eigenvalues on both sides of the imaginary axis. The eigenvalue distributions of each of the test matrices constrain the distributions of the eigenvalues of the corresponding Lanczos test matrices. For additional comments on this aspect of the test matrices see [4].

The other classes of test matrices were generated randomly, using either an exponential(2), Cauchy(4), logistic(7), Laplace(8), or arcsine(9) probability distribution. In the discussion and in the tables we use, for example P2, to denote test matrices corresponding to a particular (exponential) distribution, and RNS479 to denote a Lanczos matrix generated by a real nonsymmetric Lanczos procedure applied to 479boe14.

In contrast to the Lanczos matrices, the eigenvalues of the probability test matrices are typically distinct with distinct magnitudes. The eigenvalue and eigenvalue gap distributions vary significantly across the various probability distributions. Figures 1–2, 3–4, and 5–6 depict, respectively, computed eigenvalues and eigenvalue gap distributions for P2 ($n = 2500$), P4 ($n = 1000$), and P7 ($n = 2500$) test matrices.

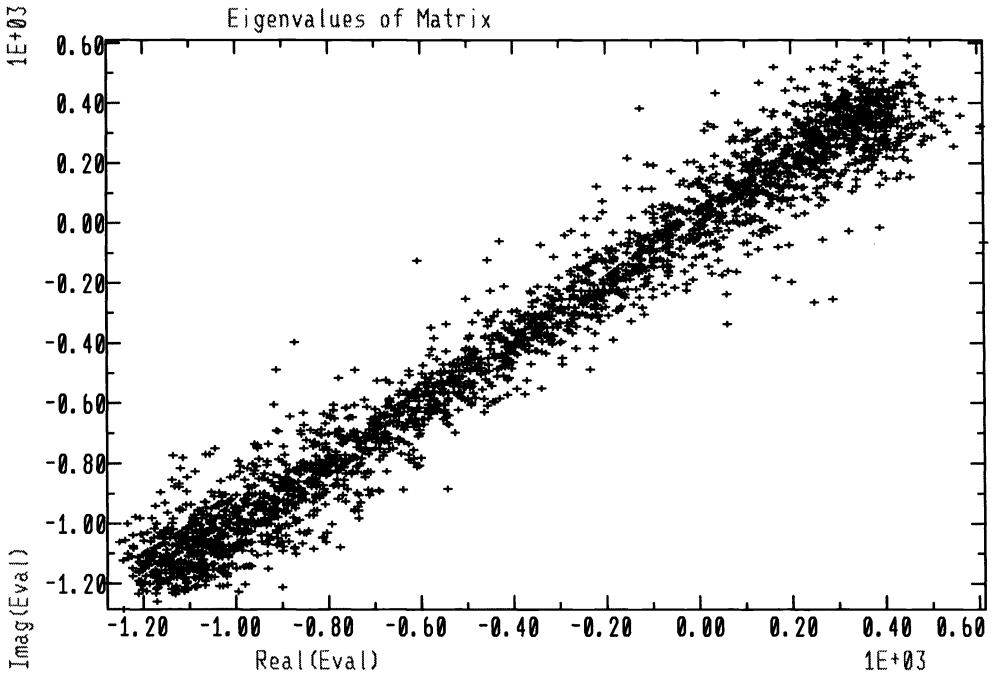


FIG. 1. *Eigenvalues computed by* CMTQL1 *for* P2 *matrix,* $n = 2500$.

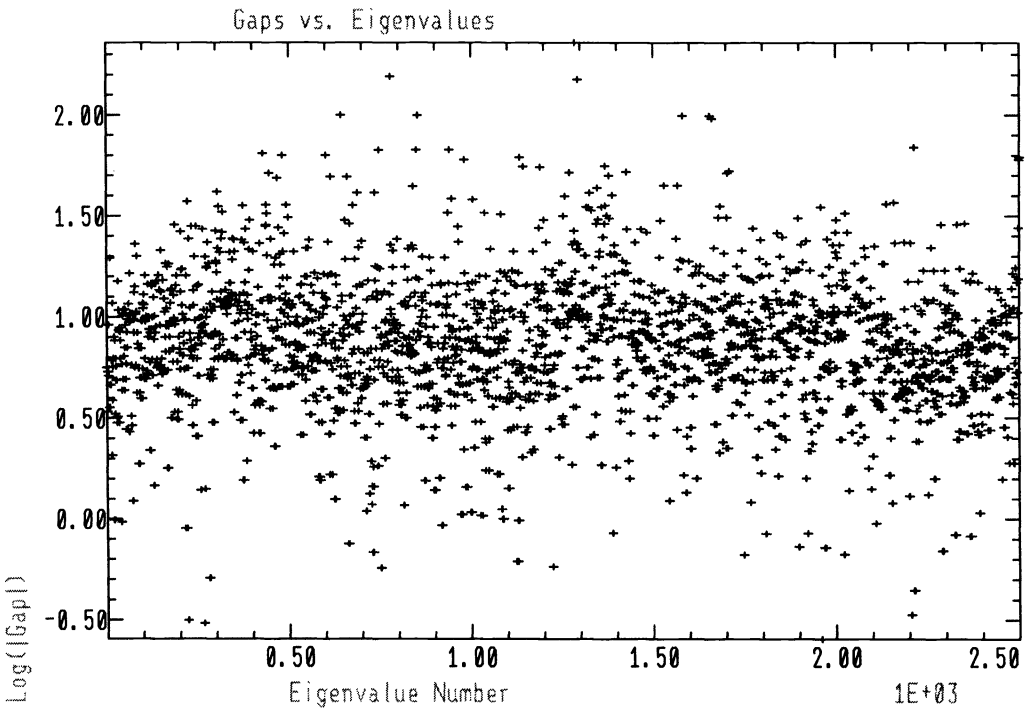


FIG. 2. *Gaps between eigenvalues computed by* CMTQL1 *for* P2 *matrix,* $n = 2500$.

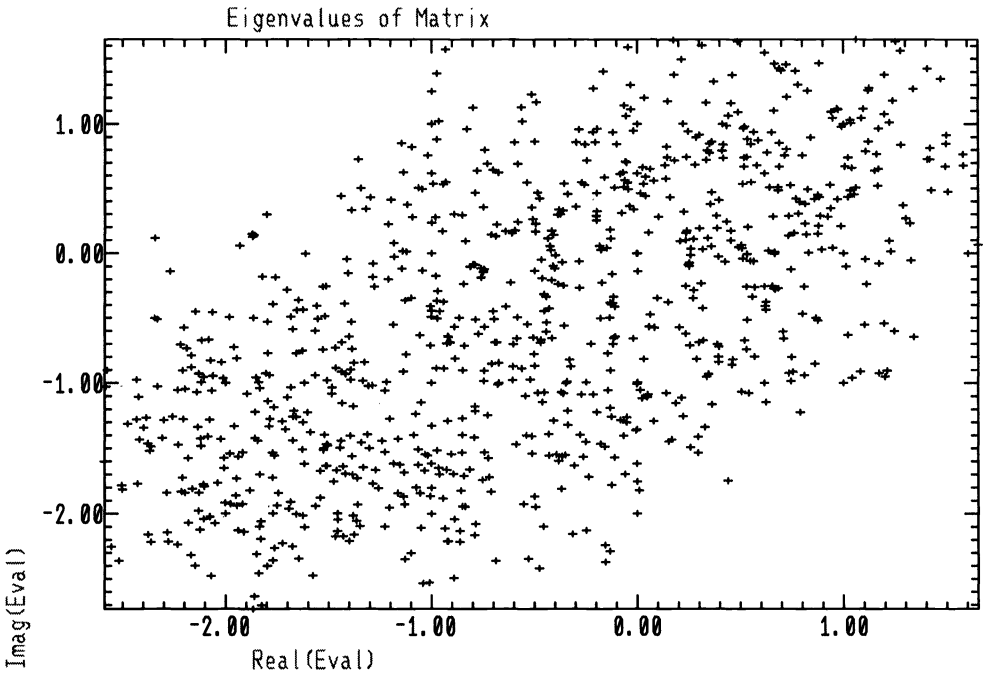


FIG. 3. *Eigenvalues computed by* CMTQL1 *for* P4 *matrix,* $n = 1000$.

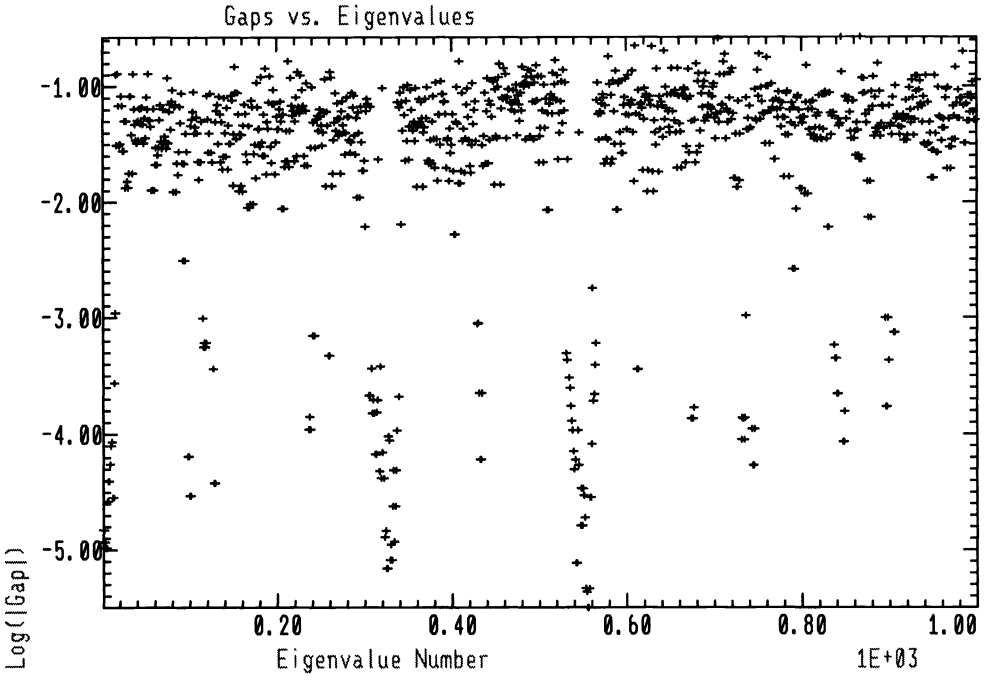


FIG. 4. *Gaps between eigenvalues computed by* CMTQL1 *for* P4 *matrix,* $n = 1000$.

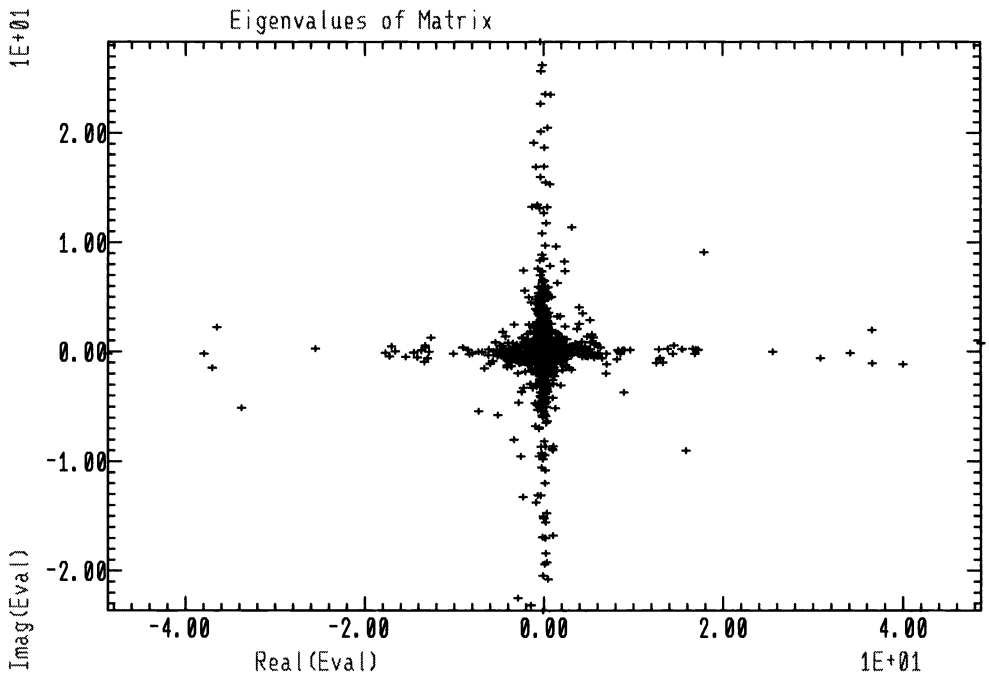


FIG. 5. *Eigenvalues computed by* CMTQL1 *for* P7 *matrix,* $n = 2500$.

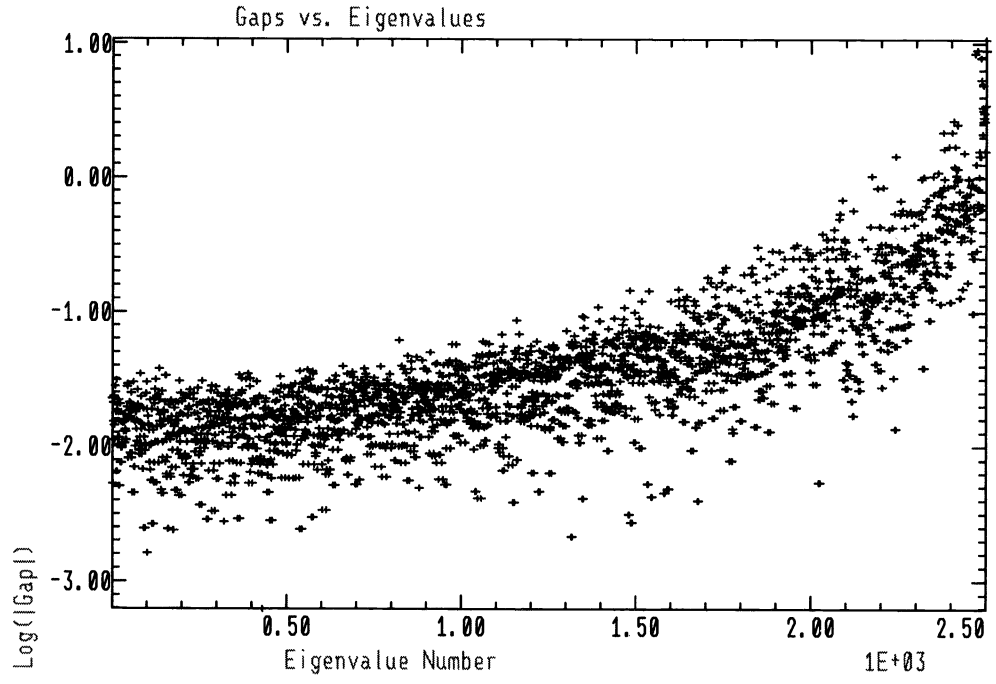


FIG. 6. *Gaps between eigenvalues computed by* CMTQL1 *for* P7 *matrix,* $n = 2500$.

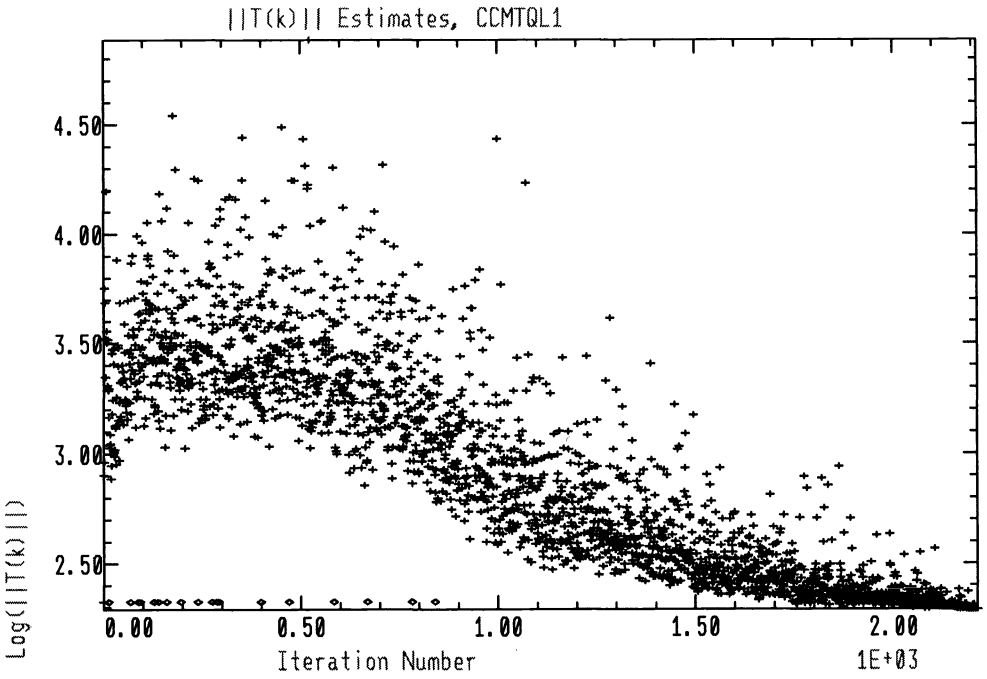


FIG. 7. *Estimates of* $\|T_k\|_2$ *in* CMTQL1 *versus iteration number* k *with restarted iterations marked by* $\diamond$ *for* P8 *matrix,* $n = 1000$.

These figures are representative for the P2, P4, and P7 test matrices. The eigenvalue distributions of P8 test matrices resemble Fig. 1. However, the eigenvalues are an order of magnitude smaller (1.1 to 195.) with the real and the imaginary parts ranging from approximately $-138.$ to $+50$. The corresponding gap distributions resemble Fig. 2 but the gaps range from .04 to 19. The eigenvalue distributions of P9 test matrices resemble Fig. 3 with the real and the imaginary parts of the eigenvalues ranging from approximately -1.0 to 2.5, and the magnitudes varying from .017 to 3.1. The corresponding gap distributions resemble Fig. 2 for P2 matrices but with somewhat smaller gaps ranging from .002 to .29. Matrices of size $n = 225, 425, 479, 1000, 2500,$ and 5000 were generated.

Tests were run using different size matrices from each set of matrices, with different choices for the exponent in the cancellation checks, and with variations in the seed used in the random number generations. Observe that by construction each complex eigenvalue of any Lanczos matrix obtained from a real nonsymmetric Lanczos procedure occurs with a complex conjugate pair. This is not, however, true of the Lanczos matrices obtained using the complex nonsymmetric procedure.

For each test problem we tracked the convergence history in terms of the number of iterations required for the convergence of each eigenvalue, the iteration numbers on which the procedure restarted, the number of partial sweeps required to achieve each restart, and maximum row sum estimates of $\|T_k\|_2$. Figure 7 is typical of the behavior of these estimates. As k increases they tend to oscillate in size but eventually level off or acquire a downward trend as the convergence proceeds. These estimates were computed using the entire kth iterate and not just its irreducible parts. They provide an indirect check on the sizes of the transformation matrices. These estimates

TABLE 1

CMTQL1 *tests on Lanczos matrices, number of sweeps and restarts required, bounds on* $\log_{10}(\|E_\mu^*\|/\|T\|^+)$.

Matrix Size	q	Nitns	NR	$\|E_\mu^*\|/\|T\|^+$	$\|T\|^+$
RNS 479					
479	4	840	1	$(-17.1, -13.4)$	2.3×10^5
1000	4	1778	14	$(-17.0, -13.2)$	5.1×10^5
1000	*	1820	*	$(-16.8, -13.0)$	
2500	5	4394	6	$(-17.3, -12.1)$	7.6×10^5
RNS 425					
425	4	876	0	$(-17.4, -13.5)$	1.7×10^4
1000	4	1887	16	$(-17.3, -11.2)$	1.7×10^4
1000	*	1999	*	$(-17.0, -13.4)$	
2500	4	4607	164	$(-17.3, -12.7)$	3.6×10^4
2500	6	4432	0	$(-17.3, -12.1)$	
CNS 479					
479	4	875	4	$(-15.3, -12.7)$	1.8×10^4
1000	4	1829	8	$(-15.9, -11.8)$	2.3×10^4
1000	6	1797	0	$(-15.6, -11.7)$	
1000	*	1902	*	$(-15.7, -11.7)$	
2500	4	4577	65	$(-16.1, -11.7)$	6.5×10^4
2500	6	4478	0	$(-15.8, -11.7)$	
CNS 425					
425	4	905	0	$(-15.4, -11.0)$	1.4×10^2
1000	4	1884	2	$(-15.3, -11.6)$	2.7×10^2
1000	6	1886	0	$(-15.3, -11.9)$	
1000	*	2051	*	$(-15.3, -10.9)$	
CNS Sher 4					
1000	4	1645	2	$(-15.7, -10.9)$	4.7×10^2
1000	*	1756	*	$(-15.0, -10.7)$	
CNS Sher 5					
1000	4	1772	5	$(-14.7, -10.3)$	1.2×10^4
1000	*	1871	*	$(-15.1, -11.1)$	

were computed and plotted for every iteration, including any restarted iterations. Restarted iterations are labeled with a $\diamond$ symbol located at an estimate of $\|T\|$. We also examined the computed eigenvalue distribution and gap distribution for each of the test matrices and checked for possible connections between the shapes of these distributions and the observed convergence. We did not identify any correlations.

Sample results from each class of matrices are included. See Tables 1 and 2 and Figs. 1–17. In Tables 1 and 2, size refers to the size of the test matrix. $q = 4$ or 6 refers to the exponent in (34). $q = *$ corresponds to a test using the original version of CMTQL1 described in [9] that did not include restarting, did not use a complex random initial shift, and only allowed single block reductions. Nitns denotes the total number of sweeps and partial sweeps required until convergence. NR denotes the total number of restarted sweeps. $\|E_\mu^*\|_2/\|T\|_2^+$ denotes the base 10 logarithms of the minimum and the maximum of the normalized backward error estimates obtained using Lemma 3.9. $\|T\|_2^+$ denotes the estimate of the norm of the test matrix. In

TABLE 2

CMTQL1 *tests on probability matrices, number of sweeps and restarts required, bounds on* $\log_{10}[\|E_\mu^*\|/\|T\|^+]$ *and on* $\log_{10}[\epsilon_\mu^{RQ}]$.

Matrix Size	q	Nitns	NR	$\|E_\mu^*\|/\|T\|^+$	ϵ_μ^{RQ}	$\|T\|^+$
P2						
1000	4	2292	12	−7.5	−9.5	1.93×10^3
1000	6	2288	0	−6.2	−7.6	
1000	*	2282	*	−8.3	−8.0	
2500	4	5859	92	−6.5	−8.7	
2500	6	5754	1	−7.0	−7.3	
2500	*	5747	*	−6.5	−6.5	
5000	4	11931	398	−5.4	−7.3	
5000	6	11579	4	−5.3	−7.2	

For P2, $\min\{\log_{10}[\|E_\mu^*\|/\|T\|^+]\} \le -13.0, \min\{\log_{10}[\epsilon^{RQ}]\} \le -15.1$.

P4						
1000	4	2411	9	−7.1	−9.8	4.0×10^0
1000	6	2397	0	−7.1	−9.6	
1000	*	2435	*	−3.5	−6.3	

For P4, $\min\{\log_{10}[\|E_\mu^*\|/\|T\|^+]\} \le -19.0, \min\{\log_{10}[\epsilon_\mu^{RQ}]\} \le -15.7.$

P7						
1000	4	2239	15	−7.5	−8.8	3.6×10^1
1000	6	2253	0	−9.2	−10.3	
1000	*	2265	*	−9.2	−9.9	
2500	4	5704	63	−8.4	−9.4	5.3×10^1
2500	6	5615	0	−8.8	−9.8	

For P7, $\min\{\log_{10}[\|E_\mu^*\|/\|T\|^+]\} \le -14.8, \min\{\log_{10}[\epsilon_\mu^{RQ}]\} \le -15.7.$

P8						
1000	4	2213	18	−7.9	−9.5	2.1×10^2
1000	6	2198	0	−8.3	−9.5	
1000	*	2200	*	−8.3	−8.9	

For P8, $\min\{\log_{10}[\|E_\mu^*\|/\|T\|^+]\} \le -13.8, \min\{\log_{10}[\epsilon_\mu^{RQ}]\} \le -15.7.$

P9						
1000	4	2584	15	−8.4	−10.2	4.0×10^0
1000	6	2564	0	−6.2	−8.3	
1000	*	2580	*	−7.7	−9.3	

For P9, $\min\{\log_{10}[\|E_\mu^*\|/\|T\|^+]\} \le -13.8, \min\{\log_{10}[\epsilon_\mu^{RQ}]\} \le -15.6.$

Table 2, ϵ_μ^{rq} denotes the base 10 logarithms of the maximum of the componentwise normalized Rayleigh quotient error estimates obtained using (36). Also in Table 2 only the maximum of the base 10 logarithms of $\|E_\mu^*\|_2/\|T\|_2^+$ are tabulated. Table 1 consists of two sections. RNS refers to test matrices generated using a real nonsymmetric Lanczos procedure. CNS refers to test matrices generated using a complex nonsymmetric Lanczos procedure.

In almost all cases the convergence appears to vary smoothly with the number of iterations, independent of the particular eigenvalue distribution and gap distributions. Figure 8 corresponding to P7 with $n = 2500$ is an exception where the convergence

accelerated markedly near the end of the process. In all of the tests fewer than 2.6 sweeps per eigenvalue were required. The larger Lanczos matrices typically required less than two sweeps per eigenvalue, presumably because of the many numerically multiple eigenvalues. The probability matrices with no numerically multiple eigenvalues required between 2.2 and 2.6 sweeps per eigenvalue. These counts are very similar to those observed for the corresponding QL procedure for real symmetric tridiagonal matrices.

We observe that restarting with random complex shifts did not typically increase the number of iterations required over the number required by the original $q = *$ procedure. We also observe that when a *large* number of restarts were used, typically these corresponded to many fewer eigenvalues. For example with a RNS 425boe13 test matrix of size 2500 and $q = 4$, restarting occurred 164 times. However, 79 of those restarts involved only 4 eigenvalues. There were several other eigenvalues which required 3 to 7 restarts, consuming 18 of the 164 restarts. The remaining restarts involved at most two attempts for any given eigenvalue.

In these tables any partial sweep is counted as a full iteration. In Table 1 we observe the uniformly small backward error estimates obtained for both the RNS and the CNS test problems. We also observe that typically there was no significant difference between the maximum backward error estimates obtained using any of the three different values of q.

The results in Table 2 for the probability test matrices are somewhat different. The maximum backward error estimates are markedly larger and in some cases significant differences between the choices of q are indicated. Detailed plots of these estimates indicate however that for $q = 4$ the maximum estimates correspond to very few eigenvalues. See, for example, Fig. 9 for P4 with $n = 1000$. With $q = *$, see Fig. 10, many of the backward error estimates increased significantly. On other problems, however, such as P2 matrices, these estimates were approximately the same size for both values of q. In Fig. 11 we plot the corresponding Rayleigh quotient differences for $q = 4$ for P4 with $n = 1000$. These differences indicate that the CMTQL1 eigenvalues are probably accurate to nine or more digits. A similar plot for $q = *$ indicates that all except three eigenvalues are accurate to eight or more digits. Of those three, the differences indicate that two of them are accurate to $7\frac{1}{2}$ digits and one to 6^+ digits.

Figures 12–13 and 14–15 correspond to a P2 matrix with $n = 2500$. Figure 12 contains the backward error estimates corresponding to $q = 4$. Twenty-one eigenvalues have error estimates larger than 10^{-8}. All are less than $10^{-6.5}$. Figure 13 is the corresponding plot for $q = *$. Figures 14 and 15 are the corresponding ϵ_μ^{RQ} plots. We observe the apparent increased accuracy achieved with the $q = 4$ procedure, as indicated by the amount of scattering in the backward error estimate plots. However, both procedures performed well.

Figures 16–17 correspond to a P7 matrix with $n = 1000$ with $q = 4$. In Fig. 16 we see that the maximum backward error estimate is associated with one eigenvalue. Moreover, for this problem only two eigenvalues have estimates larger than 10^{-9}. The corresponding ϵ_μ^{RQ} differences in Fig. 17 indicate that except for those three eigenvalues, all of the eigenvalues are computed to within 10-digit accuracy.

The tests indicate that restarting typically reduces the amount of scatter in the estimates and increases the accuracy over that obtained from the $q = *$ procedure with no restarting. However, both procedures typically work well in practice.

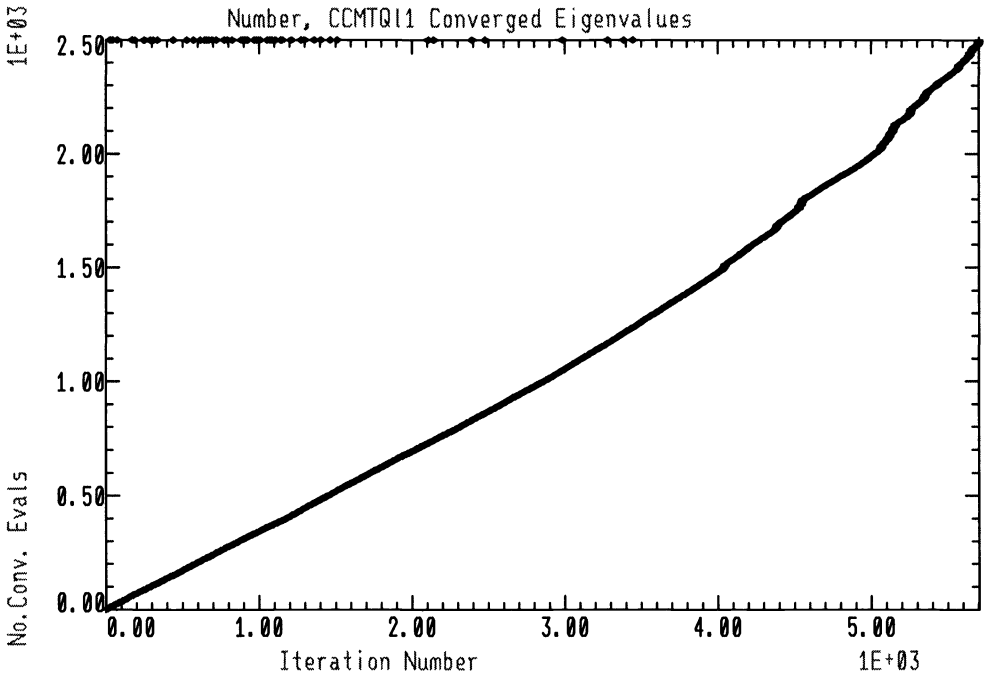


FIG. 8. *Number of computed eigenvalues versus iteration number k with restarted iterations marked by ◇ for* P7 *matrix* $n = 2500$.

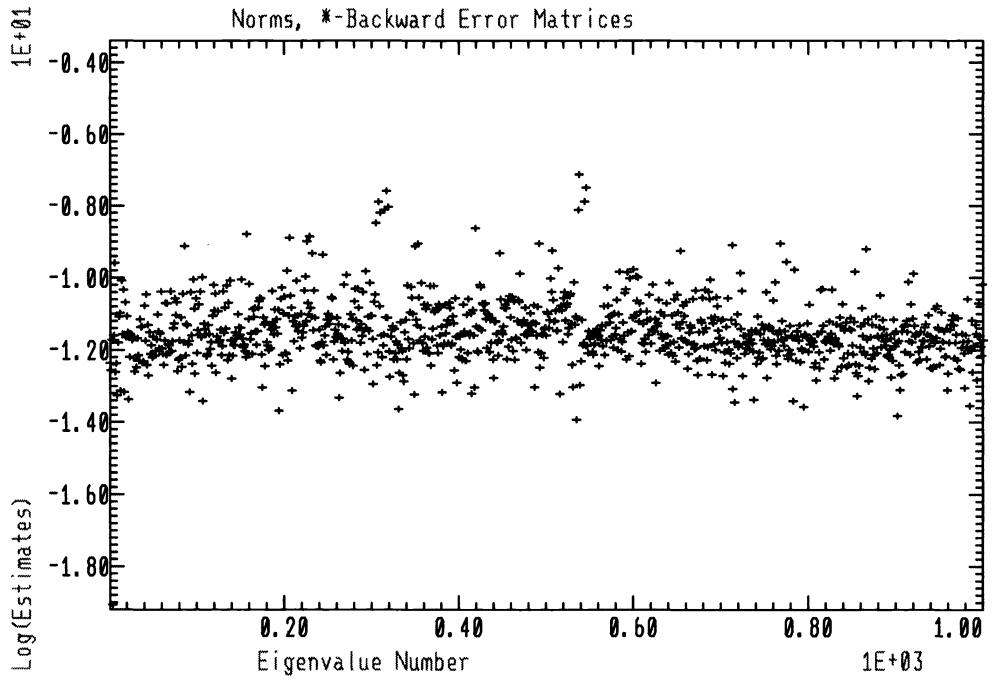


FIG. 9. *Normalized backward error estimates for* $q = 4$ *versus eigenvalue number, eigenvalues ordered by decreasing magnitude for* P4 *matrix,* $n = 1000$.

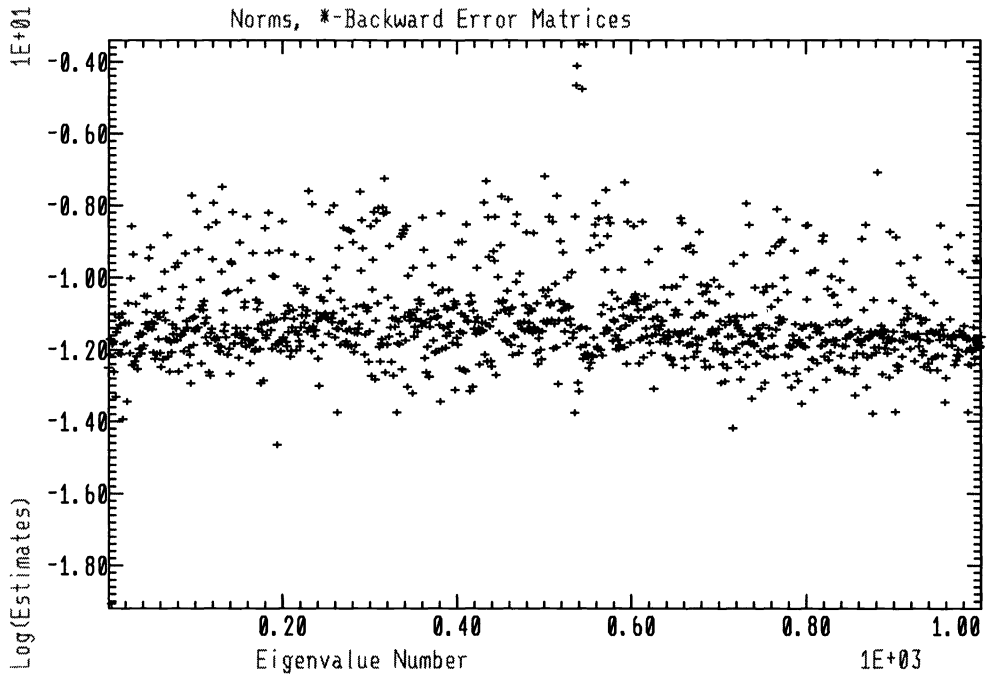


FIG. 10. *Normalized backward error estimates for* $q = *$ *versus eigenvalue number, eigenvalues ordered by decreasing magnitude for* P4 *matrix,* $n = 1000$.

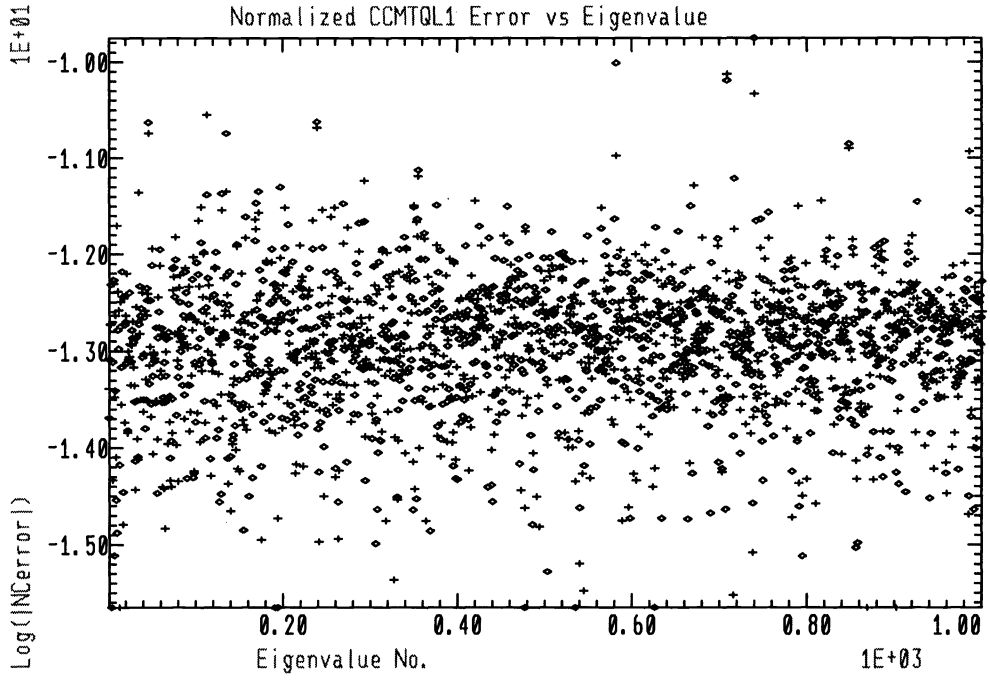


FIG. 11. *Normalized Rayleigh quotient differences for* $q = 4$ *versus eigenvalue number, eigenvalues ordered by decreasing magnitude for* P4 *matrix,* $n = 1000$.

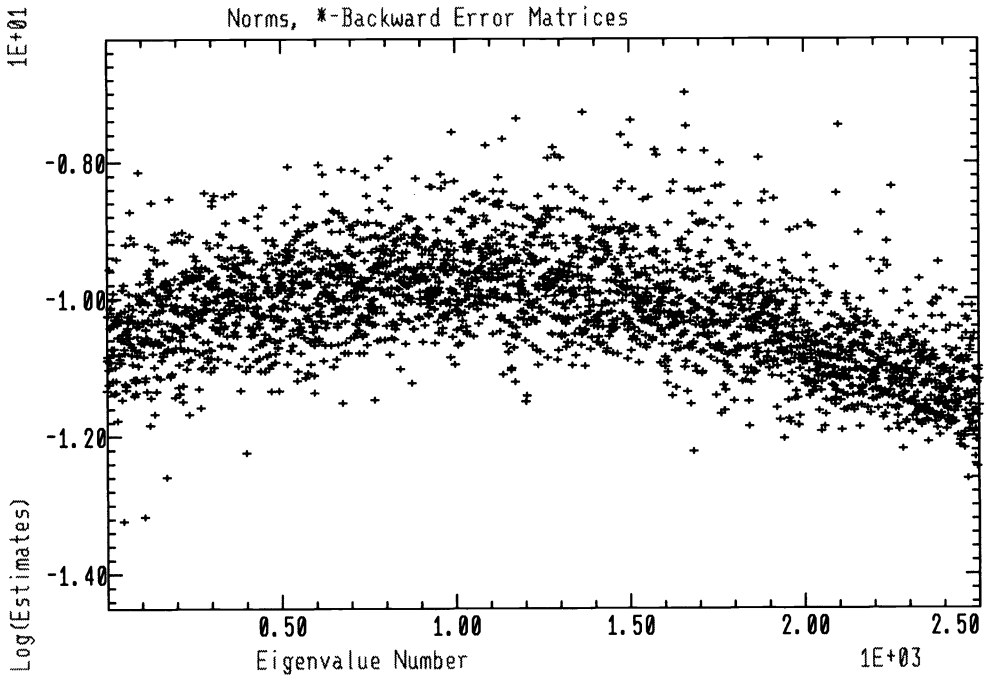


FIG. 12. *Normalized backward error estimates for* $q = 4$ *versus eigenvalue number, eigenvalues ordered by decreasing magnitude for* P2 *matrix,* $n = 2500$.

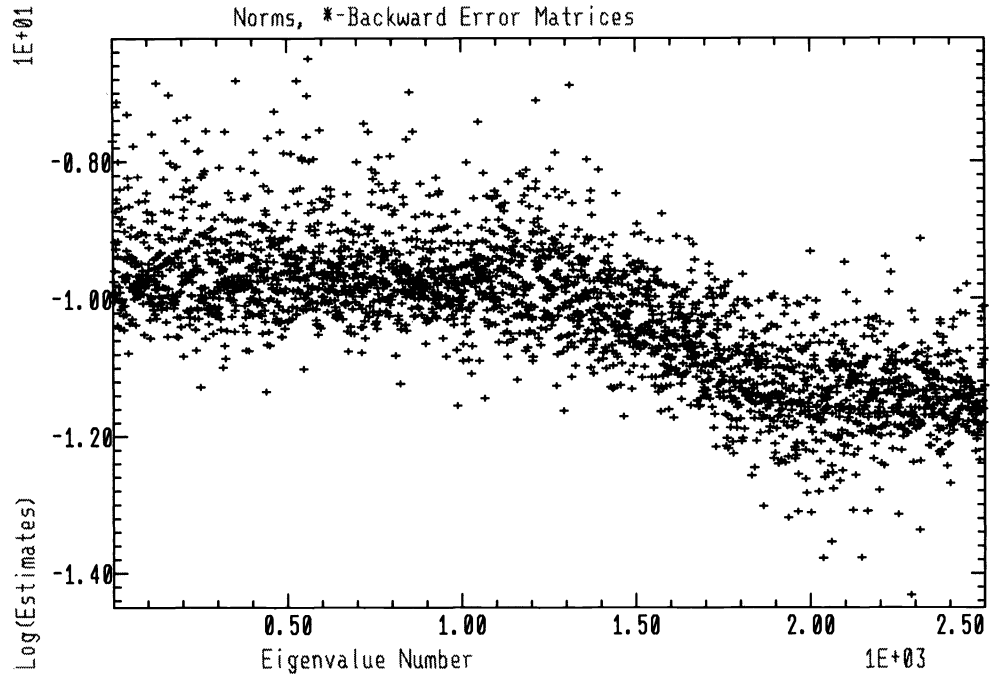


FIG. 13. *Normalized backward error estimates for* $q = *$ *versus eigenvalue number, eigenvalues ordered by decreasing magnitude for* P2 *matrix,* $n = 2500$.

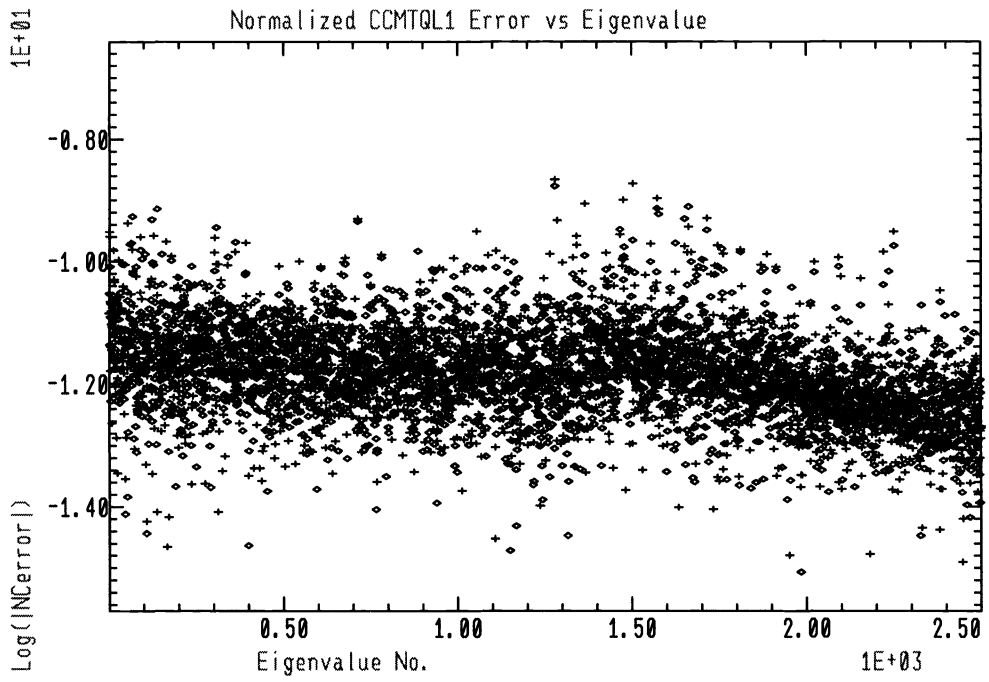


FIG. 14. *Normalized Rayleigh quotient differences for* $q = 4$ *versus eigenvalue number, eigenvalues ordered by decreasing magnitude for* P2 *matrix,* $n = 2500$.

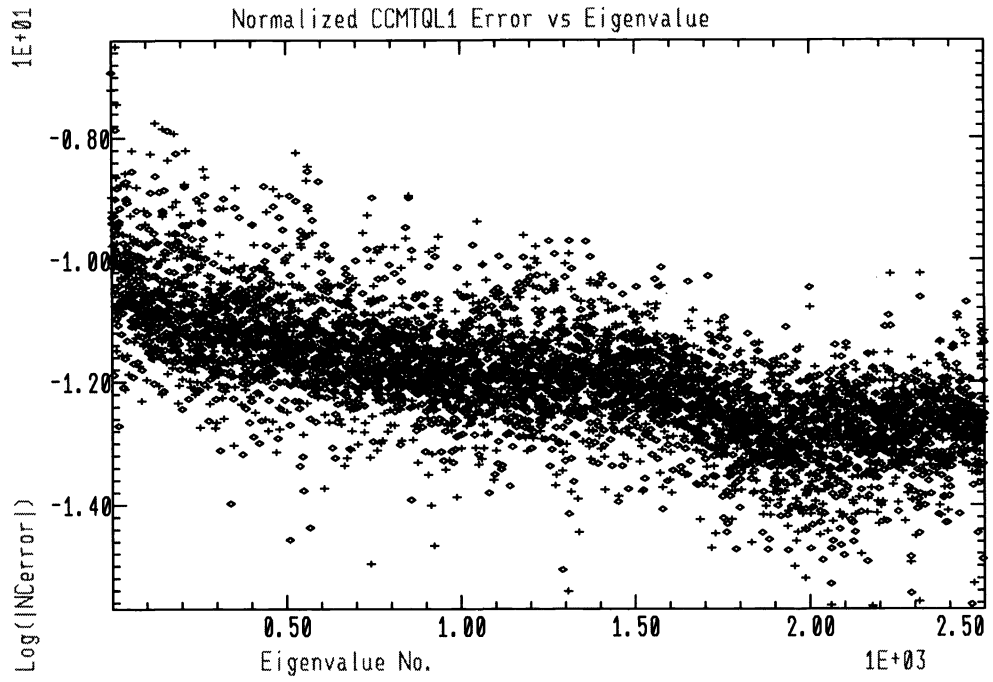


FIG. 15. *Normalized Rayleigh quotient differences for* $q = *$ *versus eigenvalue number, eigenvalues ordered by decreasing magnitude for* P2 *matrix,* $n = 2500$.

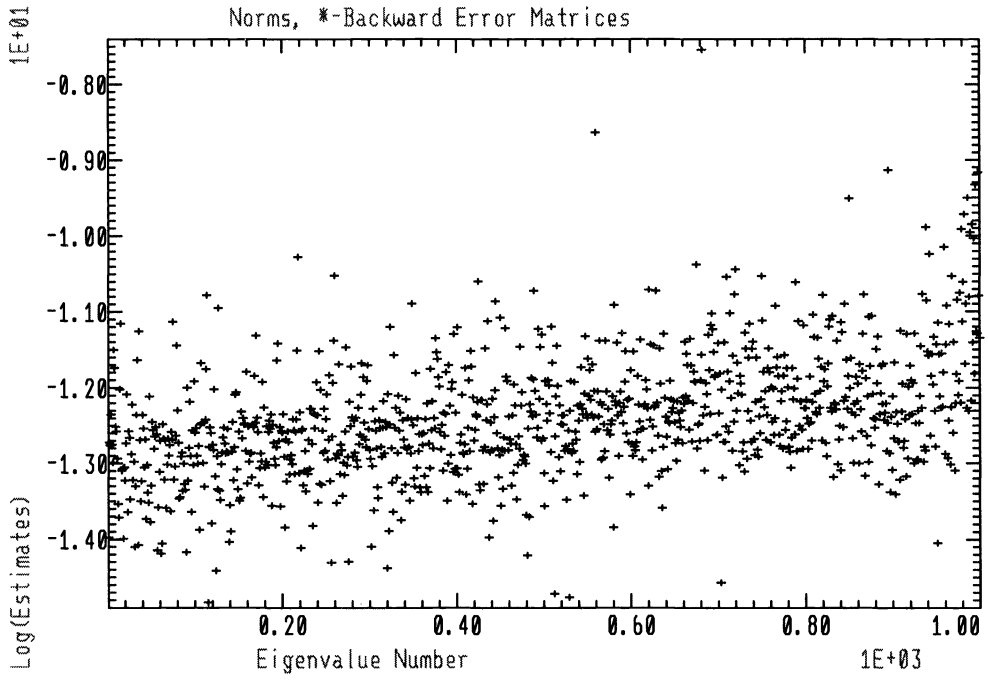


FIG. 16. *Normalized backward error estimates for* $q = 4$ *versus eigenvalue number, eigenvalues ordered by decreasing magnitude for* P7 *matrix,* $n = 1000$.

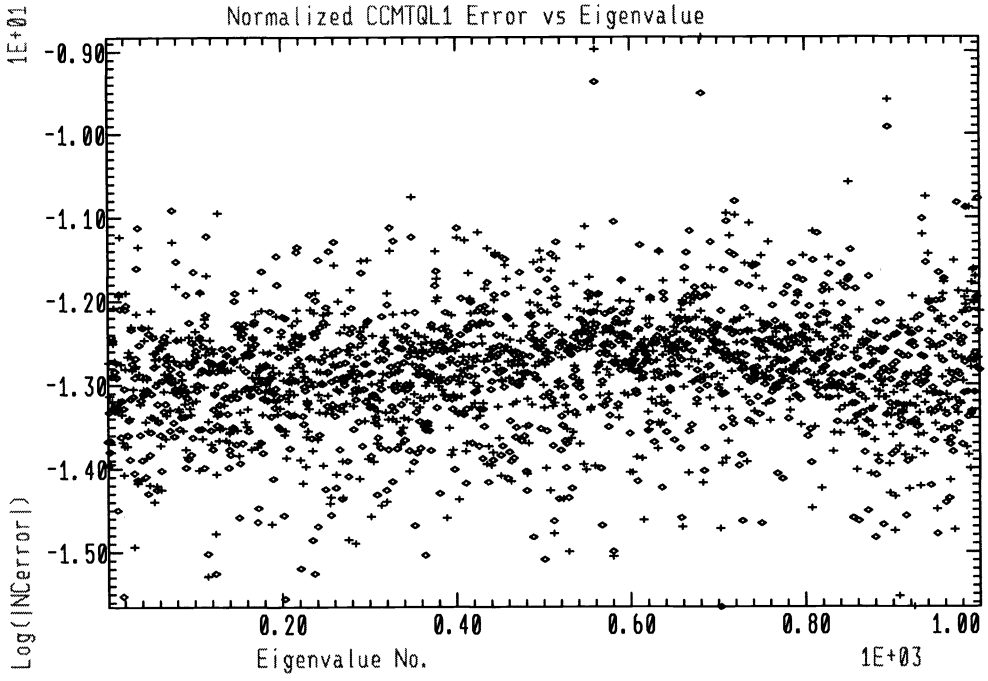


FIG. 17. *Normalized Rayleigh quotient differences for* $q = 4$ *versus eigenvalue number, eigenvalues ordered by decreasing magnitude for* P7 *matrix,* $n = 1000$.

7. Comments. After completion of this work it was pointed out to us that Gordon and Messenger [16] used a complex extension of the real symmetric band QL procedure to complex symmetric band matrices. Details of their implementation indicate that they used explicit shifting and did not incorporate any checks on numerical stability. We should also note that there is related work in [3] where algorithms for real nonsymmetric tridiagonal matrices are presented.

Others have asked why not use a complex L^TL procedure rather than a complex orthogonal QL procedure since there is no a priori guarantee that the Q transformations are well behaved. Lemma 3.6 indicates an equivalence for these decompositions. The numerical experiments indicate that the number of iterations required per eigenvalue by CMTQL1 to achieve convergence is approximately the same as that required by the corresponding real symmetric QL procedure. In that case two steps of the L^TL procedure equal one step of the QL procedure, seemingly indicating that one should not expect significant speedups in convergence by considering such procedures in the complex case. The implicit implementation of the QL procedure through Theorem 5.1 provides a mechanism for determining when these factorizations exist and a mechanism for checking local numerical stability. It is also easy to determine the condition of a complex orthogonal transformation Q since it is determined by $\|Q\|$.

REFERENCES

[1] I. BAR-ON AND V. RYABOY, *Fast diagonalization of large and dense complex symmetric matrices with applications to quantum reaction dynamics*, Tech. Report, Dept. of Chemistry, Technion, Haifa, Israel, 1994.

[2] W.E. BOYSE, D.R. LYNCH, K.D. PAULSEN, AND G.N. MINERBO, *Nodal based finite element modeling of Maxwell's equations*, IEEE Trans. Antennas Propagat., 40 (1992), pp. 642–651.

[3] A. BUNSE-GERSTNER, *Der HR-Algorithms zur numerischen Bestimmung der Eigenwerte einer Matrix*, Dissertation, University Bielefeld, Germany, 1978.

[4] J. CULLUM, *Lanczos algorithms for large scale symmetric and nonsymmetric matrix eigenvalue problems*, Proc. Lanczos International Centenary Conference, Dec. 12–17, 1993, North Carolina State University, Raleigh, Society for Industrial and Applied Mathematics, Philadelphia, 1994, pp. 11–31.

[5] J. CULLUM, W. KERNER, AND R. WILLOUGHBY, *A generalized nonsymmetric Lanczos procedure*, Comput. Phys. Comm., 53 (1989), pp. 19–48.

[6] J. CULLUM AND R. WILLOUGHBY, *Computing eigenvalues of large matrices, some Lanczos procedures and a shift and invert strategy*, in Advances in Numerical Partial Differential Equations and Optimization, S. Gomez, J. P. Hennart, and R. A. Tapia, eds., Society for Industrial and Applied Mathematics, Philadelphia, 1991, pp. 198–246.

[7] ———, *A QL procedure for complex, symmetric, tridiagonal matrices*, IBM Research Report RC 12835, IBM T. J. Watson Research Center, Yorktown Heights, NY, 1987.

[8] ———, EDS., *A practical procedure for computing eigenvalues of large sparse nonsymmetric matrices*, in Large Scale Eigenvalue Problems, North–Holland, Amsterdam, 1986, pp. 193–240.

[9] ———, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations*, Vol. 1, Theory, Progress in Scientific Computing, 3, Birkhaüser Boston, Inc., Boston, MA 1985.

[10] A. DAX AND S. KANIEL, *The* ELR *method for computing the eigenvalues of a general matrix*, SIAM J. Numer. Anal., 18 (1981), pp. 597–605.

[11] I. S. DUFF, R. G. GRIMES, AND J. G. LEWIS, *Sparse matrix test problems*, ACM Trans. Math. Software, 15 (1989), pp. 1–14.

[12] P. J. EBERLEIN, *On the diagonalization of complex symmetric matrices*, J. Instit. Math. Anal., 7 (1971), pp. 377–383.

[13] R. FREUND, *Conjugate gradient type methods for linear systems with complex symmetric coefficient matrices*, SIAM J. Sci. Statist. Comput., 13(1992), pp. 425–448.

[14] R. FREUND AND N. NACHTIGAL, QMR: *A quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 8 (1992), pp. 43–71.

[15] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, 1989.

[16] R. G. GORDON AND T. MESSENGER, *Magnetic resonance line shapes in slowly tumbling molecules*, in Electron Spin Relaxations in Liquids, L. T. Muus and P. W. Atkins, eds., Plenum Press, New York, 1972, pp. 341–375.
[17] W. KERNER, *Large-scale complex eigenvalue problems*, J. Comput. Physics, 85 (1989), pp. 1–85.
[18] N. N. LIPKIN AND C. LEFORESTIER, *A three-dimensional study of NeICl predissociation resonances by the complex-scaled discrete variable representation method*, J. Chem. Phys., 98 (1993), pp. 1888–1901.
[19] G. MORO AND J. H. FREED, *Calculation of* ESR *spectra and related Fokker–Planck forms by the use of the Lanczos algorithm*, J. Chem. Physics, 74 (1981), pp. 3757–3773.
[20] NETLIB, Public Software Library, accessible via netlib at ornl.gov.
[21] D. NEUHAUSER AND M. BAER, *The time-dependent Schroedinger equation: application of absorbing boundary condition*, J. Chem. Phys., 90 (1989), pp. 4351–4355.
[22] B. N. PARLETT, *The Symmetric Eigenproblem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[23] B. N. PARLETT, *Global convergence of the basic QR procedure on Hessenberg matrices*, Math. Comp., 22(1968), pp. 801–817.
[24] B. N. PARLETT, *Canonical decomposition of Hessenberg matrices*, Math. Comp., 21(1967), pp. 223–227.
[25] B. N. PARLETT AND W. G. POOLE, *A geometric theory for the QR, LU and power iterations*, SIAM J. Numer. Anal., 8 (1973), pp. 389–412.
[26] B. N. PARLETT, D. R. TAYLOR, AND Z.A. LIU, *A look-ahead Lanczos procedure for unsymmetric matrices*, Math. Comp., 44 (1985), pp. 105–124.
[27] Y. SAAD, *Numerical Methods for Large Eigenvalue Problems*, Manchester University Press, Manchester, UK, 1992.
[28] B. T. SMITH, J. M. GARBOW, B. S. IKEBE, V. C. KLEMA, AND C. B. MOLER, *Matrix Eigensystem Routines – EISPACK Guide*, Vols. 5, 61, Springer-Verlag, Berlin.
[29] G. W. STEWART, *Introduction to Matrix Computations*, Academic Press, New York, 1973.
[30] D. S. WATKINS, *Understanding the QR algorithm*, SIAM Rev., 24 (1982), pp. 427–440.
[31] D. S. WATKINS AND L. ELSNER, *Chasing algorithms for the eigenvalue problem*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 374–384.
[32] ———, *Convergence of algorithms of decomposition type for the eigenvalue problem*, Linear Algebra Appl., 143 (1991), pp. 19–47.
[33] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 1, pp. 110–126, January 1996

006

TOTAL LEAST NORM FORMULATION AND SOLUTION FOR STRUCTURED PROBLEMS*

J. BEN ROSEN†, HAESUN PARK‡, AND JOHN GLICK§

Abstract. A new formulation and algorithm is described for computing the solution to an overdetermined linear system, $Ax \approx b$, with possible errors in both A and b. This approach preserves any affine structure of A or $[A \mid b]$, such as Toeplitz or sparse structure, and minimizes a measure of error in the discrete L_p norm, where $p = 1, 2$, or ∞. It can be considered as a generalization of total least squares and we call it structured total least norm (STLN).

The STLN problem is formulated, the algorithm for its solution is presented and analyzed, and computational results that illustrate the algorithm convergence and performance on a variety of structured problems are summarized. For each test problem, the solutions obtained by least squares, total least squares, and STLN with $p = 1, 2$, and ∞ were compared. These results confirm that the STLN algorithm is an effective method for solving problems where A or b has a special structure or where errors can occur in only some of the elements of A and b.

Key words. data fitting, Hankel structure, least squares, linear prediction, minimization, overdetermined linear systems, Toeplitz structure, structured total least norm, total least squares, 1-norm, 2-norm, ∞-norm

AMS subject classifications. 15A99, 65F20, 65F30

1. Formulation of structured total least norm (STLN) problems.

An important data fitting technique developed over the past 15 years is that of total least squares (TLS) [7], [8]. The TLS method is a generalization of the least squares method for an overdetermined system of linear equations, $Ax \approx b$, where A is $m \times n$, with $m > n$. In the least squares solution it is assumed that the matrix A is known without error, but that the vector b is subject to error. The vector x is determined so that $\|b - Ax\|_2 = \min$.

TLS allows the possibility of error in the elements of a given (data) matrix A, so that the modified matrix is given by $A + E$, where E is an error matrix to be determined. The TLS problem can then be stated as that of finding E and x, such that

$$\|E \mid r\|_F = \min, \tag{1.1}$$

where $r = b - (A + E)x$, and $\| \cdot \|_F$ represents the Frobenius matrix norm.

A complete description of TLS is given in a recent book [14], where many applications to signal processing, system identification, and system response prediction are described. In many of these applications the matrix A has a special structure,

* Received by the editors November 23, 1993; accepted for publication (in revised form) by G. A. Watson January 27, 1995.

† Computer Science Department, University of Minnesota, Minneapolis, MN 55455, and Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093 (rosen@cs.umn.edu, jbrosen@cs.ucsd.edu). The work of the first and second authors was supported in part by National Science Foundation grant CCR-9509085. The work of the first and the third authors was supported in part by Air Force Office of Scientific Research grant AFOSR-91-0147 and the Minnesota Supercomputer Institute.

‡ Computer Science Department, University of Minnesota, Minneapolis, MN 55455 (hpark@cs.umn.edu). The work of this author was supported in part by National Science Foundation grant CCR-9209726 and also 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 and Computer Science, University of San Diego, San Diego, CA 92110 (glick@teetot.acusd.edu).

such as Toeplitz structure, or is a large, sparse matrix with relatively few nonzero elements. Furthermore, in some applications, errors occur only in a small number of the elements of A, so that while A may be dense, the matrix E could be sparse.

The generally used computational method for solving TLS is based on the singular value decomposition (SVD) of $[A \mid b]$. A complete discussion of efficient computational methods for solving TLS based on SVD is given in Chapter 4 of [14]. For applications where the matrix A has a special structure, the SVD-based methods may not always be appropriate, since they do not preserve the special structure. In fact, using the SVD approach the matrix E will typically be dense, with no special structure, even when A is Toeplitz or sparse. Thus, even those elements of E that should remain zero will typically become nonzero. Also in some situations, the use of a norm other than the Frobenius norm may be preferable. For example, if the data contains outliers, an L_1 norm might be more suitable.

A new approach, to be described, is called STLN, and it addresses these situations. The STLN formulation allows other norms, in addition to the Frobenius norm, to be used. In particular, the problem can be formulated so as to minimize the error in either the L_1 norm or the L_∞ norm, in addition to the Frobenius norm used in TLS. Another important advantage of the STLN formulation is that it permits a known structure of the matrix A and $[A \mid b]$ to be preserved in $A + E$ and $[A + E \mid b + r]$, respectively. Requirements of this kind occur in important applications. For example, a Toeplitz structure occurs in system identification problems [4], [6] and in frequency estimation [1], [11]. For other applications, see [1], [4], [14], [17].

The new approach will guarantee that, for example, E will have the same Toeplitz structure as A, or that only those elements of E which represent possible errors in A are permitted to be nonzero. In general, it can preserve any given affine structure in the computed error matrix E. In some earlier papers [3], [13], restrictions of this type have been imposed by the use of additional constraints on the problem. The use of the L_1 norm for this kind of problem has also been investigated [10], using a different method than the one presented here. This extension of the TLS solution to incorporate the algebraic pattern of the errors in A is also studied in [1] and [4] as "Constrained TLS" and "Structured TLS," respectively, both for the L_2 norm only. In [1], a complex Newton's method is utilized to solve the problem, whereas in [4], nonlinear SVD is defined and an algorithm to compute the solution is derived. For the comparison of these two algorithms, see [15]. The approach in our algorithm is different from these two and we will present the comparison of these three algorithms elsewhere.

Our formulation for solving the STLN problem takes full advantage of the special structure of a given matrix A. In particular, when $q(\leq mn)$ elements of $A \in \Re^{m \times n}$ are subject to error, a vector $\alpha \in \Re^{q \times 1}$ is used to represent the corresponding elements of the error matrix E. Note that for a sparse matrix, $q \ll mn$. Furthermore, if many elements of E must have the same value, then q is the number of *different* such elements. For example, in a Toeplitz matrix, each diagonal consists of elements with the same value, so $q \leq m + n - 1$.

The vector α and the matrix E are equivalent in the sense that given E, α is known, and vice versa. The matrix E is specified by those elements of A which may be subject to error. Each different nonzero element of E corresponds to one of the α_k, $k = 1, \ldots, q$. Also, the residual vector $r = b - (A + E)x$ is now a function of α and x, so $r = r(\alpha, x)$. Let D be a $(q \times q)$ diagonal weighting matrix that accounts for the repetition of elements of α in the matrix E. Then the STLN problem can be

stated as follows:

$$\min_{\alpha,x} \left\| \begin{array}{c} r(\alpha,x) \\ D\alpha \end{array} \right\|_p, \tag{1.2}$$

where $\|\cdot\|_p$ is the vector p-norm, for $p = 1, 2$, or ∞.

For $p = 2$, and a suitable choice for D, the problem (1.2) is equivalent to the TLS problem (1.1), with the additional requirement that the structure of A must be preserved by $A + E$.

2. STLN algorithm. An iterative algorithm for solving the STLN problem will now be described. To do this, we first explain the relationship between $E \in \Re^{m\times n}$ and $\alpha \in \Re^{q\times 1}$. Specifically, the vector Ex must be represented in terms of α. This is accomplished by defining a matrix $X \in \Re^{m\times q}$ such that

$$X\alpha = Ex. \tag{2.1}$$

The elements of X consist of the elements of $x \in \Re^{n\times 1}$, with suitable repetition, giving X a special structure. The number of nonzero elements in both E and X will be equal, so that if E is sparse, X will also be sparse. Furthermore, if the nonzero elements α_k, $k = 1, \ldots, q$, of E are properly ordered, then X will have a similar structure to E: for example, if E is a Toeplitz matrix, then X will also be a Toeplitz matrix. The construction of X from E is described in §3.

The minimization required by (1.2) is done by using a linear approximation to $r(\alpha, x)$. Let Δx represent a small change in x, and ΔE represent a small change in the variable elements of E. From (2.1), we have

$$X\Delta\alpha = (\Delta E)x, \tag{2.2}$$

where $\Delta\alpha$ represents the corresponding small change in the elements of α. Then, neglecting the second-order terms in $\|\Delta\alpha\|$ and $\|\Delta x\|$,

$$\begin{aligned} r(\alpha + \Delta\alpha, x + \Delta x) &= b - (A + E)x - X\Delta\alpha - (A + E)\Delta x \\ &= r(\alpha, x) - X\Delta\alpha - (A + E)\Delta x. \end{aligned} \tag{2.3}$$

The linearization of (1.2) now becomes

$$\min_{\Delta\alpha,\Delta x} \left\| \begin{bmatrix} X & A+E \\ D & 0 \end{bmatrix} \begin{pmatrix} \Delta\alpha \\ \Delta x \end{pmatrix} + \begin{pmatrix} -r \\ D\alpha \end{pmatrix} \right\|_p. \tag{2.4}$$

To start the iterative algorithm, the initial values of $E = 0$ and the least norm value of $x = x_{ln}$ are used, where x_{ln} is given by

$$\min_x \|b - Ax\|_p. \tag{2.5}$$

Note that the initial x for $p = 2$ is the solution to the corresponding least squares problem.

The STLN algorithm is summarized in Algorithm STLN. The computational method by which Step 2(a) is carried out depends on the value of p. For $p = 2$, the corresponding least squares problem is solved efficiently by a QR factorization of the matrix

$$M = \begin{bmatrix} X & A+E \\ D & 0 \end{bmatrix} \tag{2.6}$$

ALGORITHM STLN

Input – A Structured Total Least Norm problem (1.2), with specified matrices A, D, vector b, and tolerance ϵ.

Output – Affine structured error matrix E, vector x, and STLN error.

Begin

1. Set $E = 0$, $\alpha = 0$, compute x from (2.5) and X from x, and set $r = b - Ax$.
2. repeat
   (a) $\underset{\Delta x, \Delta\alpha}{\text{minimize}} \left\| \begin{bmatrix} X & A+E \\ D & 0 \end{bmatrix} \begin{pmatrix} \Delta\alpha \\ \Delta x \end{pmatrix} + \begin{pmatrix} -r \\ D\alpha \end{pmatrix} \right\|_p .$
   (b) Set $x := x + \Delta x$, $\alpha := \alpha + \Delta\alpha$.
   (c) Construct E from α, and X from x. Compute $r = b - (A+E)x$.
   until $(\|\Delta x\|, \|\Delta\alpha\| \leq \epsilon)$

End

when $A+E$ has full column rank, since M has full column rank in this case. In some applications, the right-hand side vector is structured as well. For example, in linear prediction, either $[A \mid b]$ or $[b \mid A]$ follows the Toeplitz or Hankel structure. In §6, we also show how the STLN algorithm can be modified to handle this situation. For $p = 1$, or $p = \infty$, Step 2(a) is solved as a linear program which takes advantage of the special structure of M (see §5). Since the matrix X has a special structure like $A+E$, the matrix M is also highly structured. To make Algorithm STLN efficient, it will be important to take advantage of the structure of M each time Step 2a is solved. For example, it is shown in [12] how a fast triangularization of M can be carried out when A is Toeplitz.

A theoretical justification for the STLN algorithm for $p = 2$ is presented in §4, and its computational performance is illustrated in §7. In the next section the construction of the matrix X, given the structure of E, is described.

3. Construction of matrix X. The matrix E is specified by those elements of A which may be subject to error. Each different nonzero element of E corresponds to one of the α_k, $k = 1, \ldots, q$, where the vector $\alpha = (\alpha_1 \cdots \alpha_q)^T$ represents $q(\leq mn)$ elements of A which are subject to error. The order in which the α_k are numbered will affect the structure of the matrix X, but for any specified ordering, the structure of X is uniquely determined.

The construction of X (starting with a zero matrix) is carried out according to the following rule.

If α_k is the (i,j)th element of E, then x_j is the (i,k)th element of X, where $i = 1, \ldots, m$, $j = 1, \ldots, n$, and $k = 1, \ldots, q$.

For example, when

$$E = \begin{pmatrix} \alpha_2 & \alpha_1 & 0 \\ \alpha_3 & \alpha_2 & \alpha_1 \\ \alpha_4 & \alpha_3 & \alpha_2 \\ 0 & \alpha_4 & \alpha_3 \end{pmatrix}, \quad \text{we have } X = \begin{pmatrix} x_2 & x_1 & 0 & 0 \\ x_3 & x_2 & x_1 & 0 \\ 0 & x_3 & x_2 & x_1 \\ 0 & 0 & x_3 & x_2 \end{pmatrix} \text{ with } \alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix},$$

where only four diagonals of the Toeplitz matrix E are subject to error. For the sparse

matrix

$$E = \begin{pmatrix} 0 & \alpha_1 & \alpha_2 \\ 0 & \alpha_3 & 0 \\ \alpha_4 & 0 & \alpha_3 \\ 0 & 0 & \alpha_5 \\ \alpha_4 & \alpha_2 & 0 \end{pmatrix}, \quad \text{we have } X = \begin{pmatrix} x_2 & x_3 & 0 & 0 & 0 \\ 0 & 0 & x_2 & 0 & 0 \\ 0 & 0 & x_3 & x_1 & 0 \\ 0 & 0 & 0 & 0 & x_3 \\ 0 & x_2 & 0 & x_1 & 0 \end{pmatrix} \quad \text{with } \alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{pmatrix},$$

where only nonzero elements of E are subject to error and the elements denoted with the same α_i are to be perturbed to have the same values.

It is also useful to define $(q \times n)$ matrices P_i, $i = 1, \ldots, m$, as follows.

If α_k is the (i,j)th element of E, then the (k,j)th element of P_i is one. All elements of P_i not equal to one are zero.

Note that at most one element of any column of P_i is a one, and many columns of P_i may consist of all zeros. See §7 for some numerical examples.

With these definitions of the matrices X and P_i, it is easy to show that:

$$E = \begin{bmatrix} \alpha^T P_1 \\ \alpha^T P_2 \\ \vdots \\ \alpha^T P_m \end{bmatrix} \quad \text{and} \quad X = \begin{bmatrix} x^T P_1^T \\ x^T P_2^T \\ \vdots \\ x^T P_m^T \end{bmatrix}. \tag{3.1}$$

The relation (2.1) follows directly from (3.1).

4. STLN optimality conditions and Newton's method. We now consider the two-norm case ($p = 2$) in more detail, since it has special properties that make a more complete theoretical analysis possible. For $p = 2$, the STLN problem (1.2) can be stated in terms of minimizing the differentiable function

$$\begin{aligned} \varphi(\alpha, x) &= \frac{1}{2}\|r(\alpha, x)\|_2^2 + \frac{1}{2}\|D\alpha\|_2^2 \\ &= \frac{1}{2} r^T r + \frac{1}{2}\alpha^T D^2 \alpha, \end{aligned} \tag{4.1}$$

where $r = b - (A + E)x$.

The first-order optimality conditions for a local optimum of $\varphi(\alpha, x)$ are the vanishing of the gradients $\nabla_\alpha \varphi$ and $\nabla_x \varphi$. Using the relations presented in the previous section these conditions become

$$\begin{aligned} \nabla_\alpha \varphi &= -X^T r + D^2 \alpha = 0, \\ \nabla_x \varphi &= -(A + E)^T r = 0. \end{aligned} \tag{4.2}$$

Now consider the least-squares solution of the $(m + q)$ equations in Step 2(a) of Algorithm STLN. The corresponding normal equations are

$$M^T M \begin{pmatrix} \Delta\alpha \\ \Delta x \end{pmatrix} = M^T \begin{pmatrix} r \\ -D\alpha \end{pmatrix} = - \begin{pmatrix} \nabla_\alpha \varphi \\ \nabla_x \varphi \end{pmatrix}, \tag{4.3}$$

where the last equality follows directly from (4.2).

When the matrix M has full rank, the matrix $M^T M$ is positive definite, and (4.3) always has a unique solution for $(\Delta\alpha^T, \Delta x^T)$. This vector will be zero if, and only if, the right-hand side of (4.3) vanishes. This means that convergence of Algorithm

STLN (i.e., $(\Delta\alpha^T, \Delta x^T) \approx 0$) is equivalent to satisfying the optimality conditions (4.2).

We now show that Step 2(a) of Algorithm STLN is essentially Newton's method applied to the gradient of $\varphi(\alpha, x)$. To simplify notation, let $y^T = (\alpha^T, x^T)$, $\varphi(y) = \varphi(\alpha, x)$, and

$$\nabla\varphi(y) = \begin{pmatrix} \nabla_\alpha\varphi(\alpha, x) \\ \nabla_x\varphi(\alpha, x) \end{pmatrix}.$$

Let $H(y)$ be the Hessian of $\varphi(y)$. We wish to find y^*, such that $\nabla\varphi(y^*) = 0$.

An iteration of Newton's method to do this is given by

$$\begin{aligned} &H(y)\Delta y = -\nabla\varphi(y), \\ &y := y + \Delta y. \end{aligned} \tag{4.4}$$

For $H(y)$ positive definite, and an initial y sufficiently close to $y*$, this Newton's method will converge to y^* at a second-order rate. See, for example, Theorem 3.1.1 in [5].

To show the relationship between Step 2(a) of Algorithm STLN and (4.4) we note that the right-hand sides of (4.3) and (4.4) are identical. Furthermore, it can be shown, using the relations in §3, that

$$H(y) = M^T M - \begin{pmatrix} 0 & P(r) \\ P^T(r) & 0 \end{pmatrix}, \tag{4.5}$$

where

$$P(r) = \sum_{i=1}^{m} r_i P_i$$

is a matrix with norm $0(\|r\|)$.

Thus Step 2(a) is, in effect, a Gauss–Newton method that uses $M^T M$ as a positive definite approximation to $H(y)$ (see, for example, §6.1 in [5]). Computational experience with the STLN algorithm (§7) demonstrates that this is an effective strategy for this type of problem.

The differentiable function $\varphi(\alpha, x)$ we wish to minimize for $p = 2$ is not a convex function of (α, x). Therefore, there is no guarantee that a point satisfying the first-order optimality conditions (4.2) is a global minimum. In fact, it could be any stationary point of $\varphi(\alpha, x)$. In general, the Gauss–Newton method will converge to the closest local minimum when the residual $r(\alpha, x)$ is sufficiently small [5]. When there is no structure imposed on E, the STLN formulation (1.2) is equivalent to the TLS problem (1.1). In our preliminary test results, we have observed that the solution produced by the STLN was always the same as that from the TLS, when no structure is imposed on E. However, there is no theoretical guarantee that Algorithm STLN will produce the global minimum solution that the TLS via the SVD produces. We do not propose Algorithm STLN for solving unstructured problems since the computational complexity will be high due to the large number of elements in α, which will be mn.

In many applications, the minimum residual is zero or very small when we have the exact data, and the global minimum value of α will be of the order of the noise in the matrix A. Therefore, the initial value $\alpha = 0$ is close to the global minimum value

and convergence to the global minimum can often be expected. Our computational results confirm this expectation (see §7).

The function being minimized by the STLN (given by (1.2)) for $p = 1, \infty$, is not differentiable, so the Gauss–Newton theory does not apply. Theoretical results on the convergence of the STLN algorithm for $p = 1, \infty$, are not known at this point, but are the subject of our continuing research.

5. STLN for $p = 1$ and $p = \infty$. For $p = 1$ or ∞, Step 2(a) is solved as a linear program (LP). To illustrate this, the linear program for $p = \infty$ is now summarized. The formulation for $p = 1$ is similar.

A scalar σ representing the maximum norm is introduced, and the corresponding linear program is then given by

$$\begin{array}{ll} \underset{\Delta\alpha,\Delta x,\sigma}{\text{minimize}} & \sigma \\ \text{subject to} & -\sigma e_m \le X\Delta\alpha + (A+E)\Delta x - r \le \sigma e_m, \\ & -\sigma e_q \le D\Delta\alpha + D\alpha \le \sigma e_q, \end{array} \tag{5.1}$$

where $e_k \in \Re^{k\times 1}$ is the vector with every element equal to one. Note that a feasible solution to this problem is easily given ($\Delta\alpha = 0$, $\Delta x = 0$, σ sufficiently large), and since $\sigma \ge 0$, an optimal solution always exists. In this form the problem has more inequality constraints, $2(m+q)$, than variables, $n+q+1$, so it is more efficient to consider (5.1) as the dual problem, and solve the equivalent primal.

The equivalent primal is

$$\begin{array}{ll} \underset{y^{(i)}\ge 0}{\text{minimize}} & r^T y^{(1)} + \alpha^T D y^{(2)} - r^T y^{(3)} - \alpha^T D y^{(4)} \\ \text{subject to} & \begin{bmatrix} M^T & -M^T \\ e_{m+q}^T & e_{m+q}^T \end{bmatrix} \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \\ y^{(4)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}. \end{array} \tag{5.2}$$

The optimal solution and basis to (5.2) immediately gives the optimal dual vector $(\Delta\alpha, \Delta x, \sigma)$. Any available LP package (Simplex or Interior) can be used to solve (5.2); however, it should be possible to use the special structure of M to solve (5.2) more efficiently. Another aspect of (5.2) that can be used to advantage is that only relatively small changes occur in the cost vector coefficients and the matrix M at each iteration of the STLN algorithm. Therefore the previous basis will often be optimal, or almost optimal, after the initial LP solution. An additional benefit obtained from the primal-dual relationship is information about the sensitivity of the STLN solution to changes in the data. Properly interpreted, the primal variables (elements of the vectors $y^{(i)}$) are measures of the change in the value of the minimum norm as a result of changes in the problem data.

It should also be noted that by the addition of one row and two columns to (5.2) a specified bound δ can be imposed, so that

$$\|D\alpha\|_\infty \le \delta. \tag{5.3}$$

With this addition, the original STLN problem (1.2), with $p = \infty$, is modified so that α is limited to values satisfying (5.3). This restriction may be important in some applications; for example, if it is known that the errors in A cannot exceed some

specified bounds. For the limiting case of $\delta = 0$, the solution to (1.2) and (5.3) will give the least norm solution (2.5) for $p = \infty$, while for sufficiently large δ it will give the STLN∞ solution to (1.2). For small values of δ, the solutions to (1.2) and (5.3) may differ significantly from the STLN∞ solution. Similar bounds can readily be imposed in the L_1 norm case.

6. STLN for structured vector *b*. In many applications, the structure is imposed not only on the data matrix A but also on the right-hand side vector b or even on $[A \mid b]$. For example, in the least squares linear prediction problem, we need to solve

$$\min_x \|Ax - b\|_2, \tag{6.1}$$

where $A \in \Re^{m\times n}$ is a Toeplitz matrix with $m \geq n$ and the right-hand side vector b follows the pattern of A, so that either $[A|b]$ is Toeplitz in backward prediction or $[b|A]$ is Toeplitz in forward prediction. For details on the least squares linear prediction problem, see [9]. In this section, we show how to modify Algorithm STLN so that it can treat possible errors in some (or all) elements of b in the same manner as errors in A are treated. We will discuss the Toeplitz structure in detail since it appears in numerous applications in signal processing, image processing, and system identification [1], [4], [9], [15]. The results presented in this section on Toeplitz structure apply to Hankel structure in a straightforward manner, since Hankel structure can be transformed to Toeplitz structure via permutations.

We introduce a vector β representing possible errors in selected elements of b. This is similar to α representing errors in A. Suppose different errors can occur in q_2 ($\leq m$) elements of b, specifically, in the elements b_{i_j}, $j = 1, \ldots, q_2$. The error vector $\beta \in \Re^{q_2\times 1}$ represents the error in b_{i_j}, $j = 1, \ldots, q_2$. The relation between β and b is given by a matrix $P_0 \in \Re^{m\times q_2}$, so that error in b is the same as $P_0\beta$. The matrix P_0 consists of only zeros and ones: the element P_{ij} of P_0 is one if β_j is the error in b_i; otherwise, it is zero. Note that every column of P_0 contains exactly one nonzero element.

Initially, E, α and β are all zero, and the new residual $\hat{r} = r = b - Ax$. In general,

$$\hat{r} = \hat{r}(\alpha, \beta, x) = (b + P_0\beta) - (A + E)x = b - (Ax + X\alpha) + P_0\beta = r + P_0\beta.$$

In ideal situations, we can impose the requirement that $\hat{r} = 0$ since $P_0\beta$ can play the role of the residual vector $r = \hat{r} - P_0\beta = b - (A+E)x$. However, this may not always be possible, due to the special structure that is imposed on E and $P_0\beta$.

When the structures imposed on E and $P_0\beta$ are such that $\hat{r}$ can be zero, the STLN solution that preserves the structure in b can be stated as

$$\min_{\hat{r}=0,\alpha,\beta,x} \left\| \begin{pmatrix} \hat{r}(\alpha, \beta, x) \\ D_1\alpha \\ D_2\beta \end{pmatrix} \right\|_p \tag{6.2}$$

for some diagonal matrices D_1 and D_2. This constrained minimization problem can be restated in different ways. We use the weighting method for the equality constrained least squares problems that transforms (6.2) into an unconstrained problem

$$\min_{\alpha,\beta,x} \left\| \begin{pmatrix} \omega\hat{r}(\alpha, \beta, x) \\ D_1\alpha \\ D_2\beta \end{pmatrix} \right\|_p, \tag{6.3}$$

ALGORITHM STLNB

Input – A Structured Total Least Norm problem (1.2), with specified matrices A, D, P_0, vector b, and tolerance ϵ.

Output – Error matrix E and error vector β with the given affine structure in $[E \mid P_0\beta]$, vector x, and STLN error.

Begin

1. Choose a large number ω.
   Set $E = 0$, $\alpha = 0$, $\beta = 0$, compute x from (2.5) and X from x, and set $\hat{r} = b - Ax$.
2. repeat
   (a) $\underset{\Delta x, \Delta\alpha, \Delta\beta}{\text{minimize}} \left\| \begin{bmatrix} \omega X & -\omega P_0 & \omega(A+E) \\ D_1 & 0 & 0 \\ 0 & D_2 & 0 \end{bmatrix} \begin{pmatrix} \Delta\alpha \\ \Delta\beta \\ \Delta x \end{pmatrix} + \begin{pmatrix} -\omega\hat{r} \\ D_1\alpha \\ D_2\beta \end{pmatrix} \right\|_p .$
   (b) Set $x := x + \Delta x$, $\alpha := \alpha + \Delta\alpha$, $\beta := \beta + \Delta\beta$.
   (c) Construct E from α, and X from x. Compute $\hat{r} = (b + P_0\beta) - (A+E)x$.
   until $(\|\Delta x\|, \|\Delta\alpha\|, \|\Delta\beta\| \leq \epsilon)$

End

where ω is a large number [2], [16]. It can be shown that for $p = 2$ when ω approaches infinity, the solution for (6.3) converges to the solution for (6.2) when the constraint $\hat{r} = 0$ can be satisfied [2], [8], [16]. Thus, by using a large weight ω, we can obtain a good approximation for the solution for (6.2) by solving the unconstrained problem (6.3). For possible numerical problems associated with large ω, see [2], [16].

The algorithm is summarized in Algorithm STLNB. When there is no structure imposed on b, we can simply choose $\beta \in \Re^{m\times 1}$ to represent the perturbation on all the elements of b, and $P_0 = I$ accordingly. Therefore, Algorithm STLNB can handle the problems that can be solved by Algorithm STLN, although Algorithm STLN will be more efficient when there is no structure on b.

For the linear prediction, where $[A \mid b]$ or $[b \mid A]$ is Toeplitz and all the diagonals are subject to error, it is always possible to find a Toeplitz perturbation $[E \mid P_0\beta]$ such that

$$b + P_0\beta \in \text{Range}(A + E).$$

Therefore, $\hat{r}$ will become zero when the solution is obtained. Also, we can reformulate (6.2) into (6.3).

We will discuss the backward prediction only since the same results hold with forward prediction as well. When we need to impose Toeplitz structure on $[A\,|\,b]$, Step 2(a) of Algorithm STLNB can be further simplified since perturbation in b can be represented using the perturbation in A, except for its first component. Specifically, if E is a Toeplitz matrix with its first column $[\alpha_n \cdots \alpha_{n+m-1}]^T$ and its first row $[\alpha_n \cdots \alpha_2\ \alpha_1]$, i.e.,

$$E = \text{Toeplitz}([\alpha_n \cdots \alpha_{n+m-1}]^T, [\alpha_n \cdots \alpha_2\ \alpha_1]) \text{ with } \alpha = [\alpha_1\ \alpha_2 \cdots \alpha_n \cdots \alpha_{n+m-1}]^T,$$
$$P_0 = I \qquad \text{with} \qquad \beta = [\beta_1\ \beta_2 \cdots \beta_m]^T,$$

then since $\beta_i = \alpha_{i-1}$, $i = 2, \ldots, m$, we have

$$\beta = \beta_1\hat{e} + \hat{P}_0\alpha, \tag{6.4}$$

where

$$\hat{P}_0 = \begin{pmatrix} 0_{1\times(m-1)} & 0_{1\times n} \\ I_{(m-1)\times(m-1)} & 0_{(m-1)\times n} \end{pmatrix} \in \Re^{m\times(m+n-1)} \text{and } \hat{e} = (1 \quad 0 \quad \cdots \quad 0)^T \in \Re^{m\times 1}.$$

From (6.4) and

$$X\Delta\alpha - P_0\Delta\beta + (A+E)\Delta x = (X-P_0)\Delta\alpha + (A+E)\Delta x - \Delta\beta_1\hat{e},$$

Step 2(a) of Algorithm STLNB is simplified to

(6.5)

$$\underset{\Delta\alpha,\Delta\beta_1,\Delta x}{\text{minimize}} \left\| \begin{pmatrix} \omega(X-\hat{P}_0) & -\omega\hat{e} & \omega(A+E) \\ D & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \Delta\alpha \\ \Delta\beta_1 \\ \Delta x \end{pmatrix} + \begin{pmatrix} -\omega\hat{r}(\alpha,\beta_1,x) \\ D\alpha \\ \beta_1 \end{pmatrix} \right\|_p,$$

where $D^2 = \text{diag}(2,3,\ldots,n+1,n+1,\ldots,3,2,1) \in \Re^{(m+n-1)\times(m+n-1)}$, when all $m+n-1$ diagonals of A are different and subject to error. Then in Step 2(b), β is modified so that $\beta_i = \alpha_{i-1}$ for $i = 2,\ldots,m$ and $\beta_1 := \beta_1 + \Delta\beta_1$.

7. Computational results. The STLN algorithm has been implemented in MATLAB in order to investigate its computational performance. Each program for a different norm is denoted with the suffix p as STLNp with $p = 1, 2, \infty$. The computational testing has included over 200 relatively small problems with $m \leq 25$ and $n \leq 21$. In all cases, A has full rank. These computational tests represent a preliminary study of the effect of structure, initial choice of x, and the magnitude of the minimum norm on the algorithm's behavior.

7.1. Convergence of STLN algorithm. The numerical results obtained were consistent in showing that for each problem the STLN algorithm converged rapidly to a minimum solution for the problem. Since the function being minimized (as given by (1.2)) is not convex, there is no guarantee of convergence to a global minimum (for $p = 2$, convergence to a local minimum is discussed in §4). Typically, the STLN algorithm starts with x as given by (2.5). Other initial values of x were also used in some cases in order to test the convergence. In every such case, the algorithm converged to a minimum value that was independent of the initial x. As a further confirmation, the Hessian matrix $H(y)$, as given by (4.5), was computed when the algorithm terminated, and was always found to be positive definite.

The convergence rate appears to be independent of problem size (over the range studied), but does depend on the size of the minimum norm. Specifically, a smaller minimum norm results in faster convergence. To illustrate the convergence, the results for two different problems will be summarized. These will be denoted as Problem I and Problem II which are defined as follows.

Problem I.

$m = 6$, $n = 4$, $q = 4$.

Matrix A = Toeplitz(col, row) :

col = [-3 7 10 -1 0 0]^T, row = [-3 0 0 0],

Matrix E = Toeplitz(col, row) :

col = [α_1 α_2 α_3 α_4 0 0]^T, row = [α_1 0 0 0],

Two values of b were used:

$b^{(1)}$ = [-12 25 62 -59 16 100]^T,

$b^{(2)}$ = [-12 25 62 -59 9 122]^T.

TABLE 7.1
Minimum norms and x for Problem I.

	$b^{(1)}$			
	LS	TLS	STLN2	STLN∞
rnorm	8.23×10^{-1}	6.14×10^{-3}	2.20×10^{-2}	0
Enorm	0	7.08×10^{-2}	1.09×10^{-1}	7.24×10^{-2}
Tnorm	8.23×10^{-1}	7.11×10^{-2}	1.11×10^{-1}	7.24×10^{-2}
x_1	4.0292	4.0292	3.9638	3.9652
x_2	0.9056	0.9058	1.0090	1.0058
x_3	-5.0122	-5.0126	-5.1025	-5.1289
x_4	9.5310	9.5314	9.5596	9.5937

	$b^{(2)}$			
	LS	TLS	STLN2	STLN∞
rnorm	10.445	5.72×10^{-2}	5.359×10^{-1}	0
Enorm	0	7.72×10^{-1}	1.432	1.136
Tnorm	10.445	7.74×10^{-1}	1.529	1.136
x_1	3.4739	3.4650	4.3948	4.2865
x_2	1.7889	1.8252	0.2927	0.0489
x_3	-6.3357	-6.3864	-5.0594	-4.994
x_4	11.157	11.221	10.924	11.017

Problem II.

$m = 9$, $n = 6$, $q = 4$.

Matrix A = Toeplitz(col, row):

$\text{col} = [\ 3\ \ -1\ \ -6\ \ 2\ \ 5\ \ 0\ \ -8\ \ -7\ \ 1\ ]^T$,

$\text{row} = [\ 3\ \ -6\ \ 2\ \ 0\ \ 8\ \ -4\ ]$,

Matrix E = Toeplitz(col, row):

$\text{col} = [\ \alpha_3\ \ \alpha_4\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ ]^T$,

$\text{row} = [\ \alpha_3\ \ \alpha_2\ \ \alpha_1\ \ 0\ \ 0\ \ 0\ ]$,

$b = [\ 62\ \ 62\ \ 5\ \ 22\ \ -65\ \ -60\ \ -10\ \ 86\ \ 101\ ]^T$.

Note that in Problem I the matrix E has the same nonzero diagonal patterns as A, whereas in Problem II, E has only four nonzero diagonals. This means that in Problem II, only those four diagonals of the Toeplitz matrix $A + E$ can change. Also as shown in Table 7.1, the vector $b^{(1)}$ is more closely approximated by the columns of A than is the case for $b^{(2)}$.

To illustrate the structure of the matrices $P_i, i = 1, \ldots, m$, the matrices P_1 and P_2 for Problem II are now given:

$$P_1 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad P_2 = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The following solutions were obtained for these problems:

- Least squares (LS),
- Total least squares (TLS),
- Structured total least norm (STLN), for $p = 2$ and $p = 1, \infty$.

The LS and TLS solutions were obtained with MATLAB using the QR decomposition and the SVD, respectively. The STLN solutions were obtained using Algorithm STLN given in §2.

For $p = 2$, Step 2(a) of the algorithm computes an LS solution. For $p = \infty$, Step 2(a) is essentially the solution of the LP (5.2) in §5. The computed results for Problem I, using these different approaches, are summarized in Figs. 7.1 and 7.2 and Table 7.1.

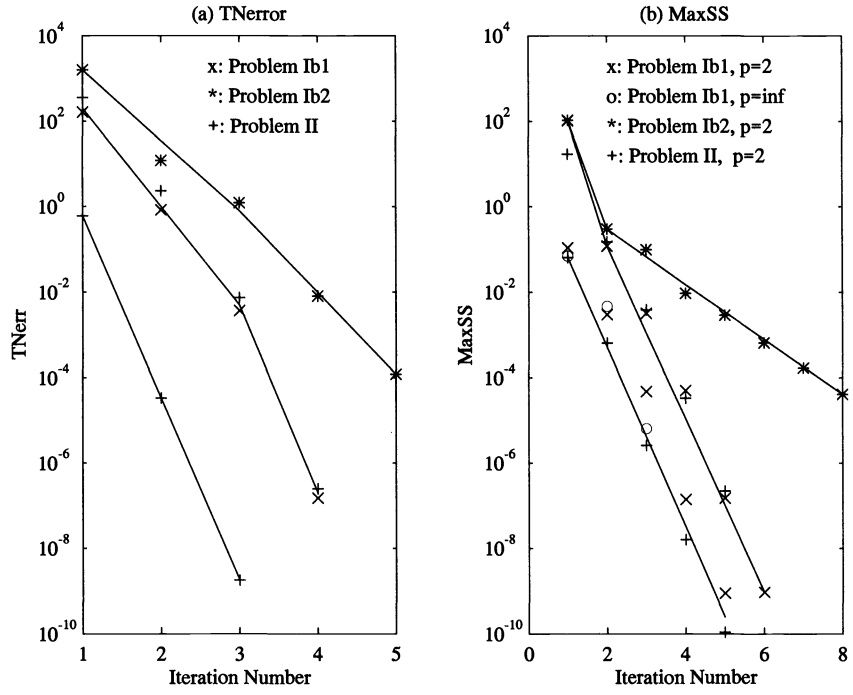


FIG. 7.1. (a) *Convergence of total norm to* STLN *with* $p = 2$. (b) *Convergence of step size.*

Figure 7.1(a) shows the convergence of the STLN algorithm to the minimum value of the total norm (TN), as given by (1.2). This minimum value computed from the STLN algorithm is the STLN. Figure 7.1(a) shows the value of the TNerr at each iteration of the STLN algorithm, where

$$\text{TNerr} = \text{TN} - \text{STLN}. \tag{7.1}$$

The convergence (using $p = 2$) is shown for four different cases: Problem Ib_1 (Problem I, with $b = b^{(1)}$), Problem Ib_2 (Problem I, with $b = b^{(2)}$), and Problem II with two different initial values of x. The initial value $\alpha = 0$ was used for all cases. These results show convergence in three iterations to TNerr $\leq 5 \times 10^{-7}$, for both Problems Ib_1 and II, even when the initial TNerr is very large (~ 100). The large initial value of TNerr is obtained by using an initial value of x very different from its converged value. The smaller initial TNerr shown for Problem II was obtained by using $x = x_{ls}$ as the initial value, where x_{ls} is the LS solution.

Convergence for Problem Ib_2 is seen to be significantly slower, with four iterations needed to obtain TNerr $\cong 10^{-4}$, from an initial TNerr $\cong 10^3$. Essentially the same convergence rate was obtained for Problem Ib_2, starting with an initial TNerr $\cong 10$, obtained with $x = x_{ls}$. To avoid complicating the figure, this last case is not included in Fig. 7.1(a).

In order to understand these convergence results, it is important to note that both the STLN and the minimum residual norm are at least ten times greater for Problem Ib_2 than they are for Problem Ib_1. These values are given in Table 7.1, to be discussed shortly.

A closely related aspect of the convergence of the STLN algorithm is given in Fig. 7.1(b), which shows the rate of decrease of the step size with iteration number. For the purposes of this graph the maximum step size (MaxSS) is defined as

$$\text{MaxSS} = \max\{\|\Delta\alpha\|_\infty, \|\Delta x\|_\infty\}, \tag{7.2}$$

and is shown as a function of the iteration number. This is a more sensitive measure of convergence, since small changes in α and x may continue even when TNerr is very small. This is most likely to occur when the minimum of TN is very flat.

The results for a total of six cases are presented in Fig. 7.1(b). These results include the four cases shown in Fig. 7.1(a) and, in addition, Problem Ib_1, with the initial value of $x = x_{ls}$, and Problem Ib_1, using the L_∞ norm with an initial value of x given by (2.5) with $p = \infty$. The results shown in Fig. 7.1 are typical of all the problems solved by the STLN algorithm. For all the problems we tested, the STLN algorithm converges to the global minimum from the chosen initial value of x, and the convergence rate is independent of the norm used (note that the data for Problems Ib_1 ($p = 2$ and $p = \infty$) and II all lie on the lowest curve of Fig. 7.1(b)). The convergence rate is second order for small residual problems, and apparently superlinear for larger residual problems (compare Problems Ib_1 and Ib_2). The minimum norm for all problems tested (except Ib_2) was similar to that in Ib_1 and II, and all these problems converged in, at most, six iterations.

These computational results are consistent with the analysis given in §4. The dependence of the convergence rate on the residual minimum norm is clearly shown by reference to Table 7.1, to be discussed below. A more complete understanding of this dependence requires further investigation, both theoretical and computational.

7.2. Comparison of STLN with LS and TLS. A direct comparison of the STLN ($p = 2$) solution with the TLS solution, for Problem II, is shown in Fig. 7.2. For this problem the matrix E is Toeplitz, with only four nonzero diagonals. The computed matrix E is shown for the TLS solution and the STLN solution. As expected for the TLS solution, all elements of E are nonzero, and it does not have a Toeplitz structure. That is, the TLS solution allows all elements of the matrix A to change. This is in contrast to the STLN solution where only the four designated diagonals are allowed to change, so that $A+E$ preserves the original Toeplitz structure of A.

Finally, the computed norms for Problems Ib_1, Ib_2, and the corresponding x vectors are given in Table 7.1. This table compares the minimum norm solutions obtained by LS, where $E = 0$; by TLS, where all elements of E can change; and by the STLN algorithm (for both $p = 2$ and $p = \infty$), where only the specified elements of E can change. For each case, the following three norms are tabulated:

$$\begin{aligned} \text{rnorm} &= \|r\|_p, \\ \text{Enorm} &= \|D\alpha\|_p, \quad \text{and} \\ \text{Tnorm} &= \left\| \begin{pmatrix} r \\ D\alpha \end{pmatrix} \right\|_p. \end{aligned} \tag{7.3}$$

It should be noted that for each problem the Tnorm satisfies the inequality

$$\text{TLS} \leq \text{STLN2} \leq \text{LS}. \tag{7.4}$$

$$E_{TLS} = \begin{pmatrix} -4.5 \times 10^{-3} & -7.7 \times 10^{-3} & 4.0 \times 10^{-6} & 7.9 \times 10^{-3} & 9.1 \times 10^{-3} & 9.1 \times 10^{-3} \\ -1.4 \times 10^{-3} & -2.4 \times 10^{-3} & 1.2 \times 10^{-6} & 2.4 \times 10^{-3} & 2.8 \times 10^{-3} & 2.8 \times 10^{-3} \\ 3.7 \times 10^{-3} & 6.4 \times 10^{-3} & -3.3 \times 10^{-6} & -6.5 \times 10^{-3} & -7.5 \times 10^{-3} & -7.5 \times 10^{-3} \\ -1.5 \times 10^{-4} & -2.6 \times 10^{-4} & 1.4 \times 10^{-7} & 2.7 \times 10^{-4} & 3.1 \times 10^{-4} & 3.1 \times 10^{-4} \\ -2.2 \times 10^{-3} & -3.8 \times 10^{-3} & 2.0 \times 10^{-6} & 3.9 \times 10^{-3} & 4.5 \times 10^{-3} & 4.5 \times 10^{-3} \\ -5.5 \times 10^{-3} & -9.5 \times 10^{-3} & 4.9 \times 10^{-6} & 9.6 \times 10^{-3} & 1.2 \times 10^{-2} & 1.1 \times 10^{-2} \\ -6.8 \times 10^{-3} & -1.2 \times 10^{-2} & 6.0 \times 10^{-6} & 1.2 \times 10^{-2} & 1.4 \times 10^{-2} & 1.4 \times 10^{-2} \\ 8.0 \times 10^{-4} & 1.4 \times 10^{-3} & -7.1 \times 10^{-7} & -1.4 \times 10^{-3} & -1.6 \times 10^{-3} & -1.6 \times 10^{-3} \\ -2.6 \times 10^{-3} & -4.5 \times 10^{-3} & 2.3 \times 10^{-6} & 4.5 \times 10^{-3} & 5.2 \times 10^{-3} & 5.2 \times 10^{-3} \end{pmatrix}$$

$$E_{STLN2} = \begin{pmatrix} -2.6 \times 10^{-2} & -1.6 \times 10^{-3} & 2.0 \times 10^{-2} & 0 & 0 & 0 \\ 6.5 \times 10^{-2} & -2.6 \times 10^{-2} & -1.6 \times 10^{-3} & 2.0 \times 10^{-2} & 0 & 0 \\ 0 & 6.5 \times 10^{-2} & -2.6 \times 10^{-2} & -1.6 \times 10^{-3} & 2.0 \times 10^{-2} & 0 \\ 0 & 0 & 6.5 \times 10^{-2} & -2.6 \times 10^{-2} & -1.6 \times 10^{-3} & 2.0 \times 10^{-2} \\ 0 & 0 & 0 & 6.5 \times 10^{-2} & -2.6 \times 10^{-2} & -1.6 \times 10^{-3} \\ 0 & 0 & 0 & 0 & 6.5 \times 10^{-2} & -2.6 \times 10^{-2} \\ 0 & 0 & 0 & 0 & 0 & 6.5 \times 10^{-2} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

FIG. 7.2. *E matrices obtained by* TLS *and* STLN2 *for Problem* II.

This is the expected result, since TLS is unconstrained (all elements of E can change), STLN is partially constrained (only specified elements of E can change), and LS is completely constrained ($E = 0$). Also, the minimum residual for Problem Ib_2 is at least ten times greater than it is for Problems Ib_1 and II, which seems to be the significant property affecting the convergence rate.

In addition to the computational convergence of the STLN algorithm, the properties of the solution obtained were investigated. In particular, the vector x and error matrix E obtained were compared for LS, TLS, and STLNp, for $p = 1, 2, \infty$. There are a number of ways in which this comparison can be made. The comparison used is based on the assumption that there exists a "correct" structured matrix A_c and vector b_c, such that

$$A_c x_c = b_c \tag{7.5}$$

for some "correct" vector x_c. In other words, error-free values exist such that the overdetermined system has a solution x_c, with zero residual. The actual data contains noise so that a perturbed (but structure preserving) matrix A_p and vector b_p are known. The objective is to get the "best" solution x_p to the perturbed system $A_p x \approx b_p$ and, to the extent possible, reconstruct the matrix A_c and vector b_c from the noisy data. Specifically, the error matrix E and residual vector r are computed so that

$$(A_p + E)x_p = b_p - r. \tag{7.6}$$

This is done by minimizing the appropriate norm of E and r.

The test problems are constructed so that A_c, b_c, and x_c are known. Then random perturbations are generated to give A_p and b_p, so that A_p and b_p preserve the same structure as A_c and b_c. The matrix E and r, x_p satisfying (7.6) are then computed via LS, TLS, and STLN.

A comparison of these errors for LS, TLS, and STLN was made for three different types of structured problems.

1. The matrix A and vector b are unstructured, but errors can occur only in certain elements.

TABLE 7.2
Solution accuracy, A and b unstructured, $m = 20$, $n = 16$, $q = 4$.

Method	b_{err}	A_{err}	x_{err}
LS	3.3e-4	6.5e-4	1.8e-3
TLS	2.4e-5	6.5e-4	1.8e-3
STLN2	2.2e-5	9.7e-5	2.0e-4
STLN1	2.4e-5	7.6e-5	2.1e-4
STLN∞	2.6e-5	1.6e-4	3.4e-4

TABLE 7.3
Solution accuracy, A Toeplitz, b unstructured, $m = 11$, $n = 6$, $q = 4$.

Method	b_{err}	A_{err}	x_{err}
LS	8.9e-3	1.2e-2	1.5e-2
TLS	5.3e-4	1.2e-2	1.5e-2
STLN2	4.3e-4	5.3e-4	7.3e-4
STLN1	5.5e-4	5.4e-4	3.5e-4
STLN∞	5.7e-4	5.2e-4	5.9e-4

TABLE 7.4
Solution accuracy, $[A \mid b]$ Toeplitz, $m = 14$, $n = 4$, $q = 17$.

Method	A_{err}	x_{err}
LS	4.4e-3	1.4e-1
TLS	4.0e-3	2.4e-2
STLN2	3.8e-3	3.3e-3
STLN1	2.2e-6	7.2e-6
STLN∞	4.7e-3	1.3e-1

2. The matrix A is Toeplitz, with b unstructured.
3. The matrix $[A \mid b]$ is Toeplitz.

To illustrate the comparison obtained with over 200 test problems, a typical case has been selected for each type of structured problem. These typical cases are presented in Tables 7.2, 7.3, and 7.4. The following quantities are tabulated to give the measure of robustness [8]:

$$
\begin{aligned}
b_{\text{pert}} &= ||b_p - b_c||_2 / ||b_c||_2, \\
A_{\text{pert}} &= ||A_p - A_c||_F / ||A_c||_F, \\
b_{\text{err}} &= ||b_p - r - b_c||_2 / ||b_c||_2, \\
A_{\text{err}} &= ||A_p + E - A_c||_F / ||A_c||_F, \\
x_{\text{err}} &= ||x_p - x_c||_2 / ||x_c||_2.
\end{aligned}
$$

Table 7.2 gives the comparison for A and b unstructured, with all elements of b and four elements of A perturbed, and $m = 20$, $n = 16$, and $q = 4$. The values of $b_{\text{pert}} = 2.4\text{e}{-5}$ and $A_{\text{pert}} = 6.5\text{e}{-4}$ were used. The matrices A_c and A_p are dense, but E is sparse with only four nonzero elements, for the STLN solutions. The matrix E is zero for LS and dense for TLS.

Table 7.3 gives the comparison for A Toeplitz, and b unstructured, with $m = 11$, $n = 6$, and $q = 4$. The matrices A_c and A_p are both Toeplitz, and E is Toeplitz with four nonzero diagonals for the STLN solutions. The matrix E is zero for LS and dense for TLS. The values of $b_{\text{pert}} = 5.2\text{e}{-4}$ and $A_{\text{pert}} = 1.2\text{e}{-2}$ were used.

Table 7.4 gives the comparison for $[A \mid b]$ Toeplitz, with $m = 14$, $n = 4$, and

$q = 17$. The problem presented was selected to illustrate the performance of the STLN algorithm with an outlier in the data. In addition to small random perturbations in each diagonal of $[A \mid b]$, a much larger error was introduced in one of the diagonals. The small random perturbations gave $A_{\text{pert}} = 2.1\text{e}{-6}$ and the exact data with outlier only gave $A_{\text{pert}} = 5.1\text{e}{-3}$. The b_{err} is included in A_{err} for $[A \mid b]$ Toeplitz. The matrix E is zero for LS and dense for TLS. The most significant result shown in Table 7.4 is that STLN1 is essentially unaffected by the outlier. Note that the STLN∞ solution is affected most by the outlier.

The results presented in Tables 7.2, 7.3, and 7.4 show that the errors in the STLN1 and STLN2 solutions are significantly less than in the LS and TLS solutions. Similar results were obtained for all structured problems of these types tested.

8. Conclusions and future work. A new algorithm has been presented for solving an important class of problems related to TLS. The main new features of this approach are that it preserves the problem structure, and also permits the minimization of error in different norms. Both the theoretical analysis and the computational results show that the STLN algorithm is an efficient computational method for problems with a special structure, or where the number of elements with possible error (in the matrix A) is not too large. The ability to minimize the error in norms other than the 2-norm is also important, since we believe this will give more robust solutions in certain cases. When the data are from the complex field, the presented algorithms can be used in a straightforward way for the 2-norm. For the 1-norm and ∞-norm with complex data, the STLN problem will require the solution of a nonlinear programming problem rather than linear programming.

In order to more fully investigate the potential of the STLN formulation and algorithm for a range of applications, a number of areas need further study. Future work on STLN will include computational testing of much larger problems arising in important applications where the matrix A has a special, or sparse structure, and theoretical and computational analysis of the effect of the magnitude of the minimum norm on the convergence rate.

Acknowledgment. We wish to thank Dr. Sabine Van Huffel for introducing us to this problem and subsequent discussions, and Prof. Philip Gill for a helpful discussion.

REFERENCES

[1] T.J. ABATZOGLOU, J.M. MENDEL, G.A. HARADA, *The constrained total least squares technique and its application to harmonic superresolution*, IEEE Trans. Signal Processing, 39 (1991), pp. 1070–1087.

[2] A.A. ANDA AND H. PARK, *Self-scaling fast rotations for stiff least squares problems*, Linear Algebra Appl., to appear.

[3] 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.

[4] B. DE MOOR, *Structured total least squares and L_2 approximation problems*, Linear Algebra Appl., special issue on Numerical Linear Algebra Methods in Control, Signals and Systems, Van Dooren et al., eds., 188-189 (1993), pp. 163–207.

[5] R. FLETCHER *Practical Methods of Optimization, Vol. 1, Unconstrained Optimization.* John Wiley, New York, 1980.

[6] E. FOGEL, *Total least squares for Toeplitz structures*, in Proc. 21st IEEE Conference on Decision and Control, Orlando, 3 (1982), pp. 1003–1004.

[7] G. H. GOLUB AND C. F. VAN LOAN, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 883–893.

[8] ———, *Matrix Computations*, second edition, Johns Hopkins University Press, Baltimore, 1989.

[9] S. HAYKIN, *Adaptive Filter Theory*, second edition, Prentice Hall, Englewood Cliffs, NJ, 1991.

[10] M. R. OSBORNE AND G. A. WATSON, *An analysis of the total approximation problem in separable norms and an algorithm for the total l_1 problem*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 410–424.

[11] M. A. RAHMAN AND K. B. YU, *Total least squares approach for frequency estimation using linear prediction*, IEEE Trans. on Acoustic Speech Signal Processing, ASSP–35 (1987), pp. 1440–1454.

[12] J.B. ROSEN, H. PARK, AND J. GLICK, *Total least norm formulation and solution for structured problems*, AHPCRC preprint 94-041, University of Minnesota, July, 1994.

[13] S. VAN HUFFEL AND J. VANDEWALLE, *Analysis and properties of the generalized total least squares problem $AX \approx B$ when some or all columns of A are subject to errors*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 294–315.

[14] ———, *The Total Least Squares Problem, Computational Aspects and Analysis*, Society for Industrial and Applied Mathematics, Philadelphia, 1991.

[15] S. VAN HUFFEL, B. DE MOOR, AND H. CHEN, *Relationships between constrained and structured TLS, with applications to signal enhancement*, Proc. Internat. Symp. Mathematical Theory for Networks and Systems, Regensburg, Germany, August 2-6, 1993, to appear.

[16] C.F. VAN LOAN, *On the method of weighting for equality constrained least squares problems*, SIAM J. Numer. Anal., 22 (1985), pp. 851–864.

[17] G.A. WATSON, *On a class of algorithms for total approximation*, J. Approx. Theory, 45 (1985), pp. 219–231.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 1, pp. 127–138, January 1996

007

SOLUTION OF VANDERMONDE-LIKE SYSTEMS AND CONFLUENT VANDERMONDE-LIKE SYSTEMS*

HAO LU†

Abstract. It is shown that the solution of Vandermonde-like systems and the solution of confluent Vandermonde-like systems can be obtained by evaluation of certain polynomials and Hermite evaluation of certain rational functions in terms of a J-match of polynomials, respectively. Based on these results, the existence of an $O(n \log^2 n)$ algorithm is shown for both Vandermonde-like systems and confluent Vandermonde-like systems for the case where the polynomials satisfy a three-term recurrence relation.

Key words. Vandermonde-like system, confluent Vandermonde-like system, J-match, computational complexity, divide and conquer

AMS subject classifications. 65F05, 65Y05, 68C25

1. Introduction.

Let $t_0, \ldots, t_p$ be $p+1$ complex numbers, $n_0, \ldots, n_p$ be $p+1$ positive integers and $\mathbf{p}(\lambda) = (p_0(\lambda), p_1(\lambda), \ldots, p_{n-1}(\lambda))^T$, where $n = \sum_{i=0}^{p} n_i$ and $P(\lambda) = \{p_0(\lambda), p_1(\lambda), \ldots, p_{n-1}(\lambda)\}$ is a basis of the linear space $\mathbb{C}_{n-1}[\lambda]$ of all complex polynomials of degree at most $n-1$. The confluent Vandermonde-like matrix (see [16]), denoted by $V_c(\mathbf{p})$, is given by

$$V_c(\mathbf{p}) = (B_0, B_1, \ldots, B_p), \tag{1}$$

where and throughout the paper the row index and the column index of an $n \times m$ matrix run from 0 to $n-1$ and from 0 to $m-1$, respectively, and B_k is an $n \times n_k$ matrix with (i,j) entry

$$\left.\frac{d^j(p_i(\lambda))}{d\lambda^j}\right|_{\lambda=t_k}.$$

In the case of $n_0 = n_1 = \cdots = n_p = 1$, $V_c(\mathbf{p})$ yields a Vandermonde-like matrix [10], [15]. We denote the Vandermonde-like matrix by $V(\mathbf{p})$. Consider Vandermonde-like systems

$$V(\mathbf{p})\mathbf{x} = \mathbf{b} \tag{2}$$

and confluent Vandermonde-like systems

$$V_c(\mathbf{p})\mathbf{x} = \mathbf{b}. \tag{3}$$

These systems are associated with the construction of quadrature formulae [2],[13], [18], [23] and the approximation of linear functionals [3], [27].

The Vandermonde-like matrix and the confluent Vandermonde-like matrix are generalizations of the well-known Vandermonde matrix and the confluent Vandermonde matrix [4], [8], [9], [20], [22], respectively. In the early 1970s, Björck, Elfving, and Pereyra presented some $O(n^2)$ algorithms for Vandermonde systems and confluent

* Received by the editors August 22, 1994; accepted for publication (in revised form) by N. Higham January 24, 1995. This work was partially supported by the Netherlands Organization for Scientific Research grant 611-302-025.

† Department of Mathematics, University of Nijmegen, Toernooiveld, 6525 ED Nijmegen, The Netherlands (na.hlu@na-net.ornl.gov).

Vandermonde systems based on forward and backward recursion [4], [5]. In 1988 and 1990, Higham considered Vandermonde-like systems and confluent Vandermonde-like systems [15], [16]. He derived some $O(n^2)$ fast algorithms for both systems for the case where the polynomials $p_0(\lambda)$, $p_1(\lambda)$, ..., $p_{n-1}(\lambda)$ satisfy a three-term recurrence relation that generalizes previous methods for $p_k(\lambda) = \lambda^k$. In 1990 and 1994, I showed that solution of Vandermonde systems and confluent Vandermonde systems can be obtained by evaluating certain polynomials and certain rational functions [19], [20], respectively. By incorporating fast polynomial arithmetic, it is shown that the number of operations for solving Vandermonde systems in [19] and confluent Vandermonde systems in [20] and [22] can be further reduced to $O(n \log^2 n)$ and $O(n \log p \log n)$, respectively. Other $O(n \log^2 n)$ algorithms for Vandermonde linear systems can be found in [7] by Canny, Kaltofen, and Yagati and in [11] by Gohberg and Olshevsky. There are a number of $O(n^2)$ fast algorithms for Vandermonde systems (see Traub [27], Tang and Golub [26]) and Chebyshev–Vandermonde systems (see Reichel and Opfer [25], Gohberg and Olshevsky [12]). For the numerical properties and stability of algorithms for solving the systems, Gautschi estimated the condition number of Vandermonde-like matrices for various choices of the polynomials $p_0(\lambda), p_1(\lambda), \ldots, p_{n-1}(\lambda)$ in 1983 [10]. Higham derived some error bounds for certain algorithms in 1987 and 1990 [14], [16]. In [17] he proposed iterative refinement to enhance the stability.

The discussion of inversion of Vandermonde matrices, confluent Vandermonde matrices, Vandermonde-like matrices, and confluent Vandermonde-like matrices can be found in various sources. Formulas for the entries of the inverse of a Vandermonde matrix were given by Macon and Spitzbart in 1958 [24] and by Traub in 1966 [27]. Later in 1988, Verde-Star considered structure and algebraic properties of the inverses of confluent Vandermonde matrices [28]. The expression of inverses of confluent Vandermonde matrices is given in [20]. A number of fast algorithms for inversion are available (see, for example, Traub [27] for Vandermonde matrices, Gohberg and Olshevsky [12] for Chebyshev–Vandermonde matrices, and Calvetti and Reichel [6] for Vandermonde-like matrices).

The purpose of this paper is to extend the structure results on solution of Vandermonde systems and confluent Vandermonde systems to Vandermonde-like systems and confluent Vandermonde-like systems, respectively, and show the existence of $O(n \log^2 n)$ asymptotically fast algorithms for both Vandermonde-like systems and confluent Vandermonde-like systems for the case where the polynomials $p_0(\lambda), p_1(\lambda)$, $\ldots, p_{n-1}(\lambda)$ satisfy a three-term recurrence relation. The rest of the paper is organized as follows. In §2, we introduce J-matches for bases of $\mathbb{C}_{n-1}[\lambda]$ and links of a basis with its J-match. The existence of a J-match for any basis of $\mathbb{C}_{n-1}[\lambda]$ is established. If the polynomials $p_0(\lambda), p_1(\lambda), \ldots, p_{n-1}(\lambda)$ satisfy a k-term recurrence relation, we show an expression of the J-match for the basis $P(\lambda) = \{p_0(\lambda), p_1(\lambda), \ldots, p_{n-1}(\lambda)\}$. In §3, it is shown that the solution of Vandermonde-like systems can be obtained by evaluating certain polynomials in terms of a J-match. By using this result it follows immediately from [20] that the solution of confluent Vandermonde-like systems can be obtained by Hermite evaluation of certain rational functions. In §4, we show the existence of an $O(n \log^2 n)$ algorithm for confluent Vandermonde-like systems for the important case where the polynomial $p_0(\lambda), p_1(\lambda), \ldots, p_{n-1}(\lambda)$ satisfy a three-term recurrence relation.

2. J-matches of polynomials. For discussion of Vandermonde-like systems and confluent Vandermonde-like systems, we introduce J-matches for bases of the linear space $\mathbb{C}_{n-1}[\lambda]$ in this section and show the existence of a J-match for any basis.

Finally, we give an expression of the J-match for the basis $\{p_0(\lambda), p_1(\lambda), \ldots, p_{n-1}(\lambda)\}$ if $p_i(\lambda)$, $i = 0, 1, \ldots, n-1$ satisfy a k-term recurrence relation.

DEFINITION. Let $P(\lambda) = \{p_0(\lambda), p_1(\lambda), \ldots, p_{n-1}(\lambda)\}$ be a basis of the linear space $\mathbb{C}_{n-1}[\lambda]$ of all complex polynomials of degree at most $n-1$. If there exists a basis $Q(\lambda) = \{q_0(\lambda), q_1(\lambda), \ldots, q_{n-1}(\lambda)\}$ of $\mathbb{C}_{n-1}[\lambda]$ and a polynomial $p(\lambda)$ such that

$$\frac{p(\lambda) - p(\mu)}{\lambda - \mu} = \sum_{i=0}^{n-1} p_i(\lambda) q_{n-i-1}(\mu), \tag{4}$$

$Q(\lambda)$ is called a J-match of $P(\lambda)$ and $p(\lambda)$ is called a link of $\{P(\lambda), Q(\lambda)\}$.

Some particular J-matches and links had been used earlier without a formal definition. Lu [19] used the J-match $\{1, \lambda, \ldots, \lambda^{n-1}\}$ with the link $p(\lambda) = \lambda^n$ for Vandermonde linear systems. Verde-Star [28], Calvetti and Reichel [6], Gohberg and Olshevsky [11] used the link $w(\lambda) = (\lambda - t_1)(\lambda - t_2) \cdots (\lambda - t_n)$ for inversion of a Vandermonde-like matrix with the nodes $t_1, t_2, \ldots, t_n$. For any basis of $\mathbb{C}_{n-1}[\lambda]$, we now show the existence of a J-match.

THEOREM 2.1. *For any basis $P(\lambda) = \{p_0(\lambda), p_1(\lambda), \ldots, p_{n-1}(\lambda)\}$ of the space $\mathbb{C}_{n-1}[\lambda]$, there exists a* J*-match $Q(\lambda) = \{q_0(\lambda), q_1(\lambda), \ldots, q_{n-1}(\lambda)\}$ of $P(\lambda)$.*

Proof. Let $p_i(\lambda) = \sum_{j=0}^{n-1} a_{ij}\lambda^j$, $q_i(\lambda) = \sum_{j=0}^{n-1} b_{ij}\lambda^j$, where a_{ij}, $i, j = 0, 1, \ldots, n-1$ are known and b_{ij}, $i, j = 0, 1, \ldots, n-1$ are unknown. Our task is to prove the existence of b_{ij}, $i, j = 0, 1, \ldots, n-1$ such that $Q(\lambda) = \{q_0(\lambda), q_1(\lambda), \ldots, q_{n-1}(\lambda)\}$ is a J-match of $P(\lambda)$. Let

$$c_{jk} = \begin{cases} \displaystyle\sum_{i=0}^{n-1} a_{ij} b_{n-i-1,k} & j, k = 0, 1, \ldots n-1, \\ 0 & j = n \text{ or } k = n. \end{cases}$$

A straightforward computation shows that

$$\begin{aligned}
&(\lambda - \mu) \sum_{i=0}^{n-1} p_i(\lambda) q_{n-i-1}(\mu) \\
&= (\lambda - \mu) \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} a_{ij}\lambda^j \sum_{k=0}^{n-1} b_{n-i-1,k}\mu^k \\
&= \sum_{j=0}^{n-1} \sum_{k=0}^{n-1} c_{jk}\lambda^{j+1}\mu^k - \sum_{j=0}^{n-1} \sum_{k=0}^{n-1} c_{jk}\lambda^j\mu^{k+1} \\
&= \sum_{j=0}^{n-1} c_{j0}\lambda^{j+1} - \sum_{k=0}^{n-1} c_{0k}\mu^{k+1} + \sum_{j=0}^{n-1} \sum_{k=1}^{n-1} c_{jk}\lambda^{j+1}\mu^k - \sum_{j=1}^{n-1} \sum_{k=0}^{n-1} c_{jk}\lambda^j\mu^{k+1} \\
&= \sum_{j=0}^{n-1} c_{j0}\lambda^{j+1} - \sum_{k=0}^{n-1} c_{0k}\mu^{k+1} + \sum_{j=1}^{n} \sum_{k=1}^{n} (c_{j-1,k} - c_{j,k-1})\lambda^j\mu^k.
\end{aligned}$$

Therefore, there exists a polynomial $p(\lambda)$ such that (4) holds if and only if

$$c_{j0} = c_{0j}, \qquad j = 0, 1, \ldots, n-1, \tag{5}$$

$$c_{j-1,k} = c_{j,k-1}, \qquad j, k = 1, 2, \ldots, n. \tag{6}$$

Let

$$A = \begin{pmatrix} a_{00} & a_{10} & \cdots & a_{n-1,0} \\ a_{01} & a_{11} & \cdots & a_{n-1,1} \\ \cdots & \cdots & \cdots & \cdots \\ a_{0,n-1} & a_{1,n-1} & \cdots & a_{n-1,n-1} \end{pmatrix}, \tag{7}$$

$$B = \begin{pmatrix} b_{n-1,n-1} & b_{n-1,n-2} & \cdots & b_{n-1,0} \\ b_{n-2,n-1} & b_{n-2,n-2} & \cdots & b_{n-2,0} \\ \cdots & \cdots & \cdots & \cdots \\ b_{0,n-1} & b_{0,n-2} & \cdots & b_{0,0} \end{pmatrix},$$

and C be an upper triangular Toeplitz matrix of the form

$$C = \begin{pmatrix} c_{n-1} & c_{n-2} & \cdots & c_0 \\ 0 & c_{n-1} & \cdots & c_1 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & c_{n-1} \end{pmatrix}.$$

It is readily seen that (5) and (6) hold if and only if B is a solution of

$$AX = C. \tag{8}$$

It is straightforward to show that $P(\lambda)$ is a basis of $\mathbb{C}_{n-1}[\lambda]$ if and only if A is nonsingular. Hence, by choosing $c_{n-1} \neq 0$ equation (8) has the nonsingular solution B. Then $Q(\lambda) = \{q_0(\lambda), q_1(\lambda), \ldots, q_{n-1}(\lambda)\}$ is a J-match of $P(\lambda)$ with a link $p(\lambda) = \sum_{i=0}^{n-1} c_i \lambda^{i+1} + d_0$, where d_0 is a constant. □

COROLLARY 2.2. *Let $Q(\lambda)$ be a J-match of a basis $P(\lambda)$ in $\mathbb{C}_{n-1}[\lambda]$ and $p(\lambda)$ be a link of $\{P(\lambda), Q(\lambda)\}$. Then* $\deg(p(\lambda)) = n$.

Proof. From the proof of Theorem 2.1 that $Q(\lambda)$ is a J-match of $P(\lambda)$ if and only if B is the solution of (8) with $c_{n-1} \neq 0$. This implies that $\deg(p(\lambda)) = n$. □

The important fact for Vandermonde-like matrices and confluent Vandermonde-like matrices is that the polynomials $p_0(\lambda), p_1(\lambda), \ldots, p_{n-1}(\lambda)$ satisfy a k-term recurrence relation. We derive an expression of the J-match for this case.

THEOREM 2.3. *Assume that the polynomials $p_i(\lambda)$ and $q_i(\lambda)$ satisfy k-term recurrence relations*

$$p_0(\lambda) = 1, \qquad p_i(\lambda) = 0, \quad i < 0, \tag{9}$$

$$p_i(\lambda) = \alpha_i(\lambda - \beta_i)p_{i-1}(\lambda) + \sum_{j=2}^{k-1} \gamma_{ij} p_{i-j}(\lambda), \quad i = 1, \ldots, n, \tag{10}$$

$$q_0(\lambda) = 1, \qquad q_i(\lambda) = 0, \quad i < 0, \tag{11}$$

$$q_i(\lambda) = \alpha_{n-i+1}(\lambda - \beta_{n-i+1})q_{i-1}(\lambda) + \sum_{j=2}^{k-1} \gamma_{n-i+j,j} q_{i-j}(\lambda), \quad i = 1, \ldots, n, \tag{12}$$

where $\alpha_i \neq 0$ *for* $i = 1, 2, \ldots, n$. *Then* $p_n(\lambda) = q_n(\lambda)$ *and*

$$\frac{p_n(\lambda) - p_n(\mu)}{\lambda - \mu} = \sum_{i=0}^{n-1} \alpha_{i+1} p_i(\lambda) q_{n-i-1}(\mu). \tag{13}$$

Proof. By using (9), (10), (11), and (12) a computation shows that

$$\begin{aligned}
&(\lambda - \mu) \sum_{i=0}^{n-1} \alpha_{i+1} p_i(\lambda) q_{n-i-1}(\mu) \\
&= \alpha_n(\lambda - \mu) p_{n-1}(\lambda) + \alpha_1(\lambda - \mu) q_{n-1}(\mu) \\
&\quad + \sum_{i=1}^{n-2} \alpha_{i+1}(\lambda - \beta_{i+1}) p_i(\lambda) q_{n-i-1}(\mu) \\
&\quad - \sum_{i=1}^{n-2} \alpha_{i+1}(\mu - \beta_{i+1}) q_{n-i-1}(\mu) p_i(\lambda) \\
&= \alpha_n(\lambda - \mu) p_{n-1}(\lambda) + \alpha_1(\lambda - \mu) q_{n-1}(\mu) \\
&\quad + \sum_{i=1}^{n-2} \Big(p_{i+1}(\lambda) - \sum_{j=2}^{k-1} \gamma_{i+1,j} p_{i-j+1}(\lambda) \Big) q_{n-i-1}(\mu) \\
&\quad - \sum_{i=1}^{n-2} \Big(q_{n-i}(\mu) - \sum_{j=2}^{k-1} \gamma_{i+j,j} q_{n-i-j}(\mu) \Big) p_i(\lambda) \\
&= \alpha_n(\lambda - \mu) p_{n-1}(\lambda) + \alpha_1(\lambda - \mu) q_{n-1}(\mu) \\
&\quad + \sum_{i=1}^{n-2} p_{i+1}(\lambda) q_{n-i-1}(\mu) - \sum_{i=1}^{n-2} p_i(\lambda) q_{n-i}(\mu) \\
&\quad + \sum_{j=2}^{k-1} \sum_{i=1}^{n-j} \gamma_{i+j,j} p_i(\lambda) q_{n-i-j}(\mu) - \sum_{j=2}^{k-1} \sum_{i=j-1}^{n-2} \gamma_{i+1,j} p_{i-j+1}(\lambda) q_{n-i-1}(\mu) \\
&= \alpha_n(\lambda - \mu) p_{n-1}(\lambda) + \alpha_1(\lambda - \mu) q_{n-1}(\mu) + q_1(\mu) p_{n-1}(\lambda) - p_1(\lambda) q_{n-1}(\mu) \\
&\quad + \sum_{j=2}^{k-1} \sum_{i=1}^{n-j} \gamma_{i+j,j} p_i(\lambda) q_{n-i-j}(\mu) - \sum_{j=2}^{k-1} \sum_{i=0}^{n-j-1} \gamma_{i+j,j} p_i(\lambda) q_{n-i-j}(\mu) \\
&= \alpha_n(\lambda - \mu) p_{n-1}(\lambda) + \alpha_1(\lambda - \mu) q_{n-1}(\mu) + \alpha_n(\mu - \beta_n) p_{n-1}(\lambda) \\
&\quad - \alpha_1(\lambda - \beta_1) q_{n-1}(\mu) + \sum_{j=2}^{k-1} \gamma_{n,j} p_{n-j}(\lambda) - \sum_{j=2}^{k-1} \gamma_{jj} q_{n-j}(\mu) \\
&= \alpha_n(\lambda - \beta_n) p_{n-1} + \sum_{j=2}^{k-1} \gamma_{n,j} p_{n-j}(\lambda) \\
&\quad - \alpha_1(\lambda - \beta_1) q_{n-1}(\mu) - \sum_{j=2}^{k-1} \gamma_{jj} q_{n-j}(\mu) \\
&= p_n(\lambda) - q_n(\mu).
\end{aligned}$$

Setting $\lambda = \mu$ shows that $p_n(\mu) = q_n(\mu)$ and (13) follows immediately. □

Denote by $T_k(\lambda)$ and $U_k(\lambda)$ Chebyshev polynomials

$$T_k(\lambda) = \cos(k\arccos(\lambda)), \qquad U_k(\lambda) = \frac{\sin((k+1)\arccos(\lambda))}{\sin(\arccos(\lambda))}$$

of the first and the second kind, respectively. It is well known that Chebyshev polynomials can also be defined by the following three-term recurrence relations:

$$T_0(\lambda) = 1, \qquad T_1(\lambda) = \lambda, \qquad T_i(\lambda) = 2\lambda T_{i-1}(\lambda) - T_{i-2}(\lambda), \qquad i \geq 2,$$

$$U_0(\lambda) = 1, \qquad U_1(\lambda) = 2\lambda, \qquad U_i(\lambda) = 2\lambda U_{i-1}(\lambda) - U_{i-2}(\lambda), \qquad i \geq 2.$$

If we choose $k = 3$, $\alpha_1 = 1$, $\alpha_i = 2$, $\gamma_{i2} = -1$ for $i = 2, \ldots, n$ and $\beta_i = 0$ for $i = 1, 2, \ldots, n$, then $p_i(\lambda) = T_i(\lambda)$ for $i = 0, 1, \ldots, n$ and $q_i(\lambda) = U_i(\lambda)$ for $i = 0, 1, \ldots, n-1$. If we choose $k = 3$, $\alpha_i = 2$, $\beta_i = 0$ for $i = 1, 2, \ldots, n$ and $\gamma_{i2} = -1$ for $i = 2, \ldots, n$, then $p_i(\lambda) = q_i(\lambda) = U_i(\lambda)$ for $i = 0, 1, \ldots, n$. Therefore, the equalities

$$\frac{T_n(\lambda) - T_n(\mu)}{\lambda - \mu} = 2\sum_{i=0}^{n-2} T_{n-i-1}(\lambda)U_i(\mu) + T_0(\lambda)U_{n-1}(\mu),$$

$$\frac{U_n(\lambda) - U_n(\mu)}{\lambda - \mu} = 2\sum_{i=0}^{n-2} U_{n-i-1}(\lambda)U_i(\mu)$$

in [12] follow from Theorem 2.3 immediately.

3. Solution of Vandermonde-like systems and confluent Vandermonde-like systems. In this section, we show that the solution of Vandermonde-like systems and confluent Vandermonde-like systems can be obtained by evaluating certain polynomials and certain rational functions, respectively.

LEMMA 3.1. *Let* $P(\lambda) = \{p_i(\lambda) = \sum_{i=0}^{n-1} a_{ij}\lambda^j,\ i = 0, 1, \ldots, n-1\}$ *be a basis of* $\mathbb{C}_{n-1}[\lambda]$. *Then the confluent Vandermonde-like matrices defined by* (1) *are nonsingular if and only if* $t_i \neq t_j$, $i \neq j$, $i, j = 0, 1, \ldots, p$ *and*

$$\det V_c(\mathbf{p}) = \det(A)\left(\prod_{i=0}^{p}\prod_{k=0}^{n_i-1} k!\right)\prod_{p\geq i>j\geq 0}(t_i - t_j)^{n_i n_j},$$

where A *is the matrix given by* (7).

Proof. A simple observation shows that

$$\mathbf{p}(\lambda) = (p_0(\lambda), p_1(\lambda), \ldots, p_{n-1}(\lambda))^T = A^T(1, \lambda, \ldots, \lambda^{n-1})^T,$$

which implies that

$$\mathbf{p}^{(j)}(\lambda) = A^T\frac{d^j}{d\lambda^j}(1, \lambda, \ldots, \lambda^{n-1})^T.$$

Therefore, $V_c(\mathbf{p}) = A^T V_c$, where V_c is the corresponding confluent Vandermonde matrix. Applying Proposition 2.1 in [20] or the determinant of a confluent Vandermonde matrix (see [1], pp. 121) shows that

$$\det(V_c(\mathbf{p})) = \det(A^T)\det(V_c) = \det(A)\left(\prod_{i=0}^{p}\prod_{k=0}^{n_i-1} k!\right)\prod_{p\geq i>j\geq 0}(t_i - t_j)^{n_i n_j}.$$

As mentioned above, A is nonsingular since $P(\lambda)$ is a basis of $\mathbb{C}_{n-1}[\lambda]$. Hence, $V_c(\mathbf{p})$ is nonsingular if and only if $t_i \neq t_j$, $i \neq j$, $i,j = 0,1,\ldots,p$. □

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 $r(x)$: $A(x) = B(x)\mathrm{quot}(A(x),B(x)) + r(x)$, in the rest of the paper. Now we show our fundamental result on solution of Vandermonde-like systems if $t_i \neq t_j$, $i \neq j$, $i,j = 0,1,\ldots,p$.

THEOREM 3.2. *Let* $P(\lambda) = \{p_0(\lambda),p_1(\lambda),\ldots,p_{n-1}(\lambda)\}$ *be a basis of* $\mathbb{C}_{n-1}[\lambda]$, $V(\mathbf{p})$ *be the Vandermonde-like matrix defined by* $\mathbf{p}(\lambda) = (p_0(\lambda),p_1(\lambda),\ldots,p_{n-1}(\lambda))^T$ *via* (1) *with* $n_0 = n_1 = \cdots = n_p = 1$ *and* $t_i \neq t_j$, $i \neq j$, $i,j = 0,1,\ldots,p$, *and*

$$l(\lambda) = (\lambda - t_0)(\lambda - t_1)\cdots(\lambda - t_{n-1}),$$

$$l_i(\lambda) = l(\lambda)/(\lambda - t_i), \qquad i = 0,1,\ldots,n-1.$$

Then the solution of the Vandermonde-like system

$$V(\mathbf{p})\mathbf{x} = \mathbf{b} \tag{14}$$

is given by

$$x_i = g(t_i)/l_i(t_i), \qquad i = 0,1,\ldots,n-1, \tag{15}$$

where

$$g(\lambda) = \mathrm{quot}(l(\lambda)b(\lambda),p(\lambda)), \qquad b(\lambda) = \sum_{i=0}^{n-1} b_i q_{n-i-1}(\lambda),$$

$Q(\lambda) = \{q_0(\lambda),\ldots,q_{n-1}(\lambda)\}$ *is a* J*-match of* $P(\lambda)$ *and* $p(\lambda)$ *is a link of* $\{P(\lambda),Q(\lambda)\}$.

Proof. Let λ_i, $i = 0,1,\ldots,n-1$ be n numbers such that $\lambda_i \neq \lambda_j$, $i \neq j$, $i,j = 0,1,\ldots,n-1$ and $\lambda_i \neq t_j$, $i,j = 0,1,\ldots,n-1$. Consider a matrix of the form

$$\widetilde{V} = \begin{pmatrix} q_{n-1}(\lambda_0) & q_{n-2}(\lambda_0) & \cdots & q_0(\lambda_0) \\ q_{n-1}(\lambda_1) & q_{n-2}(\lambda_1) & \cdots & q_0(\lambda_1) \\ \cdots & \cdots & \cdots & \cdots \\ q_{n-1}(\lambda_{n-1}) & q_{n-2}(\lambda_{n-1}) & \cdots & q_0(\lambda_{n-1}) \end{pmatrix}.$$

Applying Lemma 3.1 shows that $\widetilde{V}$ is nonsingular. Therefore, $\mathbf{x}$ is the solution of (14) if and only if $\mathbf{x}$ is the solution of the following system:

$$\widetilde{V}V(\mathbf{p})\mathbf{x} = \widetilde{V}\mathbf{b}. \tag{16}$$

Furthermore, using (4) shows that

$$\widetilde{V}V(\mathbf{p}) = \begin{pmatrix} \dfrac{p(\lambda_0)-p(t_0)}{\lambda_0-t_0} & \dfrac{p(\lambda_0)-p(t_1)}{\lambda_0-t_1} & \cdots & \dfrac{p(\lambda_0)-p(t_{n-1})}{\lambda_0-t_{n-1}} \\ \dfrac{p(\lambda_1)-p(t_0)}{\lambda_1-t_0} & \dfrac{p(\lambda_1)-p(t_1)}{\lambda_1-t_1} & \cdots & \dfrac{p(\lambda_1)-p(t_{n-1})}{\lambda_1-t_{n-1}} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{p(\lambda_{n-1})-p(t_0)}{\lambda_{n-1}-t_0} & \dfrac{p(\lambda_{n-1})-p(t_1)}{\lambda_{n-1}-t_1} & \cdots & \dfrac{p(\lambda_{n-1})-p(t_{n-1})}{\lambda_{n-1}-t_{n-1}} \end{pmatrix}.$$

Denote $\mathbf{d} = \widetilde{V}\mathbf{b} = (d_0, d_1, \ldots, d_{n-1})^T$ and

$$g(\lambda) = \sum_{i=0}^{n-1} x_i l_i(\lambda), \qquad f(\lambda) = \sum_{i=0}^{n-1} p(t_i) x_i l_i(\lambda),$$

$$F(\lambda) = p(\lambda)g(\lambda) - f(\lambda).$$

Clearly, $d_i = b(\lambda_i)$, $i = 0, 1, \ldots, n-1$. It follows from Corollary 2.2 that $\deg(p(\lambda)) = n$. Since $\deg(f(\lambda)) \leq n-1$ and $\deg(F(\lambda)) \leq 2n-1$, we have

$$g(\lambda) = \operatorname{quot}(F(\lambda), p(\lambda)).$$

On the other hand, (16) implies that

$$F(\lambda_i) = l(\lambda_i) d_i, \qquad i = 0, 1, \ldots, n-1, \tag{17}$$

and a simple calculation shows that

$$F(t_i) = p(t_i) x_i l_i(t_i) - p(t_i) x_i l_i(t_i) = 0, \qquad i = 0, 1, \ldots, n-1. \tag{18}$$

Therefore, there exists a unique polynomial $u(\lambda)$ of degree at most $n-1$ such that $F(\lambda) = l(\lambda)u(\lambda)$. It follows from (17) that

$$u(\lambda_i) = l(\lambda_i) d_i / l(\lambda_i) = d_i, \qquad i = 0, 1, \ldots, n-1.$$

Uniqueness of interpolating polynomial shows that $u(\lambda) = b(\lambda)$. Hence, we have $g(\lambda) = \operatorname{quot}(l(\lambda)b(\lambda), p(\lambda))$. Furthermore, $g(\lambda) = \sum_{i=0}^{n-1} x_i l_i(\lambda)$ shows (15). □

The result of Theorem 3.2 can be extended to confluent Vandermonde-like systems by using the approach in [20].

THEOREM 3.3. *Let $V_c(\mathbf{p})$ be a confluent Vandermonde-like matrix defined by $\mathbf{p}(\lambda) = (p_0(\lambda), p_1(\lambda), \ldots, p_{n-1}(\lambda))^T$ via (1) with $t_i \neq t_j$, $i \neq j$, $i, j = 0, 1, \ldots, p$, where $P(\lambda) = \{p_0(\lambda), p_1(\lambda), \ldots, p_{n-1}(\lambda)\}$ is a basis of $\mathbb{C}_{n-1}[\lambda]$, and*

$$r(\lambda) = (\lambda - t_0)^{n_0} (\lambda - t_1)^{n_1} \cdots (\lambda - t_p)^{n_p},$$

$$r_i(\lambda) = r(\lambda)/(\lambda - t_i)^{n_i}, \qquad i = 0, 1, \ldots, p.$$

Then the solution of confluent Vandermonde-like system (3) is given by

$$x_i = \frac{1}{(k-1)!(n_j - k)!} \left(\frac{v(\lambda)}{r_j(\lambda)} \right)^{(n_j - k)} \Bigg|_{\lambda = t_j}, \tag{19}$$

$$i = m_j + k - 1, \quad 0 \leq j \leq p, \quad 1 \leq k \leq n_j, \quad m_0 = 0, \quad m_j = \sum_{t=0}^{j-1} n_t,$$

where $v(\lambda) = \operatorname{quot}(r(\lambda)b(\lambda), p(\lambda))$, $b(\lambda)$, and $p(\lambda)$ are the same as those in Theorem 3.2.

Proof. Under the assumptions of the theorem, Lemma 3.1 shows that $V_c(\mathbf{p})$ is nonsingular. The rest of the proof is essentially the same as that of Theorem 2.2 in [20] by using Theorem 3.2 of this paper. □

Applying Theorem 3.3 shows the following structure result on inverses of confluent Vandermonde-like matrices.

COROLLARY 3.4. *Let $V_c(\mathbf{p})$ be a confluent Vandermonde-like matrix satisfying the conditions of Theorem* 3.3. *Then $V_c(\mathbf{p})$ is nonsingular and $V_c(\mathbf{p})^{-1} = (v_{ij})$, where*

$$v_{ij} = \frac{1}{(k-1)!(n_m-k)!} \left(\frac{u_j(\lambda)}{r_m(\lambda)} \right)^{(n_m-k)} \Bigg|_{\lambda=t_m}, \tag{20}$$

$$i = \sum_{r=0}^{m-1} n_r + k - 1, \quad 1 \le k \le n_m, \quad 0 \le m \le p, \quad j = 0, 1, \ldots, n-1,$$

where $u_j(\lambda) = \operatorname{quot}(q_{n-j-1}(\lambda)r(\lambda), p(\lambda))$, $p(\lambda)$, and $q_i(\lambda)$ are the same as those in Theorem 3.2 *and $r(\lambda)$ is given in Theorem* 3.3.

Proof. The proof is essentially the same as that of Corollary 2.3 in [20] by using Theorem 3.3. □

4. Computational complexity. In this section, applying Theorem 3.3 we show the existence of an $O(n \log^2 n)$ algorithm for the solution of Vandermonde-like systems as well as for the solution of confluent Vandermonde-like systems if the polynomials $p_0(\lambda), p_1(\lambda), \ldots, p_{n-1}(\lambda)$ satisfy the three-term recurrence relation (9) and (10) with $k = 3$. We consider confluent Vandermonde-like systems only. Vandermonde-like systems can be viewed as a special case of confluent one with $n_0 = n_1 = \cdots = n_p = 1$.

Define $p_n(\lambda) = \alpha_n(\lambda - \beta_n)p_{n-1}(\lambda) + \gamma_{n2}p_{n-2}(\lambda)$ with $\alpha_n \neq 0$ and any constants β_n and γ_{n2}. Let $q_0(\lambda), q_1(\lambda), \ldots, q_n(\lambda)$ be the corresponding polynomials defined by (11) and (12). Theorem 2.3 shows that $Q(\lambda) = \{\alpha_n q_0(\lambda), \alpha_{n-1}q_1(\lambda), \ldots, \alpha_1 q_{n-1}(\lambda)\}$ is a J-match of $P(\lambda) = \{p_0(\lambda), p_1(\lambda), \ldots, p_{n-1}(\lambda)\}$ and $p_n(\lambda) = q_n(\lambda)$ is a link of $\{P(\lambda), Q(\lambda)\}$. Let $\mathbf{q}_0(\lambda) = (1,0)^T$, $\mathbf{q}_i(\lambda) = (q_i(\lambda), q_{i-1}(\lambda))^T$ for $i = 1, \ldots, n$. It follows from (11) and (12) that $\mathbf{q}_i(\lambda)$ can be represented by

$$\mathbf{q}_i(\lambda) = A_i \mathbf{q}_{i-1} = A_i A_{i-1} \cdots A_1 \mathbf{q}_0, \quad i = 1, 2, \ldots, n,$$

where A_i is a λ-matrix given by

$$A_i = \begin{pmatrix} \alpha_{n-i+1}(\lambda - \beta_{n-i+1}) & \gamma_{n-i+2,2} \\ 1 & 0 \end{pmatrix}$$

and $\gamma_{n+1,2} = 0$. Denote $\widetilde{b}(\lambda) = \sum_{i=0}^{n-1} \alpha_{i+1} b_i q_{n-i-2}(\lambda)$ and $\mathbf{c}(\lambda) = (b(\lambda), \widetilde{b}(\lambda))^T$. It is easy to show that

$$\begin{aligned} \mathbf{c}(\lambda) &= \alpha_1 b_0 \mathbf{q}_{n-1}(\lambda) + \cdots + \alpha_{n-1} b_{n-2} \mathbf{q}_1(\lambda) + \alpha_n b_{n-1} \mathbf{q}_0(\lambda) \\ &= (\alpha_1 b_0 A_{n-1} \cdots A_1 + \cdots + \alpha_{n-1} b_{n-2} A_1 + \alpha_n b_{n-1} I)\mathbf{q}_0(\lambda), \end{aligned}$$

where I denotes the unit matrix of order 2. Denote

$$\begin{aligned} D_k(\lambda) &= A_k A_{k-1} \cdots A_1, \qquad k = 1, 2, \ldots, n, \\ E(\lambda) &= \alpha_1 b_0 D_{n-1} + \cdots + \alpha_{n-1} b_{n-2} D_1 + \alpha_n b_{n-1} I. \end{aligned}$$

It is readily seen that $p_n(\lambda) = q_n(\lambda)$ and $b(\lambda)$ are the (0,0) entries of $D_n(\lambda)$ and $E(\lambda)$, respectively. Assume that $n = 2^m$ for some positive integer m without loss of generality. The following algorithm converts $b(\lambda)$ and $p_n(\lambda)$ into the form $\sum h_i \lambda^i$.

ALGORITHM CONVERT

$D_{1i}(\lambda) = A_i(\lambda), \qquad E_{1i}(\lambda) = \alpha_{n-i+1} b_{n-i} I, \qquad i = 1, 2, \ldots, n$

For $i = 2 : m+1$

For $j = 1 : 2^{m-i+1}$

$D_{ij}(\lambda) = D_{i-1,2j-1}(\lambda) D_{i-1,2j}(\lambda)$

$E_{ij}(\lambda) = E_{i-1,2j}(\lambda) D_{i-1,2j-1}(\lambda) + E_{i-1,2j-1}(\lambda)$

endfor j

endfor i

It is straightforward to show that $D_n(\lambda) = D_{m+1,1}(\lambda)$ and $E(\lambda) = E_{m+1,1}(\lambda)$. Let $B(\lambda) = (f_{ij}(\lambda))_{i,j=0}^{k}$ be a polynomial matrix, where $f_{ij}(\lambda)$ is a polynomial of λ, and define $\deg(B(\lambda)) = \max_{0 \le i,j \le k} \deg(f_{ij}(\lambda))$. Clearly, $\deg(D_{ij}(\lambda)) \le 2^{i-1}$ and $\deg(E_{ij}(\lambda)) \le 2^{i-1} - 1$. If a standard fast polynomial multiplication based on the fast Fourier transform (FFT) is used in Algorithm CONVERT, the number of operations is bounded by

$$\sum_{i=1}^{m} \sum_{j=1}^{2^{m-i}} O(i 2^i) = O(n \log^2 n).$$

With a straightforward modification of Algorithm 3.2 in [20], we can solve confluent Vandermonde-like systems as follows.

ALGORITHM FSCVLS (a fast solver of confluent Vandermonde-like systems). Let $p_i(\lambda)$ be polynomials satisfying the three-term recurrence relation defined by (9) and (10) with $k = 3$ and $t_0, t_1, \ldots, t_p$ be $p+1$ distinct complex numbers. Based on Theorem 3.3, Algorithm FSCVLS computes the solution of confluent Vandermonde-like systems $V_c(\mathbf{p})\mathbf{x} = \mathbf{b}$.

Stage I: $r_{m_i+j} = t_i, \qquad m_0 = 0, \quad m_i = \sum_{k=0}^{i-1} n_k,$

$j = 1, \ldots, n_i, \quad i = 0, \ldots, p,$

$T_{0i} = \lambda - r_i, \quad i = 1, 2, \ldots, n, \qquad S_{m1} = \{0, 1, \ldots, p\}$

Compute $T_{ji} = T_{j-1,2i-1} T_{j-1,2i}$ for $i = 1, 2, \ldots, 2^{m-j}$ and $j = 1, \ldots, m$

Perform CONVERT to obtain $p_n(\lambda)$ and $b(\lambda)$.

Compute $r_{m1} = \mathrm{quot}(T_{m1} b(\lambda), p_n(\lambda))$.

$\widetilde{r}_{m1} = T_{m1}$

Stage II: For $j = m : -1 : m - \lceil \log(p+1) \rceil + 1$

For $i = 1 : 1 : 2^{m-j}$

if $S_{ji} = \{l\}$ then

call SOLUTION$(r_{ji}, \widetilde{r}_{ji}, a_l, n_l)$

$x_{m_l+k-1} = v_{n_l-k}/(k-1)!$ for $k = 1, \ldots, n_l$

$S_{ji} = \emptyset$

elseif $S_{ji} \neq \emptyset$ for $k = 2i-1, 2i$

$r_{j-1,k} = r_{ji} \,(\mathrm{mod} T_{j-1,k}), \quad \widetilde{r}_{j-1,k} = \widetilde{r}_{ji} \,(\mathrm{mod} T_{j-1,k}^2)$

endfor k

if $r_{(2i-1)2^{j-1}} = t_l$ and $l \in S$ then

call SOLUTION$(r_{ji}, \widetilde{r}_{ji}, a_l, n_l)$

$x_{m_l+k-1} = v_{n_l-k}/(k-1)!$ for $k = 1, \ldots, n_l$

$S_{j-1,2i-1} = \{t\colon\ t \in S_{ji} \text{ and } t < l\}$

$S_{j-1,2i} = \{t\colon\ t \in S_{ji} \text{ and } t > l\}$

endif

endif

endfor i
endfor j
Algorithm SOLUTION$(A(\lambda), B(\lambda), n, a)$
for $k = 0, 1, \ldots, n-1$
Compute $a_k = \frac{1}{(n+k)!}A^{(n+k)}(a)$, $b_k = \frac{1}{k!}B^{(k)}(a)$
Solve triangular Toeplitz linear system
$\text{tri}(a_0, a_1, \ldots, a_{n-1})\mathbf{v} = (b_0, b_1, \ldots, b_{n-1})^T$
end

where $\lceil x \rceil$ is the integer ceiling function of x and $\text{tri}(a_0, a_1, \ldots, a_{n-1})$ is a lower triangular Toeplitz matrix of the form

$$\text{tri}(a_0, a_1, \ldots, a_{n-1}) = \begin{pmatrix} a_0 & & & \\ a_1 & a_0 & & \\ \ddots & \ddots & \ddots & \\ a_{n-1} & \ddots & a_1 & a_0 \end{pmatrix}.$$

The proof of the correctness of FSCVLS is essentially the same as that of Algorithm 3.2 in [20]. The difference between two algorithms is that Algorithm 3.2 in [20] uses preprocessing of polynomials to compute polynomial divisions

$$r_{j-1,k} = r_{ji} \,(\text{mod} T_{j-1,k}), \quad \tilde{r}_{j-1,k} = \tilde{r}_{ji} \,(\text{mod} T_{j-1,k}^2).$$

Using preprocessing of polynomials makes polynomial divisions involved in FSCVLS fast, but we directly compute $r_{j-1,k}$ and $\tilde{r}_{j-1,k}$ by polynomial division in FSCVLS for clarity and convenience. Following the proofs of Propositions 4.1 and 4.2 in [20], we can show that the number of operations of FSCVLS, except operations of CONVERT, is $O(n \log n \log p)$ if fast polynomial multiplication and division are used. Therefore, the Algorithm FSCVLS needs $O(n \log^2 n)$ operations. Note that we can also construct an algorithm for confluent Vandermonde-like systems with a straightforward modification of Algorithm HERF in [22]. The computational complexity is also $O(n \log^2 n)$.

Many computations in Algorithm FSCVLS can be omitted in practice for some special points t_i. For example, if we use Theorem 3.2 to Chebyshev–Vandermonde linear systems with Chebyshev points, $t_i = \cos\frac{(2i+1)\pi}{2n}$ $i = 0, 1, \ldots, n-1$. The solution of the systems is deduced to FFT [21]. This yields an $O(n \log n)$ stable algorithm for this important case. The implementation and numerical examples are given in [21].

It is readily shown the existence of an $O(C(k)\, n \log^2 n)$ algorithm for the case where the polynomials $p_0(\lambda), p_1(\lambda), \ldots, p_{n-1}(\lambda)$ satisfy the k-term recurrence relation (9) and (10) with a straightforward modification of the method in this section, where $C(k)$ denotes the number of operations for multiplication of matrices of order $k-1$.

5. Conclusions. The $O(n \log^2 n)$ algorithm in §4 remains a theoretical one. Many open problems remain. From a practical point of view some $O(n^2)$ stable algorithms hopefully can be constructed for Vandermonde-like systems and inversion of Vandermonde-like matrices by using the results in §3.

Acknowledgment. I would like to thank Dr. Nicholas J. Higham and the referees for their useful comments that helped to improve the presentation.

REFERENCES

[1] A. C. AITKEN, *Determinants and Matrices*, Oliver and Boyd, Edinburgh, 1958.

[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] Å. BJÖRCK AND T. ELFVING, *Algorithms for confluent Vandermonde systems*, Numer. Math., 21 (1973), pp. 130–137.

[5] Å. BJÖRCK AND V. PEREYRA, *Solution of Vandermonde systems of equations*, Math. Comp., 24 (1970), pp. 893–903.

[6] D. CALVETTI AND L. REICHEL, *Fast inversion of Vandermonde-like matrices involving orthogonal polynomials*, BIT, 33 (1993), pp. 473–484.

[7] J. F. CANNY, E. KALTOFEN, AND L. YAGATI, *Solving systems of non-linear equations faster*, in Proc. ACM-SIGSAM 1989 Internat. Symp. Symbolic Algebraic Comput, New York, 1989, ACM, pp. 34–42.

[8] G. GALIMBERTI AND V. PEREYRA, *Solving confluent Vandermonde systems of Hermite type*, Numer. Math., 18 (1971), pp. 44–60.

[9] W. GAUTSCHI, *On inverse of Vandermonde and confluent Vandermonde matrices*, Numer. Math., 4 (1962), pp. 117–123.

[10] ———, *The condition of Vandermonde-like matrices involving orthogonal polynomials*, Linear Algebra Appl., 52/53 (1983), pp. 293–300.

[11] I. GOHBERG AND V. OLSHEVSKY, *Fast algorithms with preprocessing for matrix times vector multiplication*, J. Complexity, to appear.

[12] ———, *Fast inversion of Chebyshev–Vandermonde matrices*, Numer. Math., 67 (1994), pp. 71–92.

[13] 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.

[14] N. J. HIGHAM, *Error analysis of the Björck–Pereyra algorithm for solving Vandermonde systems*, Numer. Math., 50 (1987), pp. 613–632.

[15] ———, *Fast solution of Vandermonde-like systems involving orthogonal polynomials*, IMA J. Numer. Anal., 8 (1988), pp. 473–486.

[16] ———, *Stability analysis of algorithm for solving confluent Vandermonde-like systems*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 23–41.

[17] ———, *Iterative refinement enhances the stability of QR factorization methods for solving linear equations*, BIT, 31 (1991), pp. 447–468.

[18] J. KAUTSKY AND S. ELHAY, *Calculation of weights of interpolatory quadratures*, Numer. Math, 40 (1982), pp. 407–422.

[19] H. LU, *Computational complexity of Vandermonde linear systems*, Chinese Sci. Bull., 9 (1990), pp. 654–656. (In Chinese.)

[20] ———, *Fast solution of confluent Vandermonde linear systems*, SIAM J. Matrix Anal. Appl, 15 (1994), pp. 1277–1289.

[21] ———, *Generalized Trummer's problem with application to Vandermonde-like systems* I, SIAM J. Matrix Anal. Appl., 1995, submitted.

[22] ———, *Fast algorithms for confluent Vandermonde linear systems and generalized Trummer's problem*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 655–674.

[23] 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.

[24] N. MACON AND A. SPITZBART, *Inverses of Vandermonde matrices*, Amer. Math. Monthly, 65 (1958), pp. 95–100.

[25] L. REICHEL AND G. OPFER, *Chebyshev–Vandermonde systems*, Math. Comp., 57 (1991), pp. 703–721.

[26] W. P. TANG AND G. H. GOLUB, *The block decomposition of a Vandermonde matrix and its applications*, BIT, 21 (1981), pp. 505–517.

[27] J. F. TRAUB, *Associated polynomials and uniform methods for the solution of linear problems*, SIAM Rev., 8 (1966), pp. 277–301.

[28] L. VERDE-STAR, *Inverses of generalized Vandermonde matrices*, J. Math. Anal. Appl., 131 (1994), pp. 341–353.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 1, pp. 139–160, January 1996

008

A SCHUR METHOD FOR LOW-RANK MATRIX APPROXIMATION*

ALLE-JAN VAN DER VEEN†

Abstract. The usual way to compute a low-rank approximant of a matrix H is to take its singular value decomposition (SVD) and truncate it by setting the small singular values equal to 0. However, the SVD is computationally expensive. This paper describes a much simpler generalized Schur-type algorithm to compute similar low-rank approximants. For a given matrix H which has d singular values larger than ϵ, we find all rank d approximants $\hat{H}$ such that $H - \hat{H}$ has 2-norm less than ϵ. The set of approximants includes the truncated SVD approximation. The advantages of the Schur algorithm are that it has a much lower computational complexity (similar to a QR factorization), and directly produces a description of the column space of the approximants. This column space can be updated and downdated in an on-line scheme, amenable to implementation on a parallel array of processors.

Key words. matrix approximation, rank revealing factorizations, subspace estimation

AMS subject classifications. primary 47A58; secondary 15A60, 65F35

1. Introduction.

We consider the following problem: for a given matrix $H \in \mathbf{C}^{m\times n}$ and tolerance level $\epsilon \geq 0$, describe all matrices $\hat{H}$ such that

$$\text{(1)} \qquad \begin{array}{ll} \text{(a)} & \| H - \hat{H} \| \leq \epsilon\,, \\ \text{(b)} & \operatorname{rank}(\hat{H}) = d\,, \end{array}$$

where d is equal to the number of singular values of H that are larger than ϵ. ($\| \cdot \|$ denotes the matrix 2-norm.) Such a matrix $\hat{H}$ is a low-rank approximation of H in 2-norm. Note that there are no approximants of lower rank than d, and that we do not try to compute an approximant $\hat{H}$ of rank d that *minimizes* $\| H - \hat{H} \|$, but rather one in which the approximation error is limited. These approximants can be computed with significantly less effort.

One way to obtain an approximant that satisfies (1) is by computing a singular value decomposition (SVD) of H:

$$\text{(2)} \qquad H = U\Sigma V^* = [U_1 \quad U_2] \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^* \\ V_2^* \end{bmatrix}$$
$$(\Sigma_1)_{ii} > \epsilon\,, \quad (\Sigma_2)_{ii} \leq \epsilon\,.$$

Here, U and V are unitary matrices, and Σ is a diagonal matrix which contains the singular values σ_k of H. The matrices are partitioned such that Σ_1 contains the singular values that are larger than ϵ, and Σ_2 contains those that are smaller than ϵ. With this decomposition, a rank d approximant $\hat{H}$ is

$$\hat{H} = U_1 \Sigma_1 V_1^*\,.$$

* Received by the editors December 28, 1993; accepted for publication (in revised form) by P. van Dooren February 11, 1995.

† Delft University of Technology, Department of Electrical Engineering, 2628 CD Delft, The Netherlands. This research was performed in part while the author was on leave at Stanford University, Department of Computer Science/SCCM, Stanford, CA 94305-2140. This research was supported by the commission of the EC under the European Strategic Programme for Research and Development in Information Technology BRA program 6632(NANA2), by Advanced Research Project Agency contract F49620-91-C-0086, monitored by the Air Force Office of Scientific Research, and by National Science Foundation grant DMS-9205192.

This "truncated SVD" approximant is widely used and effectively obtained by setting the singular values that are smaller than ϵ equal to zero. It actually minimizes the approximation error: $\|H - \hat{H}\| = \sigma_{d+1} < \epsilon$, and is optimal in Frobenius norm as well. However, the SVD is expensive to compute, and much of the information that it reveals is not even used. Often, we are not so much interested in the individual singular vectors, but rather in the principal subspaces spanned by the columns of U_1 and V_1. As we show in this paper, it is indeed possible to obtain a parametrization of these subspaces and of all rank-d 2-norm approximants. All necessary information is gleaned from an implicit and nonunique factorization of $HH^* - \epsilon^2 I$ as

$$HH^* - \epsilon^2 I \;=\; BB^* - AA^*\,, \qquad [A, B] \text{ invertible}\,,$$

which is provided by a "hyperbolic" QR factorization of $[\epsilon I \;\; H]$. Such a factorization is similar to an ordinary QR factorization, except that it uses a matrix that is unitary with respect to an indefinite inner product. Under additional regularity assumptions on H, this factorization may be computed using a generalized Schur-type method, which requires only about $1/2\, m^2 n$ operations (elementary rotations) for a matrix of size $m \times n$.[1] The column span of the approximants is directly obtained from B and A: it is proven that all suitable column spans are given by the range of

$$B - AM\,, \qquad \|M\| \leq 1\,.$$

The computation of an approximant itself requires an additional $n \times n$ matrix inversion or a projection of H onto this subspace.

Continuing efforts on SVD algorithms have reduced the computational complexity of the SVD to be mainly that of reducing a matrix to a bidiagonal form, which is not much more than the complexity of a QR factorization. However, a remaining disadvantage of the SVD in demanding applications is that it is difficult to update the decomposition for a growing number of columns of H. Indeed, there are important applications in signal processing (e.g., adaptive beamforming, model identification, adaptive least squares filters) that require on-line estimation of the principal subspace for growing values of n. A number of other methods have been developed that alleviate the computational requirements, yet retain important information such as numerical rank and principal subspaces. Some of these techniques are the URV decomposition [1], which is a rank revealing form of a complete orthogonal decomposition [2], and the rank revealing QR decomposition (RRQR), [3]–[8]; see [8] for a review. The URV algorithm is iterative and requires estimates of the conditioning of certain submatrices at every step of the iteration. This is a global and data-dependent operation—not a very attractive feature. The SVD and URV decomposition can be updated [9], [1], which is still an iterative process, although it has been shown recently that a simpler scheme is feasible if the updating vectors satisfy certain stationarity assumptions [10], [11]. An initial computation of the RRQR consists of an ordinary QR, followed by an iteration that makes the decomposition rank revealing. As a one-sided decomposition, the RRQR is easier to update than an SVD, but also requires (incremental) condition estimations at each updating step.

An important aspect of the hyperbolic QR factorization is that, similar to a QR factorization, it can be updated very straightforwardly for growing n. The rank of the approximants (the dimension of the principal subspace) is updated as part of the process without any condition estimation. Nonetheless, the method provides an exact

[1] To set our mind, we usually assume that $m \leq n$, but all results remain true when $m > n$.

error bound on the subspace estimates, in terms of the associated matrix approximation error ϵ. Similar to the URV and the RRQR, the value of ϵ must be fixed beforehand. Another aspect is that the Schur method for computing the hyperbolic QR factorization is a parallel algorithm with only local and regular data dependencies and is very straightforward to implement on a systolic array of processors. The structure of the array is the same as that of the well-known Gentleman–Kung triangular array for the computation of a QR factorization using Givens rotations [12]. One negative aspect of the Schur algorithm is that it uses hyperbolic rotations, which are potentially unbounded and could make the approximation scheme less robust than the SVD or the URV. This is more a property of the implementation than of the overall technique: it occurs only if certain submatrices have a singular value close to ϵ, and in these cases, a simple local pivoting scheme suffices to virtually eliminate any risk of breakdown. Alternatively, we may derive factorization schemes that minimize the number of hyperbolic rotations.

1.1. Connections. Schur methods *an sich* are well known. Originally, Schur devised this algorithm to test whether a polynomial is bounded within the complex unit disc [13]. Schur algorithms occur in certain constrained interpolation problems (viz. [14], [15]), rational approximation by analytic functions [16], factorization and inversion of positive definite and later also indefinite Toeplitz matrices (viz. the review in [17]), and have been generalized in a number of senses to non-Toeplitz matrices. A generalization that comes close to the description here is by Dewilde and Deprettere [18] for Schur-parametrizations and band approximations of positive definite matrices, and by Diepold and Pauli [19], [20] for indefinite matrix cases. In the linear algebra community, similar generalized Schur methods, known under other names, have been used for the solution of positive definite systems [21], [22] and for the downdating of QR and Cholesky factorizations [23], although the main emphasis has been on hyperbolic Householder transformations for the same purposes [24]–[28]. The HR-decomposition in [24], later known as the hyperbolic QR factorization (e.g., [29]), is in fact precisely the tool we need.

The present application to low rank matrix approximation has been unknown so far. It is derived as a special case of a new theory for model reduction of time-varying systems [30], [31]. The time-invariant counterpart (approximation of a Hankel matrix by one of lower rank) has been known for more than a decade and is widely used in systems theory and control for model reduction and for solving H_∞-control problems (viz. [32], [33]). This theory goes back to the work of Adamjan, Arov, and Krein, in 1971, on the solution of the Schur–Takagi interpolation problem [16].

1.2. Structure of the paper. The remainder of the paper is organized as follows. Section 2 is a review of those properties of J-unitary matrices that we need in this paper, such as the existence of a hyperbolic QR factorization. In §3, this factorization is used to prove a basic version of the approximation theorem, and we introduce a parametrization of all 2-norm approximants of rank d. Some values of the parameters that lead to interesting approximants are discussed. The computation of the factorization is the topic of §4. It is shown that the factorization can be computed using a Schur-type algorithm, although certain extra conditions must be imposed on the matrix H. We derive necessary and sufficient conditions so that the algorithm does not break down, and discuss some simple pivoting schemes to alleviate these conditions. Finally, in §5, the algorithm is applied to a test case to show the behavior of some of the approximants and the effectiveness of the pivoting scheme.

1.3. Notation. The superscript $(\cdot)^*$ denotes complex conjugate transposition, $\mathcal{R}(A)$ is the column span (range) of the matrix A, I_m is the $m \times m$ identity matrix, and $0_{m\times n}$ is an $m \times n$ matrix with zero entries. $A^\dagger$ denotes the pseudoinverse of A. At some point, we will use the notation $A_{[i]}$ to be the submatrix of A consisting of its first row through its ith row, and $A_{[i,k]}$ to denote the first k columns of $A_{[i]}$.

2. J-unitary matrices. We review the definition and some properties of J-unitary matrices, most of which are well known. A matrix Θ is *J-unitary* if it satisfies

$$\Theta^* J \Theta = J\,, \qquad \Theta J \Theta^* = J\,, \qquad \text{where } J = \begin{bmatrix} I & \\ & -I \end{bmatrix}. \tag{3}$$

J is a *signature matrix*; the identity matrices need not have equal sizes. Θ is necessarily a square matrix and invertible as well: $\Theta^{-1} = J\Theta^* J$. We partition Θ according to J as

$$\Theta = \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{bmatrix}. \tag{4}$$

The J-unitarity of Θ implies among others (a.o.) $\Theta_{22}^*\Theta_{22} = I + \Theta_{12}^*\Theta_{12}$. From this expression, we derive in turn the properties

$$\begin{array}{lll} \text{(a)} & \Theta_{22} \text{ is invertible,} & \text{(c)} \quad \| \Theta_{12}\Theta_{22}^{-1} \| < 1\,, \\ \text{(b)} & \| \Theta_{22}^{-1} \| \le 1\,, & \text{(d)} \quad \| \Theta_{11} - \Theta_{12}\Theta_{22}^{-1}\Theta_{21} \| \le 1\,. \end{array} \tag{5}$$

Indeed, (a) follows because Θ_{22} is also square, (b) is obtained from

$$\Theta_{22}^{-*}\Theta_{22}^{-1} + (\Theta_{22}^{-*}\Theta_{12}^*)(\Theta_{12}\Theta_{22}^{-1}) = I\,, \tag{6}$$

so that $\Theta_{22}^{-*}\Theta_{22}^{-1} \le I$, and (c) results from the same expression because $\Theta_{22}^{-*}\Theta_{22}^{-1} > 0$. By elementary algebra, one verifies that the matrix

$$\begin{bmatrix} \Theta_{11} - \Theta_{12}\Theta_{22}^{-1}\Theta_{21} & -\Theta_{12}\Theta_{22}^{-1} \\ \Theta_{22}^{-1}\Theta_{21} & \Theta_{22}^{-1} \end{bmatrix}$$

is in fact unitary, which implies (d).

Similarly, we have

$$\begin{array}{lll} \text{(a)} & \Theta_{11} \text{ is invertible,} & \text{(c)} \quad \| \Theta_{11}^{-1}\Theta_{12} \| < 1\,, \\ \text{(b)} & \| \Theta_{11}^{-1} \| \le 1\,, & \text{(d)} \quad \| \Theta_{22} - \Theta_{21}\Theta_{11}^{-1}\Theta_{12} \| \le 1\,. \end{array} \tag{7}$$

Another important property of J-unitary matrices is the preservation of the J-inner product. Suppose that A, B, C, D are matrices, related as $[C \;\; D] = [A \;\; B]\Theta$. The J-unitarity of Θ implies

$$\begin{aligned} AA^* - BB^* &= [A \;\; B] J [A \;\; B]^* \\ &= [A \;\; B] \Theta J \Theta^* [A \;\; B]^* \\ &= CC^* - DD^*\,. \end{aligned} \tag{8}$$

Motivated by this equation, we say that J associates a positive signature to the columns of A, a negative signature to the columns of B, and likewise for C and D. We will sometimes denote this in equations by writing $+$ and $-$ over A and B.

So far, the signature matrix J in (3) was sorted: the diagonal first has all the positive entries, then the negative ones. We will at some point also need *unsorted*

signature matrices $\tilde{J}$, which is any diagonal matrix with diagonal entries equal to $+1$ or -1. As a generalization of the definition in (3), we will say that a matrix $\tilde{\Theta}$ is $(\tilde{J}_1, \tilde{J}_2)$-unitary[2] with respect to signature matrices $\tilde{J}_1, \tilde{J}_2$ if it satisfies

$$\tilde{\Theta}^* \tilde{J}_1 \tilde{\Theta} = \tilde{J}_2 , \qquad \tilde{\Theta} \tilde{J}_2 \tilde{\Theta}^* = \tilde{J}_1 . \tag{9}$$

Again, $\tilde{\Theta}$ is square and invertible: $\tilde{\Theta}^{-1} = \tilde{J}_2 \tilde{\Theta}^* \tilde{J}_1$. Sylvester's law of inertia claims that the number of positive entries in $\tilde{J}_1$ must be equal to the number of positive entries in $\tilde{J}_2$, and similarly for the negative entries. An unsorted signature matrix $\tilde{J}_1$ can always be sorted by a permutation matrix Π_1 such that $J_1 = \Pi_1 \tilde{J}_1 \Pi_1^*$ is sorted. If $J_2 = \Pi_2 \tilde{J}_2 \Pi_2^*$ is also sorted and $\tilde{\Theta}$ satisfies (9), then actually $J_1 = J_2 =: J$, and $\Theta := \Pi_1 \tilde{\Theta} \Pi_2^*$ is J-unitary in the sorted sense of (3). The permutation, and hence Θ, is not unique, but this is usually irrelevant. We work with Θ rather than $\tilde{\Theta}$ in cases where its partitioning into submatrices, as in (4), is important, but the properties (5) are independent of precisely which Θ is chosen.

A matrix A is said to be *$\tilde{J}$-nonsingular*, with respect to a certain signature matrix $\tilde{J}$, if $A\tilde{J}A^*$ is nonsingular. It is immediate that if A is $\tilde{J}_1$-nonsingular and $\tilde{\Theta}$ is a $(\tilde{J}_1, \tilde{J}_2)$-unitary matrix, then $A\tilde{\Theta}$ is $\tilde{J}_2$-nonsingular. The following basic result claims that J-nonsingular matrices can be factored.

THEOREM 2.1. *A matrix $A : m \times (m+n)$ is $\tilde{J}_1$-nonsingular if and only if there exists a signature matrix $\tilde{J}_2$ and a $(\tilde{J}_1, \tilde{J}_2)$-unitary matrix $\tilde{\Theta}$ such that*

$$A\tilde{\Theta} = [X \quad 0_{m \times n}] , \qquad X : m \times m, \text{ invertible.} \tag{10}$$

Proof. Assume that A is $\tilde{J}$-nonsingular. Then we can factor $A\tilde{J}_1 A^*$ as

$$A\tilde{J}_1 A^* = X \tilde{J}' X^* , \qquad X : m \times m, \text{ invertible,}$$

for some $m \times m$ signature matrix $\tilde{J}'$. This factorization exists and can in principle be computed from an LDU factorization with pivoting, or from an eigenvalue decomposition of $A\tilde{J}_1 A^*$. Since A is $\tilde{J}_1$-nonsingular, it is also nonsingular in the ordinary sense, so that there exists a matrix $T : (m+n) \times m$, such that $AT = X$. T is not unique. Because X is invertible, we can take

$$T = \tilde{J}_1 A^* (A \tilde{J}_1 A^*)^{-1} X .$$

Using $(A\tilde{J}_1 A^*)^{-1} = X^{-*} \tilde{J}' X^{-1}$, it is directly verified that this T satisfies $T^* \tilde{J}_1 T = \tilde{J}'$. The remainder of the proof is technical and shows that T can be extended to a square, J-unitary matrix. From $T^* \tilde{J}_1 T = \tilde{J}'$ it follows that the m columns of $\tilde{J}_1 T$ are linearly independent. Taking any basis of the orthogonal complement of the range of $\tilde{J}_1 T$ gives a matrix K, with n independent columns, such that $T^* \tilde{J}_1 K = 0$. The matrix $[T \;\; K]$ is invertible because it is square and its kernel is empty:

$$[T \quad K] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0 \;\; \Rightarrow \;\; T^* \tilde{J}_1 [T \quad K] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$
$$\Rightarrow \;\; x_1 = 0$$

and, subsequently, it also follows that $x_2 = 0$. It remains to normalize the columns of K. Put

$$[T \quad K]^* \tilde{J}_1 [T \quad K] = \begin{bmatrix} \tilde{J}' & \\ & N \end{bmatrix} .$$

[2] We will sometimes generically write J-unitary to avoid being overly detailed.

N is nonsingular because $[T \quad K]$ is invertible, and we can factor it as $N = R^* \tilde{J}'' R$. Put

$$\tilde{\Theta} = [T \quad KR^{-1}], \qquad \tilde{J}_2 = \operatorname{diag}[\tilde{J}', \tilde{J}''].$$

Then $\tilde{\Theta}$ is $(\tilde{J}_1, \tilde{J}_2)$-unitary and satisfies (10). □

A recursive application of this theorem proves that, under additional conditions, A has a triangular factorization.

COROLLARY 2.2. *Let A: $m \times (m+n)$ be $\tilde{J}_1$-nonsingular. Denote by $A_{[i]}$ the submatrix of A, consisting of its first i rows. Then there exist a signature matrix $\tilde{J}_2$ and a $(\tilde{J}_1, \tilde{J}_2)$-unitary matrix $\tilde{\Theta}$ such that*

$$A\tilde{\Theta} = [X \quad 0_{m\times n}], \qquad X\text{: } m \times m, \text{ lower triangular, invertible}$$

if and only if $A_{[i]}$ is $\tilde{J}_1$-nonsingular, for $i = 1, \ldots, m$. If the diagonal entries of X are chosen to be positive, then X is unique.

Such a factorization was proven in [24] for square matrices A and upper triangular X, but this result extends directly to the rectangular case. In [24], it was called the HR-decomposition, and it is also known as the hyperbolic QR factorization [29].

3. Approximation theory.

3.1. Central approximant. For a given $m \times n$ data matrix H and threshold ϵ, denote the SVD of H as in (2). Suppose that d singular values of H are larger than ϵ, and that none of them are equal to ϵ. Our approximation theory is based on an implicit factorization of

$$HH^* - \epsilon^2 I = BB^* - AA^*. \tag{11}$$

This is a Cholesky factorization of an indefinite Hermitian matrix. A and B are chosen to have full column rank. They are not unique, but by Sylvester's inertia law, their dimensions are well defined. Using the SVD of H, we obtain one possible decomposition as

$$HH^* - \epsilon^2 I = U_1(\Sigma_1^2 - \epsilon^2 I)U_1^* + U_2(\Sigma_2^2 - \epsilon^2 I)U_2^*,$$

where the first term is positive semidefinite and has rank d, and the second term is negative semidefinite and has rank $m - d$. Hence, B has d columns, and A has $m - d$ columns.

To obtain an implicit factorization that avoids computing HH^*, we make use of Theorem 2.1.

THEOREM 3.1. *Let $H : m \times n$ have d singular values larger than ϵ and none equal to ϵ. Then there exists a J-unitary matrix Θ such that*

$$[\epsilon I_m \quad H]\,\Theta = [A' \quad B'], \tag{12}$$

where $A' = [A \quad 0_{m\times d}]$, $B' = [B \quad 0_{m\times n-d}]$, A: $m \times (m-d)$, B: $m \times d$, and $[A \quad B]$ is of full rank.

Proof. The matrix $[\epsilon I_m \quad H]$ is J-nonsingular; by assumption, $\epsilon^2 I - HH^*$ has d negative, $m - d$ positive, and no zero eigenvalues. Hence Theorem 2.1 implies that there exists $\tilde{\Theta} : [\epsilon I_m \quad H]\tilde{\Theta} = [X \quad 0_{m\times n}]$. The columns of X are the columns of $[A, B]$, in some permuted order, where A, B correspond to columns of X that have a positive or negative signature, respectively. After sorting the columns of $[X \quad 0]$ according to their signature, (12) results. □

Note that, by the preservation of J-inner products (8), equation (12) implies (11). From the factorization (12), we can immediately derive a 2-norm approximant satisfying the conditions in (1). To this end, partition Θ according to its signature J into 2×2 blocks, such as in (4).

THEOREM 3.2. *Let $H : m \times n$ have d singular values larger than ϵ and none equal to ϵ. Define the factorization $[\epsilon I_m \;\; H]\Theta = [A' \;\; B']$ as in Theorem* 3.1. *Then*

$$\hat{H} = B'\Theta_{22}^{-1} \tag{13}$$

is a rank-d *approximant such that $\| H - \hat{H} \| < \epsilon$.*

Proof. $\hat{H}$ is well defined because Θ_{22} is invertible (5a). It has rank d because $B' = [B \;\; 0]$ has rank d. By (12), $B' = \epsilon I\Theta_{12} + H\Theta_{22}$; hence $H - \hat{H} = -\epsilon\Theta_{12}\Theta_{22}^{-1}$. It remains to use the fact that $\Theta_{12}\Theta_{22}^{-1}$ is contractive (5c). □

We mentioned in the introduction that the column span (range) of the approximant is important in signal processing applications. From Theorem 3.2, it is seen that this column span is equal to that of B: it is directly produced by the factorization. However, note that $[A \;\; B]$ in (12) is not unique: for any J-unitary matrix Θ_1, $[A_1 \;\; B_1] = [A \;\; B]\Theta_1$ also satisfies $\epsilon^2 I - HH^* = A_1A_1^* - B_1B_1^*$, and could also have been produced by the factorization. For example, for some choices of Θ_1, we will have $\mathcal{R}(B) = \mathcal{R}(U_1)$ and $\mathcal{R}(A) = \mathcal{R}(U_2)$. Using Θ_1, we can find more approximants. A parametrization of all 2-norm approximants is the topic of the following section.

3.2. Parametrization of all approximants. We will now give a formula of all possible 2-norm approximants $\hat{H}$ of H of rank equal to d; there are no approximants of rank less than d. The formula is in terms of a chain fraction description. Similar formulas frequently occur in constrained interpolation theory (see, e.g., [14], [15], and references therein).

The set of all minimal-rank 2-norm approximants will be parametrized by matrices $S_L : m \times n$, with 2×2 block partitioning as

$$S_L = \begin{matrix} \\ m-d \\ d \end{matrix} \begin{matrix} \begin{matrix} d & n-d \end{matrix} \\ \begin{bmatrix} (S_L)_{11} & (S_L)_{12} \\ (S_L)_{21} & (S_L)_{22} \end{bmatrix} \end{matrix}, \tag{14}$$

and satisfying the requirements

$$\begin{array}{ll} \text{(i)} & \text{contractive: } \| S_L \| \leq 1\,, \\ \text{(ii)} & \text{block lower: } (S_L)_{12} = 0\,. \end{array} \tag{15}$$

The first condition on S_L will ensure that $\| H - \hat{H} \| \leq \epsilon$, whereas the second condition is required to have $\hat{H}$ of rank d.

THEOREM 3.3. *With the notation and conditions of Theorem* 3.2, *all* rank-d 2-*norm approximants $\hat{H}$ of H are given by*

$$\hat{H} \;=\; (B' - A'S_L)(\Theta_{22} - \Theta_{21}S_L)^{-1}\,,$$

where S_L satisfies (i): $\| S_L \| \leq 1$ *and* (ii): $(S_L)_{12} = 0$. *The approximation error is*

$$S := H - \hat{H} \;=\; \epsilon(\Theta_{11}S_L - \Theta_{12})(\Theta_{22} - \Theta_{21}S_L)^{-1}\,. \tag{16}$$

Proof. The proof is given in the appendix. □

By this theorem, an estimate of the principal subspace of H is given by $\mathcal{R}(\hat{H}) = \mathcal{R}(B' - A'S_L) = \mathcal{R}(B - A(S_L)_{11})$, for any valid choice of S_L. Note that $(S_L)_{11}$ ranges over the set of all contractive $(m-d) \times d$ matrices, so that all suitable principal subspace estimates are given by

$$\mathcal{R}(B - AM), \qquad \| M \| \leq 1 .$$

The distance of a subspace estimate with the actual principal subspace, $\mathcal{R}(U_1)$, is measured only implicitly, in the sense that there exists an approximant $\hat{H}$ with this column span that is ϵ-close to H. Actually, for each subspace estimate there are many such approximants, since the subspace estimate only depends on $(S_L)_{11}$, whereas the approximant also depends on $(S_L)_{21}$ and $(S_L)_{22}$.

The choice of a particular approximant $\hat{H}$, or subspace estimate $\mathcal{R}(\hat{H})$, boils down to a suitable choice of the parameter S_L. Various choices are interesting.

1. The approximant $\hat{H}$ in Theorem 3.2 is obtained by taking $S_L = 0$. This approximant is the most simple to compute; the principal subspace estimate is equal to the range of B. The approximation error is given by $\epsilon \| \Theta_{12}\Theta_{22}^{-1} \|$. Note that, even if all nonzero singular values of H are larger than ϵ so that it is possible to have $\hat{H} = H$, the choice $S_L = 0$ typically does not give zero error. Hence, this simple choice of S_L could lead to "biased" estimates. This is confirmed in the simulation example in §5 and occurs in cases where σ_d is close to ϵ.

2. As the truncated SVD solution satisfies the requirements, there is an S_L that yields this particular solution and minimizes the approximation error. However, computing this S_L requires an SVD or a hyperbolic SVD [29].

3. It is sometimes possible to obtain a uniform approximation error. First write (16) in a more implicit form,

$$\begin{bmatrix} \epsilon^{-1}SG \\ -G \end{bmatrix} = \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{bmatrix} \begin{bmatrix} S_L \\ -I_n \end{bmatrix},$$

where G is an invertible $n \times n$ matrix. This equation implies

$$G^*(\epsilon^{-2}S^*S - I_n)G \;=\; S_L^*S_L - I_n .$$

Suppose $m \leq n$. If we can take S_L to be an isometry, $S_L S_L^* = I_m$, then $\operatorname{rank}(S_L^*S_L - I_n) = n - m$. It follows that $\epsilon^{-1}S$ must also be an isometry, so that all singular values of $S = H - \hat{H}$ are equal to ϵ: the approximation error is uniform. S_L can be an isometry and have $(S_L)_{12} = 0$ only if $d \geq m - d$, i.e., $d \geq m/2$. In that case, we can take, for example, $S_L = [I_m \;\; 0]$. This approximant might have relevance in signal processing applications where a singular data matrix is distorted by additive uncorrelated noise with a covariance matrix $\sigma^2 I_m$.

4. If we take $S_L = \Theta_{11}^{-1}\Theta_{12}$, then we obtain $\hat{H} = H$ and the approximation error is zero. Although this S_L is contractive (viz. (7)), it does not satisfy the condition $(S_L)_{12} = 0$, unless $d = m$ or $d = n$. Simply putting $(S_L)_{12} = 0$ might make the resulting S_L noncontractive. To satisfy both conditions on S_L, a straightforward modification sets

$$S_L = \Theta_{11}^{-1}\Theta_{12} \begin{bmatrix} I_d & \\ & 0_{n-d} \end{bmatrix} = \begin{bmatrix} (\Theta_{11}^{-1}\Theta_{12})_{11} & 0 \\ (\Theta_{11}^{-1}\Theta_{12})_{21} & 0 \end{bmatrix}. \tag{17}$$

The corresponding approximant is

$$\hat{H}^{(1)} := (B' - A'\Theta_{11}^{-1}\Theta_{12}\left[\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right])(\Theta_{22} - \Theta_{21}\Theta_{11}^{-1}\Theta_{12}\left[\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right])^{-1}, \tag{18}$$

and the corresponding principal subspace estimate is given by the range of

$$B^{(1)} := B - A(\Theta_{11}^{-1}\Theta_{12})_{11}\,. \tag{19}$$

Both the subspace estimate and the approximant itself can be computed by a Schur complement formula. The subspace estimate is "unbiased" in the sense discussed below, and is usually quite accurate when σ_d is not very close to ϵ, as is shown in simulation examples (§5). The approximation error is determined by

$$S = H - \hat{H}^{(1)} = \epsilon\Theta_{12}\begin{bmatrix} 0_d & \\ & -I_{n-d} \end{bmatrix}(\Theta_{22} - \Theta_{21}\Theta_{11}^{-1}\Theta_{12}\left[\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right])^{-1}\,. \tag{20}$$

This shows that the rank of S is at most equal to $\min(m, n-d)$. If $m = n$, then the rank of S is $m - d$, i.e., the error has the same rank as a truncated SVD solution would give.

5. In order to improve the approximation error, we propose to take $(S_L)_{11} = (\Theta_{11}^{-1}\Theta_{12})_{11}$, as in the previous item, and use the freedom provided by $(S_L)_{21}$ and $(S_L)_{22}$ to minimize the norm of the error. The subspace estimate is only determined by $(S_L)_{11}$ and is the same as before. Instead of minimizing in terms of S_L, which involves a nonlinear function and a contractivity constraint, we make use of the fact that we already know the column span of the approximant: we are looking for $\hat{H} = B^{(1)}N$, with $B^{(1)}$ given by (19) and $N : d \times n$ a minimizer of

$$\min_N \| H - B^{(1)}N \|\,.$$

A solution is given by $N = B^{(1)\dagger}H$, and the resulting approximant is

$$\begin{aligned} \hat{H} &= B^{(1)}B^{(1)\dagger}H \\ &=: \hat{H}^{(2)}\,, \end{aligned} \tag{21}$$

the projection of H onto $\mathcal{R}(B^{(1)})$. Although we do not compute the S_L to which this approximant corresponds, the residual error is guaranteed to be less than or equal to ϵ because it is at most equal to the norm of S in (20). Hence, there will be some S_L that satisfies the constraints, although we never compute it explicitly. For this S_L, the rank of the residual error is always at most equal to $m-d$, the rank of $I_m - B^{(1)}B^{(1)\dagger}$.

One other important feature of the subspace estimate $B^{(1)}$ in (19) is that it is *unbiased*, in the following sense.

LEMMA 3.4. $\mathcal{R}(B^{(1)}) \subset \mathcal{R}(H)$.

Proof. From $[(A\ \ 0)\ \ (B\ \ 0)] = [A'\ \ B'] = [\epsilon I\ \ H]\Theta$, we have

$$\begin{aligned} [A \quad 0] &= \epsilon\Theta_{11} + H\Theta_{21}, \\ [B \quad 0] &= \epsilon\Theta_{12} + H\Theta_{22}. \end{aligned}$$

Hence

$$\begin{aligned} [B^{(1)}\ \ 0] &= [B \quad 0] \ - \ [A \quad 0]\Theta_{11}^{-1}\Theta_{12}\begin{bmatrix} I & \\ & 0 \end{bmatrix} \\ &= (\epsilon\Theta_{12} + H\Theta_{22}) \ - \ (\epsilon\Theta_{11} + H\Theta_{21})\Theta_{11}^{-1}\Theta_{12}\begin{bmatrix} I & \\ & 0 \end{bmatrix} \\ &= H(\Theta_{22} - \Theta_{21}\Theta_{11}^{-1}\Theta_{12})\begin{bmatrix} I & \\ & 0 \end{bmatrix} + H\Theta_{22}\begin{bmatrix} 0 & \\ & I \end{bmatrix} + \epsilon\Theta_{12}\begin{bmatrix} 0 & \\ & I \end{bmatrix} \end{aligned}$$

so that

$$B^{(1)} = H(\Theta_{22} - \Theta_{21}\Theta_{11}^{-1}\Theta_{12}) \begin{bmatrix} I \\ 0 \end{bmatrix}. \qquad \square$$

With (7d), we also have

$$\| B^{(1)} \| \leq \| H \|. \tag{22}$$

This shows that, although J-unitary matrices may be large, this particular subspace estimate is bounded in norm by the matrix it was derived from.

4. Computation of Θ. In this section, we consider the actual construction of a J-unitary matrix Θ such that

$$[\epsilon I \quad H]\Theta = [A' \quad B'], \qquad J = \begin{bmatrix} I_m & \\ & -I_n \end{bmatrix}.$$

The proof of Theorem 2.1 provides a technique to compute Θ, but the construction is global and not really attractive. We are looking for algorithms that do not square the data and that allow easy updating of the factorization as more and more columns of H are included (growing n). Θ will be computed in two steps: $\Theta = \tilde{\Theta}\Pi$, where $\tilde{\Theta}$ is a $(J, \tilde{J}_2)$-unitary matrix with respect to J and an unsorted signature $\tilde{J}_2$ and is such that

$$\begin{matrix} + & - & & +/- & +/- \end{matrix}$$
$$[\epsilon I_m \quad H]\tilde{\Theta} = [\, X \quad 0_{m\times n}], \qquad X : m \times m. \tag{23}$$

Π is any permutation matrix such that $\Pi\tilde{J}_2\Pi^* = J$ is a sorted signature matrix. The latter factorization can be viewed as a hyperbolic QR factorization, in case X has a triangular form, and can be computed in a number of ways. Hyperbolic Householder transformations have been employed for this purpose [24], [29], zeroing full rows at each step, but the most elementary way is to use elementary rotations to create one zero entry at a time, like Givens rotations are used for QR factorizations. Such techniques are known as (generalized) Schur algorithms because of their similarity to the Schur method for Toeplitz matrices. In contrast to hyperbolic Householder transformations, they allow for straightforward updating and downdating. The main differences with the QR factorization and also with the usual definite Schur algorithms (for which $\epsilon^2 I - HH^* > 0$) are that here the basic operations are J-unitary elementary rotations of up to six different types and we must keep track of signatures to determine which type to use.

The recursive construction of Θ in this way is not always possible, unless extra conditions on the singular values of certain submatrices of H are posed. This is a well-known complication from which all indefinite Schur and hyperbolic Householder methods suffer and that, in its ultimate generality, can be treated only by global matrix operations (as in [19], [20], or as in the proof of Theorem 2.1, which uses an altogether different algorithm). The exceptions occur only for specific cases, and simple pivoting schemes (column or row permutations) are virtually always adequate to eliminate this problem. We will briefly go into these aspects in §4.5.

4.1. Elementary rotations. At an elementary level, we are looking for 2×2 matrices $\tilde{\theta}$ such that, for given scalars a, b,

$$[a \quad b]\tilde{\theta} = [x \quad 0],$$

where x is some resulting scalar. The matrices $\tilde{\theta}$ are J-unitary, but with respect to unsorted signature matrices $\tilde{j}_1, \tilde{j}_2$:

$$(\tilde{\theta})^* \tilde{j}_1 \tilde{\theta} = \tilde{j}_2 , \qquad \tilde{\theta} \tilde{j}_2 (\tilde{\theta})^* = \tilde{j}_1 .$$

The signature matrix $\tilde{j}_1$ is specified along with a, b and signifies the signature of $[a \;\; b]$; $\tilde{j}_2$ is a resulting signature matrix to be computed along with $\tilde{\theta}$ and x, and will be the resulting signature of $[x \;\; 0]$. There are two rules that determine $\tilde{j}_2$. From the J-unitarity of $\tilde{\theta}$, we have that

$$[a \;\; b] \tilde{j}_1 [a \;\; b]^* = x (\tilde{j}_2)_{11} x^*$$

$$\Rightarrow (\tilde{j}_2)_{11} = \operatorname{sign}\Big([a \;\; b] \tilde{j}_1 [a \;\; b]^* \Big) .$$

We must assume at this point that the expression in brackets is not zero, so that $(\tilde{j}_2)_{11}$ is either $+1$ or -1. The second diagonal entry of $\tilde{j}_2$ then follows from the inertia rule: by congruence, the number of positive entries in $\tilde{j}_1$ is equal to the number of positive entries in $\tilde{j}_2$, and similarly for the negative entries.

Depending on the signatures, we choose one of the following types of elementary $(\tilde{j}_1, \tilde{j}_2)$-unitary rotations (taking $|s|^2 + |c|^2 = 1$ throughout).

1. $\tilde{j}_1 = \begin{bmatrix} 1 & \\ & -1 \end{bmatrix}$, $\tilde{j}_2 = \begin{bmatrix} 1 & \\ & -1 \end{bmatrix}$ $\Rightarrow$ $\tilde{\theta} = \begin{bmatrix} 1 & -s \\ -s^* & 1 \end{bmatrix} \frac{1}{c}$.
2. $\tilde{j}_1 = \begin{bmatrix} 1 & \\ & -1 \end{bmatrix}$, $\tilde{j}_2 = \begin{bmatrix} -1 & \\ & 1 \end{bmatrix}$ $\Rightarrow$ $\tilde{\theta} = \begin{bmatrix} -s^* & 1 \\ 1 & -s \end{bmatrix} \frac{1}{c}$.
3. $\tilde{j}_1 = \begin{bmatrix} -1 & \\ & 1 \end{bmatrix}$, $\tilde{j}_2 = \begin{bmatrix} 1 & \\ & -1 \end{bmatrix}$ $\Rightarrow$ $\tilde{\theta} = \begin{bmatrix} -s^* & 1 \\ 1 & -s \end{bmatrix} \frac{1}{c}$.
4. $\tilde{j}_1 = \begin{bmatrix} -1 & \\ & 1 \end{bmatrix}$, $\tilde{j}_2 = \begin{bmatrix} -1 & \\ & 1 \end{bmatrix}$ $\Rightarrow$ $\tilde{\theta} = \begin{bmatrix} 1 & -s \\ -s^* & 1 \end{bmatrix} \frac{1}{c}$.
5. $\tilde{j}_1 = \begin{bmatrix} 1 & \\ & 1 \end{bmatrix}$, $\tilde{j}_2 = \begin{bmatrix} 1 & \\ & 1 \end{bmatrix}$ $\Rightarrow$ $\tilde{\theta} = \begin{bmatrix} c^* & -s \\ s^* & c \end{bmatrix}$.
6. $\tilde{j}_1 = \begin{bmatrix} -1 & \\ & -1 \end{bmatrix}$, $\tilde{j}_2 = \begin{bmatrix} -1 & \\ & -1 \end{bmatrix}$ $\Rightarrow$ $\tilde{\theta} = \begin{bmatrix} c^* & -s \\ s^* & c \end{bmatrix}$.

The first case is the standard elementary hyperbolic rotation. The next three cases are obtained from this case by row and column permutations. Cases 5 and 6 are not hyperbolic, but ordinary elliptic rotations; however they are $(\tilde{j}_1, \tilde{j}_2)$-unitary nonetheless. These six cases are enough to consider because every possible signature pair $(\tilde{j}_1, \tilde{j}_2)$ is covered. With $\tilde{j}_1$ and $\tilde{j}_2$ known, we select the appropriate type of rotation matrix, and the rotation parameters s and c follow subsequently from the equation $[a \;\; b]\tilde{\theta} = [x \;\; 0]$ as

Cases 1, 4 ($\|a\| > \|b\|$):	$s = b/a$,	$c = (1 - s^* s)^{1/2}$
Cases 2, 3 ($\|a\| < \|b\|$):	$s = a/b$,	$c = (1 - s^* s)^{1/2}$
Cases 5, 6:	$s = b(a^* a + b^* b)^{-1/2}$,	$c = (1 - s^* s)^{1/2}$.

4.2. Indefinite Schur algorithm. To compute the factorization (23), elementary rotations $\tilde{\theta}$ are embedded in plane rotations $\tilde{\Theta}_{(i,k)}$ which are applied to the columns of $[\epsilon I \;\; H]$ in the same way as Givens rotations are used for computing a

QR factorization. Each plane rotation produces a zero entry in H; specifically, $\tilde{\Theta}_{(i,k)}$ annihilates entry (i,k). A difference with QR is that we must keep track of the signatures associated with the columns of the matrix to determine which type of rotations to use. The general scheme, however, is as follows.

$$
\begin{array}{rcl}
[\epsilon I \quad H] & = & \begin{array}{c} + \; + \; + \quad - \; - \; - \; - \\ \left[\begin{array}{ccc|cccc} \underline{\epsilon} & & 0 & \underline{\times} & \times & \times & \times \\ & \epsilon & & \times & \times & \times & \times \\ 0 & & \epsilon & \times & \times & \times & \times \end{array}\right] \end{array} \quad \tilde{\Theta}_{(1,1)} \rightarrow \\
& & \begin{array}{c} - \; + \; + \quad + \; - \; - \; - \\ \left[\begin{array}{ccc|cccc} \times & & & 0 & \times & \times & \times \\ \times & \underline{\epsilon} & & \underline{\times} & \times & \times & \times \\ \times & & \epsilon & \times & \times & \times & \times \end{array}\right] \end{array} \quad \tilde{\Theta}_{(2,1)} \rightarrow \\
& & \begin{array}{c} - \; + \; + \quad + \; - \; - \; - \\ \left[\begin{array}{ccc|cccc} \times & & & 0 & \times & \times & \times \\ \times & \times & & 0 & \times & \times & \times \\ \times & \times & \underline{\epsilon} & \underline{\times} & \times & \times & \times \end{array}\right] \end{array} \quad \rightarrow \\
\cdots \overset{\tilde{\Theta}_{(m,n)}}{\rightarrow} & & \begin{array}{c} - \; + \; - \quad + \; + \; - \; - \\ \left[\begin{array}{ccc|cccc} \times & & & 0 & 0 & 0 & 0 \\ \times & \times & & 0 & 0 & 0 & 0 \\ \times & \times & \times & 0 & 0 & 0 & 0 \end{array}\right] \end{array} = [X \quad 0],
\end{array}
$$

$$
\tilde{\Theta} = \tilde{\Theta}_{(1,1)}\tilde{\Theta}_{(2,1)} \cdots \tilde{\Theta}_{(m,1)} \cdot \tilde{\Theta}_{(1,2)} \cdots \tilde{\Theta}_{(2,2)} \cdots \tilde{\Theta}_{(m,n)} \cdot
$$

(Except for the first matrix, the signatures of the columns in the above matrices are exemplary since they are data dependent.) The pivot elements at each step are underlined; these entries, along with the signatures of the two columns in which they appear, determine the elementary rotation $\tilde{\theta}$ that will be used at that step, as well as the resulting signature $\tilde{j}_2$. This signature is the new signature of these two columns after application of the rotation. The algorithm is summarized in Fig. 1.[3] The nulling scheme ensures that $[\epsilon I \;\; H]\tilde{\Theta} = [X \;\; 0]$, where X is a resulting lower triangular invertible matrix; it contains the columns of A and B in some permuted order. The columns of X with a positive signature are the columns of A; the columns with a negative signature are those of B. Hence, the final step (not listed in Fig. 1) is to sort these columns, such that $[X \;\; 0]\Pi = [A \;\; 0 \;\; B \;\; 0] = [A' \;\; B']$. Then $\Theta = \tilde{\Theta}\Pi$ is J-unitary with respect to J, and $[\epsilon I \;\; H]\Theta = [A' \;\; B']$.

The complexity of the algorithm is similar to that of the QR factorization—about $1/2\, m^2 n$ rotations, or $2m^2 n$ flops. The Schur algorithm has a direct implementation on a systolic array of processors. This array is entirely similar to the classical Gentleman–Kung triangular Givens array [12], except that now all data entries have a signature associated with them and the processors must perform different types of rotations, depending on these signatures. We omit the details.

4.3. Updating and downdating. The Schur method is straightforward to update as more and more columns of H become known. If $[\epsilon I \;\; H_n]\tilde{\Theta}_{(n)} = [X_n \;\; 0]$ is the

[3] As an aside, we mention that Bojanczyk et al. [23] have developed a numerically more stable implementation of the application of hyperbolic plane rotations to vectors. This is probably of relevance in the present context.

$$
\begin{array}{l}
[X \quad Y] := [\epsilon I_m \quad H] \\
\tilde{J} := \begin{bmatrix} I_m & \\ & -I_n \end{bmatrix} \\
\tilde{\Theta} = I_{m+n} \\
\text{for } k = 1 \text{ to } n \text{ and } i = 1 \text{ to } m, \\
\quad [a \quad b] := [X(i,i) \quad Y(i,k)] \\
\quad \tilde{j}_1 := \begin{bmatrix} \tilde{J}(i,i) & 0 \\ 0 & \tilde{J}(m+k, m+k) \end{bmatrix} \\
\quad \text{Compute } \tilde{\theta}, \tilde{j}_2 \text{ from } a, b, \tilde{j}_1 \text{ s.t. } [a \quad b]\tilde{\theta} = [* \quad 0] \\
\quad \text{Embed } \tilde{\theta} \text{ into } \tilde{\Theta}_{(i,k)} \\
\quad [X \quad Y] := [X \quad Y]\tilde{\Theta}_{(i,k)} \\
\quad \tilde{\Theta} := \tilde{\Theta}\tilde{\Theta}_{(i,k)} \\
\quad \tilde{J}(i,i) := (\tilde{j}_2)_{1,1} \\
\quad \tilde{J}(m+k, m+k) := (\tilde{j}_2)_{2,2} \\
\text{end} \\
\tilde{J}_2 := \tilde{J}
\end{array}
$$

FIG. 1. *Schur algorithm to compute the factorization* $[\epsilon I \quad H]\tilde{\Theta} = [X \quad 0]$ *from* H.

factorization at point n and $H_{n+1} = [H_n \quad h_{n+1}]$, then, because the algorithm works columnwise,

$$
[\epsilon I \quad H_{n+1}]\tilde{\Theta}_{(n+1)} = [X_{n+1} \quad 0] \quad \Rightarrow \quad \begin{array}{l} [X_n \quad 0 \quad h_{n+1}]\,\tilde{\theta}^{(n+1)} = [X_{n+1} \quad 0 \quad 0] \\ \tilde{\Theta}_{(n+1)} = \tilde{\Theta}_{(n)}\tilde{\theta}^{(n+1)}, \end{array}
$$

for some J-unitary matrix $\tilde{\theta}^{(n+1)}$ acting on the columns of X_n and on h_{n+1}. Hence, we can continue with the result of the factorization that was obtained at the previous step. Each update requires about $1/2m^2$ rotations.

The downdating problem involves computing the factorization for H_n with its first column h_1 removed, from a factorization of H_n. It can be converted to an updating problem, where the old column h_1 is now introduced with a positive signature,

$$
[\overset{+/-}{X_n} \quad \overset{+}{h_1}]\,\tilde{\theta}^{(n+1)} = [X_{n+1} \quad 0]\,.
$$

This is possible because, implicitly, we factor $\epsilon^2 I - H_n H_n^* + h_1 h_1^* = X_n \tilde{J} X_n^* + h_1 h_1^*$. The uniqueness of the hyperbolic QR factorization into triangular matrices with positive diagonals ([24], viz. Corollary 2.2) implies that the result X_{n+1} is precisely the same as if h_1 had never been part of H_n at all.

4.4. Breakdown. In §4.2, we had to assume that the data matrix H was such that at no point in the algorithm is $[a \quad b]\tilde{j}_1[a \quad b]^*$ equal to zero. If the expression is zero, then there is no J-unitary rotation $\tilde{\theta}$ such that $[a \quad b]\tilde{\theta} = [* \quad 0]$. In Theorem 3.1 note that the condition that none of the singular values of H equal ϵ does not preclude this case, but merely ascertains that there *exists* a $\tilde{\Theta}$ which will zero H. One simple example is obtained by taking $H = [1 \quad 1]^T$, $\epsilon = 1$. It is straightforward to show that there is no J-unitary $\tilde{\Theta}$ such that

$$
\left[\begin{array}{cc|c} 1 & & 1 \\ & 1 & 1 \end{array}\right]\tilde{\Theta} = \left[\begin{array}{cc|c} \times & 0 & 0 \\ \times & \times & 0 \end{array}\right], \tag{24}
$$

since the J-norms of the first row will not be equal. Hence Θ cannot be obtained by the recursive algorithm. However, a more general $\tilde{\Theta}$ does exist, such that

$$\begin{array}{ccc|c} + & + & - & \\ \end{array}$$

$$\left[\begin{array}{cc|c} 1 & & 1 \\ & 1 & 1 \end{array}\right]\tilde{\Theta} = \frac{1}{\sqrt{2}}\left[\begin{array}{cc|c} 1 & 1 & 0 \\ -1 & 1 & 0 \end{array}\right]$$

viz.

$$\tilde{\Theta} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 & \sqrt{2} \\ -1 & -1 & \sqrt{2} \\ 0 & 2 & -\sqrt{2} \end{bmatrix}, \qquad \tilde{J}_1 = \begin{bmatrix} 1 & & \\ & 1 & \\ & & -1 \end{bmatrix}, \qquad \tilde{J}_2 = \begin{bmatrix} 1 & & \\ & -1 & \\ & & 1 \end{bmatrix}.$$

The difference is that, in this factorization, the resulting matrix X is no longer lower triangular. Theorem 4.1 gives necessary and sufficient conditions on the singular values of H and a collection of submatrices of H, so that the Schur algorithm does not break down.

THEOREM 4.1. *Let H: $m \times n$ be a given matrix, and $\epsilon \geq 0$. Denote by $H_{[i,k]}$ the submatrix consisting of the first to the ith row and the first k columns of H. The Schur algorithm does not break down if and only if none of the singular values of $H_{[i,k]}$ is equal to ϵ, for $i = 1, \ldots, m$ and $k = 1, \ldots, n$.*

Proof. (Necessity) When processing the kth column of H by the Schur algorithm, we are in fact computing a triangular factorization of $[\epsilon I_m \;\; H_{[m,k]}]$. Corollary 2.2 claims that a suitable J-unitary operator exists if and only if $[\epsilon I_i \;\; H_{[i,k]}]$ is J-nonsingular, for $i = 1, \ldots, m$, i.e., if and only if none of the singular values of $H_{[i,k]}$ is equal to 1. The triangularization is done for $k = 1, 2, \ldots, n$ in turn.

(Sufficiency) Sufficiency at stage (i, k) follows recursively from the factorization at the previous stage and the existence and uniqueness of the factorization at the current stage. □

Similar results are known for the case where the factorization is computed via hyperbolic Householder transformations where all zeros in a row are generated at the same time. In this case there are fewer conditions [24], viz. Corollary 2.2. It should be noted that the conditions in Theorem 4.1 are quite elaborate, as only one condition (none of the singular values of H are equal to ϵ) suffices for the *existence* of Θ. Numerically, we might also run into problems if one of the singular values is close to ϵ, in which case the corresponding hyperbolic rotation has a large norm. How serious this is depends on a number of factors, and a careful numerical analysis is called for. One example where a large rotation is not fatal is the case where the singularity occurs while processing the last entry of a column ($i = m$). Although the rotation will be very large, the resulting X remains bounded and becomes singular: $X_{m,m} = 0$. Hence, the subspace information is still accurate, and X varies in a continuous way across the ϵ-boundary; only its signature is necessarily discontinuous. Pivoting schemes can be used to prevent large rotations, and are discussed in the next subsection.

4.5. Pivoting schemes. Because a breakdown occurs only for special values of the entries of H, we can in almost all cases employ a simple pivoting operation to avoid a large hyperbolic rotation. If such a rotation occurs at the zeroing of entry $h_{i,k}$, then the matrix $H_{[i,k]}$ has a singular value close to ϵ. At this point, there are a number of remedies based on the relative freedom in the order in which zero entries are created. The simplest solution is to permute the current column with the next

one, which is possible if $k < n$. We can also permute the ith row with the $i+1$-st if $i < m$. Instead of permutations, other more complicated operations are also possible, such as plane rotations of two columns or rows. Finally, if $(i,k) = (m,n)$, i.e., $h_{i,k}$ is the last entry to be zeroed, then H has a singular value equal to ϵ and there is no remedy—there is no bounded Θ. However, because it is the last rotation, X will still be bounded, but it becomes singular.

A column permutation at stage (i,k) swaps the kth column of H with the $k+1$-st, and also swaps the corresponding rows of $\tilde{\Theta}$. Before the permutation is done, the first $i-1$ entries of h_{k+1} must be made zero. Hence, a column permutation scheme is most easily implemented when entries of H are zeroed row by row, rather than column by column as in the algorithmic description in §4.2. Note that it is already sufficient to create zero entries of H in an antidiagonal fashion. This is what actually happens in a systolic array implementation, where zeros on antidiagonals of H are created in parallel. Hence, a column pivoting scheme can be readily implemented on such an array, with only one extra buffer required at each processor (to queue entries of a second column), but without sacrificing the systolic nature of the algorithm in any sense. In column permutation schemes, X stays upper triangular and, after the processing of both h_k and h_{k+1}, is the same as it would be without pivoting. $\tilde{\Theta}$ is, of course, different: it is unbounded in the first case, bounded in the second.

Row permutations are necessary, e.g., if there is no next column ($k = n$), or if columns are to be processed one at a time. It is a requirement that the first $k-1$ entries of the $i+1$-st row of H be zeroed before permuting these rows. This is automatically the case if columns are processed one by one, or requires one rotation if we use an antidiagonal zeroing scheme. Another rotation is needed to keep X lower triangular after the permutation has been performed. This makes row pivoting computationally more expensive. We also must keep track of the permutations because we are now in fact computing a factorization

$$\Pi[\epsilon I \quad H]\,\tilde{\Theta} = [X \quad 0] \quad \Leftrightarrow \quad [\epsilon I \quad H]\,\tilde{\Theta} = [\Pi^* X \quad 0] = [X' \quad 0]\,.$$

X is lower triangular, but the resulting X' generally is not. It is possible to use any other invertible transformation of the rows instead of a permutation, such as a unitary plain rotation for example. This more general approach was suggested in [29], and provides a solution even in the special cases where permutations do not lead to bounded results, such as, e.g., in the case of (24). The resulting factorization can be viewed as a hyperbolic URV decomposition. The added generality allows reduction of the number of hyperbolic rotations to one or two per column update, and leads to stable numerical implementations. (A discussion of this is relegated to future publications.)

5. Simulation results. In this section, we demonstrate some of the properties of the approximation scheme by means of a simple example. We take $H(\sigma_2) = U\Sigma(\sigma_2)\,V^*$ to be a sequence of 3×4 matrices, with U and V randomly selected as constant unitary matrices, and with singular values equal to

$$(20,\ \sigma_2,\ 0.5)\,, \qquad \sigma_2 = 0,\ 0.01,\ \ldots,\ 3.99,\ 4\,.$$

The approximation tolerance is set to $\epsilon = 1$. We compare the approximants $\hat{H}^{(0)}$ given by $S_L = 0$, $\hat{H}^{(1)}$ given by (18), $\hat{H}^{(2)}$ given by (21), and $\hat{H}^{(1)}$ when the factorization is computed with pivoting. The pivoting scheme consists of column permutations, except when processing the last column, in which case we switch to row permutations.

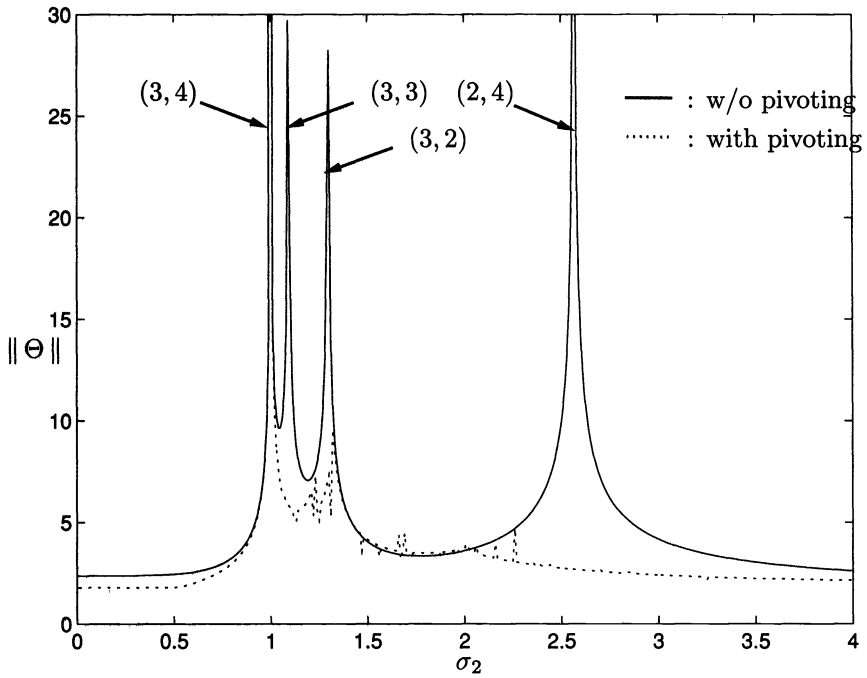


FIG. 2. *Norm of* Θ. *With pivoting,* $\| \Theta \| \to \infty$ *for certain values of* σ_2 *when the indicated entry* (i,j) *of* H *is processed. With pivoting, this only occurs when* $\sigma_2 = 1$.

The pivoting is applied in its extreme form, i.e., whenever this leads to elementary rotation matrices with a smaller norm. The approximants are compared in the following ways: (a) $\| \Theta \|$, with and without pivoting; (b) $\| H - \hat{H} \|$, for each of the mentioned approximants; (c) the accuracy of the subspace estimates, compared to the principal subspace of H (the column span of the singular vectors with corresponding singular values larger than 1). The distance between two subspaces $\mathcal{A}$ and $\mathcal{B}$ is defined as $\text{dist}(\mathcal{A}, \mathcal{B}) = \| \mathbf{P}_{\mathcal{A}} - \mathbf{P}_{\mathcal{B}} \|$, where $\mathbf{P}_{\mathcal{A}}$ is the orthogonal projection onto $\mathcal{A}$ [2].

Figure 2 shows $\| \Theta \|$ as a function of σ_2. Without pivoting, there are a number of peaks, corresponding to the values of σ_2 where one of the submatrices $H_{[i,k]}$ has a singular value equal to 1. In the range $0 \leq \sigma_2 \leq 4$, this occurred for $(i,k) = (3,4)$, $(3,3)$, $(3,2)$, and $(2,4)$, respectively. When pivoting is applied, the peak at $\sigma_2 = 1$ is, necessarily, still present, but the other peaks are mostly smoothed out. Figure 3 shows the norm of the columns of B in the scheme without pivoting. For $\sigma_2 < 1$, the rank of the approximant is 1. At $\sigma_2 = 1$, the dimension of B increases, although at first, the new column has a very small norm. For larger values of σ_2, the norm grows and the subspace becomes better defined. There is a peak at the point where $H_{[2,4]}$ has a singular value equal to 1; this peak can be removed by row pivoting but not by column pivoting. There are no peaks when $H_{(i,j)}$ has a singular value equal to 1 and $i = m$, because X becomes singular rather than unbounded when a singularity occurs at the last entry of a column. Figure 3 also shows that no peak occurs for the norm of the columns of the "improved" subspace estimate $B^{(1)}$ of (19), on which both $\hat{H}^{(1)}$ and $\hat{H}^{(2)}$ are based. This is as predicted by Lemma 3.4: $\| B^{(1)} \| \leq \| H \| = 20$. Instead of having a peak, the norm of the first column of $B^{(1)}$ dips to about 0.12.

In Fig. 4, the norm of $H - \hat{H}$ is shown for the various choices of $\hat{H}$ that we

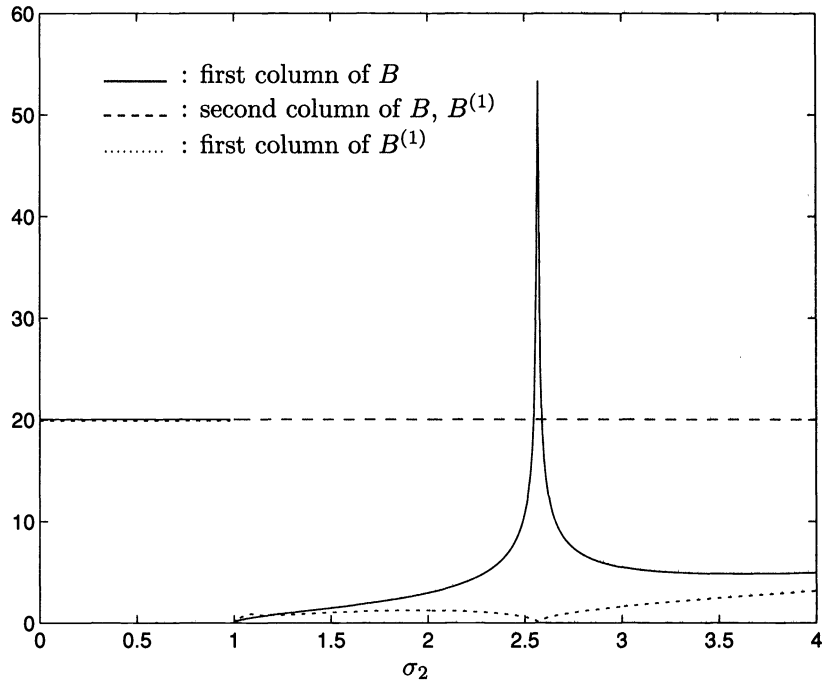


FIG. 3. *The norm of the first and second column of* B *and* $B^{(1)}$ *(no pivoting).*

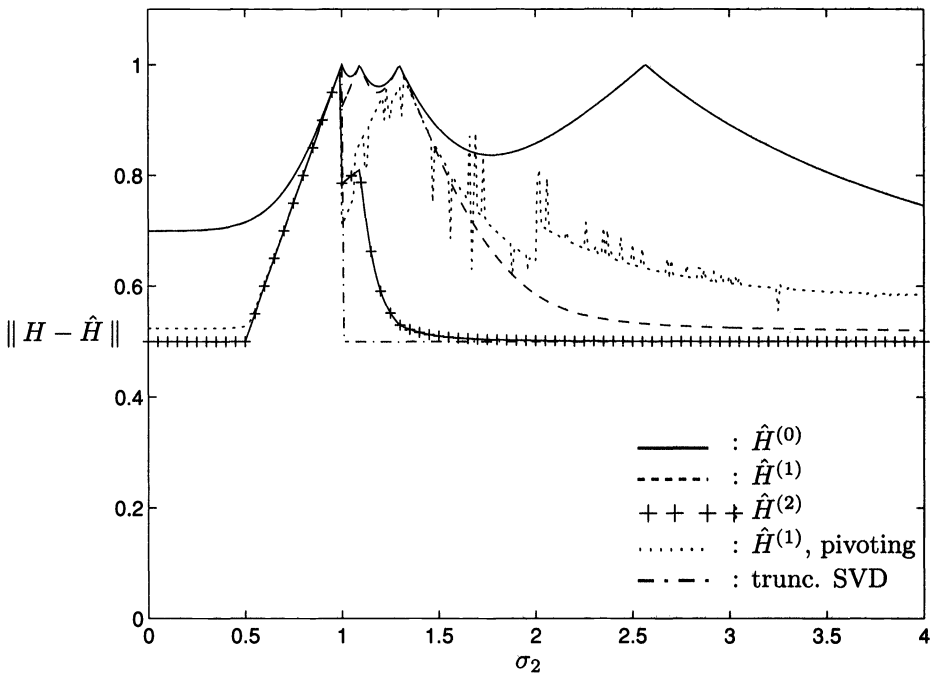


FIG. 4. *Norm of the approximation error.*

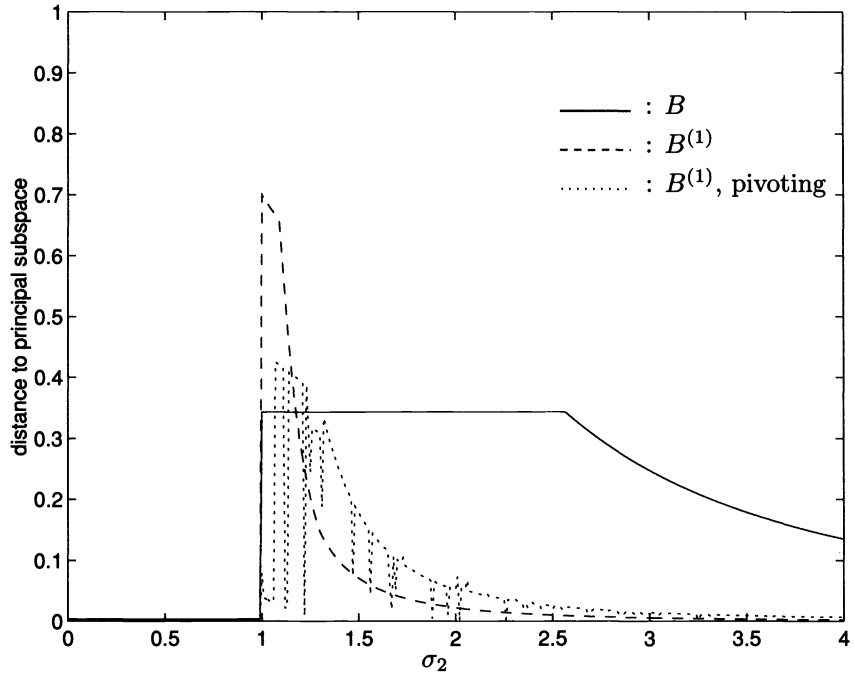


FIG. 5. *Distance between the principal and estimated subspaces.*

discussed in §3.2. The lowest line corresponds to the truncated SVD solution, which gives the lowest attainable error. It is seen that, for all approximants, the approximation error is always less than $\epsilon \equiv 1$. Of the nonpivoted schemes, the approximation error for $\hat{H}^{(0)}$ is always higher than the error for $\hat{H}^{(1)}$, $\hat{H}^{(2)}$ (but there is no *proof* that this is necessarily always the case), and the error for $\hat{H}^{(1)}$ is always higher than the error for $\hat{H}^{(2)}$, since the latter approximant minimizes this error while retaining the same subspace estimate. The approximation error for $\hat{H}^{(2)}$ is almost identically close to the theoretical minimum, except in a small region $1 \leq \sigma_2 \leq 1.5$. The errors for $\hat{H}^{(0)}$ and $\hat{H}^{(1)}$ touch a number of times on the $(\epsilon = 1)$-line. For $\hat{H}^{(0)}$ this can be explained as follows. The error for $S_L = 0$ is given by (16) as $-\epsilon\Theta_{12}\Theta_{22}^{-1}$. Because the J-unitarity of Θ implies $\Theta_{22}^{-*}\Theta_{22}^{-1} + (\Theta_{22}^{-*}\Theta_{12}^{*})(\Theta_{12}\Theta_{22}^{-1}) = I$ (viz. (6)), it follows that whenever $\| \Theta_{22} \| \to \infty$, necessarily $\| \Theta_{12}\Theta_{22}^{-1} \| \to 1$. The analysis of $\| H - \hat{H}^{(1)} \|$ from (17) is more involved and is omitted at this point.

Figure 5 depicts the distance between the principal and estimated subspaces. For $\sigma_2 < 1$, this distance is very close to zero ($< .0002$) for each of the methods. The distance jumps up when σ_2 crosses 1: the subspace increases in dimension but is at first only weakly defined. For $B^{(1)}$, the distance goes down again quickly, whereas for B, it stays constant for a while before going down.

6. Conclusions. We have derived a general formula that describes all rank-d 2-norm approximants of a given matrix H. The formula relies on a factorization that exists if none of the singular values of H is equal to ϵ, and that can be computed by a Schur-type algorithm if additional singular value conditions are satisfied. Updating and downdating is straightforward, and the algorithm is amenable to parallel implementation. It is highly suitable for adaptive subspace estimation because some of these approximants are quite close to the truncated SVD solution (as shown by a

numerical experiment), but much easier to compute. Such an application is reported in [34]. Another application is the regularization of ill-conditioned total least squares problems [35], cf. [36].

There are several open problems and remaining issues. Apart from the listed approximants, there might be other interesting choices, such as approximants that by construction have all their singular values larger than ϵ. There are applications in which an on-line computation of the approximant or its last column (instead of only its column space) is required. An integral scheme for doing this would be interesting. As a final remark, we mention that while this paper was in review, an updated version for the "improved" approximant $B^{(1)}$ was obtained. An orthonormal basis for this subspace can be updated using about twice the number of operations as the basic Schur updating algorithm, without the need for pivoting and keeping the number of hyperbolic rotations as small as possible. This will be reported elsewhere.

Appendix A. Proof of Theorem 3.3. The proof of Theorem 3.3 consists of two propositions. The first proof shows that any S_L that satisfies the constraints gives rise to a valid approximant and the second proves the converse. Without loss of generality, we take $\epsilon = 1$.

PROPOSITION A.1. *Let $H : m \times n$ be a given matrix, with d singular values larger than 1 and none equal to 1, and let S_L be a given matrix satisfying conditions* (15) (i) *and* (ii). *Define Θ, A', B' as in equation* (12). *Put*

$$S = (\Theta_{11}S_L - \Theta_{12})(\Theta_{22} - \Theta_{21}S_L)^{-1}.$$

Then S is well defined, and $\hat{H} := H - S$ is a 2-norm approximant of rank equal to d, satisfying

$$\hat{H} = (B' - A'S_L)(\Theta_{22} - \Theta_{21}S_L)^{-1}.$$

Proof. Let

$$\begin{bmatrix} -G_1 \\ G_2 \end{bmatrix} = \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{bmatrix} \begin{bmatrix} -S_L \\ I \end{bmatrix}.$$

Then

$$\begin{aligned} G_1 &= \Theta_{11}S_L - \Theta_{12}, \\ G_2 &= -\Theta_{21}S_L + \Theta_{22} = \Theta_{22}(I - \Theta_{22}^{-1}\Theta_{21}S_L). \end{aligned}$$

Because $\| \Theta_{22}^{-1}\Theta_{21} \| < 1$ and $\| S_L \| \leq 1$, G_2 is invertible, and hence $S = G_1G_2^{-1}$. The J-unitarity of Θ implies $S_L^*S_L - I = G_1^*G_1 - G_2^*G_2 = G_2^*(S^*S - I)G_2$. Since G_2 is invertible and $S_L^*S_L - I$ is negative semidefinite, it follows that the same holds for $S^*S - I$. Hence S is contractive: $\| S \| \leq 1$, and $\hat{H}$ is a 2-norm approximant of H. To derive the alternate formula for $\hat{H}$ and show that it has rank d, write

$$\begin{aligned} \hat{H} = H - S &= [I \quad H] \begin{bmatrix} -S \\ I \end{bmatrix} \\ &= [I \quad H]\Theta \begin{bmatrix} -S_L \\ I \end{bmatrix} (\Theta_{22} - \Theta_{21}S_L)^{-1} \\ &= [A' \quad B'] \begin{bmatrix} -S_L \\ I \end{bmatrix} (\Theta_{22} - \Theta_{21}S_L)^{-1}. \end{aligned}$$

Hence $\hat{H} = (B' - A'S_L)(\Theta_{22} - \Theta_{21}S_L)^{-1}$. The rank of $\hat{H}$ is equal to the rank of $B' - A'S_L$. In this expression, $B' = [B \;\; 0]$ is of full column rank d, and $A' = [A \;\; 0]$, where A is of full column rank $m - d$. Because $(S_L)_{12} = 0$, it follows that $A'S_L = [A \;\; 0]S_L = [A(S_L)_{11} \;\; 0]$ is also of rank less than or equal to d. Finally, $B' - A'S_L$ is precisely of rank d because the columns of A are linearly independent of the columns of B. □

PROPOSITION A.2. *Let H: $m \times n$ be a given matrix, with d singular values larger than 1 and none equal to 1. Define Θ, A', B' as in (12). Suppose that a matrix $\hat{H}$ satisfies*

$$\begin{aligned} &\text{(a)} \quad \| H - \hat{H} \| \leq 1\,, \\ &\text{(b)} \quad \operatorname{rank}(\hat{H}) \leq d\,. \end{aligned}$$

Then $\operatorname{rank}(\hat{H}) = d$, *and* $\hat{H} = H - S$ *where*

$$S = (\Theta_{11}S_L - \Theta_{12})(\Theta_{22} - \Theta_{21}S_L)^{-1}\,, \tag{A.25}$$

for some contractive S_L with $(S_L)_{12} = 0$.

Proof. It follows directly that S is contractive. Define matrices G_1, G_2 by

$$\begin{bmatrix} -S \\ I \end{bmatrix} = \Theta \begin{bmatrix} -G_1 \\ G_2 \end{bmatrix} \quad \Leftrightarrow \quad \begin{bmatrix} -G_1 \\ G_2 \end{bmatrix} = \Theta^{-1} \begin{bmatrix} -S \\ I \end{bmatrix}. \tag{A.26}$$

As in the proof of Proposition A.1, it follows that G_2 is invertible. The J-unitarity of Θ and the contractiveness of S implies $G_1^*G_1 \leq G_2^*G_2$. Hence $S_L := G_1G_2^{-1}$ is well defined and contractive, and (A.26) implies (A.25). The rest of the proof is technical and shows that $(S_L)_{12} = 0$. First, we define the partitionings

$$G_1 = \begin{matrix} m-d \\ d \end{matrix} \begin{bmatrix} G_{11} \\ G_{12} \end{bmatrix}, \qquad G_2 = \begin{matrix} d \\ n-d \end{matrix} \begin{bmatrix} G_{21} \\ G_{22} \end{bmatrix}, \qquad G_2^{-1} = \begin{matrix} d & n-d \end{matrix} \\ \begin{bmatrix} (G_2^{-1})_1 & (G_2^{-1})_2 \end{bmatrix},$$

which conform with the partitionings of A' and B'. Then $(S_L)_{12} = 0 \Leftrightarrow G_{11}(G_2^{-1})_2 = 0$. To prove that $G_{11}(G_2^{-1})_2 = 0$, we look at $[G_1^* \;\; G_2^*]$. The use of (A.26) and $\Theta^{-1} = J\Theta^*J$ gives

$$\begin{aligned} [G_1^* \;\; G_2^*] &= [S^* \;\; I]\Theta \\ &= [-\hat{H}^* \;\; 0]\Theta + [H^* \;\; I]\Theta\,. \end{aligned} \tag{A.27}$$

We also have

$$\begin{aligned} & [I \;\; H] = [A' \;\; B']\Theta^{-1} \\ \Leftrightarrow \quad & \begin{bmatrix} I \\ H^* \end{bmatrix} = \Theta^{-*} \begin{bmatrix} A'^* \\ B'^* \end{bmatrix} = J\Theta J \begin{bmatrix} A'^* \\ B'^* \end{bmatrix} \\ \Rightarrow \quad & 0 = [H^* \;\; I_n]J \begin{bmatrix} I \\ H^* \end{bmatrix} = [H^* \;\; I_n]\Theta J \begin{bmatrix} A'^* \\ B'^* \end{bmatrix} \\ \Rightarrow \quad & [H^* \;\; I_n]\Theta = \left[(0_{n\times(m-d)} \;\; *) \;\; (0_{n\times d} \;\; *)\right], \end{aligned}$$

where $*$ is some quantity whose precise value is not of interest. In the last step, we used the fact that $[A \;\; B]$ is of full rank. Inserting this result into (A.27) shows that

$$\mathcal{R}(G_{11}^*) \subset \mathcal{R}(\hat{H}^*)\,, \qquad \mathcal{R}(G_{21}^*) \subset \mathcal{R}(\hat{H}^*)\,.$$

G_2 is invertible; hence $\mathcal{R}(G_{21}^*)$ is of full dimension d. Since the rank of $\hat{H}$ is less than or equal to d, it follows that the rank of $\hat{H}$ is precisely equal to d, and that actually $\mathcal{R}(G_{21}^*) = \mathcal{R}(\hat{H}^*)$. This implies $\mathcal{R}(G_{11}^*) \subset \mathcal{R}(G_{21}^*)$, so that there is some matrix M such that $G_{11} = MG_{21}$. Hence

$$\begin{aligned} & G_2(G_2)^{-1} = I \\ \Leftrightarrow \quad & \begin{bmatrix} G_{21} \\ G_{22} \end{bmatrix} [(G_2^{-1})_1 \quad (G_2^{-1})_2] = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \\ \Rightarrow \quad & G_{21}(G_2^{-1})_2 = 0 \\ \Rightarrow \quad & G_{11}(G_2^{-1})_2 = MG_{21}(G_2^{-1})_2 = 0 . \qquad \square \end{aligned}$$

Acknowledgments. The author wishes to extend his warm feelings of gratitude to Professor P. Dewilde at Delft University of Technology. The approximation theory underlying the results presented in this paper stems from research in time-varying systems theory carried out in close collaboration over a number of years. The observation that the time-varying Hankel-norm approximation theory can be applied to matrices, leading to the present paper, was made by J. Götze, Technical University of Munich, while visiting Delft University in the autumn of 1993. The kind invitation by Professor G.H. Golub for a stay at Stanford University is gratefully acknowledged.

REFERENCES

[1] G.W. Stewart, *An updating algorithm for subspace tracking*, IEEE Trans. Signal Processing, 40 (1992), pp. 1535–1541.

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

[3] L.V. Foster, *Rank and null space calculations using matrix decomposition without column interchanges*, Linear Algebra Appl., 74 (1986), pp. 47–71.

[4] T.F. Chan, *Rank revealing QR factorizations*, Linear Algebra Appl., 88/89 (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. Statist. Comput., 13 (1992), pp. 727–741.

[7] C.H. Bischof and G.M. Shroff, *On updating signal subspaces*, IEEE Trans. Signal Processing, 40 (1992), pp. 96–105.

[8] S. Chandrasekaran and I.C.F. Ipsen, *On rank-revealing factorisations*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 592–622.

[9] J.R. Bunch and C.P. Nielsen, *Updating the singular value decomposition*, Numer. Math., 31 (1978), pp. 111–129.

[10] M. Moonen, P. Van Dooren, and J. Vandewalle, *An SVD updating algorithm for subspace tracking*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1015–1038.

[11] M. Moonen, P. Van Dooren, and F. Vanpoucke, *On the QR algorithm and updating the SVD and URV decomposition in parallel*, Linear Algebra Appl., 188/189 (1993), pp. 549–568.

[12] W.M. Gentleman and H.T. Kung, *Matrix triangularization by systolic arrays*, Proc. SPIE, Real Time Signal Proc. IV, 298 (1981), pp. 19–26.

[13] I. Schur, *Uber Potenzreihen, die im Innern des Einheitskreises beschränkt sind*, I, J. Reine Angew. Math., 147 (1917), pp. 205–232. Eng. Transl. *Operator Theory: Adv. Appl.*, Vol. 18, pp. 31-59, Birkhäuser-Verlag, Basel, 1986.

[14] H. Dym, *J-Contractive Matrix Functions, Reproducing Kernel Hilbert Spaces and Interpolation*, No. 71, CBMS Reg. Conf. Ser., American Math. Soc., Providence, 1989.

[15] J.A. Ball, I. Gohberg, and L. Rodman, *Interpolation of Rational Matrix Functions*, Vol. 45, Operator Theory: Advances and Applications, Birkhäuser-Verlag, Basel, 1990.

[16] V.M. Adamjan, D.Z. Arov, and M.G. Krein, *Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur–Takagi problem*, Math. USSR Sbornik, 15 (1971), pp. 31–73. (Translation of Iz. Akad. Nauk Armjan. SSR Ser. Mat. 6 (1971).)

[17] D. PAL AND T. KAILATH, *Fast triangular factorization and inversion of Hermitian, Toeplitz, and related matrices with arbitrary rank profiles*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1016–1042.

[18] P. DEWILDE AND E. DEPRETTERE, *The generalized Schur algorithm: Approximation and hierarchy*, in Operator Theory: Advances and Applications, Vol. 29, Birkhäuser-Verlag, Basel, 1988, pp. 97–116.

[19] K. DIEPOLD AND R. PAULI, *Schur parametrization of symmetric indefinite matrices based on a network theoretic model*, Archiv für Elektronik und Übertragungstechnik AEÜ, 44 (1990), pp. 92–96.

[20] ———, *A recursive algorithm for lossless embedding of non-passive systems*, in Challenges of a Generalized System Theory, P. Dewilde, ed., Essays of the Royal Dutch Academy of Sciences, North-Holland, Amsterdam, The Netherlands, 1993.

[21] PH. DELSARTE, Y. GENIN, AND Y. KAMP, *A method of matrix inverse triangularization, based on contiguous principal submatrices*, Linear Algebra Appl., 31 (1980), pp. 199–212.

[22] J.-M. DELOSME AND I.C.F. IPSEN, *Parallel solution of symmetric positive definite systems with hyperbolic rotations*, Linear Algebra Appl., 77 (1986), pp. 75–111.

[23] 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.

[24] A. BUNSE-GERSTNER, *An analysis of the* HR *algorithm for computing the eigenvalues of a matrix*, Linear Algebra Appl., 35 (1981), pp. 155–173.

[25] C.M. RADER AND A.O. STEINHARDT, *Hyperbolic Householder transformations*, IEEE Trans. Acoust. Speech, Signal Proc., 34 (1986), pp. 1589–1602.

[26] 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.

[27] G. CYBENKO AND M. BERRY, *Hyperbolic Householder algorithms for factoring structured matrices*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 499–520.

[28] C.H. BISCHOF, C.-T. PAN, AND P.T.P. TANG, *A Cholesky up and downdating algorithm for systolic and* SIMD *architectures*, SIAM J. Sci. Statist. Comput., 14 (1993), pp. 670–676.

[29] R. ONN, A.O. STEINHARDT, AND A.W. BOJANCZYK, *The hyperbolic singular value decomposition and applications*, IEEE Trans. Signal Proc., 39 (1991), pp. 1575–1588.

[30] P.M. DEWILDE AND A.-J. VAN DER VEEN, *On the Hankel-norm approximation of upper-triangular operators and matrices*, Integral Eq. Operator Theory, 17 (1993), pp. 1–45.

[31] A.-J. VAN DER VEEN AND P.M. DEWILDE, *On low-complexity approximation of matrices*, Linear Algebra Appl., 205/206 (1994), pp. 1145–1202.

[32] K. GLOVER, *All optimal Hankel norm approximations of linear multi-variable systems and their L^∞-error bounds*, Internat. J. Control, 39 (1984), pp. 1115–1193.

[33] D.J.N. LIMEBEER AND M. GREEN, *Parametric interpolation, H_∞-control and model reduction*, Internat. J. Control, 52 (1990), pp. 293–318.

[34] J. GÖTZE AND A.-J. VAN DER VEEN, *On-line subspace estimation using a Schur-type method*, IEEE Trans. Signal Proc., to appear.

[35] A.-J. VAN DER VEEN, *Schur method for low-rank matrix approximation*, in Proc. SPIE, Advanced Signal Processing Algorithms, Architectures, and Implementations V, Vol. 2296, San Diego, CA, July 1994.

[36] R.D. FIERRO, G.H. GOLUB, P.C. HANSEN, AND D.P. O'LEARY, *Regularization by truncated total least squares*, in Proc. 5th SIAM Conf. on Applied Linear Algebra, J.G. Lewis, ed., Snowbird, UT, June 1994, pp. 250–254.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 1, pp. 161–164, January 1996

009

JACOBI ANGLES FOR SIMULTANEOUS DIAGONALIZATION*

JEAN-FRANÇOIS CARDOSO† AND ANTOINE SOULOUMIAC‡

Abstract. Simultaneous diagonalization of several matrices can be implemented by a Jacobi-like technique. This note gives the required Jacobi angles in close form.

Key words. simultaneous diagonalization, Jacobi iterations, eigenvalues, eigenvectors, structured eigenvalue problem

AMS subject classifications. 65F15, 65-04

Introduction. Simultaneous diagonalization of several commuting matrices has been recently considered in [1], mainly motivated by stability and convergence concerns. Exact or approximate simultaneous diagonalization was also independently introduced as a solution to a statistical identification problem [2] (see [3] for a later paper in English). The simultaneous diagonalization algorithm described in these papers is an extension of the Jacobi technique: a joint diagonality criterion is iteratively optimized under plane rotations. The purpose of this note is to complement [1] by giving a close form expression for the optimal Jacobi angles.

1. Jacobi angles in close form. Consider a set $\mathcal{A} = \{A_k | k = 1, K\}$ of K complex $N \times N$ matrices. When the matrices in $\mathcal{A}$ are normal commuting matrices, their off-diagonal terms can be set to zero by a unitary transform, thus simultaneously diagonalizing the set $\mathcal{A}$. Define, as in [1],

$$\textbf{off}(A) \stackrel{\text{def}}{=} \sum_{1 \le i \ne j \le N} |a_{ij}|^2, \tag{1}$$

where a_{ij} denotes the (i,j)th entry of matrix A. Simultaneous diagonalization may be obtained by minimizing the composite objective $\sum_{k=1,K} \textbf{off}(UA_kU^H)$ by a unitary matrix U. The extended Jacobi technique for simultaneous diagonalization constructs U as a product of plane rotations globally applied to all the matrices in $\mathcal{A}$.

Denote $R(i,j,c,s)$ the complex rotation matrix equal to the identity matrix but for the following entries:

$$\begin{pmatrix} r_{ii} & r_{ij} \\ r_{ji} & r_{jj} \end{pmatrix} = \begin{pmatrix} c & \bar{s} \\ -s & \bar{c} \end{pmatrix} \quad \text{with} \quad c, s \in \mathbf{C} \quad \text{and} \quad |c|^2 + |s|^2 = 1. \tag{2}$$

It is desired, for each choice of $i \ne j$, to find the complex angles c and s which minimize the following objective function:

$$O(c,s) \stackrel{\text{def}}{=} \sum_{k=1,K} \textbf{off}\left(R(i,j,c,s) A_k R^H(i,j,c,s)\right). \tag{3}$$

* Received by the editors December 9, 1993; accepted for publication (in revised form) by A. Bunse-Gerstner February 23, 1995.

† Télécom Paris / CNRS URA 820 / GdR TdSI, 46 rue Barrault, 75634 Paris Cedex 13, France (cardoso@sig.enst.fr).

‡ Schlumberger Measurement and Systems / ETL, B.P. 620-05, 50 Avenue Jean Jaurès, 92542 Montrouge Cedex, France (souloum@montrouge.smr.slb.com). The work of this author was supported by Thomson-CSF/RGS/STS during this study.

For a given pair (i,j) of indices, a 3×3 real symmetric matrix G is defined as

$$G \stackrel{\text{def}}{=} \text{Real}\left(\sum_{k=1,K} h^H(A_k) h(A_k) \right), \tag{4}$$

$$h(A) \stackrel{\text{def}}{=} [a_{ii} - a_{jj} \ , \ a_{ij} + a_{ji} \ , \ \mathrm{i}(a_{ji} - a_{ij})]. \tag{5}$$

For any set $\mathcal{A}$ of $N \times N$ matrices, commuting or not, real or not, and regardless of symmetry properties such as hermitianity, unitarity, normality, etc., the following theorem allows the Jacobi angles to be computed in close form.

THEOREM 1. *Under constraint $|c|^2 + |s|^2 = 1$, the objective function $O(c,s)$ is minimized at*

$$c = \sqrt{\frac{x+r}{2r}} \qquad s = \frac{y - \mathrm{i}z}{\sqrt{2r(x+r)}} \qquad r = \sqrt{x^2+y^2+z^2}, \tag{6}$$

where $[x,y,z]^T$ is any eigenvector associated to the largest eigenvalue of G.

Proof. Let a_{ij} and a'_{ij} respectively denote the (i,j)th entry of matrices A and $A' = R(i,j,c,s) A R^H(i,j,c,s)$. Since $a'_{kk} = a_{kk}$ for $k \neq i$ and $k \neq j$ for plane rotations on the pair (i,j), the following invariance holds:

$$\mathbf{off}(A') + |a'_{ii}|^2 + |a'_{jj}|^2 = \mathbf{off}(A) + |a_{ii}|^2 + |a_{jj}|^2 \tag{7}$$

because unitary transforms preserve the norm $\sum_{kl} |a_{kl}|^2$. Hence, minimization of $\mathbf{off}(RAR^H)$ is seen, by (7), to be equivalent to *maximization* of $|a'_{ii}|^2 + |a'_{jj}|^2$. The latter, in turn, is equivalent to the maximization of $|a'_{ii} - a'_{jj}|^2$, as seen by the identity $2(|a'_{ii}|^2 + |a'_{jj}|^2) = |a'_{ii} + a'_{jj}|^2 + |a'_{ii} - a'_{jj}|^2$ and by $a'_{ii} + a'_{jj} = a_{ii} + a_{jj}$ (invariance of the trace under unitary transforms). One finds

$$a'_{ii} - a'_{jj} = (|c|^2 - |s|^2)(a_{ii} - a_{jj}) + 2cs a_{ij} + 2\bar{s}\bar{c} a_{ji}, \tag{8}$$

which is better rewritten as the inner product $a'_{ii} - a'_{jj} = h(A) v(c,s)$, between the complex 1×3 vector $h(A)$ defined in (5) and the 3×1 *real* vector $v(c,s)$ defined as

$$v(c,s)^T \stackrel{\text{def}}{=} [\ |c|^2 - |s|^2 \ , \ cs + \bar{c}\bar{s} \ , \ \mathrm{i}(cs - \bar{c}\bar{s})\ ]. \tag{9}$$

Hence, the Jacobi angles minimizing $O(c,s)$ are those maximizing

$$\sum_{k=1,K} |h(A_k) v(c,s)|^2 = v(c,s)^T \left(\sum_{k=1,K} h^H(A_k) h(A_k) \right) v(c,s). \tag{10}$$

Note that the 3×3 matrix on the right-hand side of (10) is hermitian: its imaginary part[1] is skew-symmetric and consequently contributes nothing to a quadratic form in the *real* vector $v(c,s)$. Therefore the right-hand side of (10) also is $v(c,s)^T G v(c,s)$. Next,we recognize that

$$\{v(c,s) | c,s \in \mathbf{C}, |c|^2 + |s|^2 = 1\} = \{[x,y,z]^T | x,y,z \in \mathbf{R}, x^2+y^2+z^2 = 1\}. \tag{11}$$

[1] This imaginary part is zero in the special case where $\mathcal{A}$ contains only hermitian matrices.

Thus minimization of $O(c,s)$ under the constraint $|c|^2 + |s|^2 = 1$ is equivalent to the maximization of a real 3×3 quadratic form under unit norm constraint. The solution is known to be given by any unit norm eigenvector of G associated to the (possibly degenerate) maximum eigenvalue. Now, if $[x,y,z]^T$ is a nonzero eigenvector of G, not necessarily normed to unity, associated to the largest eigenvalue, its normalization and the inversion of relation (9) yields expression (6) by choosing c real positive. This choice is possible since for any real angle ϕ, one has $v(c,s) = v(ce^{i\phi}, se^{-i\phi})$. □

2. Remarks on implementation and approximate simultaneous diagonalization. Regarding implementation, the following remarks are in order.

(i) When $\mathcal{A}$ is a set of *real* symmetric matrices, the rotation parameters c and s are real: the last component of each vector $h(A_k)$ then is zero and G can be reduced to a 2×2 matrix by deleting its last row and last column: Theorem 1 then is similar to Theorem 6.1 of [1].

(ii) For the sake of numerical stability, the Jacobi technique should be restricted to "inner rotations" [1]. In our setting, it corresponds to choosing an eigenvector with $x \geq 0$.

(iii) Since matrix G is only 3×3, its dominant eigenvector may be computed explicitly. However, lacking a close form expression with proven stability, a standard numerical eigenvalue method should be preferred for the sake of numerical stability.

(iv) It seems sensible to initialize the Jacobi algorithm for simultaneous diagonalization of a set $\mathcal{A}$ with the unitary matrix obtained as the (plain) diagonalizer of some matrix in $\mathcal{A}$. This initialization turns the spurious stationary point of the Jacobi algorithm given in (9)–(10) of [1] into a well-behaved set.

We conclude with a few words about the relevance of *approximate* simultaneous diagonalization. There is a current trend in signal and data processing of extracting information from the eigenstructure of matrices which are functions of the available data. In some cases of interest, there is a set $\mathcal{A}_T$ of matrix-valued statistics computed from a number of T available samples with the property that, almost surely, the limit set $\mathcal{A}_\infty$ contains commuting matrices; the common eigenstructure could then be computed from any member of the set $\mathcal{A}_\infty$ or from some linear combinations of matrices in $\mathcal{A}_\infty$. In practice though, only a finite number of samples is available and the matrices in $\mathcal{A}_T$ do not exactly share the same eigenstructure. Determining the eigenstructure of interest from only one matrix in $\mathcal{A}_T$ is not satisfactory because, besides relying on an arbitrary choice, it amounts to discarding the information contained in the other matrices of $\mathcal{A}_T$. Also, it may happen that each matrix of $\mathcal{A}_\infty$ has some degenerate eigenvalues but that the whole set $\mathcal{A}_\infty$ has well-determined common eigenvectors. Hence, from a statistical point of view, it is very desirable, for the sake of accuracy and robustness, to rather define the "average eigenstructure" of $\mathcal{A}_T$. Optimizing a joint diagonality criterion, possibly appropriately weighted, offers a quantitative definition of such an average eigenstructure. In this stochastic context, we note that the criterion can only be minimized but cannot generally be driven to zero: the "average eigenstructure" is well defined but corresponds only to an approximate simultaneous diagonalization.

Note. A MATLAB implementation of the extended Jacobi technique for simultaneous diagonalization is freely available upon request from cardoso@sig.enst.fr.

Acknowledgments. Comments by an anonymous referee improved the first version of this note.

REFERENCES

[1] A. BUNSE-GERSTNER, R. BYERS, AND V. MEHRMANN, *Numerical methods for simultaneous diagonalization*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 927–949.

[2] A. SOULOUMIAC AND J.-F. CARDOSO, *Comparaison de méthodes de séparation de sources*, in Proc. GRETSI, Juan les Pins, France, 1991, pp. 661–664.

[3] J.-F. CARDOSO AND A. SOULOUMIAC, *Blind beamforming for non Gaussian signals*, IEE Proceedings-F, 140 (1993), pp. 362–370.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 1, pp. 165–186, January 1996

010

APPLICATION OF ADI ITERATIVE METHODS TO THE RESTORATION OF NOISY IMAGES*

D. CALVETTI† AND L. REICHEL‡

Abstract. The restoration of two-dimensional images in the presence of noise by Wiener's minimum mean square error filter requires the solution of large linear systems of equations. When the noise is white and Gaussian, and under suitable assumptions on the image, these equations can be written as a Sylvester's equation

$$T_1^{-1}\hat{F} + \hat{F}T_2 = C$$

for the matrix $\hat{F}$ representing the restored image. The matrices T_1 and T_2 are symmetric positive definite Toeplitz matrices. We show that the ADI iterative method is well suited for the solution of these Sylvester's equations, and illustrate this with computed examples for the case when the image is described by a separable first-order Markov process. We also consider generalizations of the ADI iterative method, propose new algorithms for the generation of iteration parameters, and illustrate the competitiveness of these schemes.

Key words. Wiener filter, rational approximation, noise reduction

AMS subject classifications. 65F10, 65E05, 30E10

1. Introduction.

We describe an application of the alternating direction implicit (ADI) iteration method and a generalization thereof to the computation of the minimum mean square error estimate of a two-dimensional image in the presence of white Gaussian noise. Let F be an $N \times M$ matrix of samples from the image normalized to have zero mean, and let f be the NM-vector obtained from F by ordering its entries in lexicographic order, i.e.,

$$f := (F(1,1), \ldots, F(1,M), F(2,1), \ldots, F(N,M))^T .$$

Generally, the components of the vector f are not statistically independent, and the dependency among the various components is described by the covariance matrix of f, denoted by Φ_f. If we assume that the variability of the image in the horizontal direction is unrelated to the variability in the vertical direction, then Φ_f can be expressed as a Kronecker product

$$\Phi_f = \Phi_y \otimes \Phi_x . \tag{1.1}$$

Here Φ_x is the covariance matrix of the vector in the horizontal direction of the image and Φ_y is the covariance matrix of the vector in the vertical direction.

Assume that white Gaussian noise is added to the image, for example during transmission, i.e.,

$$g := f + \eta, \tag{1.2}$$

* Received by the editors August 17, 1994; accepted for publication (in revised form) by M. H. Gutknecht February 22, 1995.

† Stevens Institute of Technology, Department of Pure and Applied Mathematics, Hoboken, NJ 07030 (na.calvetti@na-net.ornl.gov). The research of this author was supported in part by the Design and Manufacturing Institute at Stevens Institute of Technology.

‡ Kent State University, Department of Mathematics and Computer Science, Kent, OH 44242 (reichel@mcs.kent.edu). The research of this author was supported in part by National Science Foundation grant DMS-9205531.

where $\boldsymbol{\eta}$ is the noise vector and $\boldsymbol{g}$ is the vector of the degraded image, normalized to have zero mean. A central problem in image restoration is the recovery of the original image $\boldsymbol{f}$ from the degraded image $\boldsymbol{g}$. A linear filter L is a linear operator which determines an estimate $\hat{\boldsymbol{f}}$ of the original image $\boldsymbol{f}$ from the corrupted image $\boldsymbol{g}$,

$$\hat{\boldsymbol{f}} := L\boldsymbol{g}. \tag{1.3}$$

Let $\boldsymbol{\epsilon}$ denote the error in the estimated image $\hat{\boldsymbol{f}}$,

$$\boldsymbol{\epsilon} := \boldsymbol{f} - \hat{\boldsymbol{f}}, \tag{1.4}$$

and let L be a linear filter that minimizes the mean square error. Then L is given by, see §2 for details,

$$L = \Phi_f(\Phi_f + \Phi_\eta)^{-1}, \tag{1.5}$$

where Φ_η is the covariance matrix of the noise. Therefore, the minimum mean square error estimate $\hat{\boldsymbol{f}}$ of the original image $\boldsymbol{f}$ can be computed from the degraded image $\boldsymbol{g}$ by solving the linear system of equations

$$(I + \Phi_\eta \Phi_f^{-1})\hat{\boldsymbol{f}} = \boldsymbol{g}. \tag{1.6}$$

The linear filter defined by (1.5) is often referred to as a Wiener filter. Assume that the covariance matrix Φ_f is separable and that the noise $\boldsymbol{\eta}$ is white and Gaussian with variance σ_η^2. Then, in view of (1.1), equation (1.6) can be written as

$$(I + \sigma_\eta^2 \Phi_y^{-1} \otimes \Phi_x^{-1})\hat{\boldsymbol{f}} = \boldsymbol{g}. \tag{1.7}$$

Recently, Cheong and Morgera [8] considered the solution of (1.7) for special matrices Φ_x and Φ_y by iterative methods and proposed a stationary Richardson iteration scheme. They found this scheme to be competitive with techniques based on the fast Fourier transform algorithm. This paper investigates the solution of (1.7) by the ADI iterative method and modifications thereof, and compares these methods with the conjugate gradient method. The ADI iterative methods are found to typically yield faster convergence than the conjugate gradient method, and it is well known that the latter method generally converges more rapidly than stationary Richardson iteration.

In order to describe the ADI iterative method, we introduce the matrices $\hat{F}$ and G, whose entries are the elements of $\hat{\boldsymbol{f}}$ and $\boldsymbol{g}$, respectively, stored row-wise. Then (1.7) can be written as

$$\hat{F} + \sigma_\eta^2 \Phi_y^{-1} \hat{F} \Phi_x^{-1} = G. \tag{1.8}$$

From (1.8) we obtain the Sylvester's equation

$$\hat{F}\Phi_x + \sigma_\eta^2 \Phi_y^{-1} \hat{F} = G\Phi_x. \tag{1.9}$$

The ADI iteration method for the solution of (1.9) proceeds by alternating between the solution of the two linear systems of equations,

$$\begin{aligned} F_{2k+1}(\Phi_x + \tau_{k+1} I) &= \left(\tau_{k+1} I - \sigma_\eta^2 \Phi_y^{-1}\right) F_{2k} + G\Phi_x, \\ \left(\sigma_\eta^2 \Phi_y^{-1} + \delta_{k+1} I\right) F_{2k+2} &= F_{2k+1}(\delta_{k+1} I - \Phi_x) + G\Phi_x, \end{aligned} \tag{1.10}$$

in order to determine a sequence $F_1, F_2, F_3, \ldots$ of approximants of $\hat{F}$. The matrix F_0 is an initial approximate solution; in the computed examples we let $F_0 := G$. One

seeks to choose the iteration parameters τ_k and δ_k so that the iterates F_k converge to $\hat{F}$ rapidly as k increases. The determination of suitable δ_k and τ_k is discussed in §3. There we also consider the choice of parameters δ_k and τ_k for the generalized ADI (GADI) iterative method introduced in [21]. The GADI iterative method does not require strict alternation between the equations (1.10), i.e., we allow one of the equations in (1.10) to be applied more often than the other one. The GADI iterative method can yield faster convergence than the ADI iterative method.

This paper is organized as follows. Section 2 reviews some fundamental concepts of image formation, stochastic models and linear filters, and derives the Sylvester's equation (1.9). In §3 we review results on the ADI and GADI iterative methods, and present new schemes for determining the iteration parameters τ_k and δ_k. The results of some numerical experiments in which we apply the ADI and GADI iterative methods, as well as the conjugate gradient method, to the restoration of some images are reported in §4. Concluding remarks are found in §5.

2. Image restoration. This section is divided into four subsections in which we consider image formation, stochastic models, linear filters, and the derivation of Sylvester's equation (1.9), respectively. The discussions in the first three subsections follow Andrews and Hunt [2].

2.1. Image formation. Assume that an object in the (ξ, η)-plane is illuminated by a source of radiant energy or that the object itself is a source of radiant energy. We represent this object by a radiant energy distribution function $f(\xi, \eta)$. The radiant energy reflected, transmitted, or emitted by the object propagates through space. An image formation system, e.g., a lens, intercepts the propagating radiant energy, and transforms it in such a manner that an image is formed in the (x, y)-plane. We will occasionally refer to the (ξ, η)-plane as the object plane and to the (x, y)-plane as the image plane. We represent the image by a radiant energy distribution function $g(x, y)$. Following Andrews and Hunt [2], the process of image formation is based on the following three basic principles:

1. Neighborhood processes: the image of an object point may depend on the object point and on points in a neighborhood of it.
2. Nonnegativity: the radiant energy distribution functions for the object and the image must be nonnegative, i.e.,

$$f(\xi, \eta) \geq 0\,, \qquad g(x, y) \geq 0. \tag{2.1}$$

3. Linearity: let $\tilde{h}\,(x, y, \xi, \eta, f(\xi, \eta))$ be a function that describes how the image formation system transforms energy $f(\xi, \eta)$ at the point (ξ, η) in the object plane, to energy $\tilde{h}\,(x, y, \xi, y, f(\xi, \eta))$ at the point (x, y) in the image plane. We say that the image formation system is linear if the mapping $t \to \tilde{h}(\cdot, \cdot, \cdot, \cdot, t)$ is linear, i.e., if

$$\tilde{h}\,(x, y, \xi, \eta, f(\xi, \eta)) = h(x, y, \xi, \eta) f(\xi, \eta)\,. \tag{2.2}$$

We refer to the function h as the point spread function (PSF).

For linear, as well as for nonlinear, image formation systems, the radiant energy distribution in the image plane is additive. Therefore, we can express the radiant

energy distribution in the image plane as

$$g(x,y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \tilde{h}\,(x,y,\xi,\eta,f(\xi,\eta))\,d\xi\,dy\,. \tag{2.3}$$

In particular, when $\tilde{h}$ is of the form (2.2), formula (2.3) simplifies to

$$g(x,y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h(x,y,\xi,\eta) f(\xi,\eta) d\xi\,d\eta\,. \tag{2.4}$$

When the image formation system acts uniformly across the image and object planes, i.e., when h is of the form

$$h(x,y,\xi,\eta) = \hat{h}(x-y,\xi-\eta)\,,$$

for some function $\hat{h}$, then h is said to be space invariant. We say that the PSF h is separable if it can be decomposed according to the formula

$$h(x,y,\xi,\eta) = h_1(x,y)h_2(\xi,\eta)\,.$$

Of particular interest in image restoration are point spread functions that are both space invariant and separable. Such a PSF can be written as

$$h(x,y,\xi,\eta) = \hat{h}_1(x-\xi)\hat{h}_2(y-\eta),$$

for some functions $\hat{h}_1$ and $\hat{h}_2$.

Let $G = [g_{ij}]_{1\le i\le N, 1\le j\le M}$ be an $N \times M$ matrix obtained by sampling and quantifying the radiant energy distribution function $g(x,y)$. Assume for the moment that the radiant energy distribution function $f(\xi,\eta)$ that represents the object is explicitly known, and let $F = [f_{ij}]_{1\le i\le N, 1\le j\le M}$ be an $N \times M$ matrix obtained by sampling and quantifying the function $f(\xi,\eta)$. It is convenient to represent G and F by the NM-vectors

$$\begin{aligned} \boldsymbol{g} &= [g_{11}, g_{12}, \ldots, g_{1M}, g_{21}, g_{22}, \ldots, g_{NM}]^T\,, \\ \boldsymbol{f} &= [f_{11}, f_{12}, \ldots, f_{1M}, f_{21}, f_{22}, \ldots, f_{NM}]^T\,. \end{aligned} \tag{2.5}$$

The point spread function matrix of the image formation system $H \in \mathbb{R}^{NM\times NM}$ satisfies

$$\boldsymbol{g} = H\boldsymbol{f}. \tag{2.6}$$

Properties of the PSF determine the properties of the matrix H. For instance, in view of the nonnegativity (2.1) of the radiant energy distribution, we require that

$$\begin{aligned} &f_{ij} \ge 0, \qquad g_{ij} \ge 0, \qquad 1 \le i \le N, \qquad 1 \le j \le M, \\ &h_{ij} \ge 0, \qquad\qquad\qquad 1 \le i,j \le NM\,. \end{aligned} \tag{2.7}$$

If we assume that there is no loss of radiant energy in the process of image formation, we have

$$\sum_{i=1}^{N}\sum_{j=1}^{M} f_{ij} = \sum_{i=1}^{N}\sum_{j=1}^{M} g_{ij}\,. \tag{2.8}$$

Equation (2.8) is satisfied if

$$\sum_{i=1}^{NM} h_{ij} = 1, \qquad 1 \le j \le NM . \tag{2.9}$$

If the PSF h is separable, then H can be expressed as a tensor product of two matrices H_y and H_x, i.e.,

$$H = H_y \otimes H_x .$$

If, in addition, the PSF is space invariant, then H_y and H_x are Toeplitz matrices.

2.2. Image statistical models. It is often useful in image restoration to regard a given matrix as a sample from a class of multivariate data. Let the NM-vector $\boldsymbol{g}$ be given by (2.5), and introduce its expected value vector

$$\bar{\boldsymbol{g}} := E(\boldsymbol{g})$$

and its covariance matrix Φ_g. Define the vectors

$$\boldsymbol{g}_k = [g_{k1}, g_{k2}, \ldots, g_{kM}]^T, \qquad 1 \le k \le N ,$$

and let

$$\begin{aligned} \bar{\boldsymbol{g}}_k &:= E(\boldsymbol{g}_k) , \qquad 1 \le k \le N , \\ \Phi_{kl} &:= E\left[(\boldsymbol{g}_k - \bar{\boldsymbol{g}}_k)(\boldsymbol{g}_l - \bar{\boldsymbol{g}}_l)^T\right] , \qquad 1 \le k, l \le N . \end{aligned} \tag{2.10}$$

Thus, Φ_{kl} is the covariance between rows k and l of G, and $\Phi_{kl} = \Phi_{lk}^T$. The covariance matrix Φ_g can be written in terms of the Φ_{kl} blocks

$$\Phi_g = \begin{bmatrix} \Phi_{11} & \Phi_{12} & \cdots & \Phi_{1N} \\ \Phi_{21} & \Phi_{22} & \cdots & \Phi_{2N} \\ \vdots & \vdots & & \vdots \\ \Phi_{N1} & \Phi_{N2} & \cdots & \Phi_{NN} \end{bmatrix} .$$

The vector $\boldsymbol{g}$ is said to be wide sense stationary (WSS) if it satisfies the following conditions.

(i) The expected value of $\boldsymbol{g}$ is a vector with all entries equal:

$$E(\boldsymbol{g}) = \boldsymbol{\mu} = [\mu, \mu, \ldots, \mu]^T .$$

(ii) The covariance matrix Φ_g of $\boldsymbol{g}$ is of block Toeplitz form:

$$\Phi_g = \begin{bmatrix} \Phi_{11} & \Phi_{12} & \cdots & \Phi_{1N} \\ \Phi_{12}^T & \Phi_{11} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \Phi_{12} \\ \Phi_{N1}^T & \cdots & \Phi_{12}^T & \Phi_{11} \end{bmatrix} .$$

We remark that if the mean vector $\boldsymbol{\mu}$ is known, but its entries are not all the same, then we can replace $\boldsymbol{g}$ by $\boldsymbol{g} - \boldsymbol{\mu}$. The latter vector has a zero mean vector. Thus, the form of the covariance matrix determines whether the process is stationary.

Of particular interest in image restoration are those classes of images whose covariance matrices have specific properties. We say that the covariance matrix of the image $\boldsymbol{g}$ is separable if

$$\Phi_g = \Phi_y \otimes \Phi_x \tag{2.11}$$

with

$$\Phi_z = E\{(\boldsymbol{g}_z - \bar{\boldsymbol{g}}_z)(\boldsymbol{g}_z - \bar{\boldsymbol{g}}_z)^T\}, \qquad z \in \{x, y\},$$

where $\boldsymbol{g}_z$ is the vector indicating the z-direction of the image matrix and $E\{\boldsymbol{g}_z\} = \bar{\boldsymbol{g}}_z$. If the stochastic processes $\boldsymbol{g}_x$ and $\boldsymbol{g}_y$ are WSS, then Φ_x and Φ_y are symmetric Toeplitz matrices.

A common simplifying assumption in image restoration is that the image is modeled by a separable first-order Markov process. Then the covariance matrix is of the form (2.11) with $\Phi_z = \sigma_z^2 R_z$, where

$$R_z = \begin{bmatrix} 1 & \rho_z & \rho_z^2 & \cdots & \\ \rho_z & 1 & \rho_z & \ddots & \vdots \\ \rho_z^2 & \rho_z & \ddots & \ddots & \rho_z^2 \\ \vdots & \ddots & \ddots & 1 & \rho_z \\ & \cdots & \rho_z^2 & \rho_z & 1 \end{bmatrix}, \qquad z \in \{x, y\}, \tag{2.12}$$

and ρ_z, σ_z^2 are the adjacent element correlation and variance, respectively, in the z-direction. Matrices of the form (2.12) are sometimes referred to as Kac–Murdoch–Szegő matrices; see [19] for a discussion of their properties. For future reference, we note that R_z^{-1} is tridiagonal with explicitly known entries

$$R_z^{-1} = (1-\rho_z)^{-1}(1+\rho_z)^{-1} \begin{bmatrix} 1 & -\rho_z & & & & 0 \\ -\rho_z & 1+\rho_z^2 & -\rho_z & & & \\ & -\rho_z & 1+\rho_z^2 & -\rho_z & & \\ & & \ddots & \ddots & \ddots & \\ & & & -\rho_z & 1+\rho_z^2 & -\rho_z \\ 0 & & & & -\rho_z & 1 \end{bmatrix}. \tag{2.13}$$

An application of Gershgorin's theorem, see, e.g., [30], to R_z^{-1} yields that the spectrum

of R_z^{-1} satisfies

$$\lambda(R_z^{-1}) \subset \left[\frac{1-\rho_z}{1+\rho_z}, \frac{1+\rho_z}{1-\rho_z}\right]. \tag{2.14}$$

It follows from the bounds (2.14) that

$$\lambda(R_z) \subset \left[\frac{1-\rho_z}{1+\rho_z}, \frac{1+\rho_z}{1-\rho_z}\right]. \tag{2.15}$$

2.3. Linear filters. We now assume that the model describing the image-object correspondence incorporates noise. The simplest model is expressed by the equation

$$\boldsymbol{g} = H\boldsymbol{f} + \boldsymbol{\eta}, \tag{2.16}$$

where $\boldsymbol{g}$ is the known degraded image, $\boldsymbol{f}$ is the unknown original uncorrupted image, the vector $\boldsymbol{\eta}$ contains the noise and H is the point spread function matrix. We seek to determine an approximation $\hat{\boldsymbol{f}}$ of $\boldsymbol{f}$ by applying a linear filter L to $\boldsymbol{g}$, i.e., $\hat{\boldsymbol{f}}$ is determined by (1.3). If no knowledge of the nature of the noise is assumed, a natural choice of filter might be a linear operator that minimizes

$$(H\hat{\boldsymbol{f}} - \boldsymbol{g})^T(H\hat{\boldsymbol{f}} - \boldsymbol{g}).$$

When H is nonsingular, the solution to this minimization problem is given by

$$\hat{\boldsymbol{f}} := H^{-1}\boldsymbol{g} = \boldsymbol{f} + H^{-1}\boldsymbol{\eta}.$$

The filter $L = H^{-1}$ is often referred to as the inverse filter. Computation of $\hat{\boldsymbol{f}}$ by using the inverse filter requires the solution of a linear system of equations with the $NM \times NM$ matrix H. This approach suffers from the drawback that the matrix H may be severely ill-conditioned, and, therefore, straightforward application of formula (1.3) may yield a very inaccurate estimate $\hat{\boldsymbol{f}}$, i.e., the error $\boldsymbol{\epsilon}$ given by (1.4) may be large.

The Wiener filter yields the minimum mean square error in the estimation over all possible images. This filter solves the following minimization problem:

$$\begin{aligned}\min_{\hat{f}} E\{\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}\} &= \min_{\hat{f}} E\{\operatorname{tr}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)\} \\ &= \min_{\hat{f}} E\{\operatorname{tr}(\boldsymbol{f}\boldsymbol{f}^T - L(H\boldsymbol{f}\boldsymbol{f}^T + \boldsymbol{\eta}\boldsymbol{f}^T) - (\boldsymbol{f}\boldsymbol{f}^T H^T + \boldsymbol{f}\boldsymbol{\eta}^T)L^T \\ &\qquad + L(H\boldsymbol{f}\boldsymbol{f}^T H^T + \boldsymbol{\eta}\boldsymbol{f}^T H^T + H\boldsymbol{f}\boldsymbol{\eta}^T + \boldsymbol{\eta}\boldsymbol{\eta}^T)L^T)\},\end{aligned} \tag{2.17}$$

where $\operatorname{tr}(A)$ denotes the trace of the matrix A. Under the assumption that the noise $\boldsymbol{\eta}$ and the image $\boldsymbol{f}$ are uncorrelated, we have $E(\boldsymbol{f}\boldsymbol{\eta}^T) = E(\boldsymbol{f}^T\boldsymbol{\eta}) = 0$, and the minimization problem (2.17) simplifies to

$$\min_{\hat{f}} E\{\operatorname{tr}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)\} = \min_{\hat{f}} \operatorname{tr}(\Phi_f - 2LH\Phi_f + LH\Phi_f H^T L^T + L\Phi_\eta L^T). \tag{2.18}$$

The solution of (2.18) is equivalent to the application of the filter

$$L = \Phi_f H^T(H\Phi_f H^T + \Phi_\eta)^{-1}; \tag{2.19}$$

see [2, §7.2] for details. When $H = I$, formula (2.19) simplifies to (1.5).

2.4. A Sylvester's equation. Substituting (2.19) into (1.3) yields the equation

$$(H + \Phi_\eta H^{-T} \Phi_f^{-1}) \hat{\boldsymbol{f}} = \boldsymbol{g}\,. \tag{2.20}$$

This equation reduces to (1.6) when H is the identity matrix. Assume that $H = H_y \otimes H_x$ and $\Phi_f = \Phi_y \otimes \Phi_x$, where $H_x, \Phi_x \in \mathbb{R}^{N\times N}$ and $H_y, \Phi_y \in \mathbb{R}^{M\times M}$ are symmetric matrices, and that the noise is white and Gaussian. Then (2.20) can be written as

$$\left(H_y \otimes H_x + \sigma_\eta^2 (H_y^{-1}\Phi_y^{-1}) \otimes (H_x^{-1}\Phi_x^{-1})\right) \hat{\boldsymbol{f}} = \boldsymbol{g},$$

which can be expressed in the form

$$H_y \hat{F} H_x + \sigma_\eta^2 (H_y^{-1}\Phi_y^{-1}) \hat{F} (\Phi_x^{-1} H_x^{-1}) = G\,, \tag{2.21}$$

where the entries of the $M \times N$ matrices $\hat{F}$ and G are those of the vectors $\hat{\boldsymbol{f}}$ and $\boldsymbol{g}$, respectively. Solution methods for equations of the form (2.21) are discussed in [10]. The ADI iterative methods discussed in the present paper are applicable after writing (2.21) in the form of a Sylvester's equation,

$$\hat{F} H_x^2 \Phi_x + \sigma_\eta^2 H_y^{-2} \Phi_y^{-1} \hat{F} = H_y^{-1} G H_x \Phi_x\,. \tag{2.22}$$

The computed examples of §4 consider the solution of (2.22) in the special case where H_x and H_y are identity matrices and $\Phi_z = \sigma_z^2 R_z$, $z \in \{x, y\}$, where R_z is given by (2.12).

3. ADI iteration methods. We first review the ADI iterative method for computing the solution $\hat{F} \in \mathbb{R}^{N\times M}$ of Sylvester's equation

$$FB - AF = C \tag{3.1}$$

for fairly general given matrices $A \in \mathbb{R}^{N\times N}$, $B \in \mathbb{R}^{M\times M}$ and $C \in \mathbb{R}^{N\times M}$. Equation (3.1) has a unique solution if the spectra $\lambda(A)$ and $\lambda(B)$ satisfy $\lambda(A) \cap \lambda(B) = \emptyset$; see [13]. We will assume that there are explicitly known compact sets S and T in the complex plane $\mathbb{C}$, such that

$$\lambda(A) \subset S, \quad \lambda(B) \subset T, \quad S \cap T = \emptyset\,. \tag{3.2}$$

Solution methods for (3.1) that use the structure of the linear system of equations include direct methods due to Bartels and Stewart [5] and Golub, Nash, and Van Loan [14], and iterative methods, in which the matrices A and B are reduced to Hessenberg or tridiagonal form; see [18], [26]. When sets S and T that satisfy (3.2) are explicitly known, and the matrices A and B have a structure that makes it possible to rapidly solve linear systems of equations with matrices $A - \gamma I$ and $B - \gamma I$ for certain parameters $\gamma \in \mathbb{C}$, it can be attractive to solve (3.1) by the ADI iterative method. Applications of the ADI iterative method to the solution of Sylvester's equation are reported in recent papers by Ellner and Wachspress [11], [33] and Starke [27], [28]; however, properties of the ADI iterative method were already studied in the 1950s and 1960s; see Birkhoff and Varga [6], Birkhoff, Varga, and Young [7], de Boor and Rice [9], Gaier and Todd [12], Peaceman and Rachford [25], Varga [30], and Wachspress [31], [32].

We write the ADI iterations in a manner that can be generalized later,

$$j := j+1, \quad F_{j+k} := (A-\delta_j I)^{-1}\left(F_{j+k-1}(B-\delta_j I) - C\right), \tag{3.3}$$

$$k := k+1, \quad F_{j+k} := \left((A-\tau_k I)F_{j+k-1} + C\right)(B-\tau_k I)^{-1}, \tag{3.4}$$

where we start with $j = k = 0$, and the matrix F_0 is a given initial approximate solution of (3.1). The classical ADI (CADI) iterative method proceeds by strictly alternating between formulas (3.3) and (3.4) in order to determine a sequence of approximate solutions $F_1, F_2, F_3, \ldots$ of (3.1). We would like to determine the iteration parameters δ_j and τ_k so that the matrices F_l converge rapidly to the solution $\hat{F}$ of (3.1) as l increases. A generalized ADI (GADI) iterative method, in which strict alternation between the formulas (3.3) and (3.4) is not demanded, was introduced in [21]. The analysis based on potential theory in [21] shows that the application of one of the formulas (3.3) or (3.4) more often than the other one can result in faster convergence than obtained with the CADI method. Related results have also recently been shown by Levin and Saff [22]. In §3.1 we describe the rational approximation problem that underlies the GADI iterative method. Section 3.2 presents several new algorithms suggested by the rational approximation problem for generating sequences of iteration parameters, and §3.3 shows a few computed examples with these algorithms.

3.1. A rational approximation problem. Introduce the error matrices $E_l := F_l - \hat{F}$, where $\hat{F}$ solves Sylvester's equation (3.1). Then (3.3) and (3.4) yield

$$E_{j+k} = (A-\delta_j I)^{-1} E_{j+k-1}(B-\delta_j I)$$

and

$$E_{j+k} = (A-\tau_k I)E_{j+k-1}(B-\tau_k I)^{-1},$$

respectively, and therefore

$$E_{j+k} := \prod_{p=1}^{j}(A-\delta_p I)^{-1} \prod_{q=1}^{k}(A-\tau_q I) E_0 \prod_{p=1}^{j}(B-\delta_p I) \prod_{q=1}^{k}(B-\tau_q I)^{-1}. \tag{3.5}$$

Assume that A and B are diagonalizable and have spectral decompositions $A = U_A \Lambda_A U_A^{-1}$ and $B = U_B \Lambda_B U_B^{-1}$, where Λ_A and Λ_B are diagonal matrices. Let $\|\cdot\|$ denote the spectral norm and define for any nonsingular square matrix U its condition number $\mathcal{H}(U) := \|U\|\,\|U^{-1}\|$. Then (3.5) yields the bound

$$\|E_{j+k}\| \le \max_{z\in\lambda(A)} \frac{\prod_{q=1}^{k}|z-\tau_q|}{\prod_{p=1}^{j}|z-\delta_p|} \cdot \max_{z\in\lambda(B)} \frac{\prod_{p=1}^{j}|z-\delta_p|}{\prod_{q=1}^{k}|z-\tau_q|} \cdot \|E_0\|\mathcal{H}(U_A)\mathcal{H}(U_B). \tag{3.6}$$

Introduce the rational function

$$r_{kj}(z) := \frac{\prod_{q=1}^{k}(z-\tau_q)}{\prod_{p=1}^{j}(z-\delta_p)}. \tag{3.7}$$

In view of (3.2) and (3.6), we obtain the bound

$$\|E_{j+k}\| \le \frac{\max_{z\in S}|r_{kj}(z)|}{\min_{z\in T}|r_{kj}(z)|} \cdot \|E_0\|\mathcal{H}(U_A)\mathcal{H}(U_B).$$

This bound suggests the following approximation problem:

$$\overline{\lim_{j+k\to\infty}} \inf_{\tau_q,\delta_p} \left(\frac{\max_{z\in S} |r_{kj}(z)|}{\min_{z\in T} |r_{kj}(z)|} \right)^{1/(j+k)} . \tag{3.8}$$

We refer to this limit as the asymptotically optimal rate of convergence of the GADI iterative method with respect to S and T. Special cases of the problem (3.8) have been studied thoroughly. For instance, when S and T are real intervals and $j = k$, optimal parameters δ_j and τ_k can be determined by evaluating certain elliptic functions; see [9], [12], [31], [32]. It is well known that the limit (3.8) exists for fairly general disjoint compact sets S and T in $\mathbb{C}$ when $j = k$, and equals $\exp(-\frac{1}{2C(S,T)})$, where $C(S,T)$ denotes the capacity of the condenser formed by "the plates" S and T; see Bagby [3], [4] for details. An analysis of the limit (3.8) for $j \neq k$ was first presented in [21], and more recently in [22].

3.2. Generation of iteration parameters. Bagby [4] proposed the following algorithm for determining poles δ_j and zeros τ_k of the rational function (3.7).

ALGORITHM 3.1. (Bagby Points)
Choose $\tau_1 \in S$ and $\delta_1 \in T$ such that $|\tau_1 - \delta_1| = \max_{\substack{\tau\in S \\ \delta\in T}} |\tau - \delta|$;

for $l := 1, 2, \ldots$ do
  Choose $\tau_{l+1} \in S$ such that $|r_{ll}(\tau_{l+1})| = \max_{z\in S} |r_{ll}(z)|$;
  Choose $\delta_{l+1} \in T$ such that $|r_{ll}(\delta_{l+1})| = \min_{z\in T} |r_{ll}(z)|$;
end l □

We call any sequence of points $\delta_1, \tau_1, \delta_2, \tau_2, \ldots$ in $\mathbb{C}$ determined by Algorithm 3.1 a sequence of Bagby points for the sets S and T. Bagby [4] showed that, under mild regularity conditions on S and T, the parameters δ_l and τ_l determined in this manner solve the approximation problem (3.8) for $j = k$. Bagby [4] only requires $\tau_1 \in S$ and $\delta_1 \in T$. We have found that the choice of δ_l and τ_l made in the algorithm is suitable in the sense that, typically, it makes the quotient $\max_{z\in S} |r_{ll}(z)| \,/\, \min_{z\in T} |r_{ll}(z)|$ fairly small already for small values of l.

In [21, pp. 227–228] and [22, §7] several generalizations of Bagby points are presented that solve (3.8) as $j \to \infty$ and $k = \alpha j$ for some rational constant $\alpha > 0$. Here j is chosen so that αj is an integer. Unfortunately, all of these generalized Bagby points are very cumbersome to compute. In Algorithms 3.2–3.4 we therefore describe modifications of Algorithm 3.1 for the generation of generalized Bagby points that only require a small computational effort. We illustrate the behavior of the generalized Bagby points defined by Algorithms 3.2–3.4 with computed examples in §3.3.

When the matrices A and B in (3.1) are of different sizes or have different structure it may be considerably faster to compute F_{j+k} from (3.3) than from (3.4), or vice versa. We therefore would like to generate parameters δ_j and τ_k so that the indices j and k are increased in a manner that makes the ratio k/j close to a given value α. The computational effort to determine F_l, for given $l = j+k$, depends on α and so does the rate of convergence. The tables in §3.3 illustrate the convergence for different values of α and can be helpful when determining a ratio α that minimizes the computational effort to determine an approximate solution F_l of desired accuracy. Algorithm 3.2 has been described previously in [21, §4] for the special case $\alpha = 2$.

ALGORITHM 3.2. (Generalized Bagby points $\{\delta_\ell\}_{\ell=1}^{j}$ and $\{\tau_\ell\}_{\ell=1}^{k}$ with $k/j \approx \alpha$)
Choose $\tau_1 \in S$ and $\delta_1 \in T$ such that $|\tau_1 - \delta_1| = \max_{\substack{\tau \in S \\ \delta \in T}} |\tau - \delta|$; $j := k := 1$;

for $\ell := 1, 2, \ldots$ do
  if $|\alpha - \frac{k+1}{j}| > |\alpha - \frac{k}{j+1}|$ or $|\alpha - \frac{k+1}{j}| = |\alpha - \frac{k}{j+1}|$ and $\alpha > \frac{k}{j}$ then
    Choose $\delta_{j+1} \in T$ such that $|r_{kj}(\delta_{j+1})| = \min_{z \in T} |r_{kj}(z)|$; $j := j + 1$
  else
    Choose $\tau_{k+1} \in S$ such that $|r_{kj}(\tau_{k+1})| = \max_{z \in S} |r_{kj}(z)|$; $k := k + 1$
  endif
end ℓ □

The algorithm above determines numerator degree k and denominator degree j so that the ratio k/j is close to a given value of $\alpha > 0$. Algorithm 3.3 below also seeks to determine a ratio of the numerator degree j and denominator degree k such that $\max_{z \in S} |r_{kj}(z)| / \min_{z \in T} |r_{kj}(z)|$ is small. In each step of the algorithm we determine one new zero and one new pole, and this defines $r_{k+1,j}$ and $r_{k,j+1}$, respectively. If

$$\frac{\max_{z \in S} |r_{k+1,j}(z)|}{\min_{z \in T} |r_{k+1,j}(z)|} \leq \frac{\max_{z \in S} |r_{k,j+1}(z)|}{\min_{z \in T} |r_{k,j+1}(z)|},$$

then we choose $r_{k+1,j}$ as our new rational function; otherwise we choose $r_{j,k+1}$. In this manner each step of the algorithm increases either the numerator or denominator degree.

ALGORITHM 3.3. (Generalized Bagby points $\{\delta_\ell\}_{\ell=1}^{j}$ and $\{\tau_\ell\}_{\ell=1}^{k}$ with k/j determined adaptively)
Choose $\tau_1 \in S$ and $\delta_1 \in T$ such that $|\tau_1 - \delta_1| = \max_{\substack{\tau \in S \\ \delta \in T}} |\tau - \delta|$; $j := k := 1$;

for $\ell := 1, 2, \ldots$ do
  Choose $\delta_{j+1} \in T$ such that $|r_{kj}(\delta_{j+1})| = \min_{z \in T} |r_{kj}(z)|$;
  Choose $\tau_{k+1} \in S$ such that $|r_{kj}(\tau_{k+1})| = \max_{z \in S} |r_{kj}(z)|$;
  if $\max_{z \in S} |r_{k+1,j}(z)| / \min_{z \in T} |r_{k+1,j}(z)| \leq \max_{z \in S} |r_{k,j+1}(z)| / \min_{z \in T} |r_{k,j+1}(z)|$ then
    $k := k + 1$
  else
    $j := j + 1$
  endif
end ℓ □

Computed examples of §3.3 indicate that Algorithm 3.3 may determine a sequence of rational functions that is close to optimal.

Algorithm 3.3 suggests a modification of Algorithm 3.1 for generating Bagby points. This modification is described by Algorithm 3.4 below. In this algorithm the points δ_{j+1} and τ_{j+1} are determined sequentially, i.e., we first determine δ_{j+1} and then τ_{j+1}, and vice versa, and then choose the one of the two pairs of points determined that makes the quotient $\max_{z \in T} |r_{j+1,j+1}(z)| / \min_{z \in S} |r_{j+1,j+1}(z)|$ smaller.

ALGORITHM 3.4. (Modified Bagby Points)
Choose $\tau_1 \in S$ and $\delta_1 \in T$ such that $|\tau_1 - \delta_1| = \max_{\substack{\tau \in S \\ \delta \in T}} |\tau - \delta|$;

for $j = 1, 2, \ldots$ do
Choose $\delta_{j+1} \in T$ such that $|r_{jj}(\delta_{j+1})| = \min_{z \in T} |r_{jj}(z)|$;
Choose $\tau_{j+1} \in S$ such that $|r_{j,j+1}(\tau_{j+1})| = \max_{z \in S} |r_{j,j+1}(z)|$;
$M' := \max_{z \in S} |r_{j+1,j+1}(z)| / \min_{z \in T} |r_{j+1,j+1}(z)|$;
$\delta' := \delta_{j+1}; \tau' := \tau_{j+1}$;

Choose $\tau_{j+1} \in S$ such that $|r_{jj}(\tau_{j+1})| = \max_{z \in S} |r_{jj}(z)|$;
Choose $\delta_{j+1} \in T$ such that $|r_{j+1,j}(\delta_{j+1})| = \min_{z \in T} |r_{j+1,j}(z)|$;
$M'' := \max_{z \in S} |r_{j+1,j+1}(z)| / \min_{z \in T} |r_{j+1,j+1}(z)|$;
$\delta'' := \delta_{j+1}; \tau'' := \tau_{j+1}$;
if $M' \leq M''$ then
$\delta_{j+1} := \delta'; \tau_{j+1} := \tau'$
else
$\delta_{j+1} := \delta''; \tau_{j+1} := \tau''$
endif
end j

Computed examples indicate that the sequence $\max_{z \in S} |r_{jj}(z)| / \min_{z \in S} |r_{jj}(z)|, j = 0, 1, 2, \ldots$ generated by Algorithm 3.4 typically decreases in a smoother fashion than the corresponding sequence of quotients obtained by Algorithm 3.1.

3.3. Computed examples. Several schemes for allocating zeros $\{\tau_q\}_{q=1}^{k}$ and poles $\{\delta_p\}_{p=1}^{j}$ of rational functions (3.7) with a prescribed rational quotient $k/j = \alpha$ such that

$$\overline{\lim_{\substack{j \to \infty \\ \alpha j, j \in \mathbb{N}}}} \left(\frac{\max_{z \in S} |r_{\alpha j, j}(z)|}{\min_{z \in T} |r_{\alpha j, j}(z)|} \right)^{\frac{1}{j + \alpha j}} \tag{3.9}$$

$$= \overline{\lim_{\substack{j \to \infty \\ \alpha j, j \in \mathbb{N}}}} \min_{\tau_q, \delta_p} \left(\max_{z \in S} \frac{\prod_{q=1}^{\alpha j} |z - \tau_q|}{\prod_{p=1}^{j} |z - \delta_p|} \cdot \max_{z \in T} \frac{\prod_{p=1}^{j} |z - \delta_p|}{\prod_{q=1}^{\alpha j} |z - \tau_q|} \right)^{\frac{1}{j + \alpha j}}$$

have been proposed; see [21], [22]. The allocation schemes proposed generalize Fejér points, Fekete points, and Bagby points. These generalizations share the property that they define points which are quite cumbersome to determine numerically. The allocation schemes defined by Algorithms 3.2–3.4 are easier to implement on a computer, and this is the main motivation for our interest in these allocation schemes. Moreover, Algorithm 3.3 is the only available scheme that seeks to determine adaptively a ratio k/j such that the limit superior

$$\overline{\lim_{j+k \to \infty}} \left(\frac{\max_{z \in S} |r_{kj}(z)|}{\min_{z \in T} |r_{kj}(z)|} \right)^{1/(k+j)} \tag{3.10}$$

of the computed rational functions r_{kj} is minimal for fairly general sets S and T in the complex plane. We remark that Levin and Saff [22] showed how to determine an

TABLE 3.1
$S = [-9.136, -1.095 \cdot 10^{-2}]$, $T = [3.976 \cdot 10^{-1}, 2.515]$.

	Alg. 3.1			Alg. 3.4			Alg. 3.3		
$j+k$	j	k	ϵ_{kj}	j	k	ϵ_{kj}	j	k	ϵ_{kj}
2	1	1	- 0.1	1	1	- 0.1	1	1	- 0.1
8	4	4	-2.1	4	4	-2.2	3	5	-2.4
14	7	7	-5.3	7	7	-5.1	5	9	-5.9
20	10	10	-7.9	10	10	-8.9	6	14	-8.9
26	13	13	-11	13	13	-11	9	17	-12
32	16	16	-14	16	16	-14	11	21	-16

optimal ratio $\alpha = k/j$ by using elliptic functions if S and T are disjoint real intervals. Example 3.4 below compares the rational functions obtained by Algorithm 3.3 with numerical results presented by Levin and Saff [22]. This example suggests that the rational functions determined by Algorithm 3.3 may be optimal or nearly optimal. Examples 3.1 and 3.3 use sets S and T that arise in our applications to the restoration of noisy images presented in §4. The tables show the quantities

$$\epsilon_{kj} := \log_{10}\left(\max_{z\in S} |r_{kj}(z)| / \min_{z\in T} |r_{kj}(z)|\right),$$

where r_{kj} are the computed rational functions.

Example 3.1. Let $S := [-9.136, -1.095 \cdot 10^{-2}]$ and $T := [3.976 \cdot 10^{-1}, 2.515]$. Algorithms 3.1–3.4 yield Tables 3.1 and 3.2. In Tables 3.1 and 3.2 the ϵ_{kj} are smaller for $k > j$ than for $k < j$ when $j + k$ is kept fixed. This is in agreement with the theory developed in [21], [22], because the interval T is shorter than the interval S. A comparison of the $\epsilon_{\ell\ell}$ obtained from Algorithms 3.1, 3.4, and 3.2 with $\alpha = 1$ suggests that these schemes yield $\epsilon_{\ell\ell}$ with similar asymptotic behavior as ℓ increases. However, for a fixed value of ℓ, the values $\epsilon_{\ell\ell}$ determined by the three algorithms can differ substantially. As ℓ increases, the decrease of the $\epsilon_{\ell\ell}$ determined by any one of the algorithms is generally not monotonic when the sets S and T are close. Computed examples with many sets S and T indicate that the $\epsilon_{\ell\ell}$ determined by Algorithm 3.4 typically decrease more smoothly as ℓ increases than the $\epsilon_{\ell\ell}$ obtained by Algorithms 3.1 and 3.2 with $\alpha = 1$. Algorithm 3.3 determines j and k adaptively as $j+k$ increases. For $j + k$ fixed, the ϵ_{kj} obtained by this algorithm are the smallest ones, or close to the smallest ones, when compared to the ϵ_{kj} determined by the other algorithms of this section.

Example 3.2. Let $S := [-4.740 \cdot 10^{1}, -2.110 \cdot 10^{-3}]$ and $T := [1.603 \cdot 10^{-3}, 6.249 \cdot 10^{2}]$. Algorithms 3.1–3.4 yield Tables 3.3 and 3.4. In the tables it appears that Algorithm 3.2 (with $\alpha = \frac{1}{2}$ and $\alpha = 1$) and Algorithms 3.4 and 3.3 yield the smallest values of ϵ_{kj} for fixed $j + k$. The length of T is larger than the length of S. The theory in [21], [22] shows that in order to make ϵ_{kj} decrease as rapidly as possible as $j + k$ increases, we should let j be larger than k. The entries ϵ_{kj} of Tables 3.3 and 3.4 are in agreement with this observation.

Example 3.3. Let $S := [-8.771, -1.140 \cdot 10^{-3}]$ and $T := [9.405 \cdot 10^{-3}, 1.063 \cdot 10^{2}]$. Algorithms 3.1–3.4 yield Tables 3.5 and 3.6.

The decrease of the ϵ_{kj} generated by any one of the algorithms of this section is not always monotonic when $j + k$ increases. It is therefore difficult to capture the performance of the algorithms by one table. Nevertheless, Tables 3.1–3.6 together

TABLE 3.2
$S = [-9.136, -1.095 \cdot 10^{-2}]$, $T = [3.976 \cdot 10^{-1}, 2.515]$.

	$\alpha = 4$ Alg. 3.2			$\alpha = 2$ Alg. 3.2			$\alpha = 1$ Alg. 3.2			$\alpha = \frac{1}{2}$ Alg. 3.2			$\alpha = \frac{1}{4}$ Alg. 3.2		
$j+k$	j	k	ϵ_{kj}	j	k	ϵ_{kj}	j	k	ϵ_{kj}	j	k	ϵ_{kj}	j	k	ϵ_{kj}
2	1	1	- 0.1	1	1	- 0.1	1	1	- 0.1	1	1	- 0.1	1	1	- 0.1
8	2	6	-2.3	3	5	-2.8	4	4	-2.2	5	3	-2.2	6	2	-1.7
14	3	11	-5.7	5	9	-5.9	7	7	-5.2	9	5	-4.6	11	3	-3.6
20	4	16	-7.9	7	13	-9.3	10	10	-8.4	13	7	-6.8	16	4	-5.7
26	5	21	-12	9	17	-12	13	13	-11	17	9	-9.6	21	5	-7.3
32	6	26	-15	11	21	-15	16	16	-14	21	11	-12	26	6	-9.8

TABLE 3.3
$S = [-4.740 \cdot 10^{1}, -2.110 \cdot 10^{-3}]$, $T = [1.603 \cdot 10^{-3}, 6.249 \cdot 10^{2}]$.

	Alg. 3.1			Alg. 3.4			Alg. 3.3		
$j+k$	j	k	ϵ_{kj}	j	k	ϵ_{kj}	j	k	ϵ_{kj}
2	1	1	0.0	1	1	0.0	1	1	0.0
8	4	4	1.5	4	4	0.3	4	4	0.6
14	7	7	-1.0	7	7	-0.3	7	7	0.0
20	10	10	-0.4	10	10	-1.2	10	10	-1.6
26	13	13	-1.4	13	13	-2.3	15	11	-2.3
32	16	16	-3.0	16	16	-4.0	18	14	-3.5

suggest that Algorithm 3.4 typically yields smaller values of ϵ_{kj} than Algorithm 3.1, and Algorithm 3.3 yields values of ϵ_{kj} which are never much larger, and sometimes considerably smaller, than the ϵ_{kj} (for fixed $j+k$) obtained with the other schemes.

Example 3.4. This example compares Algorithms 3.2–3.3 with numerical examples reported by Levin and Saff [22]. Instead of ϵ_{kj}, Levin and Saff [22, Table 8.1] determine the quantities

$$\mathcal{F}_{kj} := -(j+k)^{-1} \cdot \epsilon_{kj} \cdot \ln(10)$$

for $S = [-\frac{2}{3}, 1]$ and $T = [-1, -\frac{5}{6}]$. They determine the $\mathcal{F}_{kj}$ for Fejér–Walsh points τ_k and δ_j, a generalization of Fejér points discussed in [22], [28], and for points they refer to as Leja–Bagby points. These points were previously introduced in [21, Remark 3.4]. The interest in Fejér–Walsh points and Leja–Bagby points stems from the fact that for a given rational value of α, the $\mathcal{F}_{kj}$, for $k = \alpha j$, can be shown to converge to a constant $\mathcal{F}(\alpha)$ as $j \to \infty$ and j/α is an integer, and the limit $\mathcal{F}(\alpha)$ is maximal for all choices of rational functions $r_{\alpha j, j}$. The theory presented in [22] makes it possible to determine the value of α that maximizes $\mathcal{F}(\alpha)$ by evaluating certain elliptic functions. For the intervals S and T of the present example Levin and Saff [22] find that $\mathcal{F}(\alpha)$ is maximized for $\alpha = 4.28$, and they compute $\mathcal{F}_{kj}$ for $k = 4j$ for Fejér–Walsh and Leja–Bagby points.

Table 3.7 supports the conjecture that the limit $\mathcal{F}(\alpha)$ obtained by Algorithm 3.2 is the same as for Fejér–Walsh and Bagby–Leja points. The entries marked with an asterisk (*) are from [22]. The table also suggests that the $\mathcal{F}_{kj}$ determined by Algorithm 3.3 satisfy

$$\lim_{j+k \to \infty} \mathcal{F}_{kj} = \sup_{\alpha} \mathcal{F}(\alpha).$$

TABLE 3.4
$S = [-4.740 \cdot 10^1, -2.110 \cdot 10^{-3}]$, $T = [1.603 \cdot 10^{-3}, 6.249 \cdot 10^2]$.

	$\alpha = 4$ Alg. 3.2			$\alpha = 2$ Alg. 3.2			$\alpha = 1$ Alg. 3.2			$\alpha = \frac{1}{2}$ Alg. 3.2			$\alpha = \frac{1}{4}$ Alg. 3.2		
$j+k$	j	k	ϵ_{kj}	j	k	ϵ_{kj}	j	k	ϵ_{kj}	j	k	ϵ_{kj}	j	k	ϵ_{kj}
2	1	1	0.0	1	1	0.0	1	1	0.0	1	1	0.0	1	1	0.0
8	2	6	1.4	3	5	0.4	4	4	0.3	5	3	0.4	6	2	-0.1
14	3	11	1.6	5	9	-0.6	7	7	-0.5	9	5	-0.6	11	3	1.2
20	4	16	-0.1	7	13	-1.2	10	10	-1.0	13	7	-1.3	16	4	-0.2
26	5	21	-0.8	9	17	-2.9	13	13	-2.3	17	9	-2.2	21	5	0.0
32	16	16	-1.3	11	21	-2.8	16	16	-3.9	21	11	-3.6	26	6	-1.0

TABLE 3.5
$S = [-8.771, -1.140 \cdot 10^{-3}]$, $T = [9.405 \cdot 10^{-3}, 1.063 \cdot 10^2]$.

	Alg. 3.1			Alg. 3.4			Alg. 3.3		
$j+k$	j	k	ϵ_{kj}	j	k	ϵ_{kj}	j	k	ϵ_{kj}
2	1	1	0.0	1	1	0.0	1	1	0.0
8	4	4	0.8	4	4	0.0	4	4	0.8
14	7	7	-1.5	7	7	-0.4	8	6	-1.8
20	10	10	-1.4	10	10	-2.2	11	9	-1.5
26	13	13	-3.0	13	13	-2.6	12	14	-3.4
32	16	16	-4.7	16	16	-5.0	13	15	-4.8

Table 3.8 illustrates that the ϵ_{jk}, $k = \alpha j$, obtained by Algorithm 3.2 are not very sensitive to perturbations in α for $\alpha \approx \alpha^*$, where $\mathcal{F}(\alpha^*) = \sup \mathcal{F}(\alpha)$. The insensitivity of ϵ_{kj} to perturbations in α suggests that α can be chosen different from α^* in order to reduce the computational effort required for the iteration (3.3)–(3.4) without reducing the rate of convergence very much.

4. Examples of image restoration. This section considers the iterative solution of equation (2.22) derived in §2.4. As pointed out in §2, the matrices in (2.22) typically have a structure. By using this structure we can obtain rapid iterative methods for (2.22). For instance, when $H_x = H_y = I$, and R_x and R_y are positive definite symmetric Toeplitz matrices, then the CADI or GADI iterative methods can be implemented efficiently by using a superfast Toeplitz solver, such as the one developed by Ammar and Gragg [1]. A recent discussion and comparison of iterative methods is presented by Lagendijk and Biemond [20]. The construction of preconditioners for the iterative solution of (2.16) when H is a block Toeplitz matrix with Toeplitz blocks is discussed by Hanke, Nagy, and Plemmons [16], [23], [24].

The computed examples of this section consider the iterative solution of (2.22) when $H_x = H_y = I$, and R_x and R_y are Toeplitz matrices of the form (2.12). The simple form of R_x^{-1} and R_y^{-1} (see (2.13)) makes the solution of (2.22) by the CADI or GADI iterative methods attractive. Thus, we identify

$$A = -\frac{\sigma_\eta^2}{\sigma^2} R_y^{-1}, \qquad B = R_x, \tag{4.1}$$

TABLE 3.6
$S = [-8.771, -1.140 \cdot 10^{-3}]$, $T = [9.405 \cdot 10^{-3}, 1.063 \cdot 10^{2}]$.

	$\alpha = 4$ Alg. 3.2			$\alpha = 2$ Alg. 3.2			$\alpha = 1$ Alg. 3.2			$\alpha = \frac{1}{2}$ Alg. 3.2			$\alpha = \frac{1}{4}$ Alg. 3.2		
$j+k$	j	k	ϵ_{kj}	j	k	ϵ_{kj}	j	k	ϵ_{kj}	j	k	ϵ_{kj}	j	k	ϵ_{kj}
2	1	1	0.0	1	1	0.0	1	1	0.0	1	1	0.0	1	1	0.0
8	2	6	1.8	3	5	0.1	4	4	-0.3	5	3	-0.2	6	2	-0.2
14	3	11	1.2	5	9	-0.8	7	7	-1.2	9	5	-1.3	11	3	-0.3
20	4	16	0.1	7	13	-1.9	10	10	-2.2	13	7	-2.8	16	4	-1.5
26	5	21	-1.7	9	17	-2.1	13	13	-3.7	17	9	-2.9	21	5	-1.6
32	6	26	-1.3	11	21	-3.7	16	16	-5.4	21	11	-4.5	26	6	-3.9

TABLE 3.7
$S = [-\frac{2}{3}, 1]$, $T = [-1, -\frac{5}{6}]$.

	$\alpha = 4$ Fejér-Walsh			$\alpha = 4$ Bagby-Leja			$\alpha = 4$ Alg. 3.2			Alg. 3.3		
$j+k$	j	k	$\mathcal{F}_{kj}$	j	k	$\mathcal{F}_{kj}$	j	k	$\mathcal{F}_{kj}$	j	k	$\mathcal{F}_{kj}$
20	4	16	1.94*	4	16	1.48*	4	16	1.73	5	15	1.72
60	12	48	1.99*	12	48	1.81*	12	48	1.90	18	42	1.85

where $\sigma := \sigma_x \sigma_y$. Then the iterative scheme (3.3)–(3.4) becomes

$$j := j+1, \quad \left(\frac{\sigma_\eta^2}{\sigma^2} R_y^{-1} + \delta_j I\right) F_{j+k} = F_{j+k-1}\,(\delta_j I - R_x) + G R_x, \tag{4.2}$$

$$k := k+1, \quad F_{j+k}\,(R_x - \tau_k I) = -\left(\frac{\sigma_\eta^2}{\sigma^2} R_y^{-1} + \tau_k I\right) F_{j+k-1} + G R_x. \tag{4.3}$$

In view of (4.1) and (2.14)–(2.15), the sets

$$\begin{aligned} S &:= [-\tfrac{\sigma_\eta^2}{\sigma^2} \cdot \tfrac{1+\rho_y}{1-\rho_y}, -\tfrac{\sigma_\eta^2}{\sigma^2} \cdot \tfrac{1-\rho_y}{1+\rho_y}], \\ T &:= [\tfrac{1-\rho_x}{1+\rho_x}, \tfrac{1+\rho_x}{1-\rho_x}], \end{aligned} \tag{4.4}$$

satisfy (3.2). Note that the matrices $\left(\frac{\sigma_\eta^2}{\sigma^2} R_y^{-1} + \delta_j I\right)$ and $(R_x - \tau_k I)$ are positive definite for $\delta_j \in T$ and $\tau_k \in S$.

The computational work required to carry out the iterations (4.2)–(4.3) can be reduced by replacing the Toeplitz matrices $(\delta_j I - R_x)$ and $(R_x - \tau_k I)$ by tridiagonal matrices as follows. Let $\boldsymbol{w}$ be a given vector. We compute the vector $\boldsymbol{v}$ defined by

$$\boldsymbol{v} := \boldsymbol{w}(\delta_j I - R_x) \tag{4.5}$$

by first using formula (2.13) to determine

$$\boldsymbol{u} := \boldsymbol{w}(\delta_j R_x^{-1} - I) \tag{4.6}$$

and then solving

$$\boldsymbol{v} R_x^{-1} = \boldsymbol{u} \tag{4.7}$$

TABLE 3.8
$S = [-\frac{2}{3}, 1]$, $T = [-1, -\frac{5}{6}]$.

	$\alpha = 4$ Alg. 3.2			$\alpha = 2$ Alg. 3.2			Alg. 3.3			Alg. 3.1		
$j+k$	j	k	ϵ_{kj}	j	k	ϵ_{kj}	j	k	ϵ_{kj}	j	k	ϵ_{kj}
20	4	16	-15.0	7	13	-14.5	5	15	-14.9	10	10	-12.2
40	8	32	-32.8	13	27	-31.4	12	28	-31.2	20	20	-25.9
60	12	48	-49.6	20	20	-47.9	18	42	-48.1	30	30	-40.4

for $\boldsymbol{v}$ by using the Cholesky factorization of R_x^{-1}. The count of arithmetic operations for computing $\boldsymbol{v}$ in this manner grows only linearly with the size of $\boldsymbol{v}$. Similarly, given the vector $\boldsymbol{w}$, we solve

$$\boldsymbol{v}(R_x - \tau_k I) = \boldsymbol{w} \tag{4.8}$$

for $\boldsymbol{v}$ by first evaluating $\boldsymbol{u} := \boldsymbol{w}R_x^{-1}$ and then solving $\boldsymbol{v}(I - \tau_k R_x^{-1}) = \boldsymbol{u}$ for $\boldsymbol{v}$. We will assume that the CADI and GADI iteration methods have been implemented as outlined above when discussing requirements of arithmetic operations.

For comparison we also solve (2.22) by the conjugate gradient (CG) algorithm, i.e., we apply the CG algorithm to the solution of (2.20) with $H_x = H_y = I$ and initial vector $\boldsymbol{f}_0 := \boldsymbol{g}$. Each iteration by the CG method requires the computation of a matrix-vector product with the matrix $R_y^{-1} \otimes R_x^{-1}$ and the evaluation of two inner products with NM-vectors; see, e.g., [15]. Note that each iteration by the CADI and GADI methods, i.e., each application of formulas (4.2) or (4.3), requires fewer arithmetic operations than one iteration by the CG method.

We remark that iterative methods other than the ones used in our computed examples may be attractive. For instance, a preconditioned CG method with a GADI-preconditioner might perform well, but this kind of method requires further investigation. For discussions on ADI-preconditioners, see Hochbruck and Starke [17] and references therein. In these methods the structure of the linear system is used by the preconditioner, but not by the CG method. Preconditioned iterative methods in which both the preconditioner and the iterative method use the structure of the linear systems deserve further study.

The matrices of the examples of the present section have a structure that makes it possible to carry out each iteration with the CADI or GADI methods rapidly, and we doubt that much can be gained by using more complicated iterative methods for the solution of the linear systems in our examples. The study of alternative iterative methods would appear to be attractive primarily when Wiener filters with more complicated covariance matrices are used.

We turn to the description of the computed examples. Let $\tilde{F}$ be an $N \times M$ matrix that represents an image, i.e., let the entries of $\tilde{F}$ represent pixel values. In our examples each pixel is represented by eight bits, and each entry of $\tilde{F}$ is an integer between 0 and 255. Compute the mean μ of the entries of $\tilde{F}$ and form

$$F := \tilde{F} - \mu \boldsymbol{e}_N \boldsymbol{e}_M^T,$$

where $\boldsymbol{e}_j = [1, 1 \ldots, 1]^T \in \mathbb{R}^j$. Let σ^2 be the variance of the entries of F, and let ρ_x and ρ_y be the adjacent element correlation of the entries of F in the x- and y-direction, respectively. Generate white Gaussian noise with variance σ_η^2 and add to F. The matrix $G + \mu \boldsymbol{e}_N \boldsymbol{e}_M^T$ represents the noisy image, and G represents the noisy

image modified to have zero mean. The variance σ_η^2 is chosen to yield a specific signal-to-noise ratio

$$\text{SNR} := 10 \log_{10} \left(\frac{\sigma^2}{\sigma_\eta^2} \right) \text{ dB}.$$

We choose the initial approximate solution F_0 in the iterative methods to be the matrix G. Introduce the norm for $H = [h_{ij}] \in \mathbb{R}^{N \times M}$,

$$\|H\|_* = \max_{ij} |h_{ij}|.$$

We terminate the iterations as soon as the difference between two consecutive iterates F_ℓ satisfies

$$\|F_\ell - F_{\ell-1}\|_* < \frac{1}{2}. \tag{4.9}$$

Let F_q be the final iterate. The restored image $\hat{F}$ is obtained from

$$\hat{F} := \text{int } (F_q + \mu e_N e_M^T). \tag{4.10}$$

The operator int in (4.10) rounds each entry of the matrix to the closest integer, replaces negative entries by zero, and replaces entries larger than 255 by 255. The computations were carried out on an IBM RISC 6000/550 workstation in single precision arithmetic, i.e., with about seven significant digits.

Continued iteration after the criterion (4.9) was satisfied did not yield images that could be distinguished by visual inspection from the images presented. Moreover, the restored images $\hat{F}$ obtained by the different iterative schemes could not be distinguished visually. We show only one of the restored pictures and refer to it as the restored image.

FIG. 4.1. 240 × 256 *pixels*, SNR=5 dB.

Example 4.1. Let $\tilde{F}$ be a 240 × 256 matrix that represents an uncorrupted image with variance $\sigma^2 = 4.149 \cdot 10^3$ and adjacent element correlations $\rho_x = 0.4310$ and

FIG. 4.2. 240 × 256 *pixels, restored image.*

TABLE 4.1
Number of iterations required to restore Fig. 4.1.

Method	Number of iterations		
	$j+k$	k	j
CADI	10	5	5
GADI	7	4	3
CG	13		

TABLE 4.2
Number of iterations required to restore Fig. 4.3.

Method	Number of iterations		
	$j+k$	k	j
CADI	18	9	9
GADI	16	8	8
CG	51		

$\rho_y = 0.9331$. White Gaussian noise with variance $\sigma_\eta^2 = 1.825 \cdot 10^3$ is generated in order to obtain the matrix G as described above with SNR $= 5$ dB. Figure 4.1 displays the matrix given by $G + \mu \boldsymbol{e}_{240} \boldsymbol{e}_{256}^T$, where μ is the average of the entries of $\tilde{F}$. The restored image is shown by Fig. 4.2. Table 4.1 displays the number of iterations required. The sets S and T used for CADI and GADI iteration are determined by (4.4) and are those of Example 3.1. The parameters for GADI iterations are generated by Algorithm 3.3. For comparison we use parameters determined by Algorithm 3.1 for CADI iteration because the theoretical properties of this algorithm are well understood.

Example 4.2. Let $\tilde{F}$ be a 254×244 matrix that represents an uncorrupted image with variance $\sigma^2 = 5.775 \cdot 10^3$ and adjacent element correlations $\rho_x = 0.9505$ and $\rho_y = 0.9766$. We generate white Gaussian noise with variance $\sigma_\eta^2 = 1.826 \cdot 10^3$ in order to obtain the matrix G with SNR $= 5$ dB. Figure 4.3 shows the matrix $G + \mu \boldsymbol{e}_{254} \boldsymbol{e}_{244}^T$, where μ is the average of the entries of $\tilde{F}$. The restored image is shown by Fig. 4.4,

FIG. 4.3. 254 × 244 *pixels,* SNR=5 dB.

FIG. 4.4. 254 × 244 *pixels, restored image.*

and Table 4.2 displays the number of iterations necessary for restoration. The sets $S = [-2.609 \cdot 10^1, -3.747 \cdot 10^{-3}]$ and $T = [2.514 \cdot 10^{-2}, 3.977 \cdot 10^1]$ are determined by (4.4). The parameters for CADI and GADI iteration are determined in the same manner as in Example 4.1. This example illustrates that very noisy images restored by a Wiener filter require further processing in order to yield visually pleasing images; see [20] for a discussion. The difference in iteration numbers in this example and Example 4.1 depends on the sets S and T differing.

5. Conclusion. The paper presents several new schemes for generating parameters for the classical and generalized ADI iterative method. The computed examples of §4 show these iterative schemes to be competitive with the CG method. Moreover, they show the GADI iteration method to yield faster convergence than the classical

ADI iteration method.

Acknowledgments. We would like to thank Greg Ammar, Eric Grosse, and Salvatore Morgera for providing the pictures used in §4. The paper was completed during a visit to the University of Bologna. We would like to thank Fiorella Sgallari for making this visit possible and enjoyable. Finally, we would like to thank Martin Gutknecht and a referee for carefully reading the manuscript.

REFERENCES

[1] G.S. AMMAR AND W.B. GRAGG, *Numerical experience with a superfast real Toeplitz solver*, Linear Algebra Appl., 121 (1989), pp. 185–206.

[2] H.C. ANDREWS AND B.R. HUNT, *Digital Image Restoration*, Prentice-Hall, Englewood Cliffs, NJ, 1977.

[3] T. BAGBY, *The modulus of a plane condenser*, J. Math. Mech., 17 (1967), pp. 315–329.

[4] ———, *On interpolation by rational functions*, Duke J. Math., 36 (1969), pp. 95–104.

[5] R. BARTELS AND G.W. STEWART, *Algorithm* 432*: Solution of the matrix equation* $AX + XB = C$, Comm. ACM, 15 (1972), pp. 820–826.

[6] G. BIRKHOFF AND R.S. VARGA, *Implicit alternating direction methods*, Trans. Amer. Math. Soc., 92 (1959), pp. 13–24.

[7] G. BIRKHOFF, R.S. VARGA, AND D. YOUNG, *Alternating direction implicit methods*, in Advances in Computing, Vol. 3, Academic Press, New York, 1962, pp. 189–273.

[8] P.L.C. CHEONG AND S.D. MORGERA, *Iterative methods for restoring noisy images*, IEEE Trans. Acoust. Speech Signal Proc., 37 (1989), pp. 580–585.

[9] C. DE BOOR AND J.R. RICE, *Chebyshev approximation by* $a\prod \frac{x-r_i}{x+s_i}$ *with application to* ADI *iteration*, J. Soc. Indust. Appl. Math., 11 (1963), pp. 159–169.

[10] M.A. EPTON, *Methods for the solution of* $AXD - BXC = E$ *and its application in numerical solution of implicit ordinary differential equations*, BIT, 20 (1980), pp. 341–345.

[11] N.S. ELLNER AND E.L. WACHSPRESS, *Alternating direction implicit iteration for systems with complex spectra*, SIAM J. Numer. Anal., 28 (1991), pp. 859–870.

[12] D. GAIER AND J. TODD, *On the rate of convergence of optimal* ADI *processes*, Numer. Math., 9 (1967), pp. 452–459.

[13] F.R. GANTMACHER, *Matrizentheorie*, Springer, New York, 1986.

[14] G.H. GOLUB, S. NASH AND C. VAN LOAN, *A Hessenberg–Schur method for the problem* $AX + XB = C$, IEEE Trans. Automat. Control., AC-24 (1979), pp. 909–913.

[15] G.H. GOLUB AND C.F. VAN LOAN, *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, MD, 1989.

[16] M. HANKE, J. NAGY, AND R. PLEMMONS, *Preconditioned iterative regularization for ill-posed problems*, in Numer. Linear Algebra, L. Reichel, A. Ruttan, and R.S. Varga, eds., de Gruyter, Berlin, 1993, pp. 141–163.

[17] M. HOCHBRUCK AND G. STARKE, *Preconditioned Krylov subspace methods for Lyapunov matrix equations*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 156–171.

[18] D.Y. HU AND L. REICHEL, *Krylov subspace methods for the Sylvester equation*, Linear Algebra Appl., 172 (1992), pp. 283–313.

[19] M. KAC, W.L. MURDOCK, AND G. SZEGŐ, *On the eigenvalues of certain Hermitian forms*, J. Rat. Mech. Anal., 2 (1953), pp. 767–800.

[20] R.L. LAGENDIJK AND J. BIEMOND, *Iterative Identification and Restoration of Images*, Kluwer, Dordrecht, 1991.

[21] N. LEVENBERG AND L. REICHEL, *A generalized* ADI *iterative method*, Numer. Math., 66 (1993), pp. 215–233.

[22] A.L. LEVIN AND E.B. SAFF, *Optimal ray sequences of rational functions connected with the Zolotarev problem*, Constr. Approx., 10 (1994), pp. 235–273.

[23] J.G. NAGY AND R.J. PLEMMONS, *Some fast Toeplitz least squares algorithms*, in SPIE, Vol. 1566, Advanced Signal Processing Algorithms, Architectures, and Implementation II, Society of Photo-Optical Instrumentation Engineers, Bellingham, WA, 1991, pp. 35–46.

[24] ———, *Iterative image restoration using* FFT*-based preconditioners*, Report, Department of Mathematics, Southern Methodist University, Dallas, TX, 1992.

[25] D.W. PEACEMAN AND H.H. RACHFORD, *The numerical solution of parabolic and elliptic differential equations*, J. Soc. Indust. Appl. Math., 3 (1955), pp. 28–41.

[26] Y. SAAD, *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, MA, 1990, pp. 503–511.

[27] G. STARKE, *Optimal alternating direction implicit parameters for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1431–1445.

[28] ———, *Fejér–Walsh points for rational functions and their use in the* ADI *iterative method*, J. Comput. Appl. Math., 46 (1993), pp. 129–141.

[29] M. TSUJI, *Potential Theory in Modern Function Theory*, Maruzen, Tokyo, 1959.

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

[31] E.L. WACHSPRESS, *Optimum alternating-direction-implicit iteration parameters for a model problem*, J. Soc. Indust. Appl. Math., 10 (1962), pp. 339–350.

[32] ———, *Extended application of alternating direction implicit iteration model problem theory*, J. Soc. Indust. Appl. Math., 11 (1963), pp. 994–1016.

[33] ———, *Iterative solution of the Lyapunov matrix equation*, Appl. Math. Letters, 1 (1988), pp. 87–90.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 1, pp. 187–211, January 1996

011

STABILITY OF SYMMETRIC ILL-CONDITIONED SYSTEMS ARISING IN INTERIOR METHODS FOR CONSTRAINED OPTIMIZATION*

ANDERS FORSGREN†, PHILIP E. GILL‡, AND JOSEPH R. SHINNERL‡

Abstract. Many interior methods for constrained optimization obtain a search direction as the solution of a symmetric linear system that becomes increasingly ill-conditioned as the solution is approached. In some cases, this ill-conditioning is characterized by a subset of the diagonal elements becoming large in magnitude. It has been shown that in this situation the solution can be computed accurately regardless of the size of the diagonal elements.

In this paper we discuss the formulation of several interior methods that use symmetric diagonally ill-conditioned systems. It is shown that diagonal ill-conditioning may be characterized by the property of *strict t-diagonal dominance*, which generalizes the idea of diagonal dominance to matrices whose diagonals are substantially larger in magnitude than the off-diagonals. A perturbation analysis is presented that characterizes the sensitivity of t-diagonally dominant systems under a certain class of structured perturbations. Finally, we give a rounding-error analysis of the symmetric indefinite factorization when applied to t-diagonally dominant systems. This analysis resolves the (until now) open question of whether the class of perturbations used in the sensitivity analysis is representative of the rounding error made during the numerical solution of the barrier equations.

Key words. nonlinear programming, constrained optimization, interior-point methods, barrier methods, rounding-error analysis, indefinite systems, backward stability, condition number

AMS subject classifications. 49D37, 65F05, 65K05, 90C30

1. Introduction.

This paper concerns the solution of the nonlinear programming problem

$$\text{NP} \qquad \begin{array}{ll} \underset{x \in \mathbb{R}^n}{\text{minimize}} & f(x) \\ \text{subject to} & c(x) \geq 0, \end{array}$$

where $c(x)$ is an m-vector of nonlinear functions with ith component $c_i(x)$, $i = 1, \ldots, m$, and f and $\{c_i\}$ are twice-continuously differentiable. Let $g(x)$ denote the gradient of $f(x)$ and $J(x)$ the $m \times n$ Jacobian of $c(x)$.

In recent years, *interior methods* for NP have received considerable attention because of their close relationship with the "new" polynomial approaches to linear and quadratic programming. Many of these interior methods are based on the properties of the logarithmic barrier function

$$f_\mu(x) = f(x) - \mu \sum_{i=1}^{m} \ln c_i(x), \tag{1.1}$$

* Received by the editors July 6, 1994; accepted for publication (in revised form) by N. Higham January 6, 1995.

† Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology, S-100 44 Stockholm, Sweden (andersf@math.kth.se). The research of this author was supported by the Royal Swedish Academy of Sciences (Magnuson's fund, KVA) and the Swedish Natural Science Research Council (NFR).

‡ Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0012 (pgill@ucsd.edu, jshinnerl@ucsd.edu). The research of P. E. Gill was supported by Department of Energy contract DE-FG03-92ER25117, National Science Foundation grant DDI-9204547, and Office of Naval Research grant N00014-90-J-1242. The research of J. R. Shinnerl was supported by National Science Foundation grant DDI-9204547.

which is defined for each value of the positive barrier parameter μ. Fiacco and McCormick [5] give conditions under which local minimizers x_μ^* of the barrier function converge to a solution x^* of NP as $\mu \to 0$ (see also Wright [25]).

In the classical logarithmic barrier method given by Fiacco and McCormick [5], $f_\mu(x)$ is minimized for each of a sequence of decreasing values of μ using an unconstrained minimization method. If x is an approximate minimizer that lies in the strict interior of the region $\{x \mid c_i(x) \geq 0\}$, then the matrix $C(x) = \operatorname{diag}(c_1(x), c_2(x), \ldots, c_m(x))$ is nonsingular and the derivatives of $f_\mu(x)$ are given by

$$\nabla f_\mu(x) = g(x) - \mu J(x)^T C(x)^{-1} e, \tag{1.2a}$$

$$\nabla^2 f_\mu(x) = \nabla^2 f(x) - \sum_{i=1}^{m} \frac{\mu}{c_i(x)} \nabla^2 c_i(x) + \mu J(x)^T C(x)^{-2} J(x), \tag{1.2b}$$

where e is the m-vector of ones. If $f_\mu(x)$ is minimized by Newton's method with a line search, the Newton direction Δx satisfies the equations

$$\nabla^2 f_\mu(x) \Delta x = -\nabla f_\mu(x), \tag{1.3}$$

where $\nabla^2 f_\mu(x)$ and $\nabla f_\mu(x)$ are given by (1.2). Once Δx has been computed, a positive step length α is chosen so that $c_i(x+\alpha\Delta x) > 0$ for all i, and $f_\mu(x+\alpha\Delta x)$ is sufficiently lower than $f_\mu(x)$ (see, e.g., Murray and Wright [20]).

The idea of using a barrier function dates back to the mid 1950s (see Frisch [6]); however, barrier methods fell into disuse during the 1970s, mainly because of the difficulties associated with the numerical solution of the equations (1.3). These difficulties are caused by the phenomenon of inevitable ill-conditioning in the barrier Hessian $\nabla^2 f_\mu(x)$. To characterize this ill-conditioning, we need to make some assumptions concerning a solution x^* of NP. Let $\mathcal{A}^*$ denote the set of indices of constraints that are active (i.e., satisfied with equality) at x^*. Throughout this paper, we use the suffix $+$ to denote quantities associated with constraints with indices in a certain subset $\mathcal{A}_+^*$ of $\mathcal{A}^*$. For example, $J_+(x)$ is the $m_+ \times n$ submatrix formed from the rows of $J(x)$ whose indices are in $\mathcal{A}_+^*$. Sufficient conditions for a feasible point x^* to be a local solution of NP are that there exists an index set $\mathcal{A}_+^*$, and a strictly positive Lagrange multiplier vector λ_+^* such that

(1.4a) the vectors $\{\nabla c_i(x^*)\}$, $i \in \mathcal{A}^*$, are linearly independent,

(1.4b) $g(x^*) = J_+(x^*)^T \lambda_+^*$,

(1.4c) $v^T H(x^*, \lambda^*) v > 0$ for all nonzero v such that $J_+(x^*)v = 0$,

(1.4d) $\mathcal{A}_+^* = \mathcal{A}^*$,

where $H(x, \lambda)$ denotes the Hessian of the Lagrangian $\nabla^2 f(x) - \sum_{i=1}^m \lambda_i \nabla^2 c_i(x)$. Conditions (1.4) imply the existence of a locally unique differentiable trajectory of barrier minimizers x_μ^* such that $\lim_{\mu \to 0} x_\mu^* = x^*$ (see Fiacco and McCormick [5, pp. 79–82]).

Murray [19] has shown that as $\mu \to 0$, $\nabla^2 f_\mu(x_\mu^*)$ has m_+ unbounded eigenvalues corresponding to eigenvectors in the range space of $J_+(x_\mu^*)^T$, and $n - m_+$ bounded eigenvalues corresponding to eigenvectors in the null space of $J_+(x_\mu^*)$. If $0 < m_+ < n$ the barrier Hessian accordingly becomes increasingly ill-conditioned as μ is reduced. More general properties of the barrier Hessian are discussed by Wright [26].

Several approaches have been suggested for the treatment of ill-conditioning in $\nabla^2 f_\mu(x)$. If one is willing to predict the active set, the ill-conditioning can be avoided

in several ways. For example, the Newton direction can be approximated by the sum of two independent directions, one of which lies in the null-space of the matrix of predicted active constraint gradients (see Wright [24]).

More recently, with the growth of interest in the use of barrier methods as a means of avoiding the combinatorial complexity of active-set methods, there has been an emphasis on methods that do not require an estimate of the active set. One approach is to convert the inequality constraints $c(x) \geq 0$ into equalities using nonnegative slack variables (see §2.1). The resulting Newton equations for the equality-constrained problem are also ill-conditioned, but the ill-conditioning is caused by the presence of some large elements on the diagonal. This *diagonal ill-conditioning* will be defined and analyzed in §2.6. The importance of diagonal ill-conditioning was first demonstrated by Ponceleón [21], who gave a perturbation analysis of diagonally ill-conditioned systems and showed that the sensitivity of the equations under certain *structured* perturbations is independent of the large diagonals. Moreover, it was shown that as $\mu \to 0$, the sensitivity of the Newton barrier equations is identical to the sensitivity of a system whose condition reflects the condition of the original problem (see §2.5 for a definition of this system).

There are situations, however, when it is not convenient to add slack variables. In many physical and engineering applications, the constraint functions not only characterize the desired properties of the solution, but also define a region in which the problem statement is meaningful (for example, $f(x)$ or some of the constraint functions may be undefined outside the feasible region). In this situation, a barrier transformation requires the strict satisfaction of all constraints $c_i(x) \geq 0$ at the starting point and subsequent iterates. If slack variables are used to transform the inequality constraints into equalities, the barrier transformation is applied to the nonnegativity constraint on the slack variable, allowing the original inequality constraint to be violated.

If slack variables cannot be used, other methods must be used to circumvent the ill-conditioning in the unconstrained barrier equations (1.3). One possibility is to formulate the Newton equations as an *unsymmetric* system whose condition number does not go to infinity as $\mu \to 0$ (see McCormick [15]). This idea is illustrated in §2.2. As far as efficiency is concerned, there may be little disadvantage in using an unsymmetric system instead of a symmetric system (e.g., sparse matrix packages are able to exploit symmetric structure even when the numerical values are unsymmetric). However, it is not yet known how methods based on unsymmetric systems can be generalized to nonconvex problems. If the objective is not convex, even the verification of optimality requires knowledge of the inertia of the Hessian in the subspace orthogonal to the active constraint gradients. It is not at all obvious how the inertia can be estimated efficiently without utilizing symmetry.

In §2 we describe a class of *symmetric* barrier equations that allow the Newton barrier direction to be calculated accurately without the need to either add slack variables or formulate the Newton equations as an unsymmetric system. A similar approach (without analysis) is proposed by Gould [10]. The crucial feature of these symmetric systems is that any inevitable ill-conditioning is caused by some large diagonal elements. This behavior is characterized by the property of *strict t-diagonal dominance* (see §2.6), which extends the standard definition of diagonal dominance to the case where some of the diagonals are large in magnitude. If the only inequality constraints are nonnegativity constraints, our formulation is the same as the one analyzed by Ponceleón [21]. In §3 we describe the sensitivity of solutions of this class

of symmetric Newton barrier equations when the matrix is changed by a particular class of structured perturbations. The analysis suggests an effective condition number for these equations and indicates that under suitable assumptions, the solution can be computed accurately, even though the condition number of the system tends to infinity.

Finally, in §4 we give a backward rounding-error analysis of the solution of t-diagonally dominant Newton barrier equations by means of the symmetric indefinite factorization. Our analysis completely characterizes the form of the backward error and indicates that our assumptions concerning the form of the structured perturbations used in the perturbation analysis are realistic.

Our results can also be used to extend the analysis of Wright [27], who discusses the use of the LU factorization for solving ill-conditioned symmetric systems.

2. Formulation and solution of the barrier equations.

2.1. Standard form. Ponceleón [21] considers the sensitivity of solutions of linear systems arising from the application of barrier methods to quadratic programs formulated in "standard form." The standard-form equivalent of problem NP is obtained by converting each inequality constraint $c_i(x) \geq 0$ into an equality $c_i(x) - s_i = 0$ using a nonnegative slack variable s_i. This gives

$$\begin{aligned} \underset{x,s}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & c(x) - s = 0, \quad s \geq 0, \end{aligned}$$

where s is the vector of slack variables. (Since this format does not assume that the x variables are nonnegative, it is different from the problem considered by Ponceleón. We have chosen this form to demonstrate the similarities between the all-inequality and standard form approaches.) The barrier transformation is applied to the nonnegativity constraints, giving the barrier subproblem

$$\begin{aligned} \underset{x,s}{\text{minimize}} \quad & f(x) - \mu \sum_{i=1}^{m} \ln s_i \\ \text{subject to} \quad & c(x) - s = 0. \end{aligned}$$

The first-order optimality conditions of this problem are

$$\begin{aligned} g(x) - J(x)^T \lambda &= 0, \\ \lambda - \mu S^{-1} e &= 0, \\ c(x) - s &= 0, \end{aligned} \tag{2.1}$$

where S is the diagonal matrix $\operatorname{diag}(s_1, s_2, \ldots, s_m)$ and λ is the vector of Lagrange multipliers associated with the equality constraints $c(x) - s = 0$. These relations imply that $(x_\mu^*, s_\mu^*, \lambda_\mu^*)$ solve $n + 2m$ equations in the $n + 2m$ unknowns (x, s, λ).

The nonlinear equations (2.1) can also be solved using a form of Newton's method. Suppose that (x, s, λ) is an estimate of $(x_\mu^*, s_\mu^*, \lambda_\mu^*)$. Let g, c, J, and H denote the quantities $g(x)$, $c(x)$, $J(x)$, and $H(x, \lambda)$. Given (x, s, λ), the next iterate of Newton's method is $(x + \alpha \Delta x, s + \alpha \Delta s, \lambda + \alpha \Delta \lambda)$, where α is a scalar step length and $(\Delta x, \Delta s, \Delta \lambda)$ satisfies

$$\begin{pmatrix} H & 0 & -J^T \\ 0 & \mu S^{-2} & I \\ J & -I & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta s \\ \Delta \lambda \end{pmatrix} = - \begin{pmatrix} g - J^T \lambda \\ \lambda - \mu S^{-1} e \\ c - s \end{pmatrix}.$$

Simple rearrangement gives the symmetric system

$$\begin{pmatrix} H & 0 & J^T \\ 0 & \mu S^{-2} & -I \\ J & -I & 0 \end{pmatrix} \begin{pmatrix} -\Delta x \\ -\Delta s \\ \Delta\lambda \end{pmatrix} = \begin{pmatrix} g - J^T\lambda \\ \lambda - \mu S^{-1}e \\ c - s \end{pmatrix}. \tag{2.2}$$

(For details of related applications of Newton's method in the context of linear and quadratic programming, see, e.g., Kojima, Mizuno and Yoshise [13], Lustig, Marsten and Shanno [14], Megiddo [16], Mehrotra [17], Monteiro and Adler [18], and Gill et al. [7].) As $(x, \lambda) \to (x^*_\mu, \lambda^*_\mu)$ and $\mu \to 0$, this system becomes increasingly ill-conditioned, with some of the diagonals of μS^{-2} becoming unbounded. The numerical properties of this diagonal ill-conditioning have been discussed by Ponceleón [21], who showed that diagonally ill-conditioned systems can be solved accurately even when the magnitude of some diagonal elements goes to infinity. Diagonally ill-conditioned systems are considered further in §3.

In general, the sequence generated by this slack-variable form of the primal barrier method will differ from that of the unconstrained form discussed in §1. In particular, since only the slack variables are constrained to be strictly feasible, the constraints $c_i(x) > 0$ may or may not be violated. This could prove to be a disadvantage in some applications where strict feasibility of the constraints $c_i(x) \geq 0$ is required.

2.2. Unsymmetric primal formulation. An alternative to adding slack variables is to derive a system of *unsymmetric* Newton equations that is not always ill-conditioned as x converges to x^*. Consider the all-inequality form of problem NP and the resulting Newton equations (1.3). Define m auxiliary quantities $\lambda_i = \mu/c_i(x)$. The vector with components λ_i can be written as $\lambda = \mu C(x)^{-1}e$ and can be used to derive the $n + m$ conditions

$$\begin{aligned} g(x) - J(x)^T\lambda &= 0, \\ \lambda - \mu C(x)^{-1}e &= 0, \end{aligned} \tag{2.3}$$

with $C(x) = \operatorname{diag}(c_1(x), c_2(x), \ldots, c_m(x))$. To simplify the notation, we shall denote $C(x)$ by C, and define $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m)$. The relations (2.3) imply that (x^*_μ, λ^*_μ) solve $n + m$ equations in the $n + m$ unknowns (x, λ). If these nonlinear equations are solved using Newton's method as in §2.1, we obtain the Newton equations

$$\begin{pmatrix} H & -J^T \\ \mu C^{-2}J & I \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} = -\begin{pmatrix} g - J^T\lambda \\ \lambda - \mu C^{-1}e \end{pmatrix}.$$

Premultiplying the last m rows by C and performing some simple rearrangement gives

$$\begin{pmatrix} H & J^T \\ \mu C^{-1}J & -C \end{pmatrix} \begin{pmatrix} -\Delta x \\ \Delta\lambda \end{pmatrix} = \begin{pmatrix} g - J^T\lambda \\ C\lambda - \mu e \end{pmatrix}. \tag{2.4}$$

This unsymmetric system is the primal variant of the one discussed by McCormick [15] (we have presented the primal form in order to be consistent with the primal slack-variable approach of §2.1). When conditions (1.4) hold, this system is *nonsingular* and does not suffer inevitable ill-conditioning as $(x, \lambda) \to (x^*_\mu, \lambda^*_\mu)$ and $\mu \to 0$. To show this, consider the set $\mathcal{A}^*_+$ of active constraints with positive multipliers at x^*.

Let quantities associated with constraints whose indices are not in $\mathcal{A}_+^*$ be denoted by 0. If the rows and columns of the system (2.4) are reordered to match the indices in $\mathcal{A}_+^*$ and $\mathcal{A}_0^*$, then

$$\begin{pmatrix} H & J_+^T & J_0^T \\ \mu C_+^{-1} J_+ & -C_+ & \\ \mu C_0^{-1} J_0 & & -C_0 \end{pmatrix} \begin{pmatrix} -\Delta x \\ \Delta \lambda_+ \\ \Delta \lambda_0 \end{pmatrix} = \begin{pmatrix} g - J^T \lambda \\ C_+ \lambda_+ - \mu e \\ C_0 \lambda_0 - \mu e \end{pmatrix},$$

where we have assumed without loss of generality that the rows of J_+ are the leading rows of A. Under the assumptions (1.4), the following relationships hold as $(x, \lambda) \to (x_\mu^*, \lambda_\mu^*)$ and $\mu \to 0$:

$$\begin{aligned} &c_0 \to c_0^* \quad (c_0^* > 0); && c_+ \to c_+^* \quad (c_+^* = 0), \\ &\lambda_0 \to \lambda_0^* \quad (\lambda_0^* = 0); && \lambda_+ \to \lambda_+^* \quad (\lambda_+^* > 0). \end{aligned} \tag{2.5}$$

Hence, as $(x, \lambda) \to (x_\mu^*, \lambda_\mu^*)$ and $\mu \to 0$, we have $\mu C_+^{-1} \to \Lambda_+$, $\mu C_0^{-1} \to 0$, and

$$\begin{pmatrix} H & J_+^T & J_0^T \\ \mu C_+^{-1} J_+ & -C_+ & \\ \mu C_0^{-1} J_0 & & -C_0 \end{pmatrix} \to \begin{pmatrix} H & J_+^T & J_0^T \\ \Lambda_+ J_+ & 0 & 0 \\ 0 & 0 & -C_0 \end{pmatrix},$$

which is a block upper-triangular matrix whose condition number is bounded.

2.3. Symmetric primal barrier formulation. In this section we consider certain *symmetric* systems arising from the application of Newton's method to the solution of the equations (2.3). If the last m rows of (2.4) are premultiplied by the diagonal matrix $(1/\mu)C$, we obtain

$$\begin{pmatrix} H & J^T \\ J & -\frac{1}{\mu} C^2 \end{pmatrix} \begin{pmatrix} -\Delta x \\ \Delta \lambda \end{pmatrix} = \begin{pmatrix} g - J^T \lambda \\ \frac{1}{\mu} C(C\lambda - \mu e) \end{pmatrix}. \tag{2.6}$$

These symmetric primal barrier equations characterize the *primal barrier Karush–Kuhn–Tucker* (KKT) *system*, and the symmetric matrix associated with this system is known as the *primal barrier* KKT *matrix*.

The effect of symmetrizing the Newton barrier equations is to make the barrier KKT matrix ill-conditioned. If the primal barrier KKT matrix is partitioned to match the indices in $\mathcal{A}_+^*$ and $\mathcal{A}_0^*$, then

$$\begin{pmatrix} H & J_+^T & J_0^T \\ J_+ & -\frac{1}{\mu} C_+^2 & \\ J_0 & & -\frac{1}{\mu} C_0^2 \end{pmatrix} \begin{pmatrix} -\Delta x \\ \Delta \lambda_+ \\ \Delta \lambda_0 \end{pmatrix} = \begin{pmatrix} g - J^T \lambda \\ \frac{1}{\mu} C_+(C_+ \lambda_+ - \mu e) \\ \frac{1}{\mu} C_0(C_0 \lambda_0 - \mu e) \end{pmatrix}. \tag{2.7}$$

The assumptions (2.5) imply that as $(x, \lambda) \to (x_\mu^*, \lambda_\mu^*)$ and $\mu \to 0$, then the diagonals of $(1/\mu)C_+^2$ go to zero, the diagonals of $(1/\mu)C_0^2$ go to infinity, and the primal barrier KKT matrix becomes increasingly ill-conditioned.

2.4. Related symmetric formulations. Following Ye [28], Gill et al. [7], and Gonzaga [9], we note that other symmetric systems can be derived by symmetrizing the three equivalent forms of (2.3b): (i) $C(x)\lambda - \mu e = 0$, (ii) $C(x)e - \mu\Lambda^{-1}e = 0$, and (iii) $e - \mu\Lambda^{-1}C(x)^{-1}e = 0$. In case (i) we obtain the *primal-dual* system

$$\begin{pmatrix} H & J^T \\ J & -C\Lambda^{-1} \end{pmatrix} \begin{pmatrix} -\Delta x \\ \Delta\lambda \end{pmatrix} = \begin{pmatrix} g - J^T\lambda \\ c - \mu\Lambda^{-1}e \end{pmatrix}. \tag{2.8}$$

In case (ii) we obtain the *dual* system

$$\begin{pmatrix} H & J^T \\ J & -\mu\Lambda^{-2} \end{pmatrix} \begin{pmatrix} -\Delta x \\ \Delta\lambda \end{pmatrix} = \begin{pmatrix} g - J^T\lambda \\ c - \mu\Lambda^{-1}e \end{pmatrix}. \tag{2.9}$$

Finally, for case (iii) we have

$$\begin{pmatrix} H & J^T \\ J & -C\Lambda^{-1} \end{pmatrix} \begin{pmatrix} -\Delta x \\ \Delta\lambda \end{pmatrix} = \begin{pmatrix} g - J^T\lambda \\ \frac{1}{\mu}C(C\lambda - \mu e) \end{pmatrix}. \tag{2.10}$$

Each of these systems becomes ill-conditioned in the same way, namely, as $(x, \lambda) \to (x_\mu^*, \lambda_\mu^*)$ and $\mu \to 0$, some diagonal elements become infinitely large while others converge to zero. Moreover, if $(x, \lambda) = (x_\mu^*, \lambda_\mu^*)$, then $C\Lambda = \mu I$ and the matrices associated with the systems (2.6), (2.8), (2.9), and (2.10) are identical. An error analysis of the symmetric indefinite factorization is considered in §4. We note that our analysis does not treat symmetric systems that are derived by *column* symmetrization. For example, in the primal-dual case, if the matrix Λ is applied to the last m columns of (2.4), we have the symmetric system

$$\begin{pmatrix} H & J^T\Lambda \\ \Lambda J & -C\Lambda \end{pmatrix} \begin{pmatrix} -\Delta x \\ \Lambda^{-1}\Delta\lambda \end{pmatrix} = \begin{pmatrix} g - J^T\lambda \\ C\lambda - \mu e \end{pmatrix}.$$

The partitioned form of this matrix is given by

$$\begin{pmatrix} H & J_+^T\Lambda_+ & J_0^T\Lambda_0 \\ \Lambda_+J_+ & -C_+\Lambda_+ & \\ \Lambda_0J_0 & & -C_0\Lambda_0 \end{pmatrix} \begin{pmatrix} -\Delta x \\ \Lambda_+^{-1}\Delta\lambda_+ \\ \Lambda_0^{-1}\Delta\lambda_0 \end{pmatrix} = \begin{pmatrix} g - J^T\lambda \\ C_+\lambda_+ - \mu e \\ C_0\lambda_0 - \mu e \end{pmatrix}.$$

These relations imply that some of the diagonals converge to *zero* while others remain bounded. Although the analysis of systems of this type is not covered by the techniques proposed in this paper, it is possible that the symmetric indefinite factorization can also be used stably in this situation.

2.5. Solution of the symmetric barrier KKT equations. We will assume that the barrier KKT equations are solved using the *symmetric indefinite factorization* (see Bunch and Parlett [4] and Bunch and Kaufman [2]), which we refer to as the LBL^T *factorization*. If A denotes the particular barrier KKT matrix under consideration, then the LBL^T factorization defines a permutation P, a block-diagonal B, and a unit-lower triangular L such that

$$P^TAP = LBL^T, \quad \text{where} \quad B = \operatorname{diag}(B_{11}, B_{22}, \ldots, B_{ss}).$$

Each B_{jj} is either 1×1, or is 2×2 having one positive and one negative eigenvalue. The permutation P incorporates certain symmetric interchanges that are needed to preserve numerical stability.

The choice of permutation P depends on the particular pivoting strategy used (see Bunch et al. [3] for a discussion of the various strategies available). The analysis of §4 requires that the large elements on the diagonal are used to define the first sequence of 1×1 pivots. (For example, this would be the case for the pivoting strategy of Bunch and Parlett [4].) With this sequence of pivots, the symmetrically permuted A can be partitioned so that

$$P^T A P = \begin{pmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{pmatrix}, \tag{2.11}$$

where A_{11} is the matrix whose diagonals are large in magnitude. For example, suppose that the equations and variables of the primal barrier system (2.7) are reordered to give

$$\begin{pmatrix} -\frac{1}{\mu}C_0^2 & J_0 & 0 \\ J_0^T & H & J_+^T \\ 0 & J_+ & -\frac{1}{\mu}C_+^2 \end{pmatrix} \begin{pmatrix} \Delta\lambda_0 \\ -\Delta x \\ \Delta\lambda_+ \end{pmatrix} = \begin{pmatrix} \frac{1}{\mu}C_0(C_0\lambda_0 - \mu e) \\ g - J^T\lambda \\ \frac{1}{\mu}C_+(C_+\lambda_+ - \mu e) \end{pmatrix}, \tag{2.12}$$

where $-(1/\mu)C_0^2$ is the matrix whose diagonals are large in magnitude. In this case A_{11} and A_{22} are defined from the rows and columns of the matrices

$$-\frac{1}{\mu}C_0^2 \qquad \text{and} \qquad \begin{pmatrix} H & J_+^T \\ J_+ & -\frac{1}{\mu}C_+^2 \end{pmatrix}. \tag{2.13}$$

(For the purposes of this discussion, the order of the rows and columns in A_{11} and A_{22} is not important.)

When optimality conditions (1.4) hold and $(x, \lambda) \to (x_\mu^*, \lambda_\mu^*)$, then A_{22} converges to the KKT matrix

$$\begin{pmatrix} H & J_+^T \\ J_+ & \end{pmatrix}, \tag{2.14}$$

which is invariant of the particular barrier method used. The condition of this KKT matrix is determined by the singular values of J_+ and the eigenvalues of the reduced Hessian $Z_+^T H Z_+$, where the columns of Z_+ form a basis for the null space of J_+. In effect, these quantities may be considered as defining the condition of the original problem NP.

The following informal argument indicates that the sensitivity of the barrier equations principally depends upon the condition of A_{22} (and hence on the condition of the KKT matrix) rather than on the condition of A (a rigorous analysis is given in §3). Suppose that the first r rows of A are scaled by the diagonal matrix $D = \operatorname{diag}(A_{11})^{-1}$. Then $D \to 0$ as $\mu \to 0$, and

$$\begin{pmatrix} DA_{11} & DA_{21}^T \\ A_{21} & A_{22} \end{pmatrix} \to \tilde{A}, \quad \text{where} \quad \tilde{A} = \begin{pmatrix} I & 0 \\ A_{21} & A_{22} \end{pmatrix}.$$

Since the solution of a system is unaltered by a diagonal row scaling, this result leads us to expect the sensitivity of the barrier solution to be based on $\kappa(\widetilde{A}) = \|\widetilde{A}^{-1}\|\|\widetilde{A}\|$. The identity

$$\widetilde{A}^{-1} = \begin{pmatrix} I & 0 \\ -A_{22}^{-1}A_{21} & A_{22}^{-1} \end{pmatrix}$$

implies that the limiting sensitivity of the barrier solution depends upon A_{22} and A_{21}. In §3 we give a rigorous derivation of the sensitivity of barrier solutions for the case where μ is small (but nonzero).

2.6. Strict t-diagonal dominance. Each of the systems (2.2), (2.6), (2.8), (2.9), and (2.10) has the property that some diagonal elements become infinitely large in magnitude as $(x, \lambda) \to (x_\mu^*, \lambda_\mu^*)$ and $\mu \to 0$. In order to analyze the numerical properties of systems of this kind, it is important to be able to characterize the behavior of symmetric barrier matrices in a way that is invariant with the particular barrier formulation being employed. Our approach is illustrated by the matrix

$$A = \begin{pmatrix} -2t^2 & 0 & 1 \\ 0 & -2t & 1 \\ 1 & 1 & 1 \end{pmatrix}, \tag{2.15}$$

where $t \geq 1$. For all $t \geq 1$, the LBL^T factorization $A = LBL^T$ is given by

$$L = \begin{pmatrix} 1 & & \\ 0 & 1 & \\ -\dfrac{1}{2t^2} & -\dfrac{1}{2t} & 1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} -2t^2 & & \\ & -2t & \\ & & 1 + \dfrac{1}{2t} + \dfrac{1}{2t^2} \end{pmatrix}.$$

Note that the last diagonal differs from the $(3,3)$ element of A by a term of order $1/t$, which is the order of the inverse of the smallest "large" diagonal. This implies that the contribution to the rounding error in this element diminishes as $t \to \infty$. This observation is crucial in the error analysis of §4.

For large t, the magnitude of a large diagonal dominates the other elements in its column. Matrices with this property can be characterized as follows. Let A denote a symmetric $n \times n$ matrix, and let t denote a scalar such that $t \geq 1$. The jth column of A is said to be *strictly t-diagonally dominant* if

$$|a_{jj}| > t \sum_{i \neq j} |a_{ij}|.$$

If *every* column is strictly t-diagonally dominant, then A is said to be *strictly t-diagonally dominant*. If $t = 1$, our definition of t-diagonal dominance is equivalent to the usual definition of diagonal dominance (see Golub and Van Loan [8, pp. 119–120]). We also define *subsets* of t-diagonally dominant columns. A subset of r columns $j_1, \ldots, j_r$ is said to be *strictly block t-diagonally dominant* if

$$\min_{1 \leq q \leq r} |a_{j_q j_q}| > t \max_{1 \leq q \leq r} \sum_{\substack{1 \leq i \leq n \\ i \neq j_q}} |a_{i j_q}|.$$

If t is sufficiently large but some diagonals are independent of t, the matrix is ill-conditioned. We refer to this type of ill-conditioning as *diagonal ill-conditioning*. An important feature of diagonally ill-conditioned matrices is that the diagonals of t-diagonally dominant columns define a sequence of acceptable 1×1 pivots for the LBL^T factorization.

3. Perturbation analysis. This section is concerned with an analysis of systems whose matrices are strictly block t-diagonally dominant. To simplify the notation, we use $Ax = b$ to denote the system under discussion. It will be assumed throughout that the rows and columns of A may be reordered to give a matrix of the form

$$\begin{pmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{pmatrix},$$

where A_{22} is nonsingular, and the matrices A_{11} and A_{21} form a subset of r strictly t-diagonally dominant columns of A. (In terms of the barrier equations (2.11), these assumptions imply that μ is small enough to ensure that A_{22} is sufficiently close to the KKT matrix (2.14). The optimality conditions (1.4) then imply that A_{22} is nonsingular.) The dimension of A_{11} is denoted by r. A similar notation is used for the block partition of all other matrices of order n.

For the remainder of this discussion, we use $\|A\|$ to denote the spectral norm $\|A\| = \sup_{x \neq 0} \|Ax\|_2 / \|x\|_2$. Similarly, $\kappa(A)$ denotes the spectral condition number $\kappa(A) = \|A^{-1}\| \|A\|$. The matrix with elements $|a_{ij}|$ is denoted by $|A|$. The matrices consisting of the diagonal and off-diagonal elements of A are denoted by $\operatorname{diag}(A)$ and $\operatorname{offdiag}(A)$ respectively, so that $A = \operatorname{diag}(A) + \operatorname{offdiag}(A)$.

In the next theorem, we analyze the effect on the solution of $Ax = b$ of perturbing a block t-diagonally dominant matrix A by a specific class of structured perturbations. It is shown that any perturbation from this class produces a relative perturbation in x that can be of the order of the relative perturbation of A magnified by terms of the order of $\kappa(A_{22})$ and $\|A_{21}\| \|A_{22}^{-1}\|$. This result is in contrast to the standard result, which predicts that the relative perturbation in A can be magnified by at most $\kappa(A)$. It is shown in Theorem 4.4 and Corollary 4.5 of §4 that these perturbations are representative of the backward error made when solving $Ax = b$ using the LBL^T factorization of A.

THEOREM 3.1. *Let A be a symmetric $n \times n$ matrix partitioned so that*

$$A = \begin{pmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{pmatrix},$$

where A_{11} is $r \times r$ and A_{22} is nonsingular. If the first r columns of A form a strictly block t-diagonally dominant submatrix, with

$$t > 1 + r\|A_{21}\| \|A_{22}^{-1}\|, \tag{3.1}$$

then A is nonsingular. Moreover, consider the system $Ax = b$ and the perturbed system $(A + E)\widetilde{x} = b$, where

$$A + E = \begin{pmatrix} A_{11} + E_{11} & A_{21}^T + E_{12} \\ A_{21} + E_{21} & A_{22} + E_{22} \end{pmatrix},$$

and E is such that

(i) $|\operatorname{diag}(E_{11})| \le \epsilon |\operatorname{diag}(A_{11})|$;
(ii) $\|\operatorname{offdiag}(E_{11})\| \le \epsilon \|\operatorname{offdiag}(A_{11})\|$;
(iii) $\|E_{21}\| \le \epsilon \|A_{21}\|$, $\|E_{12}\| \le \epsilon \|A_{21}\|$;
(iv) $\|E_{22}\| \le \epsilon(\|A_{22}\| + \frac{1}{t}\|A_{21}\|)$.

If δ_1 and δ_2 denote constants $\delta_1 = r/(t-1)$ and $\delta_2 = \sqrt{r}(t+\sqrt{r})/(t-1)$, then

$$\frac{\|\widetilde{x} - x\|}{\|\widetilde{x}\|} \le \gamma_1 + \gamma_2 \kappa(A_{22}) + \gamma_3 \|A_{21}\| \|A_{22}^{-1}\|,$$

where $\gamma_1 = n\epsilon(\delta_1 + \delta_2)/(1 - \delta_1 \|A_{21}\| \|A_{22}^{-1}\|)$, $\gamma_2 = n\epsilon(1+\delta_1)/(1 - \delta_1 \|A_{21}\| \|A_{22}^{-1}\|)$, *and* $\gamma_3 = n\epsilon(\delta_1 + \delta_2 + (1+1/t)(1+\delta_1))/(1 - \delta_1 \|A_{21}\| \|A_{22}^{-1}\|)$.

Proof. We first show that A is nonsingular. To do this, $\|A_{11}^{-1}\|$ and $\|A_{21}\| \|A_{11}^{-1}\|$ are bounded by deriving some inequalities involving A_{21} and the diagonal and off-diagonal parts of A_{11}. Define $A_D = \operatorname{diag}(A_{11})$ and $A_O = \operatorname{offdiag}(A_{11})$, with similar definitions for E_D and E_O. The strict block t-diagonal dominance of the first r columns of A is equivalent to

$$\frac{1}{\|A_D^{-1}\|_1} > t \left\| \begin{pmatrix} A_O \\ A_{21} \end{pmatrix} \right\|_1,$$

so that

$$\|A_D^{-1}\|_1 \|A_O\|_1 < \frac{1}{t}, \qquad \text{and} \qquad \|A_D^{-1}\|_1 \|A_{21}\|_1 < \frac{1}{t}. \tag{3.2}$$

In terms of the two-norm, these bounds may be written as

$$\|A_D^{-1}\| \|A_O\| < \frac{\sqrt{r}}{t}, \qquad \text{and} \qquad \|A_D^{-1}\| \|A_{21}\| < \frac{\sqrt{r}}{t}, \tag{3.3}$$

where $\|A_D\| = \|A_D\|_1$, since A_D is diagonal. It follows from the first inequality of (3.2) that $\|A_D^{-1} A_O\|_1 < 1/t$, so that standard norm inequalities give

$$\|(I + A_D^{-1} A_O)^{-1}\| \le \sqrt{r} \|(I + A_D^{-1} A_O)^{-1}\|_1 \le \frac{t\sqrt{r}}{t-1} \tag{3.4}$$

(see, e.g., Stoer and Bulirsch [23, p. 188]). Since $A_{11} = A_D + A_O$, (3.4) implies that A_{11}^{-1} may be written as

$$A_{11}^{-1} = (I + A_D^{-1} A_O)^{-1} A_D^{-1}. \tag{3.5}$$

The second inequality of (3.3), (3.4), and (3.5) give

$$\|A_{21}\| \|A_{11}^{-1}\| \le \|A_{21}\| \|A_D^{-1}\| \|(I + A_D^{-1} A_O)^{-1}\| < \delta_1, \tag{3.6}$$

where

$$\delta_1 = \frac{r}{t-1}.$$

This inequality allows us to show that S and A are nonsingular, where S denotes the Schur complement $A_{22} - A_{21} A_{11}^{-1} A_{21}^T$. The difference between the smallest singular values of S and A_{22} may be written as

$$|\sigma_{\min}(S) - \sigma_{\min}(A_{22})| \le \|A_{21} A_{11}^{-1} A_{21}^T\| \le \|A_{21}\|^2 \|A_{11}^{-1}\|, \tag{3.7}$$

where $\sigma_{\min}$ denotes the smallest singular value (see, e.g., Golub and Van Loan [8, p. 428]). Since A_{22} is nonsingular, it holds that $\|A_{22}^{-1}\| = 1/\sigma_{\min}(A_{22})$. Substituting (3.6) in (3.7) gives

$$\sigma_{\min}(S) \geq \frac{1 - \delta_1 \|A_{21}\| \|A_{22}^{-1}\|}{\|A_{22}^{-1}\|}. \tag{3.8}$$

This inequality and the definitions of t and δ_1 imply that $\sigma_{\min}(S) > 0$, so that S is nonsingular. Since S and A_{11} have been shown to be nonsingular, it is straightforward to verify that the inverse of A may be written in the partitioned form

$$\begin{pmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{pmatrix}^{-1} = \begin{pmatrix} A_{11}^{-1} + A_{11}^{-1} A_{21}^T S^{-1} A_{21} A_{11}^{-1} & -A_{11}^{-1} A_{21}^T S^{-1} \\ -S^{-1} A_{21} A_{11}^{-1} & S^{-1} \end{pmatrix}. \tag{3.9}$$

Since A has now been shown to be nonsingular, standard analysis gives the bound

$$\frac{\|\widetilde{x} - x\|}{\|\widetilde{x}\|} \leq \|A^{-1}E\|$$

(see, e.g., Stewart and Sun [22, pp. 124–125]). Consider the matrix

$$F = \begin{pmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{pmatrix} \equiv A^{-1}E, \tag{3.10}$$

where F has been partitioned to match the partition of A. Substituting (3.9) in (3.10), we obtain F_{11}, F_{12}, F_{21}, and F_{22} as

$$\begin{aligned}
F_{11} &= (A_{11}^{-1} + A_{11}^{-1} A_{21}^T S^{-1} A_{21} A_{11}^{-1}) E_{11} - A_{11}^{-1} A_{21}^T S^{-1} E_{21}, \\
F_{12} &= (A_{11}^{-1} + A_{11}^{-1} A_{21}^T S^{-1} A_{21} A_{11}^{-1}) E_{12} - A_{11}^{-1} A_{21}^T S^{-1} E_{22}, \\
F_{21} &= -S^{-1} A_{21} A_{11}^{-1} E_{11} + S^{-1} E_{21}, \\
F_{22} &= -S^{-1} A_{21} A_{11}^{-1} E_{12} + S^{-1} E_{22}.
\end{aligned}$$

All that remains is to bound the norms of F_{11}, F_{12}, F_{21}, and F_{22}. Taking norms in each of the expressions for F_{11}, F_{12}, F_{21}, and F_{22} above, and using (3.6) in conjunction with $\|S^{-1}\| \leq \|A_{22}^{-1}\|/(1 - \delta_1 \|A_{21}\| \|A_{22}^{-1}\|)$ from (3.8) gives

$$\begin{aligned}
\|F_{11}\| &\leq \frac{\|A_{11}^{-1} E_{11}\| + \delta_1 \|A_{22}^{-1}\| \|E_{21}\|}{1 - \delta_1 \|A_{21}\| \|A_{22}^{-1}\|}, \\
\|F_{12}\| &\leq \frac{\|A_{11}^{-1}\| \|E_{12}\| + \delta_1 \|A_{22}^{-1}\| \|E_{22}\|}{1 - \delta_1 \|A_{21}\| \|A_{22}^{-1}\|}, \\
\|F_{21}\| &\leq \frac{\|A_{21}\| \|A_{22}^{-1}\| \|A_{11}^{-1} E_{11}\| + \|A_{22}^{-1}\| \|E_{21}\|}{1 - \delta_1 \|A_{21}\| \|A_{22}^{-1}\|}, \\
\|F_{22}\| &\leq \frac{\delta_1 \|A_{22}^{-1}\| \|E_{12}\| + \|A_{22}^{-1}\| \|E_{22}\|}{1 - \delta_1 \|A_{21}\| \|A_{22}^{-1}\|}.
\end{aligned}$$

It follows from (3.5) that $A_{11}^{-1} E_{11} = (I + A_D^{-1} A_O)^{-1} (A_D^{-1} E_D + A_D^{-1} E_O)$. Taking norms gives

$$\|A_{11}^{-1} E_{11}\| \leq \|(I + A_D^{-1} A_O)^{-1}\| (\|A_D^{-1} E_D\| + \|A_D^{-1} E_O\|). \tag{3.11}$$

Inequalities (3.3) and the assumptions on A_D, E_D, A_O, and E_O give

$$\|A_D^{-1}E_D\| \le \epsilon, \quad \text{and} \quad \|A_D^{-1}E_O\| \le \epsilon\|A_D^{-1}\|\|A_O\| \le \frac{\epsilon\sqrt{r}}{t}, \tag{3.12}$$

where the strict block t-diagonal dominance of the first r columns of A has been used to obtain the second group of inequalities. Collecting terms in (3.11) using (3.4) and (3.12) gives

$$\|A_{11}^{-1}E_{11}\| \le \frac{t\sqrt{r}}{t-1}\left(\epsilon + \frac{\epsilon\sqrt{r}}{t}\right) = \epsilon\delta_2, \tag{3.13}$$

where

$$\delta_2 = \frac{\sqrt{r}(t+\sqrt{r})}{t-1}.$$

Using (3.6), (3.13), and the given assumptions on t and the norms of E_{11}, E_{12}, E_{21}, and E_{22} leads to the inequalities

$$\begin{aligned}
\|F_{11}\| &\le \frac{\epsilon}{1-\delta_1\|A_{21}\|\|A_{22}^{-1}\|}\left(\delta_2 + \delta_1\|A_{21}\|\|A_{22}^{-1}\|\right),\\
\|F_{12}\| &\le \frac{\epsilon}{1-\delta_1\|A_{21}\|\|A_{22}^{-1}\|}\left(\delta_1 + \delta_1\kappa(A_{22}) + \frac{\delta_1}{t}\|A_{21}\|\|A_{22}^{-1}\|\right),\\
\|F_{21}\| &\le \frac{\epsilon}{1-\delta_1\|A_{21}\|\|A_{22}^{-1}\|}\left(\delta_2+1\right)\|A_{21}\|\|A_{22}^{-1}\|,\\
\|F_{22}\| &\le \frac{\epsilon}{1-\delta_1\|A_{21}\|\|A_{22}^{-1}\|}\left(\kappa(A_{22}) + \left(\delta_1 + \frac{1}{t}\right)\|A_{21}\|\|A_{22}^{-1}\|\right).
\end{aligned}$$

It follows that $\|F_{ij}\| \le \epsilon\delta/(1-\delta_1\|A_{21}\|\|A_{22}^{-1}\|)$, for $i,j = 1,2$, where

$$\delta = \delta_1 + \delta_2 + (1+\delta_1)\kappa(A_{22}) + \left(\delta_1 + \delta_2 + \left(1+\frac{1}{t}\right)(1+\delta_1)\right)\|A_{21}\|\|A_{22}^{-1}\|.$$

The required result now follows since the norm of F is no larger than n times the magnitude of its largest element, and hence no larger than n times any $\|F_{ij}\|$, $i,j = 1,2$. □

This theorem implies that for large t, the sensitivity of the solution depends on $\|A_{21}\|\|A_{22}^{-1}\|$ and $\kappa(A_{22})$. For t sufficiently large, we obtain $\gamma_1 \approx n\epsilon\sqrt{r}$, $\gamma_2 \approx n\epsilon$, and $\gamma_3 \approx n\epsilon(\sqrt{r}+1)$. It is important to note that the only requirement on A_{11} is that its diagonal elements are sufficiently large in magnitude. For example, the result holds even when the diagonals of A_{11} go to infinity at widely varying rates. This phenomenon may occur when the strict complementarity condition (1.4d) does not hold (see (2.15), where the diagonals are $-2t$ and $-2t^2$).

The results of Ponceleón [21] also follow from Theorem 3.1 if it is assumed that F_{11} and F_{21} can be ignored compared to F_{12} and F_{22}. This assumption implies that $\|F_{22}\|$ (i.e., $\kappa(A_{22})$) is the dominating factor in the sensitivity of the solution.

4. Rounding error analysis. In this section we give a backward error analysis of the solution of a strictly t-diagonally dominant system $Ax = b$ by means of the LBL^T factorization. The purpose is to justify the form of the perturbations used in the analysis of Theorem 3.1.

Throughout, we use the "standard model" of floating-point arithmetic in which the evaluation of an expression in floating-point arithmetic is denoted by $fl(\cdot)$, with

$$fl(a \operatorname{op} b) = (a \operatorname{op} b)(1+\delta), \quad |\delta| \leq u, \qquad \operatorname{op} = +, -, *, /. \tag{4.1}$$

(See, for example, Higham [12].) Here, u is the unit round-off associated with the particular machine being used. In the subsequent analysis it will be necessary to assume that $\max\{3n, 8\}u < 1$.

At the start of the kth stage ($k \geq 1$) of the LBL^T factorization of a matrix A we have

$$A^{(k)} = \left(a_{ij}^{(k)}\right) = P_k^T A P_k - \sum_{i=1}^{k-1} L_i B_{ii} L_i^T = \begin{pmatrix} 0 & 0 \\ 0 & A_k \end{pmatrix}, \tag{4.2}$$

where P_k is a permutation and L_i is either an $n \times 1$ or $n \times 2$ matrix consisting of the column(s) of L computed at the ith stage. The matrix A_k is called the *Schur complement*, and represents the part of A remaining to be factorized. At the kth stage, a 1×1 or symmetric 2×2 submatrix of A_k is selected as the next pivot B_{kk}. The pivot rows and columns are brought to the leading position of A_k using a symmetric interchange, which must be applied to $P_k^T A P_k$ and the rows of each L_i. The 1×1 or 2×2 pivot is then used to eliminate one or two rows and columns from the permuted Schur complement. The kth stage is completed by the computation of L_k from B_{kk}^{-1} and the pivot columns of $A^{(k)}$ (for further details, see Bunch [1]).

In all of what follows, we assume a pivoting strategy that selects the large pivots first. This implies that the r diagonals from the $r \times r$ block t-diagonally dominant submatrix of A are selected as the first r pivots, and that the matrices at the kth stage ($k \leq r+1$) satisfy

$$A^{(k)} = \left(a_{ij}^{(k)}\right) = P_k^T A P_k - \sum_{i=1}^{k-1} l_i b_{ii} l_i^T = \begin{array}{c} \\ k-1 \\ n-k+1 \end{array} \begin{array}{c} \begin{array}{cc} k-1 & n-k+1 \end{array} \\ \begin{pmatrix} 0 & 0 \\ 0 & A_k \end{pmatrix} \end{array}, \tag{4.3}$$

where $l_i^T = (0, \ldots, 0, 1, l_{i,i+1}, \ldots, l_{in})$. To simplify the notation, we assume that all necessary interchanges are done at the start of the algorithm.

Once the r dominant rows and columns have been eliminated, the LBL^T factorization continues on the remaining matrix A_{r+1} with either 1×1 or 2×2 pivots being used as necessary. If we assume that all necessary interchanges are done at the start of the algorithm, A can be written as

$$\begin{aligned} \begin{pmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{pmatrix} &= \begin{pmatrix} L_{11} \\ L_{21} \end{pmatrix} B_{11} \begin{pmatrix} L_{11}^T & L_{21}^T \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & A_{r+1} \end{pmatrix} \\ &= \begin{pmatrix} L_{11} \\ L_{21} \end{pmatrix} B_{11} \begin{pmatrix} L_{11}^T & L_{21}^T \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & L_{22} B_{22} L_{22}^T \end{pmatrix}, \end{aligned}$$

where A_{11} is $r \times r$, and $L_{22} B_{22} L_{22}^T$ is the LBL^T factorization of the $(n-r+1) \times (n-r+1)$ remaining matrix A_{r+1}. Note that B_{11} and B_{22} denote block diagonal matrices associated with the factorization

$$\begin{pmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} L_{11} & \\ L_{21} & L_{22} \end{pmatrix} \begin{pmatrix} B_{11} & \\ & B_{22} \end{pmatrix} \begin{pmatrix} L_{11}^T & L_{21}^T \\ & L_{22}^T \end{pmatrix}. \tag{4.4}$$

The vector x is now found from the linear system $LBL^Tx = b$ by solving the sequence of triangular and block-diagonal systems

$$Lz = b, \quad By = z, \quad \text{and} \quad L^Tx = y. \tag{4.5}$$

The next three lemmas concern the first r steps of the factorization and hence relate to stages associated with 1×1 pivots. Here, and throughout, a "hat" $\widehat{\ }$ is used to denote computed quantities. If a 1×1 pivot is chosen at step k of the factorization, the following quantities are computed:

$$\widehat{b}_{kk} = \widehat{a}_{kk}^{(k)}, \qquad \widehat{l}_{kk} = 1, \qquad \widehat{l}_{ik} = fl\left(\frac{\widehat{a}_{ik}^{(k)}}{\widehat{a}_{kk}^{(k)}}\right), \quad i \geq k+1; \tag{4.6a}$$

$$\widehat{a}_{ji}^{(k+1)} = \widehat{a}_{ij}^{(k+1)} = fl\left(\widehat{a}_{ij}^{(k)} - \frac{\widehat{a}_{ik}^{(k)}\widehat{a}_{jk}^{(k)}}{\widehat{a}_{kk}^{(k)}}\right), \quad i \geq j \geq k+1. \tag{4.6b}$$

Note that $\widehat{a}_{ij}^{(k+1)}$ is defined to be $\widehat{a}_{ji}^{(k+1)}$ for $j > i$.

The next lemma shows that the updated uneliminated t-diagonally dominant columns remain "almost" strictly t-diagonally dominant in the matrix remaining to be factorized. The lemma generalizes a standard result associated with diagonally dominant matrices (see, e.g., Golub and Van Loan [8, p. 120]) to cases where A is t-diagonally dominant and the computation is performed in finite-precision arithmetic.

LEMMA 4.1. *Let A be a symmetric $n \times n$ matrix whose first and jth ($j \geq 2$) columns are strictly t-diagonally dominant for some t ($t > 1$). Then, it holds that*

$$|\widehat{a}_{jj}^{(2)}| > \frac{t}{1+7u}\left(1 - 7u\frac{t^2+1}{t^2-1}\right)\sum_{\substack{i\geq 2\\ i\neq j}} |\widehat{a}_{ij}^{(2)}|,$$

where u is the unit round-off.

Proof. Since columns 1 and j are strictly t-diagonally dominant, we have

$$\frac{1}{t}|a_{11}| > \sum_{i\neq 1} |a_{i1}| = |a_{j1}| + \sum_{\substack{i\geq 2\\ i\neq j}} |a_{i1}|, \tag{4.7a}$$

$$\frac{1}{t}|a_{jj}| > \sum_{i\neq j} |a_{ij}| = |a_{1j}| + \sum_{\substack{i\geq 2\\ i\neq j}} |a_{ij}|. \tag{4.7b}$$

Using (4.1) we can write

$$\widehat{a}_{ij}^{(2)} = a_{ij}(1+\alpha_{ij}) - \frac{a_{i1}a_{j1}}{a_{11}}(1+\beta_{i1})(1+\sigma_{j1})(1+\alpha_{ij}), \tag{4.8}$$

for $i = 1, \ldots, n$, where $|\alpha_{ij}| < u$, $|\beta_{i1}| < u$, and $|\sigma_{j1}| < u$. Since we assume that $u < 1$, it is straightforward to verify that $|(1+\beta_{i1})(1+\sigma_{j1})(1+\alpha_{ij}) - 1| < 7u$. If this bound is used to majorize the sum of the absolute values of the terms in (4.8), we have

$$\sum_{\substack{i\geq 2\\ i\neq j}} |\widehat{a}_{ij}^{(2)}| \leq (1+7u)\left(\sum_{\substack{i\geq 2\\ i\neq j}} |a_{ij}| + \frac{|a_{j1}|}{|a_{11}|}\sum_{\substack{i\geq 2\\ i\neq j}} |a_{i1}|\right). \tag{4.9}$$

The second factor on the right-hand side of this inequality can be further simplified using (4.7) and the symmetry of A, i.e.,

$$\sum_{\substack{i\geq 2\\ i\neq j}} |a_{ij}| + \frac{|a_{j1}|}{|a_{11}|} \sum_{\substack{i\geq 2\\ i\neq j}} |a_{i1}| < \frac{1}{t}|a_{jj}| - |a_{j1}| + \frac{|a_{j1}|}{|a_{11}|}\left(\frac{1}{t}|a_{11}| - |a_{j1}|\right). \tag{4.10}$$

Using $t \geq 1$ in the right-hand side of this inequality gives

$$\frac{1}{t}|a_{jj}| - |a_{j1}| + \frac{|a_{j1}|}{|a_{11}|}\left(\frac{1}{t}|a_{11}| - |a_{j1}|\right) \leq \frac{1}{t}\left(|a_{jj}| - \frac{|a_{j1}|^2}{|a_{11}|}\right) \leq \frac{1}{t}\left|a_{jj} - \frac{a_{j1}^2}{a_{11}}\right|.$$

Combining this inequality with (4.9) and (4.10) provides the bound

$$\sum_{\substack{i\geq 2\\ i\neq j}} |\widehat{a}_{ij}^{(2)}| < \frac{1+7u}{t}\left|a_{jj} - \frac{a_{j1}^2}{a_{11}}\right|. \tag{4.11}$$

Letting $i = j$ in (4.8) and using the bound on the rounding errors $|(1+\beta_{i1})(1+\sigma_{j1})(1+\alpha_{ij})|$ gives

$$|\widehat{a}_{jj}^{(2)}| \geq \left|a_{jj} - \frac{a_{j1}^2}{a_{11}}\right| - 7u\left(1 + \frac{|a_{j1}|}{|a_{11}|}\frac{|a_{j1}|}{|a_{jj}|}\right)|a_{jj}|. \tag{4.12}$$

We need to bound this quantity from below by a term involving t and $|a_{jj} - a_{j1}^2/a_{11}|$. To do this, note that

$$\left|a_{jj} - \frac{a_{j1}^2}{a_{11}}\right| \geq |a_{jj}| - \frac{|a_{j1}^2|}{|a_{11}|} = \left(1 - \frac{|a_{j1}|}{|a_{11}|}\frac{|a_{j1}|}{|a_{jj}|}\right)|a_{jj}|. \tag{4.13}$$

The assumption of strict t-diagonal dominance for the first and jth columns implies that the ratios $|a_{j1}|/|a_{jj}|$ and $|a_{j1}|/|a_{11}|$ are bounded above by $1/t$. These inequalities are used in (4.12) and (4.13) to eliminate $|a_{jj}|$, giving

$$\left|\widehat{a}_{jj}^{(2)}\right| \geq \left(1 - 7u\frac{t^2+1}{t^2-1}\right)\left|a_{jj} - \frac{a_{j1}^2}{a_{11}}\right|.$$

The result now follows directly from this inequality and (4.11). □

The 3×3 example (2.15) indicates that for large t, the elements of the Schur complement A_k are very close to the corresponding elements of A. Moreover, Lemma 4.1 states that the strictly t-diagonally dominant columns of A remain "almost" strictly t-diagonally dominant in the computed Schur complements. In the next lemma we make these observations precise. In particular, it is shown that the fraction of strict t-diagonal dominance preserved in the computed Schur complements is close to one.

LEMMA 4.2. *Let A be a symmetric $n\times n$ matrix whose first r columns are strictly t-diagonally dominant, with $t \geq 4\gamma^{-r+1}$ and $\gamma = (1-8u)/(1+7u)$. Assume that the first r steps of an LBL^T factorization of A involve 1×1 pivots from A_{11}. At the start of step k ($k \leq r+1$) let A_k denote the Schur complement remaining to be factorized* (see (4.3)). *Then the first $r-k+1$ columns of the computed Schur complement $\widehat{A}_k$ are strictly $(\gamma^{k-1}t)$-diagonally dominant, and the elements of $\widehat{A}^{(k)}$ satisfy*

(a) $|\widehat{a}_{ij}^{(k)}| = |\widehat{a}_{ji}^{(k)}| \leq (1+7u)^{k-1}\left(|a_{ij}| + \left((1+\omega_k)^{k-1}-1\right)\max_{q\leq k-1}|a_{iq}|\right)$, $i \geq j \geq k$; *and*

(b) $|\widehat{a}_{jj}^{(k)}| \leq (1+7u)^{k-1}\left(1+\omega_k^2\right)^{k-1}|a_{jj}|$, $r \geq j \geq k$,

where $\omega_k = 1/(\gamma^{k-2}t)$.

Proof. The proof uses the definition of $\widehat{A}_k$ as the last $(n-k+1)\times(n-k+1)$ principal submatrix of $\widehat{A}^{(k)}$. The properties of $\widehat{A}_k$ are established by induction. Since $\widehat{A}_1 = A$, it follows that the first r columns of $\widehat{A}_1$ are strictly t-diagonally dominant, and that the elements of $\widehat{A}^{(1)}$ satisfy inequalities (a) and (b).

Now assume that the lemma is true for step k, i.e., the first $r-k+1$ columns of $\widehat{A}_k$ are strictly $(\gamma^{k-1}t)$-diagonally dominant, and the elements of $\widehat{A}_k$ satisfy inequalities (a) and (b). Recall that for $i > j$, the element $\widehat{a}_{ji}^{(k+1)}$ is defined as $\widehat{a}_{ij}^{(k+1)}$, where $\widehat{a}_{ij}^{(k+1)}$ is computed from $\widehat{A}^{(k)}$ using (4.6b). The floating-point model (4.1) gives

$$\widehat{a}_{ij}^{(k+1)} = \widehat{a}_{ij}^{(k)}(1+\alpha_{ij}^{(k)}) - \frac{\widehat{a}_{ik}^{(k)}\widehat{a}_{jk}^{(k)}}{\widehat{a}_{kk}^{(k)}}(1+\beta_{ik}^{(k)})(1+\sigma_{jk}^{(k)})(1+\alpha_{ij}^{(k)}), \tag{4.14}$$

where $|\alpha_{ij}^{(k)}| < u$, $|\beta_{ik}^{(k)}| < u$ and $|\sigma_{jk}^{(k)}| < u$.

To show that the first $r-k$ columns of $\widehat{A}_{k+1}$ are strictly $(\gamma^k t)$-diagonally dominant, we use the definition (4.6b) for the elements of $\widehat{A}^{(k+1)}$ and apply Lemma 4.1. Since $t \geq 4\gamma^{-r+1}$ with $k \leq r$ and $\gamma \leq 1$, we conclude that $\gamma^{k-1}t \geq 4\gamma^{k-r} \geq 4$. However, if $\gamma^{k-1}t \geq 4$, it is straightforward to verify that

$$\frac{(\gamma^{k-1}t)^2+1}{(\gamma^{k-1}t)^2-1} < \frac{8}{7},$$

so that Lemma 4.1 implies that, for $j = k+1, \ldots, r$,

$$|\widehat{a}_{jj}^{(k+1)}| > (\gamma^{k-1}t)\frac{1-8u}{1+7u}\sum_{\substack{i\geq k+1\\ i\neq j}}|\widehat{a}_{ij}^{(k+1)}| = (\gamma^k t)\sum_{\substack{i\geq k+1\\ i\neq j}}|\widehat{a}_{ij}^{(k+1)}|.$$

This implies that columns $1, \ldots, r-k$ of $\widehat{A}_{k+1}$ are strictly $(\gamma^k t)$-diagonally dominant, as required.

Now we show that after the kth $(k \leq r)$ step, the bound (a) applies with $k = k+1$. Let $i \geq j > k$. The strict $(\gamma^{k-1}t)$-diagonal dominance of column k of $\widehat{A}^{(k)}$ implies that $|\widehat{a}_{jk}^{(k)}/\widehat{a}_{kk}^{(k)}| \leq 1/(\gamma^{k-1}t)$. Using this inequality in (4.14) and applying the bounds $1+u \leq (1+u)^3 \leq (1+7u)$ gives

$$|\widehat{a}_{ij}^{(k+1)}| \leq (1+7u)\left(|\widehat{a}_{ij}^{(k)}| + \frac{1}{\gamma^{k-1}t}|\widehat{a}_{ik}^{(k)}|\right). \tag{4.15}$$

We use the inequality of Part (a) and the induction hypothesis to obtain the following bounds on $\widehat{a}_{ij}^{(k)}$ and $\widehat{a}_{ik}^{(k)}$:

$$|\widehat{a}_{ij}^{(k)}| \leq (1+7u)^{k-1}(|a_{ij}| + \left((1+\gamma\omega_{k+1})^{k-1}-1\right)\max_{q\leq k-1}|a_{iq}|),$$

$$|\widehat{a}_{ik}^{(k)}| \leq (1+7u)^{k-1}(|a_{ik}| + \left((1+\gamma\omega_{k+1})^{k-1}-1\right)\max_{q\leq k-1}|a_{iq}|),$$

where ω_{k+1} denotes the quantity $1/(\gamma^{k-1}t)$. If the range of the maximization is extended to include $|a_{ik}|$ and these bounds are substituted for $\widehat{a}_{ij}^{(k)}$ and $\widehat{a}_{ik}^{(k)}$ in (4.15), we obtain

$$|\widehat{a}_{ij}^{(k+1)}| \leq (1+7u)^k \left(|a_{ij}| + \left((1+\gamma\omega_{k+1})^{k-1}(1+\omega_{k+1}) - 1\right) \max_{q\leq k} |a_{iq}| \right).$$

The identity $(1+\gamma\omega_{k+1})^{k-1}(1+\omega_{k+1}) \leq (1+\omega_{k+1})^k$ gives the required result.

Now we turn to the inequality of Part (b). Consider j such that $k < j \leq r$. If $i = j$ in (4.14), then

$$|\widehat{a}_{jj}^{(k+1)}| \leq (1+u)|\widehat{a}_{jj}^{(k)}| + (1+u)^3 \frac{|\widehat{a}_{jk}^{(k)}|}{|\widehat{a}_{kk}^{(k)}|} \frac{|\widehat{a}_{jk}^{(k)}|}{|\widehat{a}_{jj}^{(k)}|} |\widehat{a}_{jj}^{(k)}|. \tag{4.16}$$

Using the strict $(\gamma^{k-1}t)$-diagonal dominance of columns k and j of $\widehat{A}^{(k)}$, and the inequalities $1+u \leq (1+u)^3 \leq 1+7u$, it follows from (4.16) that

$$|\widehat{a}_{jj}^{(k+1)}| \leq (1+7u)(1+\omega_{k+1}^2)|\widehat{a}_{jj}^{(k)}|.$$

Using the inequality of Part (b) to bound $|\widehat{a}_{jj}^{(k)}|$ gives

$$|\widehat{a}_{jj}^{(k+1)}| \leq (1+7u)^k(1+\omega_{k+1}^2)^k |a_{jj}|,$$

as required. □

In our analysis, the first r steps of the factorization are "special" in the sense that they are made with 1×1 pivots that are large in magnitude. The next lemma bounds the error in the partially computed factors defined from the first r eliminations.

LEMMA 4.3. *Assume that the LBL^T factorization of a symmetric $n \times n$ matrix A is defined so that 1×1 pivots are used during the first r steps. Then*

$$\left| a_{ij} - \sum_{k=1}^{r} \widehat{l}_{ik}\widehat{b}_{kk}\widehat{l}_{jk} - \widehat{a}_{ij}^{(r+1)} \right| \leq \frac{3(r+1)u}{1-3(r+1)u} \left(|\widehat{a}_{ij}^{(r+1)}| + \sum_{k=1}^{r} \frac{|\widehat{a}_{ik}^{(k)}\widehat{a}_{jk}^{(k)}|}{|\widehat{a}_{kk}^{(k)}|} \right), \quad i, j \geq r+1,$$

and

$$\left| a_{ij} - \sum_{k=1}^{j} \widehat{l}_{ik}\widehat{b}_{kk}\widehat{l}_{jk} \right| \leq \frac{3(j+1)u}{1-3(j+1)u} \sum_{k=1}^{j} \frac{|\widehat{a}_{ik}^{(k)}\widehat{a}_{jk}^{(k)}|}{|\widehat{a}_{kk}^{(k)}|}, \quad j \leq r, \quad i \geq j.$$

Proof. Consider the qth stage of the factorization, where $q \leq r+1$. For $i, j \geq q$ we have

$$\left| a_{ij} - \widehat{a}_{ij}^{(q)} - \sum_{k=1}^{q-1} \widehat{l}_{ik}\widehat{b}_{kk}\widehat{l}_{jk} \right| \leq \left| a_{ij} - \widehat{a}_{ij}^{(q)} - \sum_{k=1}^{q-1} \frac{\widehat{a}_{ik}^{(k)}\widehat{a}_{jk}^{(k)}}{\widehat{a}_{kk}^{(k)}} \right|$$

$$+ \sum_{k=1}^{q-1} \left| \widehat{l}_{ik}\widehat{b}_{kk}\widehat{l}_{jk} - \frac{\widehat{a}_{ik}^{(k)}\widehat{a}_{jk}^{(k)}}{\widehat{a}_{kk}^{(k)}} \right|. \tag{4.17}$$

To bound the first term of the right-hand side of this inequality, we start with (4.6b) and obtain

$$\widehat{a}_{ij}^{(q)} = fl\left(a_{ij} - \sum_{k=1}^{q-1} \frac{\widehat{a}_{ik}^{(k)}\widehat{a}_{jk}^{(k)}}{\widehat{a}_{kk}^{(k)}} \right) \quad i, j \geq q.$$

Some standard error analysis (see Stoer and Bulirsch [23, pp. 25–27] and Lemma 9.5 in Higham [11]) and the inequalities $(1+u) \le (1+u)^2 \le (1+3u)$ gives

$$\left| a_{ij} - \widehat{a}_{ij}^{(q)} - \sum_{k=1}^{q-1} \frac{\widehat{a}_{ik}^{(k)} \widehat{a}_{jk}^{(k)}}{\widehat{a}_{kk}^{(k)}} \right| \le \frac{3(q-1)u}{1-3(q-1)u} \left(|\widehat{a}_{ij}^{(q)}| + \sum_{k=1}^{q-1} \frac{|\widehat{a}_{ik}^{(k)} \widehat{a}_{jk}^{(k)}|}{|\widehat{a}_{kk}^{(k)}|} \right), \tag{4.18}$$

for all $i, j \ge q$. For the second term of the right-hand side in (4.17) we use (4.6) to give

$$\widehat{l}_{ik} \widehat{b}_{kk} \widehat{l}_{jk} = \frac{\widehat{a}_{ik}^{(k)} \widehat{a}_{jk}^{(k)}}{\widehat{a}_{kk}^{(k)}} (1 + \alpha_{ik}^{(k)})(1 + \beta_{jk}^{(k)}), \qquad i, j \ge k,$$

where $|\alpha_{ik}^{(k)}| \le u$ and $|\beta_{jk}^{(k)}| \le u$. Then, we have

$$\left| \widehat{l}_{ik} \widehat{b}_{kk} \widehat{l}_{jk} - \frac{\widehat{a}_{ik}^{(k)} \widehat{a}_{jk}^{(k)}}{\widehat{a}_{kk}^{(k)}} \right| \le 3u \frac{|\widehat{a}_{ik}^{(k)} \widehat{a}_{jk}^{(k)}|}{|\widehat{a}_{kk}^{(k)}|}, \qquad i, j \ge k. \tag{4.19}$$

A combination of (4.17), (4.18), and (4.19) now gives

$$\left| a_{ij} - \sum_{k=1}^{q-1} \widehat{l}_{ik} \widehat{b}_{kk} \widehat{l}_{jk} - \widehat{a}_{ij}^{(q)} \right| \le \frac{3qu}{1-3qu} \left(|\widehat{a}_{ij}^{(q)}| + \sum_{k=1}^{q-1} \frac{|\widehat{a}_{ik}^{(k)} \widehat{a}_{jk}^{(k)}|}{|\widehat{a}_{kk}^{(k)}|} \right), \quad i, j \ge q. \tag{4.20}$$

The first inequality of the lemma follows directly from the substitution of $q = r+1$ in this inequality.

To establish the second inequality of the lemma, consider the jth stage of the factorization, where $j \le r$. For $i \ge j$, we have

$$\left| a_{ij} - \sum_{k=1}^{j} \widehat{l}_{ik} \widehat{b}_{kk} \widehat{l}_{jk} \right| \le \left| a_{ij} - \sum_{k=1}^{j-1} \widehat{l}_{ik} \widehat{b}_{kk} \widehat{l}_{jk} - \widehat{a}_{ij}^{(j)} \right| + |\widehat{a}_{ij}^{(j)} - \widehat{l}_{ij} \widehat{b}_{jj} \widehat{l}_{jj}|.$$

A combination of (4.19) with $k = j$ and (4.20) with $q = j$ yields

$$\left| a_{ij} - \sum_{k=1}^{j} \widehat{l}_{ik} \widehat{b}_{kk} \widehat{l}_{jk} \right| \le \frac{3(j+1)u}{1-3(j+1)u} \sum_{k=1}^{j} \frac{|\widehat{a}_{ik}^{(k)} \widehat{a}_{jk}^{(k)}|}{|\widehat{a}_{kk}^{(k)}|}, \quad j \le r, \quad i \ge j,$$

as required. □

In the next theorem we show that the computed solution $\widehat{x}$ is the exact solution of $(A+E)\widehat{x} = b$, where E is a backward-error matrix whose elements are bounded in magnitude by quantities involving $1/t$ and the unit round-off u. The theorem is established by accumulating the backward error from each of the following steps: (i) the elimination of the rows and columns associated with the first r 1×1 pivots and the subsequent modification to the remaining matrix A_{r+1} (see Lemma 4.3); (ii) the LBL^T factorization of A_{r+1}; and (iii) the solution of the triangular and block-diagonal systems (4.5). In the cases (ii) and (iii), we utilize some standard bounds on the backward error derived in the literature.

THEOREM 4.4. *Let A be a symmetric $n \times n$ matrix whose first r columns are strictly t-diagonally dominant, with $t \ge 4\gamma^{-r+1}$ and $\gamma = (1-8u)/(1+7u)$. Consider*

an LBL^T factorization (4.4) *of A in which the first r stages involve 1×1 pivots from A_{11}. Assume that the computed values of A_{r+1} and its factors satisfy*

$$|\widehat{A}_{r+1} - \widehat{L}_{22}\widehat{B}_{22}\widehat{L}_{22}^T| \leq \delta \max_{\substack{p \geq r+1 \\ q \geq r+1}} |\widehat{a}_{pq}^{(r+1)}| ee^T, \tag{4.21a}$$

$$|\widehat{L}_{22}||\widehat{B}_{22}||\widehat{L}_{22}^T| \leq \eta \max_{\substack{p \geq r+1 \\ q \geq r+1}} |\widehat{a}_{pq}^{(r+1)}| ee^T, \tag{4.21b}$$

where δ is of moderate size relative to u, η is bounded and e denotes the vector of ones. If $\widehat{x}$ is the computed value of x from the triangular systems (4.5), *assume that the computed values of y and z satisfy $(\widehat{B}_{22} + \widehat{B}_{E,22})\widehat{y} = \widehat{z}$, where $\widehat{B}_{E,22}$ is a block-diagonal matrix such that*

$$|\widehat{B}_{E,22}| \leq \frac{nu}{1 - nu}|\widehat{B}_{22}|.$$

Then, $\widehat{x}$ is the exact solution of $(A + E)\widehat{x} = b$, where the elements of E satisfy

$$\begin{aligned}
|e_{jj}| &\leq \epsilon_1 |a_{jj}|, && j \leq r, \\
\max\{|e_{ji}|, |e_{ij}|\} &\leq \epsilon_2 \max_{q \leq j} |a_{iq}|, && j \leq r, \quad i > j, \\
|e_{ij}| &\leq \epsilon_3 \max_{\substack{p \geq r+1 \\ q \geq r+1}} |a_{pq}| + \frac{1}{t}\epsilon_4 \max_{\substack{p \geq r+1 \\ q \leq r}} |a_{pq}|, && i, j \geq r+1,
\end{aligned}$$

with

$$\begin{aligned}
\epsilon_1 &= \frac{7nu(2 + 3u)}{1 - 3nu}(1 + 7u)^r(1 + \omega_r^2)^r(1 + r\omega_r^2), \\
\epsilon_2 &= \frac{7nu(2 + 3u)}{1 - 3nu}(1 + 7u)^r(1 + \omega_r)^r(1 + r\omega_r), \\
\epsilon_3 &= \left(\frac{7nu(\eta + 1)}{1 - 3nu} + \delta\right)(1 + 7u)^r, \\
\epsilon_4 &= (2((1 + \omega_{r+1})^r - 1) + r\omega_r(2 + 3u)(1 + \omega_r)^r)\, t\epsilon_3,
\end{aligned}$$

$\gamma = (1 - 8u)/(1 + 7u)$, *and* $\omega_r = 1/(\gamma^{r-2}t)$.

Proof. Any backward error matrix E is of the form $E = C + F$, where C is the error from the triangular solves and F is a backward error associated with the LBL^T factorization.

First we consider the backward error resulting from the solution of the block-diagonal and triangular systems (4.5). The backward error C satisfies $(\widehat{L}\widehat{B}\widehat{L}^T + C)\widehat{x} = b$, where $\widehat{x}$ is the computed value of x from (4.5). It follows from Stoer and Bulirsch [23, p. 196] that the intermediate vectors $\widehat{z}$ and $\widehat{x}$ satisfy $(\widehat{L} + \widehat{L}_E)\widehat{z} = b$ and $(\widehat{L}^T + \widehat{U}_E)\widehat{x} = \widehat{y}$, where $\widehat{L}_E$ and $\widehat{U}_E$ have the same element-wise bound

$$|\widehat{L}_E| \leq \frac{nu}{1 - nu}|\widehat{L}| \quad \text{and} \quad |\widehat{U}_E| \leq \frac{nu}{1 - nu}|\widehat{L}^T|.$$

The assumption on $\widehat{B}_{22}$ implies that $(\widehat{B} + \widehat{B}_E)\widehat{y} = \widehat{z}$, where

$$|\widehat{B}_E| \leq \frac{nu}{1 - nu}|\widehat{B}|.$$

Combining these equations, it follows that $\hat{x}$ satisfies

$$(\hat{L} + \hat{L}_E)(\hat{B} + \hat{B}_E)(\hat{L}^T + \hat{U}_E)\hat{x} = b.$$

Collecting terms and bounding the error terms, we obtain the backward error matrix C such that $(\hat{L}\hat{B}\hat{L}^T + C)\hat{x} = b$, and

$$|C| \leq \frac{7nu}{1 - nu}|\hat{L}||\hat{B}||\hat{L}^T|.$$

If C is written in partitioned form, the bounds on each submatrix of C may be written as

$$|C_{11}| \leq \frac{7nu}{1 - nu}|\hat{L}_{11}||\hat{B}_{11}||\hat{L}_{11}^T|, \tag{4.22a}$$

$$|C_{12}| \leq \frac{7nu}{1 - nu}|\hat{L}_{11}||\hat{B}_{11}||\hat{L}_{21}^T|, \tag{4.22b}$$

$$|C_{21}| \leq \frac{7nu}{1 - nu}|\hat{L}_{21}||\hat{B}_{11}||\hat{L}_{11}^T|, \tag{4.22c}$$

$$|C_{22}| \leq \frac{7nu}{1 - nu}\left(|\hat{L}_{21}||\hat{B}_{11}||\hat{L}_{21}^T| + |\hat{L}_{22}||\hat{B}_{22}||\hat{L}_{22}^T|\right). \tag{4.22d}$$

Now we turn to the backward error associated with the LBL^T factorization itself. On completion of the factorization, we have $A + F = \hat{L}\hat{B}\hat{L}^T$, which may be written in the partitioned form

$$\begin{pmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{pmatrix} + \begin{pmatrix} F_{11} & F_{21}^T \\ F_{21} & F_{22} \end{pmatrix} = \begin{pmatrix} \hat{L}_{11} \\ \hat{L}_{21} \end{pmatrix} \hat{B}_{11} \begin{pmatrix} \hat{L}_{11}^T & \hat{L}_{21}^T \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \hat{L}_{22}\hat{B}_{22}\hat{L}_{22}^T \end{pmatrix}$$

(cf. (4.4)). First, consider the elements of F_{11} and F_{21}, i.e., elements f_{ij} for $j \leq r$. Lemma 4.3 provides symmetric bounds on $|F_{11}|$ and $|F_{21}|$. Without loss of generality, let $i \geq j$. Then, Lemma 4.3 gives

$$|f_{ij}| \leq \frac{3(j+1)u}{1 - 3(j+1)u} \sum_{k=1}^{j} \frac{|\hat{a}_{ik}^{(k)}\hat{a}_{jk}^{(k)}|}{|\hat{a}_{kk}^{(k)}|}, \quad j \leq r, \quad i \geq j.$$

The bounds on E_{11}, E_{12}, and E_{21} now follow by combining the bounds (4.22a)–(4.22c) on C and this bound on F. Using the componentwise statement of $|\hat{L}||\hat{B}||\hat{L}^T|$ for $j \leq r$ and $i \geq j$, it follows that

$$|e_{ij}| \leq \frac{7nu}{1 - nu} \sum_{k=1}^{j} |\hat{l}_{ik}\hat{b}_{kk}\hat{l}_{jk}| + \frac{3(j+1)u}{1 - 3(j+1)u} \sum_{k=1}^{j} \frac{|\hat{a}_{ik}^{(k)}\hat{a}_{jk}^{(k)}|}{|\hat{a}_{kk}^{(k)}|}.$$

This inequality may be simplified further using (4.19), viz.,

$$|e_{ij}| \leq \frac{7nu(2+3u)}{1 - 3nu} \sum_{k=1}^{j} \frac{|\hat{a}_{ik}^{(k)}\hat{a}_{jk}^{(k)}|}{|\hat{a}_{kk}^{(k)}|}, \quad j \leq r, \quad i \geq j. \tag{4.23}$$

In particular, when $i = j$, this inequality bounds the diagonals of E_{11} as

$$|e_{jj}| \leq \frac{7nu(2+3u)}{1 - 3nu}\left(|\hat{a}_{jj}^{(j)}| + \sum_{k=1}^{j-1} \frac{|\hat{a}_{jk}^{(k)}|}{|\hat{a}_{kk}^{(k)}|} \frac{|\hat{a}_{jk}^{(k)}|}{|\hat{a}_{jj}^{(k)}|}|\hat{a}_{jj}^{(k)}|\right), \quad j \leq r. \tag{4.24}$$

Lemma 4.2 now gives the required bound on $|e_{jj}|$ in terms of $\gamma = (1-8u)/(1+7u)$ and $\omega_r = 1/(\gamma^{r-2}t)$. A similar argument using inequality (4.23) and Lemma 4.2 gives the required bounds on the elements of E_{21} and the off-diagonal elements of E_{11}.

The backward error E_{22} involves the error in the formation of $\widehat{A}_{r+1}$ and the error in the computed factors $\widehat{L}_{22}$ and $\widehat{B}_{22}$. We can write

$$-F_{22} = A_{22} - \widehat{L}_{21}\widehat{B}_{11}\widehat{L}_{21}^T - \widehat{A}_{r+1} + \widehat{A}_{r+1} - \widehat{L}_{22}\widehat{B}_{22}\widehat{L}_{22}^T,$$

which in conjunction with (4.22d) gives $E_{22} = G_{22} + H_{22}$, where

$$|G_{22}| \leq \frac{7nu}{1-nu}|\widehat{L}_{21}||\widehat{B}_{11}||\widehat{L}_{21}^T| + |A_{22} - \widehat{L}_{21}\widehat{B}_{11}\widehat{L}_{21}^T - \widehat{A}_{r+1}|, \tag{4.25a}$$

$$|H_{22}| \leq \frac{7nu}{1-nu}|\widehat{L}_{22}||\widehat{B}_{22}||\widehat{L}_{22}^T| + |\widehat{A}_{r+1} - \widehat{L}_{22}\widehat{B}_{22}\widehat{L}_{22}^T|. \tag{4.25b}$$

To bound the elements of G_{22}, we use Lemma 4.3 to give

$$|g_{ij}| \leq \frac{7nu}{1-nu}\sum_{k=1}^{r}|\widehat{l}_{ik}\widehat{b}_{kk}\widehat{l}_{jk}| + \frac{3(r+1)u}{1-3(r+1)u}\left(|\widehat{a}_{ij}^{(r+1)}| + \sum_{k=1}^{r}\frac{|\widehat{a}_{ik}^{(k)}\widehat{a}_{jk}^{(k)}|}{|\widehat{a}_{kk}^{(k)}|}\right), \quad i, j \geq r+1.$$

It follows from this bound and (4.19) that

$$|g_{ij}| \leq \frac{7nu}{1-3nu}\left(|\widehat{a}_{ij}^{(r+1)}| + (2+3u)\sum_{k=1}^{r}\frac{|\widehat{a}_{ik}^{(k)}\widehat{a}_{jk}^{(k)}|}{|\widehat{a}_{kk}^{(k)}|}\right), \quad i, j \geq r+1.$$

Lemma 4.2 can now be used to give

$$|g_{ij}| \leq \epsilon_3^g|a_{ij}| + \frac{1}{t}\epsilon_4^g \max_{1\leq q\leq r}|a_{iq}|, \quad i, j \geq r+1, \tag{4.26}$$

where

$$\epsilon_3^g = \frac{7nu}{1-3nu}(1+7u)^r \quad \text{and} \quad \epsilon_4^g = ((1+\omega_{r+1})^r - 1 + r\omega_r(2+3u)(1+\omega_r)^r)t\epsilon_3^g.$$

To bound the elements of H_{22} in (4.25b), we use the assumptions (4.21) to give

$$|h_{ij}| \leq \left(\frac{7nu\eta}{1-nu} + \delta\right)\max_{\substack{p\geq r+1\\ q\geq r+1}}|\widehat{a}_{pq}^{(r+1)}|, \quad i, j \geq r+1.$$

This bound and Lemma 4.2 now gives

$$|h_{ij}| \leq \epsilon_3^h \max_{\substack{p\geq r+1\\ q\geq r+1}}|a_{pq}| + \frac{1}{t}\epsilon_4^h \max_{\substack{p\geq r+1\\ q\leq r}}|a_{pq}|, \quad i, j \geq r+1, \tag{4.27}$$

where

$$\epsilon_3^h = \left(\frac{7nu\eta}{1-nu} + \delta\right)(1+7u)^r \quad \text{and} \quad \epsilon_4^h = ((1+\omega_{r+1})^r - 1)t\epsilon_3^h.$$

A combination of (4.26) and (4.27) gives

$$|e_{ij}| \leq \epsilon_3 \max_{\substack{p\geq r+1\\ q\geq r+1}}|a_{pq}| + \frac{1}{t}\epsilon_4 \max_{\substack{p\geq r+1\\ q\leq r}}|a_{pq}|, \quad i, j \geq r+1,$$

where

$$\epsilon_3 = \left(\frac{7nu(\eta+1)}{1-3nu} + \delta\right)(1+7u)^r$$

and

$$\epsilon_4 = (2((1+\omega_{r+1})^r - 1) + r\omega_r(2+3u)(1+\omega_r)^r)\, t\epsilon_3. \qquad \square$$

If (i) t is large; (ii) $\max\{3n, 8\}u \ll 1$; (iii) δ is small; and (iv) η is not too large, then the ϵ_i, $i = 1, \ldots, 4$ of Theorem 4.4 are "small", i.e., of the order of u. In this situation, we obtain $\epsilon_1 \approx 14nu$, $\epsilon_2 \approx 14nu$, $\epsilon_3 \approx 7nu(\eta+1) + \delta$ and $\epsilon_4 \approx 4r(7nu(\eta+1)+\delta)$. This theorem gives *componentwise* bounds on E. Corollary 4.5 below provides a straightforward conversion to norms that conform to Theorem 3.1.

COROLLARY 4.5. *The matrix E of Theorem* 4.4 *satisfies*

(i) $|\operatorname{diag}(E_{11})| \le \epsilon |\operatorname{diag}(A_{11})|$;
(ii) $\|\operatorname{offdiag}(E_{11})\| \le \epsilon \|\operatorname{offdiag}(A_{11})\|$;
(iii) $\|E_{21}\| \le \epsilon \|A_{21}\|$, $\|E_{12}\| \le \epsilon \|A_{21}\|$;
(iv) $\|E_{22}\| \le \epsilon(\|A_{22}\| + \frac{1}{t}\|A_{21}\|)$,

where

$$\epsilon = \max\{\epsilon_1, \epsilon_2 r, \epsilon_2\sqrt{r(n-r)}, \epsilon_3(n-r), \epsilon_4(n-r)\}.$$

Proof. For any $m \times n$ matrix C, it holds that

$$\frac{1}{\sqrt{mn}}\|C\| \le \max_{i,j}|c_{ij}| \le \|C\|$$

(see, e.g., Golub and Van Loan [8, p. 57]). Repeated use of these inequalities on the partitioned submatrices of E and A gives the result. $\square$

We have shown that if ϵ_1, ϵ_2, ϵ_3, and ϵ_4 of Theorem 4.4 are small, then the relative perturbation ϵ of Theorem 3.1 is also small, thereby justifying the form of the perturbations used in Theorem 3.1.

5. Conclusions. We have considered some numerical issues that arise when solving the nonlinear programming problem by minimizing a sequence of logarithmic barrier functions. A class of Newton barrier methods has been proposed that requires a symmetric indefinite system of linear equations to be solved at each iteration. These symmetric systems are ill-conditioned, but the ill-conditioning is caused by some diagonal elements becoming large in magnitude. The numerical implications of diagonal ill-conditioning were first considered by Ponceleón [21], who discussed the accuracy of the Newton barrier equations for problems in standard form. Ponceleón has shown that the sensitivity of the solution under certain structured perturbations is independent of the large diagonals.

It has been shown that diagonal ill-conditioning can be exploited without the need to add slack variables. A diagonally ill-conditioned matrix is an example of a *strictly t-diagonally dominant* matrix, where the definition of a t-diagonally dominance generalizes the idea of diagonal dominance to the situation where the diagonals of a matrix are substantially larger in magnitude than the off-diagonals. A perturbation analysis has been presented that describes the sensitivity of t-diagonally dominant systems under a class of structured perturbations that includes those of Ponceleón.

Finally, we have given a rounding-error analysis of the symmetric indefinite factorization when applied to t-diagonally dominant systems. This analysis indicates that the class of perturbations used in the sensitivity analysis is representative of the errors made during the numerical solution of barrier systems.

Acknowledgments. We are grateful to Margaret Wright for making several suggestions that improved the presentation. We also thank the two referees for their helpful and perceptive comments. In particular, we thank one referee for suggesting the scaling argument that simplified the informal analysis of §2.5.

REFERENCES

[1] J. R. BUNCH, *Partial pivoting strategies for symmetric matrices*, SIAM J. Numer. Anal., 11 (1974), pp. 521–528.

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

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

[4] J. R. BUNCH AND B. N. PARLETT, *Direct methods for solving symmetric indefinite systems of linear equations*, SIAM J. Numer. Anal., 8 (1971), pp. 639–655.

[5] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley and Sons, Inc., New York, London, Sydney, Toronto, 1968.

[6] K. R. FRISCH, *The logarithmic potential method of convex programming*, Memo May 13, University Institute of Economics, Oslo, Norway, 1955.

[7] P. E. GILL, W. MURRAY, D. B. PONCELEÓN, AND M. A. SAUNDERS, *Solving reduced* KKT *systems in barrier methods for linear and quadratic programming*, Report SOL 91-7, Department of Operations Research, Stanford University, Stanford, CA, 1991.

[8] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, second ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

[9] C. C. GONZAGA, *Path-following methods for linear programming*, SIAM Rev., 34 (1992), pp. 167–224.

[10] N. I. M. GOULD, *On the accurate determination of search directions for simple differentiable penalty functions*, IMA J. Numer. Anal., 6 (1986), pp. 357–372.

[11] N. J. HIGHAM, *Analysis of the Cholesky decomposition of a semi-definite matrix*, in Reliable Numerical Computation, M. G. Cox and S. Hammarling, eds., Oxford University Press, Oxford, 1990, pp. 161–185.

[12] ———, *Accuracy and Stability of Numerical Algorithms (provisional title)*, manuscript, 1995.

[13] M. KOJIMA, S. MIZUNO, AND A. YOSHISE, *A primal-dual interior point algorithm for linear programming*, in Progress in Mathematical Programming: Interior Point and Related Methods, N. Megiddo, ed., Springer-Verlag, New York, 1989, pp. 29–47.

[14] 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.

[15] G. P. MCCORMICK, *The superlinear convergence of a nonlinear primal-dual algorithm*, Report T-550/91, Department of Operations Research, The George Washington University, Washington, DC, 1991.

[16] N. MEGIDDO, *Pathways to the optimal set in linear programming*, in Progress in Mathematical Programming: Interior Point and Related Methods, N. Megiddo, ed., Springer-Verlag, New York, 1989, pp. 131–158.

[17] S. MEHROTRA, *On the implementation of a primal-dual interior point method*, SIAM J. Optim., 2 (1992), pp. 575–601.

[18] 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.

[19] W. MURRAY, *Analytical expressions for the eigenvalues and eigenvectors of the Hessian matrices of barrier and penalty functions*, J. Optim. Theory Appl., 7 (1971), pp. 189–196.

[20] W. MURRAY AND M. H. WRIGHT, *Line search procedures for the logarithmic barrier function*, SIAM J. Optim., 4 (1994), pp. 229–246.

[21] D. B. PONCELEÓN, *Barrier methods for large-scale quadratic programming*, Report SOL 91-2, Department of Operations Research, Stanford University, Stanford, CA, 1991.

[22] G. W. STEWART AND J. SUN, *Matrix Perturbation Theory*, Academic Press, New York, 1990.
[23] J. STOER AND R. BULIRSCH, *Introduction to Numerical Analysis*, second ed., Springer-Verlag, New York, Heidelberg, Berlin, 1993.
[24] M. H. WRIGHT, *Numerical methods for nonlinearly constrained optimization*, Ph.D. thesis, Department of Computer Science, Stanford University, Stanford, CA, 1976.
[25] ———, *Interior methods for constrained optimization*, in Acta Numerica 1992, A. Iserles, ed., Cambridge University Press, New York, 1992, pp. 341–407.
[26] ———, *Some properties of the Hessian of the logarithmic barrier function*, Math. Programming, 67 (1994), pp. 265–295.
[27] S. WRIGHT, *Stability of linear equations solvers in interior-point methods*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1287–1307.
[28] Y. YE, *Line searches in potential reduction algorithm for linear programming*, manuscript, 1989.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 1, pp. 212–221, January 1996

012

NUMERICAL METHODS FOR NEARLY SINGULAR CONSTRAINED MATRIX SYLVESTER EQUATIONS*

ALI R. GHAVIMI† AND ALAN J. LAUB†

Abstract. A recently published result describes a numerical procedure for solving a matrix Sylvester equation that is subject to certain constraints. It is quite possible that this Sylvester equation, or another intermediate one in the solution process, is nearly singular. As a result, certain computed parameters can have unexpectedly large norms and be very inaccurate. This paper incorporates an implicit deflation method for nearly singular matrix Sylvester equations to implement a reliable version of the published algorithm.

Key words. matrix Sylvester equations, nearly singular equations, implicit deflation, reduced-order observers

AMS subject classifications. 15A12, 65F05, 65F30, 65F35, 93B40

1. Introduction.

In this paper, an effective numerical procedure is presented for solving certain constrained matrix Sylvester equations. These equations frequently arise in control theory in connection with the design of reduced-order observers for linear time-invariant (LTI) systems that achieve precise loop transfer recovery when possible [1], [12]. Another application of constrained Sylvester-like equations is the design of minimum norm state feedback matrices via a pole placement method as in, for example, [10].

In finite precision floating-point arithmetic, it is quite possible that the corresponding Sylvester equations are nearly singular [4], [6]. Consequently, solutions of these matrix equations computed using general-purpose algorithms are usually of large norm and generally inaccurate. Such ill conditioning is normally an unknown function of the data and cannot be predicted in advance. Therefore, countermeasures should be taken to overcome this undesirable effect.

This paper modifies an existing method for solving constrained matrix equations in [1] resulting in a more reliable version of the published algorithm. The following describes the problem in detail. Let

$$\dot{x} = Ax + Bu, \tag{1}$$

$$y = Cx \tag{2}$$

denote a state model of an LTI system with n states, m inputs, and p outputs, that is, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, and $C \in \mathbb{R}^{p\times n}$. Under the usual observability condition, it is possible to design a reduced-order state observer $z(t) \in \mathbb{R}^{n-p}$ for $x(t)$ to satisfy

$$\dot{z} = Fz + (TB)u + Ly, \tag{3}$$

where $F \in \mathbb{R}^{(n-p)\times(n-p)}$ is stable and $L \in \mathbb{R}^{(n-p)\times p}$ and $T \in \mathbb{R}^{(n-p)\times n}$ are design parameters. It can be shown that if $TB = 0$, the resulting observer-based state feedback control has the same robustness as that of the direct state feedback system

* Received by the editors July 29, 1994; accepted for publication (in revised form) by K. Sigmon April 3, 1995. This research was supported by Air Force Office of Scientific Research grant F49620-94-1-0104DEF and by National Science Foundation grant ECS-9120643.

† Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106-9560 (laub@ece.ucsb.edu). Correspondence concerning this paper should be addressed to Dr. Laub.

[12]. In this respect, precise loop transfer recovery can be characterized [1], [12] by a constrained matrix Sylvester equation of the form

$$TA - FT = LC, \tag{4}$$

$$TB = 0, \tag{5}$$

$$\text{and } \begin{bmatrix} T \\ C \end{bmatrix} \text{ is full rank.} \tag{6}$$

In [1], existence conditions, along with a newly developed algorithm, are presented for the solution to (4)–(6) under the assumptions that $n > p > m$ and $\operatorname{rank}(CB) = \operatorname{rank}(B) = m$. The algorithm is as follows.

1.1. Algorithm for constrained matrix Sylvester equation (Barlow et al.)**.** The following algorithm is taken from [1] and finds a solution of (4)–(6).

1. Denote a QR factorization of B by

$$B = W \begin{bmatrix} S \\ 0 \end{bmatrix},$$

where $S \in \mathbb{R}^{m\times m}$ is full rank and upper triangular, $W = [W_1 \quad W_2] \in \mathbb{R}^{n\times n}$ is orthogonal, and $W_1 \in \mathbb{R}^{n\times m}$, $W_2 \in \mathbb{R}^{n\times(n-m)}$.

2. Set

$$\begin{aligned} A_1 &= W_2^T A W_1 \in \mathbb{R}^{(n-m)\times m}, \\ A_2 &= W_2^T A W_2 \in \mathbb{R}^{(n-m)\times(n-m)}, \\ C_1 &= C W_1 \in \mathbb{R}^{p\times m}, \\ C_2 &= C W_2 \in \mathbb{R}^{p\times(n-m)}. \end{aligned}$$

3. Denote a QR factorization of C_1 by

$$C_1 = [Q_1 \quad Q_2] \begin{bmatrix} R \\ 0 \end{bmatrix},$$

where $Q_1 \in \mathbb{R}^{p\times m}$, $Q_2 \in \mathbb{R}^{p\times(p-m)}$, and $R \in \mathbb{R}^{m\times m}$ is full rank and upper triangular.

4. Define $E \in \mathbb{R}^{p\times(n-m)}$ by

$$E = \begin{bmatrix} E_1 \\ E_2 \end{bmatrix} = Q^T C_2,$$

where $E_1 \in \mathbb{R}^{m\times(n-m)}$ and $E_2 \in \mathbb{R}^{(p-m)\times(n-m)}$.

5. Let $\hat{A} = A_2 - A_1 R^{-1} E_1 \in \mathbb{R}^{(n-m)\times(n-m)}$. Solve the reduced Sylvester equation

$$Z\hat{A} - FZ = L_2 E_2, \tag{7}$$

where the solution $Z \in \mathbb{R}^{(n-p)\times(n-m)}$ and $L_2 \in \mathbb{R}^{(n-p)\times(p-m)}$ is chosen randomly. Note that $J = A_1 R^{-1}$ is computed by solving the triangular system $JR = A_1$.

6. Set $L_1 = ZJ \in \mathbb{R}^{(n-p)\times m}$, $L = [L_1 \quad L_2]Q^T$, and $T = ZW_2^T$.

Steps 1–4 and 6 of the above algorithm depend on certain matrix factorizations and multiplications. As long as R is not too ill conditioned with respect to inversion, these operations can be performed reliably using stable algorithms and thus pose no particular difficulty. Note from Steps 1–3 that the condition of R is intimately related to rank properties of C and B. Obviously, R is ill conditioned if C is nearly rank deficient. In the context of observer design such rank deficiency can result from either poor scaling of the state variables or near redundancy of measurements. These issues are usually resolved prior to any attempt to do an observer or a controller design. It is generally the case that C is "robustly" full rank. Similar comments apply to the regularity assumption that CB be of full rank and to the assumption that B is of full rank.

Step 5, however, involves solving the Sylvester equation (7) which may be nearly singular. Unfortunately, there is no straightforward relationship between the condition of a Sylvester operator and the condition of its defining coefficient matrices, although some relationships are known. For example, it is shown in [4] that a Sylvester operator is nearly singular when both coefficient matrices are ill conditioned with respect to inversion. In this respect, it is thus worth noting that this sufficient condition often obtains in the case of large-scale dynamical systems whose matrix sizes often lead to an undesired increase in ill conditioning.

When (7) is ill conditioned, the computed solution Z and, as a consequence, L and T, are generally of large norm and numerically inaccurate. The problem is worsened if, by the random choice of L_2, the right-hand side L_2E_2 is rich in the direction of the approximate (numerical) left null space [3], [6], [7] of the Sylvester equation (7). It may be thought, at first glance, that large-norm computed solutions T and L of (4) can be scaled properly so that (4) and (6) are still satisfied. However, note that the near singularity of the system in (7) is intimately related to the overall ill conditioning of the system in (4). Furthermore, the conditioning of either (4) or (7) is scale-invariant with respect to any choice of the pair $(T,\ L)$ or $(Z,\ L_2)$, respectively. And finally, the conditioning of the matrix in (6) may even be worsened if T is scaled by a large amount. This problem can be easily and accurately resolved by choosing L_2 as a random matrix such that L_2E_2 is restricted to the approximate (numerical) range space of the Sylvester operator [6], i.e., a consistent set of equations. Furthermore, the resulting equation can be solved by the implicit deflation method for nearly singular Sylvester equations in [4].

The rest of this paper is organized as follows. Section 2 outlines the computation of approximate null vectors of nearly singular Sylvester equations. Section 3 provides a summary of the implicit deflation algorithm for ill-conditioned Sylvester equations. Section 4 describes the contribution of this paper for solving the ill-conditioned set of constrained matrix Sylvester equations (4)–(6). Section 5 presents numerical examples to illustrate the results.

2. Simultaneous inverse iteration algorithm. Suppose that the Sylvester equation

$$AX + XB = C \tag{8}$$

with $X \in \mathbb{R}^{n\times m}$ is nearly rank-ν deficient. That is, the Kronecker form of the associated Sylvester operator

$$S = I_m \otimes A + B^T \otimes I_n \in \mathbb{R}^{mn\times mn} \tag{9}$$

has ν small singular values of the order of the relative machine precision. Here I_k denotes the $k \times k$ identity matrix and $\otimes$ denotes the Kronecker product [9]. The following algorithm employs inverse iteration, together with an acceleration scheme, to construct approximations of the left and right numerical null vectors of S by the orthogonal columns of $U \in \mathbb{R}^{mn \times \nu}$ and $V \in \mathbb{R}^{mn \times \nu}$, respectively, one vector at a time.

- For nullity $\nu \geq 1$, choose $U_0 \in \mathbb{R}^{mn \times \nu}$ with orthonormal columns.
- Loop until convergence
  1. Solve $S\tilde{V} = U_0$ for $\tilde{V}$
  2. $\tilde{V} = VT$ (QL factorization)
  3. Solve $S^T\tilde{U} = V$ for $\tilde{U}$
  4. $\tilde{U} = UR$ (QL factorization)
  5. $G = U^T S V$
  6. $G = M\Lambda N^T$ (SVD)
  7. $U \longleftarrow UM$ and $V \longleftarrow VN$
  8. If $\lambda_j >$ `tol`, a user-specified tolerance, for $1 \leq j \leq \nu$,
     a. $U \longleftarrow U(:, j+1:\nu)$ (truncate the first j columns)
     b. $V \longleftarrow V(:, j+1:\nu)$ (truncate the first j columns)
     c. $\nu = j$
     d. Stop

     Else
     e. $\nu \longleftarrow \nu + 1$
     f. $U_0 = [U \;\; \tilde{U}]$ such that $U_0^T U_0 = I_\nu$
     g. Return to Innerstep 1.

Innersteps 1 and 3 of the above algorithm involve solving linear systems with coefficient matrices S and S^T, respectively. However, these systems are never formed explicitly. Instead, equivalent Sylvester equations are solved with appropriate choice of the right-hand-side vectors. The QL factorization in the above algorithm is of the form $Z = QL$, where $Q \in \mathbb{R}^{n \times m}$ has orthonormal columns and $L \in \mathbb{R}^{m \times m}$ is lower triangular with positive diagonal elements. Finally, note that in our calculations, without any reason to do otherwise, we generally achieve satisfactory results by choosing `tol` to be somewhat larger than the relative machine precision. For the examples in §5, we used `tol` $= 10^{-14}$. Further details concerning convergence, as well as certain practical aspects, have been studied extensively in [6].

The algorithm in §3 describes deflated solutions of nearly singular Sylvester equations.

3. Implicit deflation method for nearly singular Sylvester equations. Suppose that (8) is nearly singular and the associated Sylvester operator S in (9) has ν small singular values. Assume further that the numerical or approximate left and right null vectors of S are denoted by the orthonormal columns of $\tilde{U} = (\mathrm{vec}U_1, \ldots, \mathrm{vec}U_\nu), \tilde{V} = (\mathrm{vec}V_1, \ldots, \mathrm{vec}V_\nu) \in \mathbb{R}^{mn \times \nu}$, respectively, and have been determined by the inverse iteration algorithm in [6]. Note that the vec of an $m \times n$ matrix M is the mn-vector formed by successively stacking the n columns of M on top of each other [9]. Let Q be an orthogonal projection denoted by $Q = I_{mn} - \tilde{V}\tilde{V}^T$. Then the SVD-based deflated solution, i.e., the minimum norm least-squares solution of (8), is obtained as follows.

- Start with any X_0 that satisfies $Q\mathrm{vec}(X_0) = \mathrm{vec}(X_0)$.
- Set $k = 0$.
- Do while improvement

1. $R_k = C - AX_k - X_kB$
2. $R_k \longleftarrow R_k - \sum_{j=1}^{\nu} \mathrm{tr}(U_j^T R_k)U_j$ ($\mathrm{tr}(\cdot)$ denotes trace of a matrix)
3. Solve $AD_k + D_kB = R_k$ for D_k
4. $X_{k+1} = X_k + D_k - \sum_{j=1}^{\nu} \mathrm{tr}(V_j^T D_k)V_j$
5. Set $k \longleftarrow k+1$.

A choice of the initial condition X_0 is obtained by setting $\mathrm{vec}(X_0) = Q\mathrm{vec}(Z)$, where $Z \in \mathbb{R}^{n\times m}$ is a random matrix. Since Q is a projection, then the constraint $Q\mathrm{vec}(X_0) = \mathrm{vec}(X_0)$ is satisfied. Because of the large size of Q, however, the following manipulations are needed to avoid explicit formation of Q. Note that $Q\mathrm{vec}(Z)$ can be written as

$$Q\mathrm{vec}(Z) = Z - \tilde{V}\tilde{V}^T\mathrm{vec}Z = Z - \tilde{V}\begin{bmatrix} (\mathrm{vec}V_1)^T\mathrm{vec}(Z) \\ (\mathrm{vec}V_2)^T\mathrm{vec}(Z) \\ \vdots \\ (\mathrm{vec}V_\nu)^T\mathrm{vec}(Z) \end{bmatrix} = \mathrm{vec}\left(Z - \sum_{j=1}^{\nu}\mathrm{tr}(V_j^T Z)V_j\right).$$

Therefore, X_0 can be formed by $X_0 = Z - \sum_{j=1}^{\nu}\mathrm{tr}(V_j^T Z)V_j$ for any arbitrary choice of Z.

Conditions for the convergence of the above algorithm have been examined in [4]. In particular, it is shown that convergence is guaranteed if the computed singular vectors have sufficient accuracy. In practice, this poses no further difficulty since the singular subspaces obtained by the inverse iteration algorithm are quite accurate. This particular issue has been studied extensively in [6]. As far as the complexity of the algorithm is concerned, a simple operation count reveals that the innersteps 1, 2, and 4 require only a small fraction of the floating-point operations needed for the Sylvester solve in the innerstep 3. The Sylvester equations can be solved using the Bartels–Stewart algorithm [2] or the Hessenberg–Schur method [8]. These methods require certain relatively expensive initial factorizations (Schur or Hessenberg) of the coefficient matrices A and B. Therefore, overall computational savings can be realized since these factorizations need not be computed more than once.

4. Appropriate choice of L_2 in (7). As was noted earlier, a critical step in the algorithm of [1] involves solving the Sylvester equation in (7),

$$Z\hat{A} - FZ = L_2E_2,$$

which is repeated here for convenience. This equation can be ill conditioned for a number of reasons. For example, if $\hat{A}$ and F have any nearly equal eigenvalues, then $S = \hat{A}^T \otimes I - I \otimes F$ is nearly singular, or equivalently, (7) is ill conditioned. This can be seen as follows. Let $(\mu_j,\ y_j)$ and $(\nu_k,\ x_k)$ denote left and right eigenvalue/eigenvector pairs for $\hat{A}$ and F, respectively. Specifically, suppose $\hat{A}^T y_j = \mu_j y_j$ and $Fx_k = \nu_k x_k$. Then $p_i = y_j \otimes x_k = \mathrm{vec}(x_k y_j^T)$ is an eigenvector of S associated with the eigenvalue $\lambda_i = \mu_j - \nu_k$. This can be seen easily from $S(y_j \otimes x_k) = \hat{A}^T y_j \otimes x_k - y_j \otimes Fx_k = (\mu_j - \nu_k)(y_j \otimes x_k)$. Moreover, if y_j and x_k are of unit norm,

$$\|S(y_j \otimes x_k)\|_2 = \|(\mu_j - \nu_k)(y_j \otimes x_k)\|_2 = |\mu_j - \nu_k|(y_j^T y_j \otimes x_k^T x_k)^{1/2} = |\mu_j - \nu_k|.$$

Thus when $|\mu_j - \nu_k|$ is small, $P_i = x_k y_j^T$ can be considered as an approximate null vector of (7). In other words, the numerical or approximate null space of S is nonzero, or equivalently, S is nearly singular. Therefore, in this sense, the nearness of the

eigenvalues of the matrices $\hat{A}$ and F provides a priori information on the near singularity of the Sylvester operator S or the ill conditioning of the Sylvester equation in (7). However, the reverse implication is not true. That is, S may be nearly singular regardless of such an eigenvalue constraint.

The near singularity of S can best be defined in terms of its singular value decomposition. Denote an SVD of $S = \hat{A}^T \otimes I - I \otimes F$ by $S = U\Sigma V^T$, where

$$U = [\text{vec}(U_1),\ \text{vec}(U_2),\ \ldots, \text{vec}(U_k)] \in \mathbb{R}^{k\times k},$$
$$V = [\text{vec}(V_1),\ \text{vec}(V_2),\ \ldots,\ \text{vec}(V_k)] \in \mathbb{R}^{k\times k}$$

are orthogonal, $\Sigma = \text{diag}(\sigma_1,\ \sigma_2,\ \ldots,\ \sigma_k)$, and $k = (n-m)(n-p)$. Moreover, the singular vectors U_i and V_i satisfy $S\text{vec}(V_i) = \sigma_i \text{vec}(U_i)$ and $S^T\text{vec}(U_i) = \sigma_i \text{vec}(V_i)$. Equivalently, these singular vector equations can be written in matrix terms as $AV_i + V_iB = \sigma_i U_i$ and $A^TU_i + U_iB^T = \sigma_i V_i$ where $U_i, V_i \in \mathbb{R}^{(n-p)\times(n-m)}$. It should be noted that there is no straightforward relationship between the singular vectors of S and those of the parameter matrices $\hat{A}$ and F.

The solution Z can be written in terms of the singular vectors as

$$\text{vec}(Z) = V\Sigma^{-1}U^T\text{vec}(L_2E_2)$$

or equivalently,

$$Z = \sum_{i=1}^{k} \frac{\text{tr}(U_i^T L_2 E_2)}{\sigma_i} V_i. \tag{10}$$

From the above expression for Z, it is seen that for small values of the σ_i's, i.e., when the (approximate) numerical null space of S is nonzero, the solution Z is greatly magnified in the direction of the corresponding V_i's. In other words, if $\text{vec}(L_2E_2)$ is rich in the direction of the approximate null vectors, a large norm solution should be expected.

In light of the above, a suitable choice for L_2 can then be established by choosing a random matrix L_2 such that $\text{vec}(L_2E_2)$ is orthogonal to the approximate left null space of S. This ensures that $\text{vec}(L_2E_2)$ is restricted to the approximate range space of S and in turn, Z is purged of the approximate null vectors. In this respect, the following provides a characterization of appropriate values of L_2. Suppose that (7) is nearly singular and its ν-dimensional left singular subspace associated with the small singular values is given by the columns of

$$\tilde{U} = (\text{vec}U_1,\ \text{vec}U_2,\ \ldots,\ \text{vec}U_\nu) \in \mathbb{R}^{(n-m)(n-p)\times\nu},$$

where $U_i \in \mathbb{R}^{(n-p)\times(n-m)}$ for $i = 1,\ \ldots,\ \nu$. It is then required that

$$\begin{aligned} 0 = \tilde{U}^T\text{vec}(L_2E_2) &= \tilde{U}^T\left(E_2^T \otimes I_{(m-p)}\right)\text{vec}(L_2) \\ &= \left[\left(E_2 \otimes I_{(m-p)}\right)\tilde{U}\right]^T \text{vec}(L_2) \\ &= H\text{vec}(L_2), \end{aligned}$$

where

$$H = \left[\text{vec}(U_1E_2^T),\ \ \text{vec}(U_2E_2^T), \ldots,\ \ \text{vec}(U_\nu E_2^T)\right]^T \in \mathbb{R}^{\nu\times(n-p)(n-m)}.$$

Therefore, $\text{vec}(L_2) \in \mathcal{N}(H)$ (the null space of H) from which a choice for L_2 is easily obtained.

Next, (7) is solved using the implicit deflation algorithm for nearly singular Sylvester equations. This step is needed because L_2 or the right-hand side of (7) can only be computed approximately. Therefore, $\text{vec}(L_2E_2)$ may not be exactly consistent with the numerical range space of the corresponding Sylvester operator. As a result, slight deviation of a proper right-hand side, together with roundoff errors, may magnify the solution in the direction of null vectors and in turn degrade the accuracy of the final result. Instead, the solution generated by the implicit deflation algorithm provides an accurate representation of the actual solution for any proper choice of L_2. The following examples illustrate the underlying idea.

5. Numerical results. In this section, two numerical examples are presented to demonstrate further the necessity of modifying the algorithm in [1]. These examples are intended for the sake of illustration. However, examples of ill-conditioned observer design can be found in a variety of problems in control theory, especially in connection with the control of large-scale dynamical systems. It is well known that an increase in problem size often results in a concomitant increase in ill conditioning. The work in [11] provides a simple but dramatic example of this phenomenon. This is aside from the fact that such an undesirable effect is also likely to occur even for problems with moderate sizes when performing operations in floating-point arithmetic with finite precision. All calculations described below were performed using Matlab 4.2b for Windows implemented on a Texas Instruments 486 DX2 laptop computer.

Example 1. Consider designing a reduced-order observer that achieves precise loop transfer recovery for the linear system (1)–(2) with

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad \text{and} \quad C = \begin{bmatrix} 2 & 0 & 0 & 0 & 2 \\ 0 & 2 & 0 & 2 & 0 \end{bmatrix}.$$

Assume further that the desired eigenvalues of the observer matrix F are $\{-\frac{1}{\sqrt{2}} \pm j\frac{1}{\sqrt{2}}, -5\}$, and the complex conjugate pair is subject to a random perturbation of the order of 10^{-12} (here $j = \sqrt{-1}$). Without loss of generality, F can be chosen as

$$F = \begin{bmatrix} -\frac{1}{\sqrt{2}} + \alpha & \frac{1}{\sqrt{2}} + \beta & 1 \\ -\frac{1}{\sqrt{2}} + \gamma & -\frac{1}{\sqrt{2}} + \delta & 1 \\ 0 & 0 & -5 \end{bmatrix}$$

with α, β, γ, and δ of the order of 10^{-12}. The particular values of the perturbations are insignificant in this example. However, in the interest of completeness, we give the specific values used, rounded to five significant figures, as $\alpha = 2.1896e - 12$, $\beta = 4.7045e - 13$, $\gamma = 6.7886e - 12$, and $\delta = 6.7930e - 12$.

Note. All numbers quoted in the remainder of this example and in Example 2 are given to five significant figures only. The full 16 significant figures are available from the authors.

It is clear in this case that A and F have no eigenvalues in common, and all the existence conditions of the algorithm in [1] are satisfied. Furthermore, the magnitudes of the eigenvalues of the Sylvester operator associated with (4)

$$S = A^T \otimes I_3 - I_5 \otimes F$$

are the same as those of F. Therefore, the eigenvalues are well separated from the origin. Moreover, the singular values of S range from 0.27586 to 6.0641. Thus, the condition number $\kappa_2(S)$ is approximately 21.982 and the Sylvester operator is well conditioned with respect to inversion. Therefore, it is expected that the solutions T and L have reasonable norms and accurately solve (4). For this problem, direct application of the method in [1] gives the following results:

$$T = \begin{bmatrix} -4.8557e+10 & -4.7227e+10 & 1.1535e+11 & -1.1590e+11 & 0 \\ 1.1535e+11 & -1.1590e+11 & 4.8557e+10 & 4.7227e+10 & 0 \\ 2.8160e-03 & 1.2997e-02 & -2.5994e-03 & 1.4080e-02 & 0 \end{bmatrix},$$

$$L = \begin{bmatrix} -5.7948e+10 & 5.0932e-01 \\ 2.3614e+10 & 8.1481e-01 \\ 7.0399e-03 & 3.3900e-02 \end{bmatrix}, \quad TB = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

where $\|T\|_F \approx 2.5030e+11$ and $\|L\|_F \approx 6.2575e+10$. The residual norm of (4) is $\|LC - TA + FT\|_F \approx 1.3059e{-}04$. It is worth mentioning that the associated relative residual is of the order of the relative machine precision; however, this quantity is not a good measure of the solution accuracy [5]. In particular, this quantity is meaningless if the Sylvester equation is ill conditioned. In this case, (6) is full rank and its condition number is approximately $1.4953e+14$.

The above results are somewhat unexpected since at first glance the problem appears to be well conditioned by construction. However, further analysis indicates that the method is certain to produce such results since the Sylvester operator associated with (7) is nearly singular. The reason for such ill conditioning arises from the fact that $\hat{A}$ and F in (7) have two nearly equal eigenvalues, namely, two of the eigenvalues of F are $O(10^{-12})$ perturbations of the stable eigenvalues of $\hat{A}$, i.e., $-\frac{1}{\sqrt{2}} \pm j\frac{1}{\sqrt{2}}$. This shows that a priori conditions on the eigenstructures of A and F in (4) and the conditioning of S are not sufficient to guarantee stability of the proposed method. As was stated in §1, a simple scaling of T or L is not sufficient to satisfy the required conditions, e.g., a well-conditioned matrix in (6). It is easy to verify that if T is scaled by its norm, then the condition number of the resulting matrix in (6), $[\frac{T^T}{\|T\|_F}, \ C^T]^T$, is about $6.0768e+14$.

While in this example the near singularity occurs as a result of nearly zero eigenvalues of the Sylvester operator associated with (7), it is generally an unknown function of the data and cannot be predicted in advance. As was suggested previously, this problem can be solved easily by choosing L_2 so that $\mathrm{vec}(L_2E_2)$ is orthogonal to the left null vectors of (7). The solutions of the resulting equation are then obtained by applying the implicit algorithm to the Sylvester equation (7). The results are as follows:

$$T = \begin{bmatrix} -1.5039e-01 & 4.2592e-01 & 4.2206e-02 & 1.3619e-01 & 0 \\ 2.1561e-01 & 2.4361e-01 & 1.2891e-01 & 4.4111e-01 & 0 \\ -3.9499e-01 & -1.8230e+00 & 3.6461e-01 & -1.9750e+00 & 0 \end{bmatrix},$$

$$L = \begin{bmatrix} 6.8094e-02 & 9.0078e-01 \\ 2.2055e-01 & 1.2560e+00 \\ -9.8748e-01 & -4.7551e+00 \end{bmatrix}, \quad TB = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

and $\|LC - TA + FT\|_F = 8.4754e{-}15$. The matrix in (6) is full rank and its condition number is about 47.366; hence, it is well conditioned with respect to inversion. With $\|T\|_F = 2.8380$, $\|L\|_F = 5.1018$, and the small absolute residual norm, it is clear that the latter results are both more accurate and efficient relative to those obtained directly by the algorithm in [1].

The above example also suggests that it may even be possible to arrive at a solution when A and F have several almost equal eigenvalues. That is, some of the poles of the observer system F are approximately the same as stable poles of the system matrix A. In this case, (4) and, consequently, (7) are nearly singular. Therefore, direct application of the algorithm in [1] produces large norm solutions; as a result, even if (5) is satisfied, the matrix in (6) is nearly rank deficient. This problem can be resolved once again by the approach proposed earlier and the following example demonstrates further the effectiveness of the method in this respect.

Example 2. Consider the problem in the previous example with

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & -\sqrt{2} \end{bmatrix}$$

and B, C, and F as before. The eigenvalues of A are

$$\Lambda(A) = \left\{0,\ 0,\ 0,\ -\frac{1}{\sqrt{2}} \pm j\frac{1}{\sqrt{2}}\right\}.$$

Therefore, A and F have two almost equal eigenvalues and hence S is nearly singular. Using the method of [1] yields

$$T = \begin{bmatrix} -5.2329e+10 & -6.2465e+09 & 6.1163e+10 & -8.0251e+10 & 0 \\ 6.1163e+10 & -8.0251e+10 & 5.2329e+10 & 6.2465e+09 & 0 \\ 4.7760e-02 & 2.2043e-01 & -4.4086e-02 & 2.3880e-01 & 0 \end{bmatrix},$$

$$L = \begin{bmatrix} -4.0125e+10 & 4.5799e-02 \\ 3.1233e+09 & 4.5378e-01 \\ 1.1940e-01 & 5.7495e-01 \end{bmatrix}, \quad TB = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

with $\|T\|_F \approx 1.6099e{+}11$, $\|L\|_F \approx 4.0247e{+}10$, and $\|LC - TA + FT\|_F = 8.7364e{-}05$. The matrix in (6) is full rank, and its condition number is about $5.6985e+12$. Therefore, it is ill conditioned with respect to inversion. Furthermore, if T is scaled to have unit Frobenius norm, the condition number of the resulting matrix in (6) is about $2.3045e+13$. By contrast, the use of the deflation method in the algorithm of [1] produces the same smaller norm T, L, and residual as in Example 1. (This is because of the nature of T and the fact that $\hat{A}$ and F are the same in the reduced Sylvester equation (7) and independent of the choice of L_2.)

Again, the condition number of the matrix in (6) is about 47.366; therefore, the full rank condition is satisfied. As before, a comparison between the norms of the solutions and the residuals suggests that the deflation method more accurately

solves the equations of interest. In this example S is nearly singular since A and F have two nearly equal eigenvalues. However, as has been noted in [6], S can be ill conditioned regardless of such an eigenvalue property. In any case, the proposed modified algorithm is capable of producing accurate and efficient results.

6. Conclusions. This paper modifies the algorithm in [1] to efficiently implement a solution of the constrained matrix Sylvester equation (4)–(6). The method incorporates the use of a newly developed implicit deflation algorithm for nearly singular matrix Sylvester equations to account for the case that (4)–(6) are ill conditioned. The need for this modification is compelling since a priori information on system parameters is not sufficient to guarantee the stability of the algorithm in [1]. Two numerical examples have been presented to illustrate the underlying idea.

REFERENCES

[1] J. B. Barlow, M. M. Monahemi, and D. P. O'Leary, *Constrained matrix Sylvester equations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1–9.

[2] R. H. Bartels and G. W. Stewart, *Algorithm* 432: *Solution of the matrix equation* $AX + XB = C$, Commun. ACM, 15 (1972), pp. 820–826.

[3] L. V. Foster, *Rank and null space calculations using matrix decomposition without column interchanges*, Linear Algebra Appl., 74 (1986), pp. 47–71.

[4] A. R. Ghavimi and A. J. Laub, *An implicit deflation method for ill-conditioned Sylvester and Lyapunov equations*, Internat. J. Control, 61 (1995), pp. 1119–1141.

[5] ———, *Backward error, sensitivity, and refinement of computed solutions of algebraic Riccati equations*, Numer. Linear Algebra Appl., 2 (1995), pp. 29–49.

[6] ———, *Computation of approximate null vectors of Sylvester and Lyapunov equations*, IEEE Trans. Automat. Control, 40 (1995), pp. 387–391.

[7] G. H. Golub and C. F. Van Loan, *Matrix Computations*, second edition, Johns Hopkins University Press, Baltimore, 1989.

[8] G. H. Golub, S. Nash, and C. F. Van Loan, *A Hessenberg–Schur method for the problem* $AX + XB = C$, IEEE Trans. Automat. Control, AC-24 (1979), pp. 909–913.

[9] A. Graham, *Kronecker Products and Matrix Calculus with Applications*, Wiley, New York, 1981.

[10] L. H. Keel, J. A. Fleming, and S. P. Bhattacharyya, *Minimum norm pole assignment via Sylvester's equation*, in Linear Algebra and Its Role in Systems Theory, pp. 265–272, Amer. Math. Soc., 1985, Vol. 47, Contemporary Math. Series.

[11] C. S. Kenney and A. J. Laub, *Controllability and stability radii for companion form systems*, Math. Control Signals Systems, 1 (1988), pp. 239–256.

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

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 2, pp. 223–247, April 1996

RELATIONS BETWEEN GALERKIN AND NORM-MINIMIZING ITERATIVE METHODS FOR SOLVING LINEAR SYSTEMS*

JANE CULLUM† AND ANNE GREENBAUM‡

Abstract. Several iterative methods for solving linear systems $Ax = b$ first construct a basis for a Krylov subspace and then use the basis vectors, together with the Hessenberg (or tridiagonal) matrix generated during that construction, to obtain an approximate solution to the linear system. To determine the approximate solution, it is necessary to solve either a linear system with the Hessenberg matrix as coefficient matrix or an extended Hessenberg least squares problem. In the first case, referred to as a *Galerkin* method, the residual is orthogonal to the Krylov subspace, whereas in the second case, referred to as a *norm-minimizing* method, the residual (or a related quantity) is minimized over the Krylov subspace. Examples of such pairs include the full orthogonalization method (FOM) (Arnoldi) and generalized minimal residual (GMRES) algorithms, the biconjugate gradient (BCG) and quasi-minimal residual (QMR) algorithms, and their symmetric equivalents, the Lanczos and minimal residual (MINRES) algorithms. A relationship between the solution of the linear system and that of the least squares problem is used to relate the residual norms in Galerkin processes to the norms of the quantities minimized in the corresponding norm-minimizing processes. It is shown that when the norm-minimizing process is converging rapidly, the residual norms in the corresponding Galerkin process exhibit similar behavior, whereas when the norm-minimizing process is converging very slowly, the residual norms in the corresponding Galerkin process are significantly larger. This is a generalization of the relationship established between Arnoldi and GMRES residual norms in P. N. Brown, *A theoretical comparison of the Arnoldi and GMRES algorithms*, SIAM J. Sci. Statist. Comput., 12, 1991, pp. 58–78. For MINRES and Lanczos, and for two nonsymmetric bidiagonalization procedures, we extend the arguments to incorporate the effects of finite precision arithmetic.

Key words. GMRES, Arnoldi, biconjugate gradients, QMR, iterative methods

AMS subject classifications. 65F10, 65F15

1. Introduction.

The Arnoldi algorithm [1] (also known as the full orthogonalization method or FOM [22]) and the generalized minimal residual (GMRES) algorithm [23] are two recently developed Krylov methods for solving nonsymmetric linear systems

$$Ax = b. \tag{1}$$

In Brown [3], a theoretical comparison of the two methods is presented. Brown exhibits connections between the singularity of the Hessenberg matrices generated in the Arnoldi algorithm and the stagnation of the corresponding iterates in the GMRES algorithm. From this he infers a relationship between the stagnation (the plateaus) observed in GMRES and near-singularity of these Hessenberg matrices. He also obtains relationships between the norms of the residuals generated by the Arnoldi algorithm and the norms of the residuals generated by the GMRES algorithm which, when combined with a relationship in [23], can be used to infer that if the iterates in both methods are well defined, then if one of the methods performs very well on a

* Received by the editors April 2, 1993; accepted for publication (in revised form) by R. Freund April 11, 1995.

† IBM T. J. Watson Research Center, Yorktown Heights, NY 10598. Part of this research was supported by National Science Foundation grant GER-9450081 while this author was visiting the Department of Computer Science, University of Maryland, College Park, MD.

‡ Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012 (greenbau@nyu.edu). Part of this research was performed while this author was visiting IBM T. J. Watson Research Center and was supported by National Science Foundation grant 25968-5375 and DOE contract DEFG0288ER25053.

particular problem the other one will also, and if one method performs very poorly, then so will the other one.

In this paper we obtain a slightly different formulation of these GMRES/Arnoldi residual norm relationships from which it is easier to see this interdependence. This formulation explains the general correspondence between peaks in the plot of residual norms in the Arnoldi algorithm and plateaus in the GMRES algorithm, as observed in [4] for a different pair of iterative methods. Moreover, we demonstrate, by numerical example, that plateaus (stagnation) need not be associated with near-singularity of the Hessenberg matrices. In fact the proof given in [3] connecting stagnation of the GMRES iterates to singularity of the Hessenberg matrices is not applicable to the nearly singular case.

The biconjugate gradient (BCG) algorithm [7] and the quasi-minimal residual (QMR) algorithm [8] are another pair of Krylov methods for solving linear systems. In [8], relationships are established between these two algorithms that are very similar to the relationships obtained in [3] for the Arnoldi and GMRES residuals. In this case, however, the appropriate quantities to compare are the BCG residuals and what we refer to as the QMR *quasi-residuals*, the vectors whose norms are actually minimized at each step of the QMR algorithm. From these relationships one can infer that if the norm of the QMR quasi-residual is greatly reduced at a step, then the norm of the BCG residual at that step is approximately equal to the norm of the QMR quasi-residual, whereas if the QMR quasi-residual norm remains nearly constant, the BCG residual norm is significantly greater than the QMR quasi-residual norm. The norms of the actual QMR residuals may be somewhat larger than those of the quasi-residuals (by as much as a factor $\sqrt{k+1}$ at step k [8]) or they may be somewhat smaller, but it is, in some sense, a happy accident if the actual residual norms turn out to be much smaller than the quasi-residual norms. No attempt is made to produce the sort of cancellation that is needed to make this happen and in practice it seldom occurs. Thus the relationship between BCG residuals and QMR quasi-residuals can be thought of as an approximate relationship between BCG residuals and QMR residuals. Roughly speaking, then, for a given problem these algorithms will either both perform well or both perform poorly.

The above statements assume exact arithmetic. For *real symmetric* problems, similar relationships are shown to hold in finite precision arithmetic as well. For real symmetric problems, and in exact arithmetic, both of these pairs of algorithms reduce to the MINRES [20] and Lanczos [17] algorithms. In this case, the Lanczos vectors are orthogonal and the quasi-residual norms are the same as the residual norms, so the relationship between BCG residual norms and QMR quasi-residual norms becomes a relationship between Lanczos residual norms and MINRES residual norms. In finite precision computations, the computed Lanczos vectors may be far from orthogonal, yet numerical experiments reported in [4] suggest that for certain types of problems these relationships hold to a close approximation, even after orthogonality of the Lanczos vectors is lost. The report [4] describes a series of numerical experiments using a pair of bidiagonalization procedures, denoted by SQMR and BLanczos, for solving nonsymmetric systems of equations. These procedures replace the original nonsymmetric problem by a larger symmetric problem, and then use MINRES and Lanczos on the associated symmetric iteration matrices.

We prove that for certain classes of real symmetric problems, the relationships between the residuals demonstrated for GMRES/Arnoldi in exact arithmetic hold approximately for the MINRES and Lanczos residuals even in finite precision. We then

prove that if a nonsymmetric problem (1) is well conditioned, then the real symmetric problems generated in the nonsymmetric bidiagonalization algorithm considered in [4] belong to this class. We therefore obtain the result that the residual relationship, as demonstrated numerically in [4], is theoretically valid in finite precision arithmetic for that pair of bidiagonalization algorithms. These proofs rely on the fact that the tridiagonal matrices generated by a finite precision Lanczos computation are the same as those that would be generated by the exact Lanczos algorithm applied to a certain larger matrix with nearby eigenvalues, and that components of the computed Lanczos vectors can be related to corresponding groups of components in the exact Lanczos vectors associated with this larger matrix [12].

The exact arithmetic results obtained in this paper are easy consequences of relations between linear system solutions and extended least squares solutions that have been established in a number of places. See, for example, [3, 8, 9, 15, 20, 28]. It is indeed surprising that the precise relation between Galerkin and norm-minimizing methods was not recognized and explicitly stated much earlier!

In §2, the theorem relating the solution of a linear system and an extended least squares problem is established. In §3 the Arnoldi and GMRES algorithms are described and the theorem of §2 is used to establish the relationship between Arnoldi and GMRES residual norms. In §4 the BCG and QMR algorithms are described and the analogous relationship between the BCG residual norms and the QMR quasi-residual norms is derived. In both sections, numerical examples are given to demonstrate the relationships.

In §5 we first define the real symmetric Lanczos and MINRES algorithms and then define the two nonsymmetric bidiagonalization algorithms considered in [4]. In §6 we focus on the real symmetric case in finite precision arithmetic and establish the analogous relationship between the norms of computed Lanczos and MINRES residuals for a class of real symmetric problems. We then prove in §7 that if a nonsymmetric problem (1) is well conditioned, then the real symmetric problems generated by the two bidiagonalization algorithms described in §5 belong to this class, so that in fact the residual relationship holds for these nonsymmetric bidiagonalization algorithms in finite precision arithmetic.

2. Relation between linear systems and least squares. Let H_k, $k = 1, 2, \ldots$, denote a family of upper Hessenberg matrices, where H_k is k by k and H_{k-1} is the $k-1$ by $k-1$ principal submatrix of H_k. For each k, define the $k+1$ by k matrix $H_k^{(e)}$ by

$$H_k^{(e)} = \begin{pmatrix} H_k \\ h_{k+1,k} e_k^T \end{pmatrix},$$

where $e_k^T = (0, \ldots, 0, 1)$.

The matrix $H_k^{(e)}$ can be factored in the form $Q_k^* R_k^{(e)}$, where Q_k is a $k+1$ by $k+1$ unitary matrix and $R_k^{(e)}$ is a $k+1$ by k matrix whose top k by k block, denoted R_k, is upper triangular and whose last row consists of zeros. This factorization can be performed using plane rotations:

$$(F_k \cdots F_1)\, H_k^{(e)} = R_k^{(e)}, \quad \text{where } F_i = \begin{pmatrix} I_{i-1} & & & \\ & c_i & -s_i & \\ & s_i & c_i & \\ & & & I_{k-i} \end{pmatrix}.$$

Note that the first $k-1$ sines and cosines, s_i, c_i, $i = 1, \ldots, k-1$, in the Givens rotations used to factor $H_k^{(e)}$ are the same as those used to factor $H_{k-1}^{(e)}$.

Let $\beta > 0$ and let e_1 denote the first unit vector, either a k-vector or a $(k+1)$-vector, depending on the context. Assume that H_k is nonsingular, and let $\check{y}_k$ denote the solution of the linear system $H_k y = \beta e_1$. Let y_k denote the solution of the least squares problem $\min_y \|H_k^{(e)} y - \beta e_1\|$. Finally, let

$$\check{\nu}_k = H_k^{(e)} \check{y}_k - \beta e_1, \qquad \nu_k = H_k^{(e)} y_k - \beta e_1.$$

The following result is established in a slightly different form in [9] and is implicit in a number of other works [3, 8, 15, 20, 28].

THEOREM 1. *Using the above notation, the norms of ν_k and $\check{\nu}_k$ are related to the sines and cosines of the Givens rotations by*

$$\|\nu_k\| = \beta |s_1 s_2 \cdots s_k| \quad \textit{and} \quad \|\check{\nu}_k\| = \beta \frac{1}{|c_k|} |s_1 s_2 \cdots s_k|. \tag{2}$$

It follows that

$$\|\check{\nu}_k\| = \frac{\|\nu_k\|}{\sqrt{1 - (\|\nu_k\| / \|\nu_{k-1}\|)^2}}, \tag{3}$$

or, equivalently,

$$\left(\frac{\|\nu_k\|}{\|\check{\nu}_k\|} \right)^2 + \left(\frac{\|\nu_k\|}{\|\nu_{k-1}\|} \right)^2 = 1. \tag{4}$$

Proof. Let $Q_k = F_k \cdots F_1$ be the $k+1$ by $k+1$ unitary matrix reducing $H_k^{(e)}$ to $R_k^{(e)}$. The least squares problem can be written in the form

$$\min_y \|H_k^{(e)} y - \beta e_1\| = \min_y \|Q_k (H_k^{(e)} y - \beta e_1)\| = \min_y \|R_k^{(e)} y - \beta Q_k e_1\|.$$

The solution y_k is determined by solving the upper triangular linear system with coefficient matrix R_k and right-hand side equal to the first k entries of $\beta Q_k e_1$. The remainder $R_k^{(e)} y_k - \beta Q_k e_1$ is therefore zero except for the last entry, which is just the last entry of $-\beta Q_k e_1 = -\beta (F_k \cdots F_1) e_1$, which is easily seen to be $-\beta s_1 \cdots s_k$. This establishes the first equality in (2).

For the linear system solution $\check{y}_k = H_k^{-1} \beta e_1$, we have

$$\check{\nu}_k = H_k^{(e)} H_k^{-1} \beta e_1 - \beta e_1,$$

which is zero except for the last entry, which is $\beta h_{k+1,k}$ times the $(k,1)$ entry of H_k^{-1}. Now H_k can be factored in the form $\tilde{Q}_k^* \tilde{R}_k$, where $\tilde{Q}_k = \tilde{F}_{k-1} \cdots \tilde{F}_1$, and $\tilde{F}_i$ is the k by k principal submatrix of F_i. The matrix $H_k^{(e)}$, after applying the first $k-1$ plane rotations, has the form

$$(F_{k-1}\cdots F_1)\,H_k^{(e)} = \begin{pmatrix} x & x & \dots & x \\ & x & \dots & x \\ & & \ddots & \vdots \\ & & & r \\ & & & h \end{pmatrix},$$

where r is the (k,k) entry of the upper triangular matrix $\tilde{R}_k$ and $h = h_{k+1,k}$. The kth rotation is chosen to annihilate the nonzero entry in the last row:

$$c_k = \frac{r}{\sqrt{r^2+h^2}}, \quad s_k = -\frac{h}{\sqrt{r^2+h^2}}.$$

Note that r and hence c_k is nonzero since H_k is nonsingular. Now we have $H_k^{-1} = \tilde{R}_k^{-1}\tilde{Q}_k$, and the $(k,1)$ entry of this is $1/r$ times the $(k,1)$ entry of $\tilde{Q}_k = \tilde{F}_{k-1}\cdots\tilde{F}_1$, and this is just $s_1\cdots s_{k-1}$. It follows that the nonzero entry of $\check{\nu}_k$ is $\beta(h_{k+1,k}/r)s_1\cdots s_{k-1}$. Finally, using the fact that $|s_k/c_k| = |h/r| = |h_{k+1,k}/r|$, we obtain the second equality in (2).

From (2), it is clear that

$$\frac{\|\nu_k\|}{\|\nu_{k-1}\|} = |s_k|, \quad \frac{\|\nu_k\|}{\|\check{\nu}_k\|} = |c_k|.$$

The results (3) and (4) follow from the fact that $|c_k|^2 + |s_k|^2 = 1$. □

3. The Arnoldi and GMRES algorithms. Consider a system of linear equations $Ax = b$, where A is an N by N nonsingular matrix and b is a given N-vector. For ease of notation we will assume that the matrix A and the vectors involved in the solution algorithms are real, but our results here and in other sections are easily modified for complex matrices. Given an initial guess x_0 for the solution, the Arnoldi and GMRES algorithms construct approximate solutions x_k, $k = 1, 2, \dots$, of the form

$$x_k = x_0 + t_k, \quad t_k \in K_k(A, r_0) \equiv \operatorname{span}\{r_0, Ar_0, \dots, A^{k-1}r_0\}, \tag{5}$$

where $r_0 \equiv b - Ax_0$ is the initial residual and $K_k(A, r_0)$ is referred to as the kth Krylov space. The residual $r_k \equiv b - Ax_k$ is given by

$$r_k = r_0 - At_k.$$

The two methods differ in how the approximate solutions are chosen from the space (5). For the Arnoldi method (see [23]), the kth residual vector, denoted r_k^A, satisfies

$$r_k^A \perp K_k(A, r^0), \tag{6}$$

while the kth GMRES residual vector, denoted r_k^G, satisfies

$$r_k^G \perp AK_k(A, r_0). \tag{7}$$

From (7) it follows that of all vectors x_k of the form (5), the GMRES approximation x_k^G has the residual of smallest Euclidean norm:

$$\|r_k^G\| = \min_{x_k \in\, x_0 + K_k(A,r_0)} \|b - Ax_k\|. \tag{8}$$

The properties (5) and (6) completely characterize the Arnoldi iterates, while the properties (5) and either (7) or (8) completely define the GMRES iterates.

In order to generate these approximate solutions, both algorithms construct an orthonormal basis for the Krylov space $K_k(A, r_0)$. This can be accomplished by using the modified Gram–Schmidt procedure, for example:

Modified Gram–Schmidt procedure:

1. Compute $r_0 = b - Ax_0$ and set $u_1 = r_0/\|r_0\|$.
2. For $j = 1, \ldots, k$ do:
   $u_{j+1} := Au_j$
   for $i = 1, \ldots, j$ do:

$$h_{ij} := u_i^T u_{j+1}, \quad u_{j+1} := u_{j+1} - h_{ij} u_i$$

$$h_{j+1,j} := \|u_{j+1}\|, \quad u_{j+1} := u_{j+1}/h_{j+1,j}.$$

Other methods have also been proposed for computing these basis vectors [27], but we will not be concerned with the particular implementation used. Note that it is necessary to save all of the basis vectors and at each step to orthogonalize the new vector against each of the previous basis vectors.

Let U_k denote the N by k matrix whose columns are the orthonormal basis vectors $u_1, \ldots, u_k$, and let H_k denote the k by k upper Hessenberg matrix whose nonzero entries are the scalars h_{ij}. The above recurrence can be written in matrix form as

$$AU_k = U_k H_k + h_{k+1,k} u_{k+1} e_k^T, \tag{9}$$

where u_{k+1} is the $(k+1)$st normalized basis vector and e_k is the kth unit k-vector $(0, \ldots, 0, 1)^T$. If H_k is nonsingular, then it can be seen from expression (9) and the definition (5), (6) of the Arnoldi iterates that the kth Arnoldi iterate x_k^A is of the form

$$x_k^A = x_0 + U_k y_k^A, \tag{10}$$

where y_k^A satisfies

$$H_k y_k^A = \beta e_1, \quad \beta = \|r_0\|. \tag{11}$$

If H_k is singular, then it is shown in [3] that the kth Arnoldi iterate does not exist and, moreover, that the kth GMRES iterate does not improve.

Defining the $k+1$ by k matrix $H_k^{(e)}$ to be

$$H_k^{(e)} = \begin{pmatrix} H_k \\ 0 \cdots 0 \; h_{k+1,k} \end{pmatrix},$$

equation (9) can be written in the form

$$AU_k = U_{k+1} H_k^{(e)}. \tag{12}$$

Using this equation and the characterization (5), (8) of the GMRES iterates, it is shown in [23] that the kth GMRES iterate x_k^G is of the form

$$x_k^G = x_0 + U_k y_k^G, \tag{13}$$

where y_k^G satisfies the least squares problem

$$\|\beta e_1 - H_k^{(e)} y_k^G\| = \min_y \|\beta e_1 - H_k^{(e)} y\|. \tag{14}$$

The following theorem is an immediate consequence of Theorem 1. Let s_i and c_i, $i = 1, \ldots, k$, be the sines and cosines of the Givens rotations used to factor $H_k^{(e)}$. Relations between GMRES and Arnoldi residuals and the sines and cosines of Givens rotations were established in [23] and [3], but the direct relation between GMRES and Arnoldi residuals was never stated explicitly.

THEOREM 2. *In exact arithmetic, if $c_k \neq 0$ at iteration k, then the Arnoldi and GMRES residuals are related by*

$$\|r_k^A\| = \frac{\|r_k^G\|}{\sqrt{1 - (\|r_k^G\|/\|r_{k-1}^G\|)^2}}. \tag{15}$$

Proof. From (10), (13), and (12), it follows that the Arnoldi and GMRES residuals can each be written in the form

$$\begin{aligned} r_k^{A,G} &= r_0 - AU_k y_k^{A,G} \\ &= r_0 - U_{k+1} H_k^{(e)} y_k^{A,G} \\ &= U_{k+1}(\beta e_1 - H_k^{(e)} y_k^{A,G}). \end{aligned} \tag{16}$$

Since the columns of U_{k+1} are orthonormal, it follows that

$$\|r_k^{A,G}\| = \|\beta e_1 - H_k^{(e)} y_k^{A,G}\|, \tag{17}$$

and the desired relation (15) now follows from Theorem 1 and the definitions (11) and (14) of y_k^A and y_k^G. □

Note that for the Arnoldi method, relation (17) follows from (16), even if the columns of U_{k+1} are not orthonormal. It requires only that $\|u_{k+1}\| = 1$, since the quantity $\beta e_1 - H_k^{(e)} y_k^A$ has only the last component nonzero.

Theorem 2 shows that if the GMRES residual norm is reduced by a significant factor at step k, then the Arnoldi residual norm will be approximately equal to the GMRES residual norm at step k since the denominator in the right-hand side of (15) will be close to 1. If the GMRES residual norm remains almost constant, however, then the denominator in the right-hand side of (15) is close to 0 and the Arnoldi residual norm will be much larger. Table 1 shows the relation between the GMRES residual norm reduction and the ratio of Arnoldi to GMRES residual norm. Note that the GMRES residual norm must be *very* flat before the Arnoldi residual norm is orders of magnitude larger than the GMRES residual norm.

To illustrate these results, we consider a real matrix A of the following form:

$$A = \Sigma V^T \Lambda V \Sigma^{-1}, \tag{18}$$

where Σ is a diagonal matrix with positive entries, V is an orthogonal matrix, and Λ is a real block diagonal matrix, consisting of at most two by two blocks, each corresponding to a complex conjugate pair of eigenvalues of A. We note that since any of the 2×2 blocks in Λ can be diagonalized by a 2×2 unitary transformation, Σ

TABLE 1
Relation between GMRES residual norm reduction and ratio of Arnoldi to GMRES residual norm.

$\|r_k^G\|/\|r_{k-1}^G\|$	$\|r_k^A\|/\|r_k^G\|$
.5	1.2
.9	2.3
.99	7.1
.9999	70.7
.999999	707

specifies the singular values of an eigenvector matrix of A. Every real diagonalizable matrix B is unitarily similar to a matrix of the form (18), since if $B = X\Lambda X^{-1}$ for some real matrix X and $X = U\Sigma V^T$ is a singular value decomposition of X, then we have $B = UAU^T$. Since the iterative methods we consider are invariant under unitary similarity transformations—the residual norms at each step of the algorithm for solving $Ax = b$ are the same as those at each step of the algorithm for solving $UAU^Ty = Ub$, $x = U^Ty$—it follows that all possible residual norm plots corresponding to diagonalizable real matrices can be obtained by considering matrices of the form (18).

For our example, we set $N = 111$. We chose Σ to have one *small* singular value (.8), two *large* singular values (10 and 10.3), and the remaining singular values ranging from 2.6 upward with a uniform spacing of .02 between successive singular values. The matrix Λ was defined by specifying three randomly generated complex eigenvalues of magnitudes .02, .1, and 10 and a real eigenvalue of magnitude 1, generating the remainder of the spectrum randomly as complex numbers in the box $1 \le x \le 3$, $2 \le y \le 4$, and then defining a 2×2 real block in Λ for each corresponding complex conjugate pair of eigenvalues. The V matrix was set equal to the permutation matrix which for $1 \le j \le N-1$ maps each coordinate vector e_j into e_{j+1} and maps e_N into e_1. The solution was set equal to the vector whose components are all 1, and the initial guess was the zero vector. The convergence tolerance was 10^{-13}, as measured by the ratio of the norm of the residual at iteration k to the norm of the initial residual.

Figure 1 shows a plot of the logarithms of the Arnoldi and GMRES residual norms versus iteration number for this example problem. The solid line is the GMRES convergence curve and the dashed line is the Arnoldi curve. The norm of the starting residual was 92.7.

Observe that for the specified convergence tolerance these algorithms converged simultaneously in 95 iterations. From iterations 63 to 95, the GMRES convergence is basically fast and, as the picture indicates, on that portion of the curve and in fact on similar steep portions of the GMRES curve, the Arnoldi norms converge in a similar fashion.

Also observe the matching of the peaks in the Arnoldi residual norm plot with the plateaus in the GMRES residual norm plot. The double peak corresponding approximately to iterations 22 to 36 coincides with the rough recognition of the members of the conjugate pair of size 10^{-1} as eigenvalues in the spectra of the associated Hessenberg matrices in (9). The second double peak from approximately iterations 42 to 62 corresponds to the identification of the members of the conjugate pair of size 2×10^{-2}. In test problems with smaller eigenvalues, these double peaks are more clearly visible and may be either overlapping or split apart. We note that in [25, 26] connections between the appearance of certain eigenvalues in the spectra of the GMRES/Arnoldi Hessenberg matrices and subsequent speedups in the convergence of GMRES were

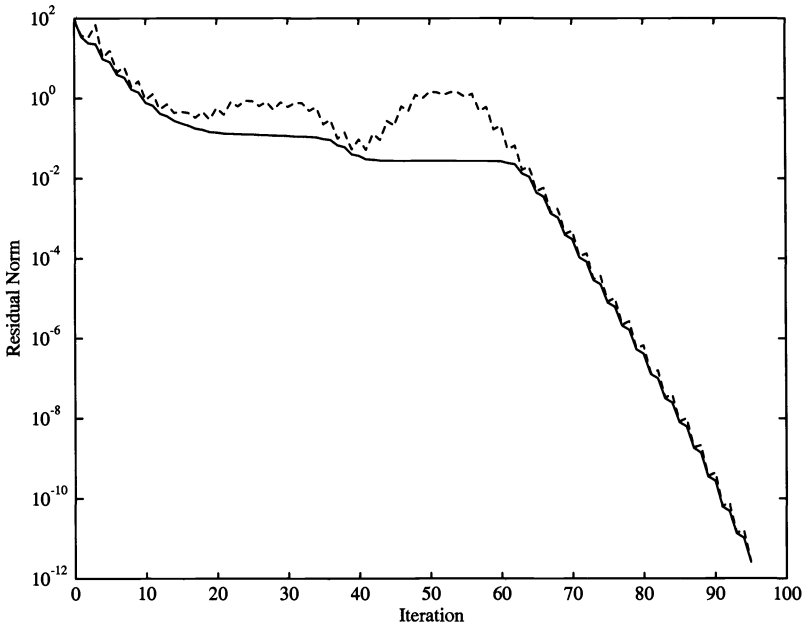


FIG. 1. *GMRES (solid) and Arnoldi (dashed) residual norms.*

observed.

On the scale of Fig. 1, it is difficult to see the precise correlation between the degree of flatness of the GMRES curve and the height of the Arnoldi curve above the GMRES curve. However, for example, at iterations 24 and 50 within the double peaks in the residual norm curve for the Arnoldi iterations, the norm of the GMRES residual was reduced, respectively, by factors of .989529 and .999825, and the corresponding ratios of the Arnoldi to GMRES residual norms were approximately 6.93 and 53.4. These values are as predicted by Theorem 2.

In this example, across the first double peak/plateau the condition numbers of the corresponding Hessenberg matrices vary from 261 to 2351 and are less than the condition number 4625.8 of the original iteration matrix A. Across the second double peak/plateau, however, this is not the case. Over iterations 42 to 62 the condition numbers of the Hessenberg matrices range from 965 to 106,146. After iteration 62, these condition numbers settle down to the condition number of A.

4. The BCG and QMR algorithms. The BCG [7] and QMR [8] algorithms also construct approximate solutions $x_k, k = 1, 2, \ldots$, of the form (5). They differ from the Arnoldi and GMRES algorithms in that, instead of constructing an orthonormal basis for the Krylov space $K_k(A, r_0)$, they construct sequences of biorthogonal vectors spanning the spaces $K_k(A, r_0)$ and $K_k(A^T, \hat{r}_0)$, where $\hat{r}_0$ is an arbitrary vector, often chosen equal to r_0. This can be accomplished using the nonsymmetric Lanczos algorithm and requires two simple three-term recurrences:

Nonsymmetric Lanczos algorithm:

1. Set $v_1 = r_0/\|r_0\|$ and $w_1 = \hat{r}_0/\|\hat{r}_0\|$. Set $\rho_1 = 1$ and $\xi_1 = 1$ and $v_0 = w_0 = 0$.

2. For $j = 1, \ldots, k$ do

$$\alpha_j = w_j^T A v_j / w_j^T v_j,$$
$$\beta_1 = 0, \quad \beta_j = \xi_j w_j^T v_j / w_{j-1}^T v_{j-1}, \text{ if } j > 1,$$
$$\nu_{j+1} = A v_j - \alpha_j v_j - \beta_j v_{j-1},$$
$$\rho_{j+1} = \|\nu_{j+1}\|, \quad v_{j+1} = \nu_{j+1}/\rho_{j+1},$$
$$\omega_{j+1} = A^T w_j - \alpha_j w_j - (\beta_j \rho_j / \xi_j) w_{j-1},$$
$$\xi_{j+1} = \|\omega_{j+1}\|, \quad w_{j+1} = \omega_{j+1}/\xi_{j+1}.$$

Here we have used the nonsymmetric Lanczos formulation that scales by setting the norms of the Lanczos vectors to unity. Note that each step of the nonsymmetric Lanczos algorithm requires matrix vector multiplications by A and A^T but does not require saving and orthogonalizing against all previous basis vectors, as is required by the Arnoldi/GMRES methods.

Unfortunately, the nonsymmetric Lanczos algorithm can break down. If $w_j^T v_j = 0$ for some j, the coefficients in the above algorithm are undefined. If $w_j = 0$ or $v_j = 0$, then this means an invariant subspace for either A^T or A has been found, but if $w_j^T v_j = 0$ when neither w_j nor v_j is zero, then this is referred to as a *serious* breakdown. While an exact breakdown is unlikely, near breakdowns can cause numerical instabilities. To avoid such problems, various look-ahead strategies have been proposed; see, e.g., [2, 8, 10, 21]. The relationship we establish between BCG and QMR residuals holds in exact arithmetic provided the same look-ahead steps are used in the underlying Lanczos recurrence for each algorithm.

Let V_k denote the N by k matrix whose columns are the basis vectors $v_1, \ldots, v_k$ generated for the space $K_k(A, r_0)$ and let W_k denote the N by k matrix whose columns are the basis vectors $w_1, \ldots, w_k$ generated for the space $K_k(A^T, \hat{r}_0)$. Let the tridiagonal matrix T_k be defined by

$$T_k = \begin{pmatrix} \alpha_1 & \beta_2 & & & \\ \rho_2 & \alpha_2 & \beta_3 & & \\ & \rho_3 & \ddots & \ddots & \\ & & \ddots & \alpha_{k-1} & \beta_k \\ & & & \rho_k & \alpha_k \end{pmatrix}.$$

If no look-ahead steps are performed, then the α's, β's, and ρ's are numbers. If look-ahead steps have been performed then T_k can still be written in this form but now the entries are matrices of size determined by the number of look-ahead steps necessary before a regular Lanczos vector can be produced. For details see, for example, [8].

The above recurrences can be written in matrix form as

$$AV_k = V_k T_k + \rho_{k+1} v_{k+1} e_k^T, \tag{19}$$
$$A^T W_k = W_k \Gamma_k^{-1} T_k \Gamma_k + \xi_{k+1} w_{k+1} e_k^T,$$

where

$$\Gamma_k = \operatorname{diag}(\gamma_1, \ldots, \gamma_k), \quad \gamma_1 = 1, \quad \gamma_j = \gamma_{j-1} \rho_j / \xi_j, \quad j > 1.$$

The kth BCG iterate x_k^B is chosen so that the residual r_k^B satisfies

$$r_k^B \perp K_k(A^T, \hat{r}_0). \tag{20}$$

This is somewhat analogous to the condition (6) defining the Arnoldi iterates and, like (6), this condition may be impossible to satisfy with an iterate of the form (5). Using expression (19), condition (20) can be written in the form

$$x_k^B = x_0 + V_k y_k^B, \tag{21}$$

where y_k^B satisfies

$$T_k y_k^B = \beta e_1, \quad \beta = \|r_0\|. \tag{22}$$

It is shown in [8] that this equation has a solution if and only if the (block) tridiagonal matrix T_k is nonsingular. For the remainder of this discussion we will assume that the matrices $T_1, \ldots, T_k$ are nonsingular. Here again look-ahead strategies can be used to deal with near-singularity of the tridiagonal matrices.

The QMR algorithm is derived in much the same way as the GMRES algorithm described in the previous section. Define the $k+1$ by k matrix $T_k^{(e)}$ by

$$T_k^{(e)} = \begin{pmatrix} T_k \\ 0 \ldots 0 \ \rho_{k+1} \end{pmatrix}. \tag{23}$$

Equation (19) can be written in the form

$$AV_k = V_{k+1} T_k^{(e)}. \tag{24}$$

The kth QMR iterate x_k^Q is of the form

$$x_k^Q = x_0 + V_k y_k^Q,$$

so the kth QMR residual r_k^Q is of the form

$$r_k^Q = r_0 - AV_k y_k^Q = r_0 - V_{k+1} T_k^{(e)} y_k^Q = V_{k+1}(\beta e_1 - T_k^{(e)} y_k^Q), \tag{25}$$

where $\beta = \|r_0\|$ and e_1 is the first unit $(k+1)$-vector. Ideally, one would like to choose y_k^Q to minimize $\|r_k^Q\|$, but since the columns of V_{k+1} are not orthogonal, this would not be practical. Instead, the QMR iterate is defined by taking y_k^Q to minimize the quantity in the parentheses in (25). That is, y_k^Q satisfies the least squares problem

$$\|\beta e_1 - T_k^{(e)} y_k^Q\| = \min_y \|\beta e_1 - T_k^{(e)} y\|. \tag{26}$$

We refer to the vector $\beta e_1 - T_k^{(e)} y_k^Q$ as the QMR *quasi-residual* and denote it z_k^Q. The actual QMR residual is

$$r_k^Q = V_{k+1} z_k^Q. \tag{27}$$

In [8] a more general definition of the QMR iterate is given, allowing for an arbitrary diagonal scaling of the least squares problem (26). It is not clear how this diagonal scaling should be chosen, however, and here we consider only the scaling inherent in (26) with the right and left Lanczos vectors each having norm one.

Since the columns of V_{k+1} are not orthonormal, the norms of the true residuals are not the same as those of the quasi-residuals. One can give upper and lower bounds

on the ratios of these norms, however. Since the columns of V_{k+1} each have norm one, it is shown in [8] that

$$\|r_k^Q\| \leq \sqrt{k+1}\, \|z_k^Q\|. \tag{28}$$

A lower bound on $\|r_k^Q\|$ is given by

$$\|r_k^Q\| \geq \sigma_{\min}(V_{k+1})\|z_k^Q\|, \tag{29}$$

where $\sigma_{\min}(V_{k+1})$ denotes the smallest singular value of V_{k+1}. While it is possible that $\sigma_{\min}(V_{k+1})$ could be very small (especially in finite precision arithmetic, where this is usually the case!), it is unlikely, in such cases, that the inequality (29) will be a near equality, since the approximate solution x_k^Q is chosen to satisfy (26) without regard to the matrix V_{k+1}.

The following theorem is an immediate consequence of Theorem 1 from §2. Let $\hat{s}_i$ and $\hat{c}_i$, $i = 1, \ldots, k$, be the sines and cosines of the Givens rotations used to factor $T_k^{(e)}$. Relations between BCG residuals and QMR quasi-residuals and the sines and cosines of the Givens rotations were established in [8], but the direct relation between BCG residuals and QMR quasi-residuals was never explicitly stated.

THEOREM 3. *In exact arithmetic, if $\hat{c}_k \neq 0$ at iteration k, then the BCG residual and the QMR quasi-residual are related by*

$$\|r_k^B\| = \frac{\|z_k^Q\|}{\sqrt{1 - (\|z_k^Q\|/\|z_{k-1}^Q\|)^2}}. \tag{30}$$

Proof. From (21) and (24), it follows that the BCG residual can be written in the form

$$\begin{aligned} r_k^B &= r_0 - AV_k y_k^B \\ &= r_0 - V_{k+1} T_k^{(e)} y_k^B \\ &= V_{k+1}(\beta e_1 - T_k^{(e)} y_k^B). \end{aligned} \tag{31}$$

From the definition (22) of y_k^B it follows that the quantity in parentheses in (31) has a nonzero entry only in the $(k+1)$st component, and since $\|v_{k+1}\| = 1$, we have

$$\|r_k^B\| = \|\beta e_1 - T_k^{(e)} y_k^B\|. \tag{32}$$

Using relation (32) and the definition (26) of the QMR quasi-residual, the desired result now follows from Theorem 1. □

Note that while the choice of the starting vector $\hat{r}_0$ affects the tridiagonal matrix that is generated and hence affects the sines and cosines of the Givens transformations, it does not affect the relationship (30). This relationship holds provided only that the same vector $\hat{r}_0$ is used for both the BCG and QMR computations.

Figure 2 shows a plot of the logarithms of the norms of the BCG residuals (dashed line), the QMR residuals (dotted line), and the QMR quasi-residuals (solid line) versus iteration number for the same example described in the previous section. Observe that for the convergence tolerance used, both algorithms converged in 101 iterations if the norm of the true residuals is used to measure convergence. Note that on the log plot it can be seen that the QMR residual norm and the quasi-residual norm are of the same order of magnitude, although they are not identical.

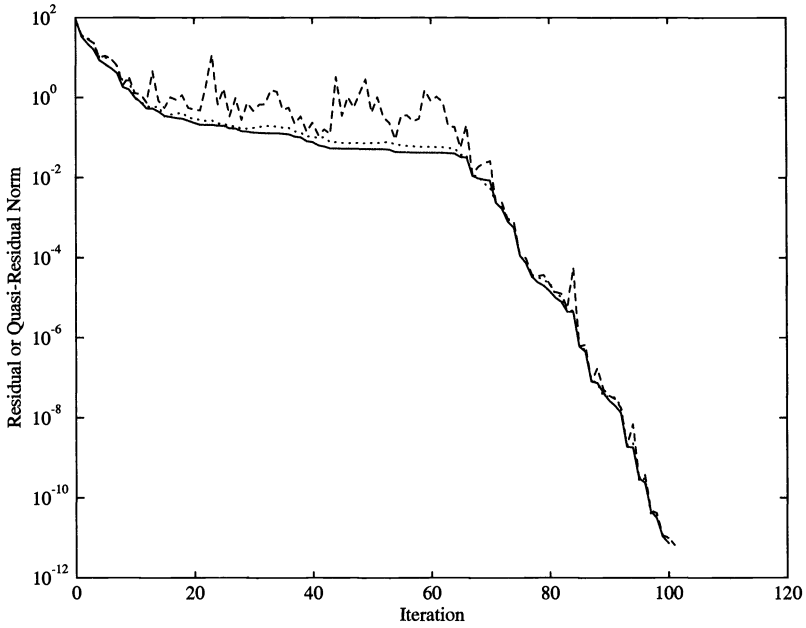


FIG. 2. *QMR quasi-residual norm (solid), BCG residual norm (dashed), and QMR residual norm (dotted).*

On the scale of Fig. 2 it is again difficult to see the precise correlation between the degree of flatness of the QMR quasi-residual norm curve and the height of the corresponding BCG curve above the QMR curve. However, for example, at iterations 33 and 59 in the plot of the residual norms of the BCG iterates, the QMR quasi-residual norms were reduced, respectively, by factors of .99647 and .99962. The corresponding ratios of the BCG norms to the QMR quasi-residual norms were 11.9 and 36.4. These values fit well with the predictions from Theorem 3. The corresponding ratios of the BCG residual norms to the QMR residual norms were 7.95 and 26.04.

The BCG peaks covering approximately iterations 20 to 40 correspond to the appearance in the spectra of the associated tridiagonal matrices in (22) of the conjugate pair of eigenvalues of magnitude 10^{-1}. The next and more recognizable two peaks, from approximately iterations 43 to 64, correspond to the identification of the members of the conjugate pair of magnitude 2×10^{-2}.

In this example, the condition numbers of the tridiagonal matrices converged more or less monotonically to 62,609, a factor of almost 15 times greater than the condition number of A. On iterations 20 to 40 the condition numbers ranged from 473 to 34,310, and from iterations 43 to 64 they ranged from 33,500 to 62,609.

5. Bidiagonalization/SQMR/BLanczos and symmetric Lanczos.

5.1. Real symmetric Lanczos algorithm. The Lanczos algorithm for constructing an orthonormal basis for the Krylov space $K_k(A, r_0)$, where A is a real symmetric matrix, can be written as follows.

Real symmetric Lanczos algorithm:

1. Set $v_1 = r_0/\|r_0\|$. Set $\rho_1 = 1$ and $v_0 = 0$.

2. For $j = 1, \ldots, k$ do

$$\alpha_j = v_j^T(Av_j - \rho_j v_{j-1}),$$
$$\nu_{j+1} = Av_j - \alpha_j v_j - \rho_j v_{j-1},$$
$$\rho_{j+1} = \|\nu_{j+1}\|, \quad v_{j+1} = \frac{\nu_{j+1}}{\rho_{j+1}}.$$

If T_k denotes the symmetric tridiagonal matrix

$$T_k = \begin{pmatrix} \alpha_1 & \rho_2 & & \\ \rho_2 & \ddots & \ddots & \\ & \ddots & \alpha_{k-1} & \rho_k \\ & & \rho_k & \alpha_k \end{pmatrix}$$

and $T_k^{(e)}$ the extended matrix (23), then formulas (19) and (24) express this recurrence in matrix form. The MINRES and Lanczos algorithm iterates are defined using equations (26) and (22), respectively.

For real symmetric problems, assuming exact arithmetic, the Lanczos vectors are orthonormal and the norm of the quasi-residual and the actual residual defined in (27) are the same. In this case relation (30) becomes

$$\|r_k^L\| = \frac{\|r_k^M\|}{\sqrt{1 - (\|r_k^M\|/\|r_{k-1}^M\|)^2}}. \tag{33}$$

5.2. Bidiagonalization of nonsymmetric systems. Any nonsymmetric system (1) can be solved by solving a larger symmetrized version of the problem. The use of bidiagonalization to symmetrize a nonsymmetric problem was suggested in [17] and was subsequently used to compute singular values of A [11] and to solve (1) and associated least squares problems [19] and [20]. Simple bidiagonalization replaces (1) by the following $2N \times 2N$ real symmetric but indefinite system

$$B\bar{x} = \bar{b}, \text{ where } B \equiv \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}, \ \bar{x} \equiv \begin{pmatrix} y \\ x \end{pmatrix}, \ \bar{b} \equiv \begin{pmatrix} b \\ 0 \end{pmatrix}, \tag{34}$$

whose solution contains the desired solution. We have the following lemma relating the eigenvalues of B to the singular values of A and the eigenvectors of B to concatenations of left and right singular vectors of A.

LEMMA 1 (see [5]). *Let A be any real nonsymmetric $N \times N$ matrix with singular value decomposition $A = X\Sigma Y^T$, where $\Sigma = \text{diag}\ \{\sigma_1, \sigma_2, \ldots, \sigma_N\}$ and $Y^TY = X^TX = I$. Then*

$$BZ = Z \begin{pmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{pmatrix}, \text{ where } Z = \frac{1}{\sqrt{2}} \begin{pmatrix} X & X \\ Y & -Y \end{pmatrix}. \tag{35}$$

Both bidiagonalization procedures SQMR and BLanczos map (1) into (34) and then use the real symmetric Lanczos recursion to map (34) into simple tridiagonal problems. Specifically, if we apply the real symmetric Lanczos recursions to B with starting vector $w_1 = \bar{r}_0/\|\bar{r}_0\|$, where $\bar{r}_0 = -B\bar{x}_0 + \bar{b}$ with $\bar{x}_0 = (0, x_0^T)^T$, then in exact arithmetic we obtain the matrix recursion

$$BW_k = W_k T_k + \rho_{k+1} w_{k+1} e_k^T, \tag{36}$$

where

$$T_k = \begin{pmatrix} 0 & \rho_2 & & & \\ \rho_2 & 0 & \rho_3 & & \\ & \rho_3 & 0 & \ddots & \\ & & \ddots & \ddots & \rho_k \\ & & & \rho_k & 0 \end{pmatrix}. \tag{37}$$

Because of the special structure of B and of w_1, all of the $\alpha_j \equiv 0$. Theoretically, the $W_k \equiv \{w_1, \ldots, w_k\}$ are orthonormal, and $T_k = W_k^T B W_k$ is an orthogonal projection of B onto the Krylov subspace $K_k(B, w_1)$.

If $\rho_j \neq 0$, $2 \leq j \leq 2k$, then each T_{2j} is nonsingular and its eigenvalues occur in $\pm$ pairs. Each T_{2j-1} is singular. For details see [5]. Therefore, the Lanczos iterates are defined only on even-numbered steps. We denote the corresponding Lanczos and MINRES iterates by $\bar{x}_k^L$ and $\bar{x}_k^M$ and the corresponding residuals by $\bar{r}_k^L$ and $\bar{r}_k^M$. In Lemmas 2 and 3 we use these quantities to define the BLanczos and SQMR iterates (and residuals) for (1). In exact arithmetic, the kth Lanczos iterate is obtained by solving

$$T_{2k}\bar{z} = \rho e_1, \quad \rho = \|\bar{r}_0\| \tag{38}$$

and forming

$$\bar{x}_k^G = \bar{x}_0 + W_{2k}\bar{z}, \text{ where } \bar{x}_0 = \begin{pmatrix} 0 \\ x_0 \end{pmatrix}. \tag{39}$$

In exact arithmetic, the kth MINRES iterate $\bar{x}_k^M$ is obtained by solving the least squares problem

$$\min_{\bar{z}} \|T_{2k}^{(e)}\bar{z} - \rho e_1\|, \text{ where } T_{2k}^{(e)} = \begin{pmatrix} T_{2k} \\ \rho_{2k+1} e_{2k}^T \end{pmatrix} \tag{40}$$

and forming

$$\bar{x}_k^M = \bar{x}_0 + W_{2k}\bar{z}. \tag{41}$$

Lemmas 2 and 3 extract the corresponding BLanczos and SQMR iterates from the above relationships.

LEMMA 2 (see [4]). *If we apply the Lanczos method to* (34), *then all odd-numbered components of* $\bar{z}$ *in* (38) *are zero. Furthermore, if we specify the* k*th BLanczos iterate* x_k^{BL} *to consist of the last* N *components of* $\bar{x}_k^L$, *then* $r_k^{BL} \equiv b - Ax_k^{BL}$ *consists of the first* N *components of* $\bar{r}_k^L$, *and* $x_k^{BL} = x_0 + V_k\bar{z}^*$, *where* $*$ *denotes the even-numbered components of* $\bar{z}$, $V_k \equiv \{v_1, \ldots, v_k\}$ *and each* v_j *consists of the last* N *components of* w_{2j}. *In addition, in exact arithmetic,*

$$x_k^{BL} = x_{k-1}^{BL} + \bar{z}(2k)\ v_k \quad \textit{and} \quad \|r_k^{BL}\| = |\rho_{2k+1}\bar{z}(2k)|, \tag{42}$$

where $\bar{z}(2k) = (-1)^{k+1} \prod_{j=1}^{k} \rho_{2j-1} / \prod_{j=1}^{k} \rho_{2j}$.

Now consider the SQMR iterates x_k^{SQ}. The least squares problem in (40) can be solved by successively applying Givens transformations $F_j(c_j, s_j)$ to $T_{2k}^{(e)}$ to obtain

$$(F_{2k} \cdots F_1)\, T_{2k}^{(e)} = R_{2k}^{(e)} = \begin{pmatrix} \bar{R}_{2k} \\ 0 \end{pmatrix}, \tag{43}$$

where $\bar{R}_{2k}$ is $2k \times 2k$ and upper triangular. For each j, $c_{2j-1} = 0$ and $s_{2j-1} = 1$, and we therefore use c_j, s_j to denote the cosine and sine which define F_{2j}. If we set $\delta_j = \bar{R}_{2k}(j,j)$ and $\bar{P}_{2k} = W_{2k}\bar{R}_{2k}^{-1}$, then

$$p_k = \left[v_k - \rho_{2k-1}\rho_{2k}\delta_{2k-2}^{-1}p_{k-1}\right]\delta_{2k}^{-1}, \tag{44}$$

where p_k is the N-vector consisting of the last N components of the $2k$th column of $\bar{P}_{2k}$.

LEMMA 3. *If we apply MINRES to* (34), *then all odd-numbered components of* $\bar{z}$ *in* (40) *are zero. Furthermore, if we specify the kth SQMR iterate* x_k^{SQ} *to consist of the last N components of* $\bar{x}_k^M$, *then* $r_k^{SQ} \equiv b - Ax_k^{SQ}$ *consists of the first N components of* $\bar{r}_k^M$, *and* $x_k^{SQ} = x_0 + P_k\bar{z}^*$, *where* $*$ *denotes the even-numbered components of* $\bar{z}$, $P_k \equiv \{p_1, \ldots, p_k\}$ *and* p_j *consists of the last N components of* $\bar{p}_{2j}$. *In addition, in exact arithmetic,*

$$\begin{aligned} x_k^{SQ} &= x_{k-1}^{SQ} - c_k\|r_{k-1}^{SQ}\|p_k, \\ \|r_{k-1}^{SQ}\| &= \min_{\bar{z}}\|T_{2k-2}^e\bar{z} - \rho e_1\| = \left|\prod_{j=1}^{k-1} s_j\right| \cdot \|r_0\|, \end{aligned} \tag{45}$$

where c_j, s_j *define the* $2j$*th Givens transformations which were used in the factorization of* $T_{2k}^{(e)}$.

Proofs of Lemmas 2 and 3 are in [4]. If we were defining only SQMR then there is no apparent reason not to consider the $T_{2k-1}^{(e)}$. However from Brown [3] we know, at least in exact arithmetic, that since T_{2k-1} is singular, $T_{2k-1}^{(e)}$ and $T_{2k}^{(e)}$ would yield the same SQMR iterate. In the tests presented in [4] there was no reorthogonalization of any Lanczos vectors.

In §6 we consider the real symmetric Lanczos procedures, Lanczos and MINRES, in finite precision arithmetic and demonstrate that a relationship analogous to (33) exists for a certain class of symmetric problems. In §7 we then show that if a nonsymmetric problem (1) is well conditioned, then the real symmetric problems generated by the bidiagonalization algorithms defined in this section are in that class. Then, using Lemmas 2 and 3, we obtain a relationship analogous to (33) which is valid for these bidiagonalization methods in finite precision arithmetic.

6. Finite precision arithmetic, Lanczos/MINRES. Quantities generated by the Lanczos and the MINRES algorithms will be denoted with superscripts L and M, respectively. Finite precision quantities will be denoted with tildes. In finite precision computations, the matrix equations (19) and (24) are replaced by

$$A\tilde{V}_k = \tilde{V}_k\tilde{T}_k + \tilde{\rho}_{k+1}\tilde{v}_{k+1}e_k^T + F_k = \tilde{V}_{k+1}\tilde{T}_k^{(e)} + F_k. \tag{46}$$

For standard implementations of the real symmetric Lanczos algorithm it is shown in [18] that the Frobenius norm of the roundoff matrix F_k satisfies

$$\|F_k\|_F \leq c\sqrt{k}\,\epsilon\,\|A\|_F, \tag{47}$$

where ϵ is the machine precision and c is a moderate size constant. We will not use this bound explicitly but will simply express error estimates in terms of $\|F_k\|$.

Suppose the computed Lanczos and MINRES approximations satisfy

$$\tilde{x}_k^L = x_0 + \tilde{V}_k\tilde{y}_k^L + g_k^L, \quad \tilde{x}_k^M = x_0 + \tilde{V}_k\tilde{y}_k^M + g_k^M, \tag{48}$$

where $\tilde{y}_k^L$ is the exact solution to the tridiagonal system

$$\tilde{T}_k y = \beta e_1 \tag{49}$$

and $\tilde{y}_k^M$ is the exact solution to the least squares problem

$$\min_y \|\tilde{T}_k^{(e)} y - \beta e_1\|. \tag{50}$$

If the tridiagonal matrix T_k is not too badly conditioned, then the error due to the inexact solution of the linear system or least squares problem will be small. The Lanczos and MINRES residuals for the computed quantities satisfy

$$\begin{aligned} \tilde{r}_k^{L,M} &= r_0 - A\tilde{V}_k \tilde{y}_k^{L,M} - Ag_k^{L,M} \\ &= r_0 - \tilde{V}_{k+1}\tilde{T}_k^{(e)} \tilde{y}_k^{L,M} - F_k \tilde{y}_k^{L,M} - Ag_k^{L,M} \\ &= \tilde{V}_{k+1}(\beta e_1 - \tilde{T}_k^{(e)} \tilde{y}_k^{L,M}) - F_k \tilde{y}_k^{L,M} - Ag_k^{L,M}. \end{aligned} \tag{51}$$

In finite precision computations, the columns of Lanczos vectors $\tilde{V}_{k+1}$ in (51) frequently lose orthogonality. Yet numerical experiments in [4] suggest that relation (33) holds to a close approximation, even after orthogonality of the Lanczos vectors is lost. We now show why this is to be expected, assuming that the terms $-F_k\tilde{y}_k^{L,M} - Ag_k^{L,M}$ in (51) are small compared to $\|\tilde{r}_k^L\|$ and $\|\tilde{r}_k^M\|$.

It is shown in [12] that for any given K, the tridiagonal matrices generated by K steps of a finite precision Lanczos recurrence with a real symmetric matrix A are the same as those that would be generated by the exact Lanczos algorithm applied to a larger real symmetric matrix $\bar{A}$ whose eigenvalues all lie within tiny intervals about the eigenvalues of A. The size of the intervals depends on the machine precision and on an upper bound K for the number of steps that will be run. It is further shown that components of the computed Lanczos vectors associated with a particular eigenvector of A are related to the components of the corresponding exact Lanczos vectors associated with the eigenvectors of $\bar{A}$ whose eigenvalues lie in the interval about this eigenvalue of A, in the following way:

$$(\tilde{v}_k(i))^2 \doteq \sum_\ell (\bar{v}_k(i_\ell))^2. \tag{52}$$

Here $\tilde{v}_k(i)$ represents the component of the computed Lanczos vector $\tilde{v}_k$ in the direction of the ith eigenvector of A. If $\bar{v}_k$ is the corresponding exact Lanczos vector obtained from the recurrence with $\bar{A}$, then $\bar{v}_k(i_\ell)$, $\ell = 1, \ldots$ denotes the components of $\bar{v}_k$ in the directions of the eigenvectors of $\bar{A}$ whose eigenvalues lie in the tiny interval about the ith eigenvalue of A.

Let $\bar{r}_k^L$ and $\bar{r}_k^M$ denote the corresponding exact arithmetic residual vectors obtained by applying the Lanczos and the MINRES algorithms to a linear system $\bar{A}\bar{x} = \bar{b}$, with an initial guess $\bar{x}_0$ such that $\bar{r}_0 = \bar{b} - \bar{A}\bar{x}_0$ has the same norm as r_0 and is parallel to the first Lanczos vector $\bar{v}_1$. The following lemmas relate the norms of the computed residual vectors $\tilde{r}_k^L$ and $\tilde{r}_k^M$ to the norms of the exact residual vectors $\bar{r}_k^L$ and $\bar{r}_k^M$.

LEMMA 4. *Let $\tilde{r}_k^L$ satisfy* (51) *and let $\bar{r}_k^L$ be the residual vector obtained after k steps of the exact Lanczos algorithm applied to the linear system $\bar{A}\bar{x} = \bar{b}$, as described above. Then*

$$\|\tilde{r}_k^L\| = \|\bar{r}_k^L\| + h_k^L, \quad \text{where } |h_k^L| \le \|F_k\| \cdot \|\tilde{y}_k^L\| + \|A\| \cdot \|g_k^L\|. \tag{53}$$

Proof. Since the exact Lanczos recurrence generates the same tridiagonal matrix $\tilde{T}_k$ and the same parameter $\tilde{\rho}_{k+1}$ (since this is an element of $\tilde{T}_{k+1}$) as the finite precision computation, the exact residual vector $\bar{r}_k^L$ satisfies

$$\bar{r}_k^L = -\tilde{\rho}_{k+1}\bar{v}_{k+1}\beta(e_k^T\tilde{T}_k^{-1}e_1).$$

It follows from (51) that

$$\tilde{r}_k^L = -\tilde{\rho}_{k+1}\tilde{v}_{k+1}\beta(e_k^T\tilde{T}_k^{-1}e_1) - F_k\tilde{y}_k^L - Ag_k^L,$$

since the quantity in parentheses in (51) has only its last component nonzero. Since $\|\bar{v}_{k+1}\| = \|\tilde{v}_{k+1}\| = 1$, the desired result (53) follows. □

LEMMA 5. *Let $\tilde{r}_k^M$ satisfy* (51) *and let $\tilde{V}_{k+1}$ satisfy*

$$A\tilde{V}_{k+1} = \tilde{V}_{k+2}\tilde{T}_{k+1}^{(e)} + F_{k+1}. \tag{54}$$

Then $A\tilde{r}_k^M$ is given by

$$A\tilde{r}_k^M = \tilde{v}_{k+1}\gamma_{k+1} + \tilde{v}_{k+2}\gamma_{k+2} + F_{k+1}\tilde{z}_k^M - AF_k\tilde{y}_k^M - A^2g_k^M, \tag{55}$$

where $\tilde{z}_k^M = \beta e_1 - \tilde{T}_k^{(e)}\tilde{y}_k^M$ and the coefficients γ_{k+1} and γ_{k+2} depend only on the elements of the extended tridiagonal matrix $\tilde{T}_{k+1}^{(e)}$.

Proof. Note that since $\tilde{y}_k^M$ minimizes $\|\beta e_1 - \tilde{T}_k^{(e)}y\|$, the remainder $\tilde{z}_k^M = \beta e_1 - \tilde{T}_k^{(e)}\tilde{y}_k^M$ is orthogonal to the columns of $\tilde{T}_k^{(e)}$. Note also that the extended tridiagonal matrix $\tilde{T}_{k+1}^{(e)}$ can be written in the form

$$\tilde{T}_{k+1}^{(e)} = \begin{pmatrix} \tilde{T}_k^{(e)^T} & \\ 0\ldots0\ \tilde{\rho}_{k+1} & \tilde{\alpha}_{k+1} \\ 0\ldots0\ \ 0 & \tilde{\rho}_{k+2} \end{pmatrix},$$

where the elements $\tilde{\alpha}_{k+1}$ and $\tilde{\rho}_{k+1}$, $\tilde{\rho}_{k+2}$ are those generated by the symmetric Lanczos recurrence. Multiplying (54) by $\tilde{z}_k^M$ on the right gives

$$A\tilde{V}_{k+1}\tilde{z}_k^M = \tilde{v}_{k+1}\gamma_{k+1} + \tilde{v}_{k+2}\gamma_{k+2} + F_{k+1}\tilde{z}_k^M, \tag{56}$$

where the coefficients

$$\gamma_{k+1} = \tilde{\rho}_{k+1}\tilde{z}_k^M(k) + \tilde{\alpha}_{k+1}\tilde{z}_k^M(k+1), \quad \gamma_{k+2} = \tilde{\rho}_{k+2}\tilde{z}_k^M(k+1)$$

are functions of the elements of the extended tridiagonal matrix $\tilde{T}_{k+1}^{(e)}$. Using (51) to substitute for $\tilde{V}_{k+1}\tilde{z}_k^M$ in (56) gives

$$A(\tilde{r}_k^M + F_k\tilde{y}_k^M + Ag_k^M) = \tilde{v}_{k+1}\gamma_{k+1} + \tilde{v}_{k+2}\gamma_{k+2} + F_{k+1}\tilde{z}_k^M,$$

from which the result (55) follows. □

LEMMA 6. *The exact arithmetic residual vector $\bar{r}_k^M$ satisfies*

$$\bar{A}\bar{r}_k^M = \bar{v}_{k+1}\gamma_{k+1} + \bar{v}_{k+2}\gamma_{k+2}, \tag{57}$$

where γ_{k+1} and γ_{k+2} are the same coefficients as in (55).

Proof. Since the exact Lanczos vectors satisfy

$$\bar{A}\bar{V}_{k+1} = \bar{V}_{k+2}\tilde{T}_{k+1}^{(e)},$$

where $\tilde{T}_{k+1}^{(e)}$ is the same tridiagonal matrix as in (54), the result follows by the same arguments as used in Lemma 5. □

We wish to show that $\|\tilde{r}_k^M\| \approx \|\bar{r}_k^M\|$. To see this, it is necessary to translate into bases in which A and $\bar{A}$ are diagonal. That is, suppose $A = W\Lambda W^T$ and $\bar{A} = \bar{W}\bar{\Lambda}\bar{W}^T$, where Λ and $\bar{\Lambda}$ are diagonal and W and $\bar{W}$ are orthogonal matrices. Let $\tilde{v}(i)$ denote the ith component of a vector $W^T\tilde{v}$ associated with the finite precision computation for Λ and let $\bar{v}(i_\ell)$ denote the i_ℓth component of a vector $\bar{W}^T\bar{v}$ associated with the exact calculation for $\bar{\Lambda}$. The index i_ℓ will range over all eigenvalues $\bar{\lambda}_{i_\ell}$ of $\bar{\Lambda}$ that lie in the interval about eigenvalue λ_i of Λ. There is no loss in generality in making this transformation. The arguments used need only the fact that the error term in (54) is small and the size of that term is independent of the orthogonal matrix W.

Since the bound proved in [12] on the size of the intervals containing the eigenvalues of $\bar{\Lambda}$ appears to be a large overestimate of their actual size [13], it is not very enlightening to include this bound in our estimates. Instead we will simply assume

$$\max_\ell |\lambda_i - \bar{\lambda}_{i_\ell}| \le \xi \quad \forall i. \tag{58}$$

The bound in [12] on the difference between the left- and right-hand sides in (52) is also an overestimate. Therefore, we will simply assume

$$\left| (\tilde{v}_j(i))^2 - \sum_\ell (\bar{v}_j(i_\ell))^2 \right| \le \delta \quad \forall i, j. \tag{59}$$

The following lemma establishes one more relation between the components $W^T\tilde{v}_k$ of the computed Lanczos vectors and the components $\bar{W}^T\bar{v}_k$ of the exact Lanczos vectors.

LEMMA 7. *The following relation holds between components of the computed Lanczos vectors for Λ and those of the exact Lanczos vectors for $\bar{\Lambda}$:*

$$\left| \tilde{v}_{k+1}(i)\tilde{v}_k(i) - \sum_\ell \bar{v}_{k+1}(i_\ell)\bar{v}_k(i_\ell) \right| \le \frac{1}{\tilde{\rho}_{k+1}} \left[\delta \cdot \sum_{j=1}^{k} |\lambda_i - \tilde{\alpha}_j| \right.$$

$$\left. + \xi \cdot \sum_{j=1}^{k} \sum_\ell (\bar{v}_j(i_\ell))^2 + f \cdot \sum_{j=1}^{k} |\tilde{v}_j(i)| \right], \tag{60}$$

where $f = \max_{i,j} |F_k(i,j)|$ and ξ and δ are as defined in (58) *and* (59).

Proof. Writing the three-term Lanczos recurrences for the relevant components we have

$$\tilde{\rho}_{k+1}\tilde{v}_{k+1}(i) = (\lambda_i - \tilde{\alpha}_k)\tilde{v}_k(i) - \tilde{\rho}_k\tilde{v}_{k-1}(i) - F_k(i,k),$$

$$\tilde{\rho}_{k+1}\bar{v}_{k+1}(i_\ell) = (\bar{\lambda}_{i_\ell} - \tilde{\alpha}_k)\bar{v}_k(i_\ell) - \tilde{\rho}_k\bar{v}_{k-1}(i_\ell).$$

Multiplying the first of these equations by $\tilde{v}_k(i)$ and multiplying the second by $\bar{v}_k(i_\ell)$ and summing over ℓ gives

$$\tilde{\rho}_{k+1}\tilde{v}_{k+1}(i)\tilde{v}_k(i) = (\lambda_i - \tilde{\alpha}_k)(\tilde{v}_k(i))^2 - \tilde{\rho}_k\tilde{v}_k(i)\tilde{v}_{k-1}(i) - F_k(i,k)\tilde{v}_k(i), \tag{61}$$

$$\tilde{\rho}_{k+1}\sum_\ell \bar{v}_{k+1}(i_\ell)\bar{v}_k(i_\ell) = (\lambda_i - \tilde{\alpha}_k)\sum_\ell(\bar{v}_k(i_\ell))^2 - \tilde{\rho}_k\sum_\ell \bar{v}_k(i_\ell)\bar{v}_{k-1}(i_\ell)$$

$$+\sum_\ell(\bar{\lambda}_{i_\ell} - \lambda_i)(\bar{v}_k(i_\ell))^2. \tag{62}$$

Subtract (62) from (61) to obtain

$$\tilde{\rho}_{k+1}d_{k+1,i} = (\lambda_i - \tilde{\alpha}_k)\delta_{k,i} - \tilde{\rho}_k d_{k,i} - F_k(i,k)\tilde{v}_k(i) - \xi_{k,i}\sum_\ell(\bar{v}_k(i_\ell))^2, \tag{63}$$

where we have defined

$$d_{j,i} \equiv \tilde{v}_j(i)\tilde{v}_{j-1}(i) - \sum_\ell \bar{v}_j(i_\ell)\bar{v}_{j-1}(i_\ell),$$

$$\delta_{j,i} \equiv (\tilde{v}_j(i))^2 - \sum_\ell(\bar{v}_j(i_\ell))^2,$$

$$\xi_{j,i} \ : \ \sum_\ell(\bar{\lambda}_{i_\ell} - \lambda_i)(\bar{v}_j(i_\ell))^2 = \xi_{j,i}\sum_\ell(\bar{v}_j(i_\ell))^2.$$

Clearly, $|\delta_{j,i}| \le \delta$ and $|\xi_{j,i}| \le \xi$ for all i and j. Applying formula (63) recursively gives

$$\tilde{\rho}_{k+1}d_{k+1,i} = \sum_{j=1}^{k}(-1)^{k-j}\left[\delta_{j,i}(\lambda_i - \tilde{\alpha}_j) - F_k(i,j)\tilde{v}_j(i) - \xi_{j,i}\sum_\ell(\bar{v}_j(i_\ell))^2\right].$$

Dividing by $\tilde{\rho}_{k+1}$, taking absolute values on each side, and bounding the quantities $|\delta_{j,i}|$, $|F_k(i,j)|$, and $|\xi_{j,i}|$ on the right-hand side gives the desired result (60). □

LEMMA 8. *The residual vectors $\tilde{r}_k^M$ and $\bar{r}_k^M$ are related by*

$$\|\tilde{r}_k^M\| = \|\bar{r}_k^M\| + h_k^M,$$

where

$$|h_k^M| \le \|\tilde{r}_k^M\|\,\frac{1}{\lambda_{\min}^2}\left[\frac{1}{2}N\cdot\delta + \frac{1}{2}d + |\lambda_{\min}|\cdot\frac{\|\zeta\|}{\|\tilde{r}_k^M\|} + |\lambda_{\min}|\cdot\xi\right] + O(\Delta^2), \tag{64}$$

where $\lambda_{\min}$ is the eigenvalue of A of smallest absolute value, δ and ξ are defined by (59) *and* (58), *ζ is given by*

$$\zeta = F_{k+1}\tilde{z}_k^M - AF_k\tilde{y}_k^M - A^2g_k^M,$$

and d satisfies

$$d = \sum_{i=1}^{N}|d_{k+2,i}|,$$

$$d \le \frac{1}{\tilde{\rho}_{k+2}}\left[\delta\cdot 2(k+1)N|\lambda_{\max}| + \xi\cdot(k+1) + f\cdot(k+1)\sqrt{N}\right] \tag{65}$$

with $\lambda_{\max}$ the eigenvalue of A of largest absolute value. The term $O(\Delta^2)$ denotes higher-order terms in δ, ξ, $\|\zeta\|$, and d.

Proof. Equation (55) can be written in component form as

$$\lambda_i \tilde{r}_k^M(i) = \gamma_{k+1}\tilde{v}_{k+1}(i) + \gamma_{k+2}\tilde{v}_{k+2}(i) + \zeta(i), \tag{66}$$

and equation (57) becomes

$$\bar{\lambda}_{i_\ell} \bar{r}_k^M(i_\ell) = \gamma_{k+1}\bar{v}_{k+1}(i_\ell) + \gamma_{k+2}\bar{v}_{k+2}(i_\ell). \tag{67}$$

Squaring both sides in (66) gives

$$\begin{aligned}\lambda_i^2(\tilde{r}_k^M(i))^2 = {} & \gamma_{k+1}^2(\tilde{v}_{k+1}(i))^2 + \gamma_{k+2}^2(\tilde{v}_{k+2}(i))^2 + 2\gamma_{k+1}\gamma_{k+2}\tilde{v}_{k+1}(i)\tilde{v}_{k+2}(i) \\ & + 2\lambda_i \tilde{r}_k^M(i)\zeta(i) - (\zeta(i))^2,\end{aligned} \tag{68}$$

and squaring both sides and summing over ℓ in (67) gives

$$\begin{aligned}\lambda_i^2 \sum_\ell (\bar{r}_k^M(i_\ell))^2 = {} & \gamma_{k+1}^2 \sum_\ell (\bar{v}_{k+1}(i_\ell))^2 + \gamma_{k+2}^2 \sum_\ell (\bar{v}_{k+2}(i_\ell))^2 \\ & + 2\gamma_{k+1}\gamma_{k+2} \sum_\ell \bar{v}_{k+1}(i_\ell)\bar{v}_{k+2}(i_\ell) + \sum_\ell (\lambda_i - \bar{\lambda}_{i_\ell})(\lambda_i + \bar{\lambda}_{i_\ell})(\bar{r}_k^M(i_\ell))^2.\end{aligned} \tag{69}$$

Finally, subtracting (69) from (68) gives

$$\begin{aligned}\lambda_i^2 \left[(\tilde{r}_k^M(i))^2 - \sum_\ell (\bar{r}_k^M(i_\ell))^2 \right] = {} & \gamma_{k+1}^2\delta_{k+1,i} + \gamma_{k+2}^2\delta_{k+2,i} + 2\gamma_{k+1}\gamma_{k+2}d_{k+2,i} \\ & + 2\lambda_i \tilde{r}_k^M(i)\zeta(i) - (\zeta(i))^2 - (\lambda_i - \eta_{k,i})(\lambda_i + \eta_{k,i}) \sum_\ell (\bar{r}_k^M(i_\ell))^2,\end{aligned} \tag{70}$$

where $\eta_{k,i}$ satisfies

$$\sum_\ell (\lambda_i - \bar{\lambda}_{i_\ell})(\lambda_i + \bar{\lambda}_{i_\ell})(\bar{r}_k^M(i_\ell))^2 = (\lambda_i - \eta_{k,i})(\lambda_i + \eta_{k,i}) \sum_\ell (\bar{r}_k^M(i_\ell))^2.$$

By the mean value theorem we have $|\lambda_i - \eta_{k,i}| \le \xi$. Divide each side of (70) by λ_i^2, sum over i, and use the bounds on $|\delta_{j,i}|$, $|d_{j,i}|$, and $|\lambda_i - \eta_{j,i}|$ to obtain

$$\begin{aligned}| \, \|\tilde{r}_k^M\|^2 - \|\bar{r}_k^M\|^2 \, | \le {} & \frac{1}{\lambda_{\min}^2} \left[(\gamma_{k+1}^2 + \gamma_{k+2}^2)n\delta + 2|\gamma_{k+1}\gamma_{k+2}|d \right. \\ & \left. + 2|\lambda_{\min}| \, \|\tilde{r}_k^M\| \, \|\zeta\| + 2|\lambda_{\min}| \, \xi \|\bar{r}_k^M\|^2 \right] + O(\Delta^2).\end{aligned} \tag{71}$$

Use the fact that $\gamma_{k+1}^2 + \gamma_{k+2}^2 = \|\bar{r}_k^M\|^2$ and $2|\gamma_{k+1}\gamma_{k+2}| \le \|\bar{r}_k^M\|^2$ and divide each side in (71) by $\|\tilde{r}_k^M\| + \|\bar{r}_k^M\|$ to obtain the desired result (64). The bound (65) is obtained by summing over i in expression (60). □

THEOREM 4. *The computed Lanczos and MINRES residuals are related by*

$$\begin{aligned}\|\tilde{r}_k^L\| = {} & \frac{\|\tilde{r}_k^M\|}{\sqrt{1 - (\|\tilde{r}_k^M\|/\|\tilde{r}_{k-1}^M\|)^2}} - \frac{h_k^M - h_{k-1}^M(\|\tilde{r}_k^M\|/\|\tilde{r}_{k-1}^M\|)^3}{[1 - (\|\tilde{r}_k^M\|/\|\tilde{r}_{k-1}^M\|)^2]^{3/2}} \\ & + h_k^L + O(\Delta^2).\end{aligned} \tag{72}$$

Proof. The exact arithmetic residual vectors associated with $\bar{A}$ satisfy

$$\|\bar{r}_k^L\| = \frac{\|\bar{r}_k^M\|}{\sqrt{1 - (\|\bar{r}_k^M\|/\|\bar{r}_{k-1}^M\|)^2}},$$

so from Lemmas 4 and 8 we have

$$\|\tilde{r}_k^L\| - h_k^L = \frac{\|\tilde{r}_k^M\| - h_k^M}{\sqrt{1 - ((\|\tilde{r}_k^M\| - h_k^M)/(\|\tilde{r}_{k-1}^M\| - h_{k-1}^M))^2}}.$$

Manipulating this expression gives the result (72). □

Note that Theorem 4 implies that relation (33) holds to a close approximation in finite precision arithmetic, provided the roundoff terms h_k^L, h_k^M, and h_{k-1}^M are much smaller than $\|\tilde{r}_k^M\|$ and provided the reduction factor $\|\tilde{r}_k^M\|/\|\tilde{r}_{k-1}^M\|$ is not too close to one.

7. Finite precision arithmetic, BLanczos/SQMR. If the tridiagonal matrices generated by the Lanczos algorithm are well conditioned, then one can expect the roundoff terms h_k^L and h_k^M in Theorem 4 to be small since the roundoff term F_k in (46) is tiny and g_k^L and g_k^M in (48) will be small if the tridiagonal systems are solved accurately. In this section we show that the even-order tridiagonal matrices generated by the BLanczos and SQMR algorithms described in §5 are essentially as well conditioned as the original matrix A in (1).

For each tridiagonal matrix T_k generated by the BLanczos and SQMR algorithms, let $\lambda_i^{(k)}$, $1 \le i \le k$ denote the eigenvalues of T_k. The proof that the even-order T_{2k} are as well conditioned as A uses the interlacing property of eigenvalues of tridiagonal matrices. (See, for example, [24, p. 46] or, later, [16].) This property says that if T_k is any principal submatrix of a symmetric tridiagonal matrix T, then between each pair of eigenvalues of T_k is at least one eigenvalue of T. Also, there is an eigenvalue of T that is less than the smallest eigenvalue of T_k and an eigenvalue of T that is greater than the largest eigenvalue of T_k. We also need the following lemma.

LEMMA 9 (see [4]). *Each unreduced, even-ordered tridiagonal matrix T_{2k} defined by* (37) *is nonsingular and has eigenvalues that occur in ± pairs. Each odd-order T_{2k+1} is singular and has a simple zero eigenvalue.*

Using these properties and results from [12] relating the tridiagonal matrices generated by a finite precision computation to those that would be generated by an exact calculation for a certain larger matrix with nearby eigenvalues, we are able to prove the following theorem.

THEOREM 5. *Let A be any real nonsymmetric matrix with singular values $0 < \sigma_N \le \sigma_{N-1} \le \cdots \le \sigma_1$. Let $\tilde{T}_{2k}$, $k = 1, 2, \ldots, K$ be the even-ordered tridiagonal matrices generated by applying either BLanczos or SQMR to* (1) *in finite precision arithmetic. Then for all $1 \le j \le K$, the eigenvalues of $\tilde{T}_{2j}$ lie in the intervals $[-\sigma_1 - \xi, -\sigma_N + \xi] \cup [\sigma_N - \xi, \sigma_1 + \xi]$, where ξ is a bound on the distance between the eigenvalues of B in* (34) *and those of a corresponding exact arithmetic matrix $\bar{B}$, as described in the previous section.*

Proof. Since bidiagonalization is an application of the real symmetric Lanczos procedure, the results in [12] are applicable. Therefore, there exists a matrix $\bar{B}$ whose eigenvalues lie in tiny intervals about the eigenvalues of B such that the exact Lanczos algorithm applied to $\bar{B}$ generates tridiagonal matrices $\bar{T}_{2j}$ identical to the tridiagonal

matrices $\tilde{T}_{2j}$, $j = 1, \ldots, K$ generated by the finite precision bidiagonalization process applied to B in (34). Let ξ be a bound on the width of these intervals. Since the Lanczos computation for $\bar{B}$ is exact, there exists some $M \geq 2K$ such that the eigenvalues of $\bar{T} \equiv \bar{T}_M$ are the eigenvalues of $\bar{B}$. It follows from the interlacing theorem that between each pair of eigenvalues of $\tilde{T}_{2j}$ is an eigenvalue of $\bar{T}$ and hence of $\bar{B}$. Additionally, all eigenvalues of $\tilde{T}_{2j}$ must lie between the smallest and largest eigenvalues of $\bar{B}$, $[-\sigma_1 - \xi, \sigma_1 + \xi]$. From Lemma 9 the eigenvalues of T_{2j} occur in $\pm$ pairs. Therefore, if for some j, $\tilde{T}_{2j}$ had an eigenvalue in the interval $(-\sigma_N + \xi, \sigma_N - \xi)$, then it would necessarily have a pair of eigenvalues in this interval and hence $\bar{B}$ would have to have an eigenvalue in this interval. This is a contradiction, and therefore the eigenvalues of each $\tilde{T}_{2j}$ must be contained in the intervals given in the theorem. □

From Theorem 5 it follows that if the original matrix A in (1) is well conditioned, then the error term in Theorem 4 will be small. Therefore, using Lemmas 2 and 3 we get that the BLanczos and the SQMR residual norms will satisfy an approximate relationship of the form (72).

8. Conclusions. In exact arithmetic we have derived a precise relation between the sizes of the Arnoldi and GMRES residuals at any iteration k and between the sizes of the BCG residual and the QMR quasi-residual at any iteration k. This relation implies roughly that if the Galerkin iterates are well defined and if one member of either pair of algorithms converges very well, then the other member of the pair will also converge very well, and if one member performs very poorly then the other member will also perform poorly. While the residual (or a related quantity) in the norm-minimizing method cannot grow, as it can in a Galerkin method, it is no more useful to have a near constant residual norm than it is to have a growing one. If one prefers to see a (weakly) monotonically decreasing convergence curve, one can always plot the norm of the smallest residual obtained so far.

While those proofs assumed exact arithmetic, the relation between GMRES and Arnoldi residual norms can be expected to hold to a close approximation in finite precision arithmetic as well, since orthogonality of the Arnoldi vectors is maintained, or can be maintained, with a sufficiently careful implementation of the algorithm [6].

For the QMR and BCG algorithms, precise details of the implementation and use of look-ahead steps will determine whether or not the relation continues to hold in finite precision arithmetic. If the BCG iterate produced at some step has a very large norm, then future iterates updated from this one may never approach the true solution [14]. This situation can be avoided through the use of look-ahead procedures or by storing certain intermediate quantities and using these to generate the BCG approximations. For example, the BCG iterates can be generated from the QMR iterates [8]. There seems to be little if any reason, however, to choose the Galerkin variant over the norm-minimizing variant when each can be generated with essentially the same amount of work and storage. A possible exception may be the case of very ill conditioned symmetric problems, where it was observed in [20] that the SYMMLQ implementation of the Lanczos algorithm (the symmetric equivalent of BCG, but implemented in such a way that large intermediate iterates are not used to generate future iterates) sometimes attained a higher level of accuracy than the MINRES algorithm (the symmetric equivalent of QMR).

For real symmetric problems with well-conditioned tridiagonal matrices we have shown that, despite the loss of orthogonality of the Lanczos vectors, the relation between the Lanczos and MINRES residuals holds to a close approximation in finite precision arithmetic. We then used this result to prove that if a nonsymmetric

problem (1) is well conditioned, then the residuals generated by the nonsymmetric bidiagonalization algorithms, BLanczos and SQMR, also satisfy these relationships to a close approximation in finite precision arithmetic.

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] C. Brezinski, M. Redivo Zaglia, and H. Sadok, *A breakdown-free Lanczos type algorithm for solving linear systems*, Numer. Math., 63 (1992), pp. 29–38.

[3] P. N. Brown, *A theoretical comparison of the Arnoldi and GMRES algorithms*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 58–78.

[4] J. K. Cullum, *Peaks and plateaus in Lanczos methods for solving nonsymmetric systems of equations $Ax = b$*, IBM research report RC 18084, T. J. Watson Research Center, Yorktown Heights, NY, 1992.

[5] J. K. Cullum and R. A. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations*, Vol. 1, Theory, Progress in Scientific Computing Vol. 3, S. Abarbanel et al. eds., Birkhäuser, Basel, Switzerland, 1985.

[6] J. Drkošová, A. Greenbaum, M. Rozložník, and Z. Strakoš, *Numerical stability of GMRES*, BIT, 35 (1995), pp. 309–330.

[7] R. Fletcher, *Conjugate gradient methods for indefinite systems*, in Numerical Analysis Dundee 1975, G. A. Watson, ed., Lecture Notes in Mathematics 506, Springer-Verlag, Berlin, 1976, pp. 73–89.

[8] 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.

[9] R. W. Freund, *A transpose-free quasi-minimal residual algorithm for non-Hermitian linear systems*, SIAM J. Sci. Comput., 14 (1993), pp. 470–482.

[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] 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.

[12] A. Greenbaum, *Behavior of slightly perturbed Lanczos and conjugate gradient recurrences*, Lin. Algebra Appl., 113 (1989), pp. 7–63.

[13] A. Greenbaum and Z. Strakos, *Predicting the behavior of finite precision Lanczos and conjugate gradient computations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 121–137.

[14] A. Greenbaum, *Accuracy of computed solutions from conjugate-gradient-like methods*, in PCG '94: Advances in Numerical Methods for Large Sparse Sets of Linear Equations, M. Natori and T. Nodera, eds., Keio University, Yokahama, Japan, 1994, pp. 126–138.

[15] M. H. Gutknecht, *Changing the norm in conjugate gradient type algorithms*, SIAM J. Numer. Anal., 30 (1993), pp. 40–56.

[16] R. O. Hill, Jr. and B. N. Parlett, *Refined interlacing properties*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 239–247.

[17] 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.

[18] C. C. Paige, *Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix*, J. Inst. Math. Appl., 18 (1976), pp. 341–349.

[19] ———, *Bidiagonalization of matrices and solution of linear equations*, SIAM J. Numer. Anal., 11 (1974), pp. 197–209.

[20] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.

[21] 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.

[22] Y. Saad, *Krylov subspace methods for solving unsymmetric linear systems*, Math. Comp., 37 (1981), pp. 105–126.

[23] 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.

[24] G. Szego, *Orthogonal Polynomials*, American Mathematical Society Colloquium Publications, Volume XXIII, New York, 1959.

[25] H. A. van der Vorst and C. Vuik, *The superlinear convergence behavior of GMRES*, J. Comp. Appl. Math., 48 (1993), pp. 327–341.

[26] A. VAN DER SLUIS AND H. A. VAN DER VORST, *The convergence behavior of Ritz values in the presence of close eigenvalues*, Lin. Algebra Appl. 88/89 (1987), pp. 651–694.

[27] H. F. WALKER, *Implementation of the GMRES method using Householder transformations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 152–163.

[28] R. WEISS, *Convergence behavior of generalized conjugate gradient methods*, Ph.D. thesis, University of Karlsruhe, Karlsruhe, Germany, 1990.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 2, pp. 248–267, April 1996

COMPUTATION OF NUMERICAL PADÉ–HERMITE AND SIMULTANEOUS PADÉ SYSTEMS I: NEAR INVERSION OF GENERALIZED SYLVESTER MATRICES*

STAN CABAY†, ANTHONY R. JONES‡, AND GEORGE LABAHN§

Abstract. We present new formulae for the "near" inverses of striped Sylvester and mosaic Sylvester matrices. The formulae assume computation over floating-point rather than exact arithmetic domains. The near inverses are expressed in terms of numerical Padé–Hermite systems and simultaneous Padé systems. These systems are approximants for the power series determined from the coefficients of the Sylvester matrices. The inverse formulae provide good estimates for the condition numbers of these matrices and serve as primary tools in a companion paper for the development of a fast, weakly stable algorithm for the computation of Padé–Hermite and simultaneous Padé systems and, thereby, also for the numerical inversion of striped and mosaic Sylvester matrices.

Key words. striped Sylvester inverses, mosaic Sylvester inverses, Padé–Hermite approximants, simultaneous Padé approximants, numerical stability

AMS subject classifications. 41A21, 65F05, 65G05

1. Introduction. Let $n = [n_0, \ldots, n_k]$ be a vector of integers with $n_\beta \geq 0$, $0 \leq \beta \leq k$. A striped Sylvester matrix of order $\|n\|$ is given by

$$\mathcal{M}_n = \left[\begin{array}{ccc|c|ccc} a_0^{(0)} & & & & a_k^{(0)} & & \\ & \ddots & & & & \ddots & \\ \vdots & & a_0^{(0)} & \cdots & \vdots & & a_k^{(0)} \\ & & \vdots & & & & \vdots \\ a_0^{(\|n\|-1)} & \cdots & a_0^{(\|n\|-n_0)} & & a_k^{(\|n\|-1)} & \cdots & a_k^{(\|n\|-n_k)} \end{array}\right], \tag{1}$$

where $\|n\| = n_0 + \cdots + n_k$ and where the coefficients $a_\beta^{(\ell)} \in \mathcal{R}$, the field of real numbers. Assume that $a_0^{(0)} \neq 0$. In this paper, we present a formula for the inverse of $\mathcal{M}_n$ expressed in terms of Padé–Hermite and simultaneous Padé systems.

Padé–Hermite and simultaneous Padé systems [6, 8] are approximants for the vector $A^t(z) = [a_0(z), \ldots, a_k(z)]$ of power series associated with the coefficients of $\mathcal{M}_n$; namely,

$$a_\beta(z) = \sum_{\ell=0}^{\|n\|-1} a_\beta^{(\ell)} z^\ell, \quad \text{with } 0 \leq \beta \leq k.$$

They provide necessary and sufficient conditions for $\mathcal{M}_n$ to be nonsingular and generalize the notions of Padé–Hermite and simultaneous Padé approximants. Padé–Hermite and simultaneous Padé approximants were introduced by Hermite in the last

* Received by the editors May 31, 1994; accepted for publication (in revised form) by M. Gutknecht May 2, 1995.

† Department of Computing Science, University of Alberta, Edmonton, Alberta T6G 2H1, Canada (cabay@cs.ualberta.ca). The research of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant A8035.

‡ Bell Northern Research, P. O. Box 3511, Station C, Ottawa, Ontario K1Y 4H7, Canada (anthonyj@bnr.ca).

§ Department of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada (glabahn@daisy.uwaterloo.ca). The research of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant FS1525C.

century and have been widely studied by several authors (for a bibliography, see, for example [22, 1, 2, 3, 14]).

The inverse formula given here is intended to be applied in a numerical setting; it accommodates errors that may have been made in the computation of Padé–Hermite and simultaneous Padé systems. That is, the formula gives the "near" inverse for $\mathcal{M}_n$ because it expresses the inverse in terms of Padé–Hermite and simultaneous Padé systems which are computed using floating-point arithmetic. There are other closed formulae (cf. [11, 15, 17, 18, 19]) for $\mathcal{M}_n^{-1}$. The formula given here is different in that it expresses the inverse directly in terms of numerical Padé–Hermite and simultaneous Padé systems.

The near inverse formula depends on the computation of both Padé systems. It is possible to determine a simultaneous Padé system from its "dual" Padé–Hermite system via the adjoint operation [5, 14]. In a numerical setting, however, this does not provide enough control over the resulting floating-point errors [13]. Instead, simultaneous Padé systems can be computed independently. Whereas a Padé–Hermite system can be obtained by solving a system of linear equations with $\mathcal{M}_n$ as the coefficient matrix, a simultaneous Padé system can be similarly and independently obtained with a coefficient matrix that is a specialized mosaic Sylvester matrix. This specialized mosaic Sylvester matrix of order $k\|n\|$ is defined to be

$$\mathcal{M}_n^* = \begin{bmatrix} \mathcal{S}_{0,1}^* & \cdots & \mathcal{S}_{0,k}^* \\ \vdots & & \vdots \\ \mathcal{S}_{k,1}^* & \cdots & \mathcal{S}_{k,k}^* \end{bmatrix}, \tag{2}$$

where $\mathcal{S}_{\alpha,\beta}^*$ are matrices of size $(\|n\| - n_\alpha) \times \|n\|$, with

$$\mathcal{S}_{0,\beta}^* = -\begin{bmatrix} a_\beta^{(0)} & & \cdots & a_\beta^{(\|n\|-1)} \\ & \ddots & & \vdots \\ & & a_\beta^{(0)} \quad \cdots & a_\beta^{(n_0)} \end{bmatrix},$$

$$\mathcal{S}_{\beta,\beta}^* = \begin{bmatrix} a_0^{(0)} & & \cdots & a_0^{(\|n\|-1)} \\ & \ddots & & \vdots \\ & & a_0^{(0)} \quad \cdots & a_0^{(n_\beta)} \end{bmatrix},$$

for $1 \le \beta \le k$, and with the remaining $\mathcal{S}_{\alpha,\beta}^* = 0$. The matrix $\mathcal{M}_n^*$ is closely related to $\mathcal{M}_n$. Indeed, in the special case when $k = 1$, the matrix $\mathcal{M}_n$ and the transpose of $\mathcal{M}_n^*$ coincide up to a sign and a permutation of the rows and columns. Similarly, when $k \ge 1$ and $a_0(z) = 1$, the matrix $\mathcal{M}_n$ and the transpose of $\mathcal{M}_n^*$ are both obtained by a suitable block extension of the same matrix. In this paper, we also provide a "near" inverse formula for the matrix $\mathcal{M}_n^*$, again in terms of numerical Padé–Hermite and simultaneous Padé systems.

The inverse formulae provide "good" estimates for the condition numbers of $\mathcal{M}_n$ and $\mathcal{M}_n^*$ and these enable the stable numerical computation of Padé–Hermite and simultaneous Padé systems described in the companion paper [5]. In return, the accurate computation of Padé–Hermite and simultaneous Padé systems by the algorithm in [5] enables the effective inversion of generalized Sylvester matrices by the formulae given in this paper, as well as the solution of linear systems of equations with these as the coefficient matrices.

This paper is organized as follows. Preliminary definitions and basic facts about Padé–Hermite and simultaneous Padé systems are given in the next two sections. Section 3 gives a near commutativity relationship between these two systems in floating-point arithmetic, while §§4 and 5 give the approximate inversion formulae for striped and mosaic Sylvester matrices. The final section gives a summary and some conclusions.

We conclude this section by defining some norms which are used in §§4 and 5. Let

$$a(z) = \sum_{\ell=0}^{\infty} a^{(\ell)} z^{\ell} \in \mathcal{R}[[z]],$$

where $\mathcal{R}[[z]]$ is the domain of power series with coefficients from $\mathcal{R}$, and define the bounded power series

$$\mathcal{R}^B[[z]] = \left\{ a(z) \;\middle|\; a(z) \in \mathcal{R}[[z]],\ \sum_{\ell=0}^{\infty} |a^{(\ell)}| < \infty \right\}.$$

A norm for $a(z) \in \mathcal{R}^B[[z]]$ is

$$\|a(z)\| = \sum_{\ell=0}^{\infty} |a^{(\ell)}|.$$

$\mathcal{R}^B[[z]]$ includes the domain of polynomials $\mathcal{R}[z]$. So, for

$$s(z) = \sum_{\ell=0}^{\partial} s^{(\ell)} z^{\ell} \in \mathcal{R}[z],$$

we use the norm

$$\|s(z)\| = \sum_{\ell=0}^{\partial} |\, s^{(\ell)} \,|.$$

For vectors and matrices over $\mathcal{R}^B[[z]]$, we use the 1-norm unless otherwise specified. So, for example, the norm for $A^t(z)$ is

$$\|A^t(z)\| = \max_{0 \le \beta \le k} \{\|a_\beta(z)\|\}$$

and the norm for $S(z) \in \mathcal{R}_{(k+1)\times(k+1)}[z]$ is

$$\|S(z)\| = \max_{0 \le \beta \le k} \left\{ \sum_{\alpha=0}^{k} \|S_{\alpha,\beta}(z)\| \right\}.$$

It is easy to verify that various compatibility conditions are satisfied. For example,

$$\|A^t(z) \cdot S(z)\| \le \|A^t(z)\| \cdot \|S(z)\|$$

and

$$\|a(z) \cdot b(z)\| \le \|a(z)\| \cdot \|b(z)\|,$$

where $b(z)$ is also a bounded power series. In addition, for $S^*(z) \in \mathcal{R}_{(k+1)\times(k+1)}[z]$ and $A^*(z) \in \mathcal{R}^B_{(k+1)\times k}[[z]]$,

$$\|S^*(z) \cdot A^*(z)\| \leq \|S^*(z)\| \cdot \|A^*(z)\|,$$

$$\|S(z) \cdot S^*(z)\| \leq \|S(z)\| \cdot \|S^*(z)\|.$$

In the subsequent development, we also make use of the inequality

$$\|a(z) \pmod{z^{\|n\|+1}}\| \leq \|a(z)\|,$$

where

$$a(z) \pmod{z^{\|n\|+1}} = \sum_{\ell=0}^{\|n\|} a^{(\ell)} z^\ell + \sum_{\ell=\|n\|+1}^{\infty} 0 \cdot z^\ell \in \mathcal{R}^B[[z]].$$

2. Multidimensional Padé systems. In this section, we introduce the notion of Padé–Hermite and simultaneous Padé systems. Let $n = [n_0, \ldots, n_k]$ and define $\|n\| = n_0 + \cdots + n_k$. Also let

$$A^t(z) = [a_0(z), \ldots, a_k(z)],$$

where

$$a_\beta(z) = \sum_{\ell=0}^{\infty} a_\beta^{(\ell)} z^\ell, \qquad \beta = 0, \ldots, k,$$

with $a_\beta^{(\ell)} \in \mathcal{R}$, the field of real numbers. Assume that $a_0^{(0)} \neq 0$, which means that $a_0^{-1}(z)$ exists. Assume also that $A^t(z)$ is scaled so that $\|a_\beta(z) \pmod{z^{\|n\|+1}}\| = 1$, $0 \leq \beta \leq k$.

The $(k+1) \times (k+1)$ matrix of polynomials

$$S(z) = \left[\begin{array}{c|c} zp(z) & U^t(z) \\ \hline zQ(z) & V(z) \end{array}\right] = \left[\begin{array}{c|ccc} zp(z) & u_1(z) & \cdots & u_k(z) \\ \hline zq_1(z) & v_{1,1}(z) & \cdots & v_{1,k}(z) \\ \vdots & \vdots & & \vdots \\ zq_k(z) & v_{k,1}(z) & \cdots & v_{k,k}(z) \end{array}\right] \tag{3}$$

is a numerical Padé–Hermite system (NPHS) [8] of type n for $A(z)$ if the following conditions are satisfied.

I (*Degree conditions*). For $1 \leq \alpha, \beta \leq k$,

$$p(z) = \sum_{\ell=0}^{n_0-1} p^{(\ell)} z^\ell, \qquad u_\beta(z) = \sum_{\ell=0}^{n_0} u_\beta^{(\ell)} z^\ell, \tag{4}$$

$$q_\alpha(z) = \sum_{\ell=0}^{n_\alpha-1} q_\alpha^{(\ell)} z^\ell, \qquad v_{\alpha,\beta}(z) = \sum_{\ell=0}^{n_\alpha} v_{\alpha,\beta}^{(\ell)} z^\ell.$$

II (*Order condition*).

$$A^t(z)S(z) = z^{\|n\|}T^t(z) + \delta T^t(z), \tag{5}$$

where $T^t(z) = [r(z), zW^t(z)]$ with $W^t(z) = [w_1(z), \ldots, w_k(z)]$ is the residual, and where $\delta T^t(z) = [z\,\delta r(z), \delta W^t(z)]$ is the residual error, with $\delta W^t(z) = [\delta w_1(z), \ldots, \delta w_k(z)]$ and with

$$\delta r(z) = \sum_{\ell=0}^{\|n\|-2} \delta r^{(\ell)}\, z^\ell, \quad \delta w_\beta(z) = \sum_{\ell=0}^{\|n\|} \delta w_\beta^{(\ell)}\, z^\ell.$$

If $\delta T^t(z) = 0$, then $S(z)$ is an exact, rather than a numerical, Padé–Hermite system.

III (*Nonsingularity condition*). The constant term of $V(z)$ is a diagonal matrix,

$$V(0) = \text{diag}\ [\gamma_1, \ldots, \gamma_k], \tag{6}$$

and

$$\gamma \equiv (a_0^{(0)})^{-1} \prod_{\alpha=0}^{k} \gamma_\alpha \neq 0, \tag{7}$$

where $\gamma_0 = r(0)$.

Remark 1. The nonsingularity condition III is equivalent to the condition that $r(0) \neq 0$ and that $V(0)$ be a nonsingular diagonal matrix.

Remark 2. The NPHS is said to be *normalized* [8] if the nonsingularity condition III is replaced by $r(0) = 1$ and $V(0) = I_k$. This can be achieved by multiplying $S(z)$ on the right by Γ^{-1}, where

$$\Gamma = \text{diag}\,[\gamma_0, \ldots, \gamma_k]. \tag{8}$$

The NPHS is said to be *scaled* [13] if each column of $S(z)$ has norm equal to 1 and if, in addition, $\gamma_\beta > 0$, $0 \leq \beta \leq k$. Here, also, scaling an NPHS is accomplished by multiplying it on the right by an appropriate diagonal matrix.

Remark 3. The nonsingularity condition III, namely $\gamma \neq 0$, refers to the nonsingularity of the associated striped Sylvester matrix $\mathcal{M}_n$ defined in (1); in [8] it is shown that an exact NPHS exists iff $\mathcal{M}_n$ is nonsingular.

Remark 4. In [4, 5, 8], rather than $S(z)$, the Padé-Hermite system is defined to be $S(z) \cdot \text{diag}[z, 1, \ldots, 1]$. The notation used here is consistent with that of [15] and simplifies the presentation of some of the results.

A Padé–Hermite system gives an approximation to a vector of formal power series using matrix multiplication on the right. A simultaneous Padé system is a similar approximation using matrix multiplication on the left and with degree constraints that can be thought of as being "dual" to the degree constraints of a Padé–Hermite system.

Let[1]

$$A^*(z) = \left[\begin{array}{ccc} -a_1(z) & \cdots & -a_k(z) \\ \hline a_0(z) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & a_0(z) \end{array}\right] \tag{9}$$

be a $(k+1) \times k$ matrix of power series. The $(k+1) \times (k+1)$ matrix of polynomials

$$S^*(z) = \left[\begin{array}{c|c} v^*(z) & U^{*t}(z) \\ \hline zQ^*(z) & zP^*(z) \end{array}\right] = \left[\begin{array}{c|ccc} v^*(z) & u_1^*(z) & \cdots & u_k^*(z) \\ \hline zq_1^*(z) & zp_{1,1}^*(z) & \cdots & zp_{1,k}^*(z) \\ \vdots & \vdots & & \vdots \\ zq_k^*(z) & zp_{k,1}^*(z) & \cdots & zp_{k,k}^*(z) \end{array}\right] \tag{10}$$

is a numerical simultaneous Padé system (NSPS) [6, 8] of type n for $A^*(z)$ if the following conditions are satisfied.

I (*Degree conditions*). For $1 \leq \alpha, \beta \leq k$,

$$\begin{aligned} v^*(z) &= \sum_{\ell=0}^{\|n\|-n_0} v^{*(\ell)} z^\ell, & u_\beta^*(z) &= \sum_{\ell=0}^{\|n\|-n_\beta} u_\beta^{*(\ell)} z^\ell, \\ q_\alpha^*(z) &= \sum_{\ell=0}^{\|n\|-n_0-1} q_\alpha^{*(\ell)} z^\ell, & p_{\alpha,\beta}^*(z) &= \sum_{\ell=0}^{\|n\|-n_\beta-1} p_{\alpha,\beta}^{*(\ell)} z^\ell. \end{aligned} \tag{11}$$

II (*Order condition*).

$$S^*(z)A^*(z) = z^{\|n\|}T^*(z) + \delta T^*(z), \tag{12}$$

where $T^{*t}(z) = [z\, W^*(z) | R^{*t}(z)]$ with $R^*(z)$ a $k \times k$ matrix is the residual, and where $\delta T^{*t}(z) = [\delta W^*(z) | z\, \delta R^{*t}(z)]$ is the residual error, with $\delta R^*(z)$ a $k \times k$ matrix and

$$\delta w_\beta^*(z) = \sum_{\ell=0}^{\|n\|} \delta w_\beta^{*(\ell)}\, z^\ell, \quad \delta r_{\alpha,\beta}^*(z) = \sum_{\ell=0}^{\|n\|-2} \delta r_{\alpha,\beta}^{*(\ell)}\, z^\ell.$$

If $\delta T^*(z) = 0$, then $S^*(z)$ is an exact NSPS.

III (*Nonsingularity condition*). The constant term of $R^*(z)$ is a diagonal matrix

$$R^*(0) = \operatorname{diag}\, [\gamma_1^*, \ldots, \gamma_k^*], \tag{13}$$

and

$$\gamma^* \equiv (a_0^{(0)})^{-1} \prod_{\alpha=0}^{k} \gamma_\alpha^* \neq 0, \tag{14}$$

[1] More generally,

$$A^*(z) = \left[\begin{array}{ccc} a_{0,1}^*(z) & \cdots & a_{0,k}^*(z) \\ \hline a_{1,1}^*(z) & \cdots & a_{1,k}^*(z) \\ \vdots & & \vdots \\ a_{k,1}^*(z) & \cdots & a_{k,k}^*(z) \end{array}\right] = \left[\begin{array}{c} B^{*t}(z) \\ \hline C^*(z) \end{array}\right]$$

with $C^*(0)$ nonsingular, but the restriction to (9), which algebraically is made without loss of generality, permits us to establish in §3 a duality relationship between Padé–Hermite systems and simultaneous Padé systems.

where $\gamma_0^* = v^*(0)$.

Remark 5. The NSPS is said to be *normalized* [6] if the nonsingularity condition III is replaced by $v^*(0) = 1$ and $R^*(0) = I_k$. This can be achieved by multiplying $S^*(z)$ on the left by Γ^{*-1}, where

$$\Gamma^* = \operatorname{diag}[\gamma_0^*, \ldots, \gamma_k^*]. \tag{15}$$

The NSPS is said to be *scaled* when each row of $S^*(z)$ has norm equal to 1 and if, in addition, $\gamma_\alpha^* > 0$, $0 \le \alpha \le k$. Here, also, scaling a NSPS is accomplished by multiplying it on the left by an appropriate diagonal matrix.

Remark 6. The nonsingularity condition III, namely $\gamma^* \neq 0$, refers to the nonsingularity of the associated mosaic Sylvester matrix $\mathcal{M}_n^*$ defined in (2); in [6] it is shown that an exact NSPS exists iff $\mathcal{M}_n^*$ is nonsingular.

Remark 7. In [4, 5, 8], rather than $S^*(z)$, the simultaneous Padé system is defined to be $\operatorname{diag}[1, z, \ldots, z] \cdot S^*(z)$. This is for notational convenience only.

3. Duality. Theorem 1 below gives a relationship between Padé–Hermite and simultaneous Padé systems which is crucial to the results of the subsequent sections. It generalizes earlier results of Mahler and their extensions to block matrices [9, 14, 16, 20].

THEOREM 1. *If $S(z)$ is an NPHS of type n for $A(z)$ and $S^*(z)$ is an NSPS of type n for $A^*(z)$, then*

$$S^*(z) \cdot S(z) = z^{\|n\|}(a_0^{(0)})^{-1}\Gamma^*\Gamma + \theta_I(z), \tag{16}$$

where

$$\theta_I(z) = a_0^{-1}(z)\left\{\begin{bmatrix} v^*(z) \\ zQ^*(z) \end{bmatrix} \delta T^t(z) + \delta T^*(z) \begin{bmatrix} zQ(z) \mid V(z) \end{bmatrix}\right\} \pmod{z^{\|n\|+1}}.$$

Proof. The theorem (in the case that $\delta T(z) = 0$ and $\delta T^*(z) = 0$) follows from [14]. The arguments used in the following proof, however, are considerably simpler. Let

$$B^t(z) = [a_1(z), \ldots, a_k(z)].$$

Then, using (5) and (12),

$$\begin{aligned}
&a_0(z)\, S^*(z) \cdot S(z) \\
&= a_0(z)\left\{\begin{bmatrix} v^*(z) \\ zQ^*(z) \end{bmatrix} \begin{bmatrix} zp(z) \mid U^t(z) \end{bmatrix} + \begin{bmatrix} U^{*t}(z) \\ zP^*(z) \end{bmatrix} \begin{bmatrix} zQ(z) \mid V(z) \end{bmatrix}\right\} \\
&= \begin{bmatrix} v^*(z) \\ zQ^*(z) \end{bmatrix} \{a_0(z) \begin{bmatrix} zp(z) \mid U^t(z) \end{bmatrix} + B^t(z) \begin{bmatrix} zQ(z) \mid V(z) \end{bmatrix}\} \\
&\quad + \left\{a_0(z) \begin{bmatrix} U^{*t}(z) \\ zP^*(z) \end{bmatrix} - \begin{bmatrix} v^*(z) \\ zQ^*(z) \end{bmatrix} B^t(z)\right\} \begin{bmatrix} zQ(z) \mid V(z) \end{bmatrix} \\
&= \begin{bmatrix} v^*(z) \\ zQ^*(z) \end{bmatrix} A^t(z)S(z) + S^*(z)A^*(z) \begin{bmatrix} zQ(z) \mid V(z) \end{bmatrix} \\
&= z^{\|n\|}\left\{\begin{bmatrix} v^*(z) \\ zQ^*(z) \end{bmatrix} \begin{bmatrix} r(z) \mid W^t(z) \end{bmatrix} + \begin{bmatrix} zW^{*t}(z) \\ R^*(z) \end{bmatrix} \begin{bmatrix} zQ(z) \mid V(z) \end{bmatrix}\right\} \\
&\quad + \begin{bmatrix} v^*(z) \\ zQ^*(z) \end{bmatrix} \delta T^t(z) + \delta T^*(z) \begin{bmatrix} zQ(z) \mid V(z) \end{bmatrix}.
\end{aligned} \tag{17}$$

But, from (4) and (11), the degrees of $S^*(z)S(z)$ are bounded componentwise by $\|n\|$. It then follows from (17) that

$$\begin{aligned} S^*(z)S(z) &= z^{\|n\|}(a_0^{(0)})^{-1}\left[\begin{array}{c|c} v^*(0)r(0) & 0 \\ \hline 0 & R^*(0)V(0) \end{array}\right] + \theta_I(z) \\ &= z^{\|n\|}(a_0^{(0)})^{-1}\Gamma^*\Gamma + \theta_I(z), \end{aligned}$$

which is (16). □

COROLLARY 2. *If $S(z)$ is a* normalized *NPHS of type n for $A(z)$ and $S^*(z)$ is a* normalized *NSPS of type n for $A^*(z)$, then for sufficiently small*[2] *$\delta T(z)$ and $\delta T^*(z)$*

$$S(z)\cdot S^*(z) = z^{\|n\|}(a_0^{(0)})^{-1}I_{k+1} + \theta_{II}(z), \tag{18}$$

where

$$\theta_{II}(z) = S(z)\,\theta_I(z)\,[z^{\|n\|}(a_0^{(0)})^{-1}I_{k+1} + \theta_I(z)]^{-1}\,S^*(z).$$

Proof. If $\theta_I(z)$ is so small, that is, if $\delta T(z)$ and $\delta T^*(z)$ are so small, that $z^{\|n\|}(a_0^{(0)})^{-1}I_{k+1} + \theta_I(z)$ is nonsingular, then from (16),

$$S^{*-1}(z) = S(z)\cdot[z^{\|n\|}(a_0^{(0)})^{-1}I_{k+1} + \theta_I(z)]^{-1}.$$

So,

$$\begin{aligned} &S(z)S^*(z) - z^{\|n\|}(a_0^{(0)})^{-1}I_{k+1} \\ &= \left\{S(z) - z^{\|n\|}(a_0^{(0)})^{-1}S^{*-1}(z)\right\}S^*(z) \\ &= S(z)\left\{I_{k+1} - z^{\|n\|}(a_0^{(0)})^{-1}[z^{\|n\|}(a_0^{(0)})^{-1}I_{k+1} + \theta_I(z)]^{-1}\right\}S^*(z) \\ &= S(z)\,\theta_I(z)\,[z^{\|n\|}(a_0^{(0)})^{-1}I_{k+1} + \theta_I(z)]^{-1}\,S^*(z). \qquad \square \end{aligned}$$

COROLLARY 3. *The residuals $T(z)$ for a* normalized *NPHS of type n for $A(z)$ and $T^*(z)$ for a* normalized *NSPS of type n for $A^*(z)$ satisfy*

$$T^t(z)\,S^*(z) = (a_0^{(0)})^{-1}\,A^t(z) + \theta_{III}^t(z), \tag{19}$$

where

$$\theta_{III}^t(z) = \left\{A^t(z)\theta_{II}(z) - \delta T^t(z)S^*(z)\right\}/z^{\|n\|}.$$

Proof. From (5) and (18), it follows that

$$\begin{aligned} \left\{z^{\|n\|}\,T^t(z) + \delta T^t(z)\right\}S^*(z) &= A^t(z)\,S(z)\,S^*(z) \\ &= A^t(z)\left\{z^{\|n\|}\,(a_0^{(0)})^{-1}I_{k+1} + \theta_{II}(z)\right\} \end{aligned}$$

and so (19) is true. □

[2] It is adequate, for example, that condition (34) of Corollary 6 be satisfied.

4. The inverse of a striped Sylvester matrix. In this section, a formula is given for the inverse of $\mathcal{M}_n$ expressed in terms of both $S(z)$ and $S^*(z)$. This enables estimating the condition number of $\mathcal{M}_n$ without explicitly computing $\mathcal{M}_n^{-1}$.

Associated with the NPHS $S(z)$, define the order $\|n\|$ matrices

$$\mathcal{P} = \left[\begin{array}{ccc|ccc|c|ccc} p^{(0)} & \cdots & p^{(n_0-1)} & q_1^{(0)} & \cdots & q_1^{(n_1-1)} & & q_k^{(0)} & \cdots & q_k^{(n_k-1)} \\ \vdots & \iddots & 0 & \vdots & \iddots & 0 & & \vdots & \iddots & 0 \\ p^{(n_0-1)} & \iddots & & q_1^{(n_1-1)} & \iddots & & \cdots & q_k^{(n_k-1)} & \iddots & \\ 0 & & \vdots & 0 & & \vdots & & 0 & & \vdots \\ \vdots & & & \vdots & & & & \vdots & & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & & 0 & \cdots & 0 \end{array}\right] \tag{20}$$

and, for $\beta = 1, 2, \ldots, k$,

$$\mathcal{U}_\beta = \left[\begin{array}{ccc|ccc|c|ccc} u_\beta^{(1)} & \cdots & u_\beta^{(n_0)} & v_{1,\beta}^{(1)} & \cdots & v_{1,\beta}^{(n_1)} & & v_{k,\beta}^{(1)} & \cdots & v_{k,\beta}^{(n_k)} \\ \vdots & \iddots & 0 & \vdots & \iddots & 0 & & \vdots & \iddots & 0 \\ u_\beta^{(n_0)} & \iddots & & v_{1,\beta}^{(n_1)} & \iddots & & \cdots & v_{k,\beta}^{(n_k)} & \iddots & \\ 0 & & \vdots & 0 & & \vdots & & 0 & & \vdots \\ \vdots & & & \vdots & & & & \vdots & & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & & 0 & \cdots & 0 \end{array}\right]. \tag{21}$$

Also, for any power series $a(z) = \sum_{\ell=0}^{\infty} a^{(\ell)} z^\ell$, and any integer function $f(i,j), i, j = 1, 2, \ldots$, let $\left[a^{(f(i,j))}\right]$ denote a matrix of order $\|n\|$ whose element in position (i,j) is $a^{(f(i,j))}$.

The main result of this section is Theorem 4 which gives the inverse of $\mathcal{M}_n$ in terms of the NPHS $S(z)$ and the NSPS $S^*(x)$ of types n for $A(z)$.

THEOREM 4. *In terms of the* normalized *NPHS $S(z)$ and the* normalized *NSPS $S^*(x)$ of types n for $A(z)$, the inverse of $\mathcal{M}_n$ satisfies*

$$\mathcal{M}_n^{-1}\left\{\left[a_0^{(i-j)}\right] + \theta_{IV}\right\} = a_0^{(0)}\left\{\mathcal{P}^t\left[v^{*(\|n\|-i-j+1)}\right] + \sum_{\beta=1}^{k} \mathcal{U}_\beta^t\left[q_\beta^{*(\|n\|-i-j)}\right]\right\}, \tag{22}$$

where

$$\theta_{IV} = a_0^{(0)}\left\{\left[(\theta_{III})_0^{(i-j)}\right] - \sum_{\alpha=0}^{k}\left[a_\alpha^{(\|n\|+i-j)}\right]\left[(\theta_{II})_{\alpha,0}^{(i-j)}\right] + \left[\delta r^{(i+j-2)}\right]\left[v^{*(\|n\|-i-j+1)}\right] + \sum_{\beta=1}^{k}\left[\delta w_\beta^{(i+j-1)}\right]\left[q_\beta^{*(\|n\|-i-j)}\right]\right\}.$$

Proof. The coefficient of z^{i+j-2}, for $i, j = 1, 2, \ldots, \|n\|$, in the first component of (5), namely,

$$a_0(z)\, p(z) + \sum_{\alpha=1}^{k} a_\alpha(z)\, q_\alpha(z) = z^{\|n\|-1} r(z) + \delta r(z),$$

is

$$\sum_{\ell=0}^{n_0} a_0^{(i+j-\ell-2)} p^{(\ell)} + \sum_{\alpha=1}^{k} \sum_{\ell=0}^{n_\alpha - 1} a_\alpha^{(i+j-\ell-2)} q_\alpha^{(\ell)} = r^{(-\|n\|+i+j-1)} + \delta r^{(i+j-2)}.$$

This is the (i,j)th component of

$$\left[r^{(-\|n\|+i+j-1)}\right] + \left[\delta r^{(i+j-2)}\right] = \left[a_0^{(\|n\|+i-j)}\right] \left[p^{(-\|n\|+i+j-2)}\right] + \sum_{\alpha=1}^{k} \left[a_\alpha^{(\|n\|+i-j)}\right] \left[q_\alpha^{(-\|n\|+i+j-2)}\right] + \mathcal{M}_n \mathcal{P}^t. \tag{23}$$

Similarly, the coefficient of z^{i+j-1}, for $i,j = 1,2,\ldots,\|n\|$, in the $(\beta+1)$st component, $\beta = 1,\ldots,k$, of (5), namely,

$$a_0(z)\, u_\beta(z) + \sum_{\alpha=1}^{k} a_\alpha(z)\, v_{\alpha,\beta}(z) = z^{\|n\|+1} w_\beta(z) + \delta w_\beta(z),$$

is

$$\sum_{\ell=0}^{n_0} a_0^{(i+j-\ell-1)} u_\beta^{(\ell)} + \sum_{\alpha=1}^{k} \sum_{\ell=0}^{n_\alpha} a_\alpha^{(i+j-\ell-1)} v_{\alpha,\beta}^{(\ell)} = w_\beta^{(-\|n\|+i+j-2)} + \delta w_\beta^{(i+j-1)}.$$

This is the (i,j)th component of

$$\left[w_\beta^{(-\|n\|+i+j-2)}\right] + \left[\delta w_\beta^{(i+j-1)}\right] = \left[a_0^{(\|n\|+i-j)}\right] \left[u_\beta^{(-\|n\|+i+j-1)}\right] + \sum_{\alpha=1}^{k} \left[a_\alpha^{(\|n\|+i-j)}\right] \left[v_{\alpha,\beta}^{(-\|n\|+i+j-1)}\right] + \mathcal{M}_n \mathcal{U}_\beta^t. \tag{24}$$

Next, the coefficient of z^{i-j-1} for $i,j = 1,\ldots,\|n\|$, in the first row and first column of (18) for a normalized NPHS and a normalized NSPS, namely,

$$p(z)\, v^*(z) + \sum_{\beta=1}^{k} u_\beta(z) q_\beta^*(z) = z^{\|n\|-1} (a_0^{(0)})^{-1} + z^{-1} (\theta_{II})_{0,0}(z),$$

is

$$\sum_{\ell=0}^{n_0-1} v^{*(i-j-\ell-1)}\, p^{(l)} + \sum_{\beta=1}^{k} \sum_{\ell=0}^{n_0} q_\beta^{*(i-j-\ell-1)} u_\beta^{(\ell)} = (\theta_{II})_{0,0}^{(i-j)}.$$

This is the (i,j)th component of

$$\left[p^{(-\|n\|+i+j-2)}\right] \left[v^{*(\|n\|-i-j+1)}\right] + \sum_{\beta=1}^{k} \left[u_\beta^{(-\|n\|+i+j-1)}\right] \left[q_\beta^{*(\|n\|-i-j)}\right] = \left[(\theta_{II})_{0,0}^{(i-j)}\right]. \tag{25}$$

The coefficient of z^{i-j-1} in the first column and the $(\alpha+1)$st row $\alpha = 1,\ldots,k$ of (18), namely,

$$q_\alpha(z)\, v^*(z) + \sum_{\beta=1}^{k} v_{\alpha,\beta}(z) q_\beta^*(z) = z^{-1} (\theta_{II})_{\alpha,0}(z),$$

is

$$\sum_{\ell=0}^{n_\alpha} v^{*(i-j-\ell-1)}\, q_\alpha^{(l)} + \sum_{\beta=1}^{k}\sum_{\ell=0}^{n_\alpha} q_\beta^{*(i-j-\ell-1)} v_{\alpha,\beta}^{(\ell)} = (\theta_{II})_{\alpha,0}^{(i-j)}.$$

This is the (i,j)th component of

$$\left[q_\alpha^{(-\|n\|+i+j-2)}\right]\left[v^{*(\|n\|-i-j+1)}\right] + \sum_{\beta=1}^{k}\left[v_{\alpha,\beta}^{(-\|n\|+i+j-1)}\right]\left[q_\beta^{*(\|n\|-i-j)}\right] = \left[(\theta_{II})_{\alpha,0}^{(i-j)}\right]. \tag{26}$$

Also, the coefficient of z^{i-j}, for $i,j = 1,\ldots,\|n\|$ in the first component of (19) for a normalized NPHS and NSPS, namely,

$$r(z)v^*(z) + z^2\sum_{\beta=1}^{k} w_\beta(z)q_\beta^*(z) = (a_0^{(0)})^{-1}\, a_0(z) + (\theta_{III})_0(z)$$

is the (i,j)th component of

$$(a_0^{(0)})^{-1}\left[a_0^{(i-j)}\right] + \left[(\theta_{III})_0^{(i-j)}\right] = \left[r^{(-\|n\|+i+j-1)}\right]\left[v^{*(\|n\|-i-j+1)}\right] + \sum_{\beta=1}^{k}\left[w_\beta^{(-\|n\|+i+j-2)}\right]\left[q_\beta^{*(\|n\|-i-j)}\right]. \tag{27}$$

We are finally ready to prove the theorem. From (23)–(27),

$$\begin{aligned}
&\mathcal{M}_n\left\{\mathcal{P}^t\left[v^{*(\|n\|-i-j+1)}\right] + \sum_{\beta=1}^{k}\mathcal{U}_\beta^t\left[q_\beta^{*(\|n\|-i-j)}\right]\right\}\\
&= \left\{\left[r^{(-\|n\|+i+j-1)}\right] + \left[\delta r^{(i+j-2)}\right] - \left[a_0^{(\|n\|+i-j)}\right]\left[p^{(-\|n\|+i+j-2)}\right]\right.\\
&\quad \left. - \sum_{\alpha=1}^{k}\left[a_\alpha^{(\|n\|+i-j)}\right]\left[q_\alpha^{(-\|n\|+i+j-2)}\right]\right\}\left[v^{*(\|n\|-i-j+1)}\right]\\
&\quad + \sum_{\beta=1}^{k}\left\{\left[w_\beta^{(-\|n\|+i+j-2)}\right] + \left[\delta w_\beta^{(i+j-1)}\right] - \left[a_0^{(\|n\|+i-j)}\right]\left[u_\beta^{(-\|n\|+i+j-1)}\right]\right.\\
&\quad \left. - \sum_{\alpha=1}^{k}\left[a_\alpha^{(\|n\|+i-j)}\right]\left[v_{\alpha,\beta}^{(-\|n\|+i+j-1)}\right]\right\}\left[q_\beta^{*(\|n\|-i-j)}\right]\\
&= \left[r^{(-\|n\|+i+j-1)}\right]\left[v^{*(\|n\|-i-j+1)}\right] + \sum_{\beta=1}^{k}\left[w_\beta^{(-\|n\|+i+j-2)}\right]\left[q_\beta^{*(\|n\|-i-j)}\right]\\
&\quad + \left[\delta r^{(i+j-2)}\right]\left[v^{*(\|n\|-i-j+1)}\right] + \sum_{\beta=1}^{k}\left[\delta w_\beta^{(i+j-1)}\right]\left[q_\beta^{*(\|n\|-i-j)}\right]\\
&\quad - \sum_{\alpha=0}^{k}\left[a_\alpha^{(\|n\|+i-j)}\right]\left[(\theta_{II})_{\alpha,0}^{(i-j)}\right]\\
&= (a_0^{(0)})^{-1}\left[a_0^{(i-j)}\right] + \theta_{IV}.
\end{aligned}$$

The result (22) follows. □

Corollary 5 drops the requirement in Theorem 4 that $S(z)$ and $S^*(z)$ be normalized. In particular, the corollary is valid in the case that $S(z)$ and $S^*(z)$ are scaled.

COROLLARY 5. *In terms of the (unnormalized) NPHS $S(z)$ of type n for $A(z)$ and the (unnormalized) NSPS $S^*(z)$ of type n for $A^*(z)$, the inverse of $\mathcal{M}_n$ is given by*

$$\mathcal{M}_n^{-1}\left\{\left[a_0^{(i-j)}\right]+\ddot{\theta}_{IV}\right\} = a_0^{(0)}\left\{(\gamma_0\gamma_0^*)^{-1}\mathcal{P}^t\left[v^{*(\|n\|-i-j+1)}\right]+\sum_{\beta=1}^{k}(\gamma_\beta\gamma_\beta^*)^{-1}\mathcal{U}_\beta^t\left[q_\beta^{*(\|n\|-i-j)}\right]\right\}, \tag{28}$$

where

$$\begin{aligned}\ddot{\theta}_{IV} = a_0^{(0)}\Bigg\{&\left[(\ddot{\theta}_{III})_0^{(i-j)}\right]-\sum_{\alpha=0}^{k}\left[a_\alpha^{(\|n\|+i-j)}\right]\left[(\ddot{\theta}_{II})_{\alpha,0}^{(i-j+1)}\right]\\&+(\gamma_0\gamma_0^*)^{-1}\left[\delta r^{(i+j-2)}\right]\left[v^{*(\|n\|-i-j+1)}\right]\\&+\sum_{\beta=1}^{k}(\gamma_\beta\gamma_\beta^*)^{-1}\left[\delta w_\beta^{(i+j-1)}\right]\left[q_\beta^{*(\|n\|-i-j)}\right]\Bigg\},\end{aligned} \tag{29}$$

$$\ddot{\theta}_{III}^t(z) = \left\{A^t(z)\ddot{\theta}_{II}(z)-\delta T^t(z)(\Gamma^*\Gamma)^{-1}S^*(z)\right\}/z^{\|n\|}, \tag{30}$$

$$\begin{aligned}\ddot{\theta}_{II}(z) = {}& S(z)\,(\Gamma^*\Gamma)^{-1}\ddot{\theta}_I(z)\,(\Gamma^*\Gamma)^{-1}\\&\cdot[z^{\|n\|}(a_0^{(0)})^{-1}I_{k+1}+\ddot{\theta}_I(z)(\Gamma^*\Gamma)^{-1}]^{-1}S^*(z),\end{aligned} \tag{31}$$

$$\begin{aligned}\ddot{\theta}_I(z) = a_0^{-1}(z)\Bigg\{&\begin{bmatrix} v^*(z)\\ zQ^*(z)\end{bmatrix}\delta T^t(z)\\&+\delta T^*(z)\left[\;zQ(z)\;\middle|\;V(z)\;\right]\Bigg\}\ (\mathrm{mod}\ z^{\|n\|+1}).\end{aligned} \tag{32}$$

Proof. The normalized NPHS is obtained from an unnormalized one by multiplying it on the right by the diagonal matrix $\mathrm{diag}[\gamma_0^{-1},\dots,\gamma_k^{-1}]$. Similarly, the normalized NSPS is obtained from an unnormalized one by multiplying it on the left by the diagonal matrix $\mathrm{diag}[\gamma_0^{*-1},\dots,\gamma_k^{*-1}]$. The result now follows directly from (22). □

Let

$$\kappa = \sum_{\beta=0}^{k}(\gamma_\beta\gamma_\beta^*)^{-1}. \tag{33}$$

In the corollary below, we give a bound for $\mathcal{M}_n^{-1}$ in terms of κ.

COROLLARY 6. *If the residual errors $\delta T^t(z)$ and $\delta T^*(z)$ associated with the scaled $S(z)$ and the scaled $S^*(z)$ are not too large, so that*

$$\begin{aligned}&\left[(\kappa+1)(k+2)|a_0^{(0)}|(\|a_0^{-1}(z)\ (\mathrm{mod}\ z^{\|n\|+1})\|+1)\right]^2\\&\left[(k+2)\|\delta T^t(z)\|+\|\delta T^*(z)\|\right]\le 1/8,\end{aligned} \tag{34}$$

then

$$\|\mathcal{M}_n^{-1}\|_1 \leq 2\kappa \cdot |a_0^{(0)}| \cdot \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\|. \tag{35}$$

Proof. We use Corollary 5 with $S(z)$ and $S^*(z)$ scaled. We begin by finding a bound for $\ddot{\theta}_{IV}$ appearing in the inverse formula (28) for $\mathcal{M}_n$. A bound for $\ddot{\theta}_{IV}$ depends on bounds for $\ddot{\theta}_I(z)$, $\ddot{\theta}_{II}(z)$, and $\ddot{\theta}_{III}(z)$. From (16),

$$\|\ddot{\theta}_I(z)\| \leq \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\| \cdot \left\{(k+1)\|\delta T^t(z)\| + \|\delta T^*(z)\|\right\},$$

because $\|S(z)\| = 1$ and $\|S^*(z)\| \leq k+1$. From (16) and (32), note that $\ddot{\theta}_I(z)$ is a matrix polynomial of at most degree $\|n\|$ and so, using (34),

$$\|a_0^{(0)} z^{\|n\|} \ddot{\theta}_I(z^{-1})(\Gamma^*\Gamma)^{-1}\| = \|a_0^{(0)} \ddot{\theta}_I(z)(\Gamma^*\Gamma)^{-1}\| \leq \kappa \cdot |a_0^{(0)}| \cdot \|\ddot{\theta}_I(z)\| \leq 1/2,$$

because $\|(\Gamma^*\Gamma)^{-1}\| \leq \kappa$. So as in [21, p. 187], the inverse of $(a_0^{(0)})^{-1} I_{k+1} + z^{\|n\|}\ddot{\theta}_I(z^{-1}) \cdot (\Gamma^*\Gamma)^{-1}$ exists and

$$\begin{aligned}\left\| \left\{ (a_0^{(0)})^{-1} I_{k+1} + z^{\|n\|} \ddot{\theta}_I(z^{-1})(\Gamma^*\Gamma)^{-1} \right\}^{-1} \right\| &\leq \frac{|a_0^{(0)}|}{1 - \|a_0^{(0)} z^{\|n\|} \ddot{\theta}_I(z^{-1})(\Gamma^*\Gamma)^{-1}\|} \\ &\leq 2|a_0^{(0)}|.\end{aligned}$$

To determine a bound for $\ddot{\theta}_{II}(z)$ in (31), let $N = \max_{0 \leq \beta \leq k}\{n_\beta\}$ and observe from (18) that $\ddot{\theta}_{II}(z)$ is also a matrix polynomial, now of degree at most $\|n\| + N$. Consequently,

$$\begin{aligned}\|\ddot{\theta}_{II}(z)\| &= \|z^{\|n\|+N} \ddot{\theta}_{II}(z^{-1})\| \\ &= \left\| \left\{ z^N S(z^{-1}) \right\} (\Gamma^*\Gamma)^{-1} \left\{ z^{\|n\|} \ddot{\theta}_I(z^{-1}) \right\} (\Gamma^*\Gamma)^{-1} \right. \\ &\quad \left. \cdot \left\{ (a_0^{(0)})^{-1} I_{k+1} + z^{\|n\|} \ddot{\theta}_I(z^{-1})(\Gamma^*\Gamma)^{-1} \right\}^{-1} [z^{\|n\|} S^*(z^{-1})] \right\| \\ &\leq \kappa^2 \|z^N S(z^{-1})\| \cdot \|z^{\|n\|} \ddot{\theta}_I(z^{-1})\| \cdot \|z^{\|n\|} S^*(z^{-1})\| \\ &\quad \cdot \left\| \left\{ (a_0^{(0)})^{-1} I_{k+1} + z^{\|n\|} \ddot{\theta}_I(z^{-1})(\Gamma^*\Gamma)^{-1} \right\}^{-1} \right\| \\ &\leq 2\kappa^2 |a_0^{(0)}| \cdot \|S(z)\| \cdot \|\ddot{\theta}_I(z)\| \cdot \|S^*(z)\| \\ &\leq 2\kappa^2 (k+1) \cdot |a_0^{(0)}| \cdot \|\ddot{\theta}_I(z)\|.\end{aligned} \tag{36}$$

In addition, it now follows that a bound for $\ddot{\theta}_{III}(z)$ in (30) is given by

$$\|\ddot{\theta}_{III}^t(z)\| \leq 2\kappa^2(k+1) \cdot |a_0^{(0)}| \cdot \|\ddot{\theta}_I(z)\| + \kappa(k+1) \cdot \|\delta T^t(z)\|.$$

We are now ready to give a bound for $\ddot{\theta}_{IV}$ appearing in the inverse formula (28). But first observe that

$$\left\| \left[(\ddot{\theta}_{III})_0^{(i-j)} \right] \right\|_1 \leq \|\ddot{\theta}_{III}^t(z)\|$$

and that

$$\left\| \sum_{\alpha=0}^{k} \left[a_\alpha^{(\|n\|+i-j)} \right] \left[(\ddot{\theta}_{II})_{\alpha,0}^{(i+j)} \right] \right\|_1 \leq \sum_{\alpha=0}^{k} \|(\ddot{\theta}_{II})_{\alpha,0}(z)\| \leq \|(\ddot{\theta}_{II})(z)\|.$$

Thus,

$$\begin{aligned}
\|\ddot{\theta}_{IV}\|_1 &= \left\| a_0^{(0)} \left\{ \left[(\ddot{\theta}_{III})_0^{(i-j)} \right] - \sum_{\alpha=0}^{k} \left[a_\alpha^{(\|n\|+i-j)} \right] \left[(\ddot{\theta}_{II})_{\alpha,0}^{(i-j)} \right] \right.\right. \\
&\quad + (\gamma_0\gamma_0^*)^{-1} \left[\delta r^{(i+j-2)} \right] \left[v^{*(\|n\|-i-j+1)} \right] \\
&\quad \left.\left. + \sum_{\beta=1}^{k} (\gamma_\beta\gamma_\beta^*)^{-1} \left[\delta w_\beta^{(i+j-1)} \right] \left[q_\beta^{*(\|n\|-i-j)} \right] \right\} \right\|_1 \\
&\le |a_0^{(0)}| \cdot \left\{ \|\ddot{\theta}_{III}^t(z)\| + \|\ddot{\theta}_{II}(z)\| + (\gamma_0\gamma_0^*)^{-1}\|\delta T^t(z)\| + \sum_{\beta=1}^{k} (\gamma_\beta\gamma_\beta^*)^{-1}\|\delta T^t(z)\| \right\} \\
&\le |a_0^{(0)}| \left\{ \|\ddot{\theta}_{III}^t(z)\| + \|\ddot{\theta}_{II}(z)\| + \kappa\|\delta T^t(z)\| \right\} \\
&\le |a_0^{(0)}| \left\{ \kappa(k+1)\|\delta T^t(z)\| + 4\kappa^2(k+1)|a_0^{(0)}| \cdot \|\ddot{\theta}_I(z)\| + \kappa\|\delta T^t(z)\| \right\} \\
&\le \kappa(k+2) \cdot |a_0^{(0)}| \left\{ \|\delta T^t(z)\| \right. \\
&\qquad \left. + 4\kappa|a_0^{(0)}| \cdot \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\| \left[(k+1)\|\delta T^t(z)\| + \|\delta T^*(z)\| \right] \right\} \\
&\le 4\|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\| \left[(\kappa+1)(k+2)|a_0^{(0)}| \right]^2 \left[(k+2)\|\delta T^t(z)\| + \|\delta T^*(z)\| \right].
\end{aligned}$$

It then follows from (34) that

$$\left\| \left[a_0^{(i-j)} \right]^{-1} \ddot{\theta}_{IV} \right\|_1 \le \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\| \cdot \|\ddot{\theta}_{IV}\|_1 \le 1/2,$$

and so $I_{\|n\|} + [a_0^{(i-j)}]^{-1}\ddot{\theta}_{IV}$ is invertible. In addition,

$$\begin{aligned}
\left\| \left\{ \left[a_0^{(i-j)} \right] + \ddot{\theta}_{IV} \right\}^{-1} \right\|_1 &\le \left\| \left\{ I_{\|n\|} + \left[a_0^{(i-j)} \right]^{-1} \ddot{\theta}_{IV} \right\}^{-1} \left[a_0^{(i-j)} \right]^{-1} \right\|_1 \\
&\le \frac{\| \left[a_0^{(i-j)} \right]^{-1} \|_1}{1 - \| \left[a_0^{(i-j)} \right]^{-1} \ddot{\theta}_{IV} \|_1} \\
&\le 2\| \left[a_0^{(i-j)} \right]^{-1} \|_1 \\
&\le 2\|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\|.
\end{aligned}$$

Therefore, a bound for $\mathcal{M}_n^{-1}$ in (28) is given

$$\begin{aligned}
\|\mathcal{M}_n^{-1}\|_1 \le \left\| \left\{ \left[a_0^{(i-j)} \right] + \ddot{\theta}_{IV} \right\}^{-1} \right\| \cdot \left\| a_0^{(0)} \left\{ (\gamma_0\gamma_0^*)^{-1} \mathcal{P}_n^t \left[v^{*(\|n\|-i-j+1)} \right] \right.\right. \\
\left.\left. + \sum_{\beta=1}^{k} (\gamma_\beta\gamma_\beta^*)^{-1} \mathcal{U}_{n,\beta}^t \left[q_\beta^{*(\|n\|-i-j)} \right] \right\} \right\|_1 \\
\le 2\kappa|a_0^{(0)}| \cdot \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\|. \qquad \square
\end{aligned}$$

From (35), it follows that a bound for the 1-norm condition number of $\mathcal{M}_n$ is

$$\|\mathcal{M}_n\|_1 \cdot \|\mathcal{M}_n^{-1}\|_1 \leq 2\kappa |a_0^{(0)}| \cdot \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\|$$

because it is assumed that each $a_\beta(z)$ is scaled.

5. The inverse of a mosaic Sylvester matrix. In this section, a formula is given for the inverse of $\mathcal{M}_n^*$ expressed in terms of both $S(z)$ and $S^*(z)$. This enables estimating the condition number of $\mathcal{M}_n^*$ without explicitly computing $\mathcal{M}_n^{*-1}$.

Associated with the NPHS $S(z)$ and the NSPS $S^*(z)$, for $\beta = 1, 2, \ldots, k$, define the $\|n\| \times k\|n\|$ matrices

$$\mathcal{V}_\beta = \left[\begin{array}{ccc|c|ccc} v_{1,\beta}^{(\|n\|-1)} & \cdots & v_{1,\beta}^{(0)} & & v_{k,\beta}^{(\|n\|-1)} & \cdots & v_{k,\beta}^{(0)} \\ \vdots & \iddots & & \cdots & \vdots & \iddots & \\ v_{1,\beta}^{(0)} & & & & v_{k,\beta}^{(0)} & & \end{array}\right],$$

$$\mathcal{Q} = \left[\begin{array}{ccccc|c|ccccc} q_1^{(\|n\|-2)} & \cdots & q_1^{(0)} & 0 & & & q_k^{(\|n\|-2)} & \cdots & q_k^{(0)} & 0 & \\ \vdots & \iddots & & & & \cdots & \vdots & \iddots & & & \\ q_1^{(0)} & \iddots & & & & & q_k^{0)} & \iddots & & & \\ 0 & & & & & & 0 & & & & \end{array}\right],$$

$$\mathcal{V}^* = \left[\begin{array}{ccc|ccc|c|ccc} v^{*(1)} & \cdots & v^{*(\eta_0)} & u_1^{*(1)} & \cdots & u_1^{*(\eta_1)} & & u_k^{*(1)} & \cdots & u_k^{*(\eta_k)} \\ \vdots & \iddots & 0 & \vdots & \iddots & 0 & & \vdots & \iddots & 0 \\ v^{*(\eta_0)} & \iddots & & u_1^{*(\eta_1)} & \iddots & & \cdots & u_k^{*(\eta_k)} & \iddots & \\ 0 & & \vdots & 0 & & \vdots & & 0 & & \vdots \\ \vdots & & & \vdots & & & & \vdots & & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & & 0 & \cdots & 0 \end{array}\right],$$

and

$$\mathcal{Q}_\beta^* = \left[\begin{array}{ccc|ccc|c|ccc} q_\beta^{*(0)} & \cdots & q_\beta^{*(\eta_0-1)} & p_{\beta,1}^{*(0)} & \cdots & p_{\beta,1}^{*(\eta_1-1)} & & p_{\beta,k}^{*(0)} & \cdots & p_{\beta,k}^{*(\eta_k-1)} \\ \vdots & \iddots & 0 & \vdots & \iddots & 0 & & \vdots & \iddots & 0 \\ q_\beta^{*(\eta_0-1)} & \iddots & & p_{\beta,1}^{*(\eta_1-1)} & \iddots & & \cdots & p_{\beta,k}^{*(\eta_k-1)} & \iddots & \\ 0 & & \vdots & 0 & & \vdots & & 0 & & \vdots \\ \vdots & & & \vdots & & & & \vdots & & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & & 0 & \cdots & 0 \end{array}\right],$$

where $\eta_\beta = \|n\| - n_\beta$. For $\beta = 1, 2, \ldots, k$, also define the $\|n\| \times k\|n\|$ residual error matrices

$$\delta W_\beta = [\delta \bar{W}_\beta, \mathbf{0}_{n_1}, \ldots, \mathbf{0}_{n_k}]$$

and

$$\delta R = [\delta \bar{R}, \mathbf{0}_{n_1}, \ldots, \mathbf{0}_{n_k}],$$

where

$$\delta\bar{W}_\beta = \begin{bmatrix} \delta w_\beta^{(\|n\|-1)} & \cdots & \delta w_\beta^{(n_0)} \\ & & \vdots \\ \vdots & & \delta w_\beta^{(0)} \\ & \iddots & \\ \delta w_\beta^{(0)} & & \end{bmatrix}, \quad \delta\bar{R} = \begin{bmatrix} \delta r^{(\|n\|-2)} & \cdots & \delta r^{(n_0-1)} \\ & & \vdots \\ \vdots & & \delta r^{(0)} \\ & \iddots & 0 \\ \delta r^{(0)} & \iddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix},$$

and $\mathbf{0}_{n_\beta}$ is an $\|n\| \times \|n\| - n_\beta$ matrix of zeros. Also, let

$$\theta = \begin{bmatrix} \theta_{0,0} & \cdots & \theta_{0,k} \\ \vdots & & \vdots \\ \theta_{k,0} & \cdots & \theta_{k,k} \end{bmatrix},$$

where each $\theta_{\alpha,\beta}$ is an $(\|n\| - n_\alpha) \times (\|n\| - n_\beta)$ matrix given by

$$\theta_{\alpha,\beta} = \begin{bmatrix} (\theta_{II})_{\alpha,\beta}^{(\|n\|+1)} & \cdots & (\theta_{II})_{\alpha,\beta}^{(2\|n\|-n_\beta)} \\ \vdots & & \vdots \\ (\theta_{II})_{\alpha,\beta}^{(n_\alpha+2)} & \cdots & (\theta_{II})_{\alpha,\beta}^{(\|n\|+n_\alpha-n_\beta+1)} \end{bmatrix}$$

with $\theta_{II}(z)$ the error appearing in (18). Finally, let $[a_0^{(i-j)}]$ denote an order $\|n\|$, lower triangular, matrix as in §4.

The main result of this section is Theorem 7 below which gives the inverse of $\mathcal{M}_n^*$ in terms of the NPHS $S(z)$ and the NSPS $S^*(z)$ of types n for $A(z)$.

THEOREM 7. *In terms of the* normalized *NPHS* $S(z)$ *and the* normalized *NSPS* $S^*(x)$ *of types* n *for* $A(z)$, *the inverse of* $\mathcal{M}_n^*$ *satisfies*

$$\mathcal{M}_n^{*-1}\left\{(a_0^{(0)})^{-1} I_{k\|n\|} + \theta_{IV}^*\right\} = \mathcal{Q}^t \left[a_0^{(i-j)}\right]^{-1} \mathcal{V}^* + \sum_{\beta=1}^{k} \mathcal{V}_\beta^t \left[a_0^{(i-j)}\right]^{-1} \mathcal{Q}_\beta^*, \tag{37}$$

where

$$\theta_{IV}^* = \theta - \delta R^t \left[a_0^{(i-j)}\right]^{-1} \mathcal{V}^* - \sum_{\beta=1}^{k} \delta W_\beta^t \left[a_0^{(i-j)}\right]^{-1} \mathcal{Q}_\beta^*. \tag{38}$$

Proof. Let

$$\bar{\mathcal{Q}} = \left[\begin{array}{ccc|ccc|c|ccc} p^{(\|n\|-2)} & \cdots & p^{(n_0-1)} & q_1^{(\|n\|-2)} & \cdots & q_1^{(n_1-1)} & & q_k^{(\|n\|-2)} & \cdots & q_k^{(n_k-1)} \\ & & \vdots & & & \vdots & & & & \vdots \\ \vdots & & p^{(0)} & \vdots & & q_1^{(0)} & \cdots & \vdots & & q_k^{(0)} \\ & \iddots & 0 & & \iddots & 0 & & & \iddots & 0 \\ p^{(0)} & \iddots & & q_1^{(0)} & \iddots & & & q_k^{(0)} & \iddots & \\ 0 & & & 0 & & & & 0 & & \end{array}\right].$$

Then, the order condition (5) for an NPHS implies that

$$\mathcal{M}_n^* \cdot \mathcal{Q}^t = \bar{\mathcal{Q}}^t \cdot \left[a_0^{(i-j)}\right] - \delta R^t. \tag{39}$$

To see this, note the (i,j)th component, $1 \le i \le \|n\| - n_0$, $1 \le j \le \|n\|$, of (39) is the coefficient of $z^{\|n\|-i-j}$ in

$$a_0(z)\, p(z) + \sum_{\alpha=1}^{k} a_\alpha(z)\, q_\alpha(z) = z^{\|n\|-1} r(z) + \delta r(z).$$

The remaining components of (39) are obvious identities.

Similarly, for $1 \le \beta \le k$, let

$$\bar{\mathcal{V}}_\beta = \left[\begin{array}{ccc|ccc|c|ccc} u_\beta^{(\|n\|-1)} & \cdots & u_\beta^{(n_0)} & v_{1,\beta}^{(\|n\|-1)} & \cdots & v_{1,\beta}^{(n_1)} & & v_{k,\beta}^{(\|n\|-1)} & \cdots & v_{k,\beta}^{(n_k)} \\ & & \vdots & & & \vdots & & & & \vdots \\ \vdots & & u_\beta^{(0)} & \vdots & & v_{1,\beta}^{(0)} & \cdots & \vdots & & v_{k,\beta}^{(0)} \\ & \iddots & & & \iddots & & & & \iddots & \\ u_\beta^{(0)} & & & v_{1,\beta}^{(0)} & & & & v_{k,\beta}^{(0)} & & \end{array}\right].$$

Then, the coefficient of $z^{\|n\|-i-j+1}$, $1 \le i \le \|n\| - n_0$, $1 \le j \le \|n\|$, in the order condition (5) for an NPHS, namely,

$$a_0(z)\, u_\beta(z) + \sum_{\alpha=1}^{k} a_\alpha(z)\, v_{\alpha,\beta}(z) = z^{\|n\|+1} w_\beta(z) + \delta w_\beta(z),$$

gives the (i,j)th component of

$$\mathcal{M}_n^* \cdot \mathcal{V}_\beta^t = \bar{\mathcal{V}}_\beta^t \cdot \left[a_0^{(i-j)}\right] - \delta W^t. \tag{40}$$

The remaining components of (40) are easy to verify.

Next, observe that Theorem 1 and Corollary 2 imply that

$$\bar{\mathcal{Q}}^t \cdot \mathcal{V}^* + \sum_{\beta=1}^{k} \bar{\mathcal{V}}_\beta^t \cdot \mathcal{Q}_\beta^* = (a_0^{(0)})^{-1} I_{k\|n\|} + \theta. \tag{41}$$

Combining (39), (40), and (41), we obtain the result (37). □

Corollary 8 drops the requirement in Theorem 7 that $S(z)$ and $S^*(z)$ be normalized. In particular, the results of the corollary apply when $S(z)$ and $S^*(z)$ are scaled.

COROLLARY 8. *In terms of the NPHS $S(z)$ (unnormalized) of type n for $A(z)$ and the NSPS $S^*(z)$ (unnormalized) of type n for $A^*(z)$, the inverse of $\mathcal{M}_n^*$ is given by*

$$\begin{aligned} &\mathcal{M}_n^{*-1} \left\{(a_0^{(0)})^{-1} I_{k\|n\|} + \ddot{\theta}_{IV}^*\right\} \\ &\quad = (\gamma_0 \gamma_0^*)^{-1} \mathcal{Q}^t \left[a_0^{(i-j)}\right]^{-1} \mathcal{V}^* + \sum_{\beta=1}^{k} (\gamma_\beta \gamma_\beta^*)^{-1} \mathcal{V}_\beta^t \left[a_0^{(i-j)}\right]^{-1} \mathcal{Q}_\beta^*, \end{aligned} \tag{42}$$

where

$$\ddot{\theta}^*_{IV} = \ddot{\theta} - (\gamma_0\gamma_0^*)^{-1}\delta R^t \left[a_0^{(i-j)}\right]^{-1} \mathcal{V}^* - \sum_{\beta=1}^{k} (\gamma_\beta\gamma_\beta^*)^{-1}\delta W_\beta^t \left[a_0^{(i-j)}\right]^{-1} \mathcal{Q}_\beta^*$$

and

$$\ddot{\theta} = \begin{bmatrix} \ddot{\theta}_{0,0} & \cdots & \ddot{\theta}_{0,k} \\ \vdots & & \vdots \\ \ddot{\theta}_{k,0} & \cdots & \ddot{\theta}_{k,k} \end{bmatrix}$$

with

$$\ddot{\theta}_{\alpha,\beta} = \begin{bmatrix} (\ddot{\theta}_{II})^{(\|n\|+1)}_{\alpha,\beta} & \cdots & (\ddot{\theta}_{II})^{(2\|n\|-n_\beta)}_{\alpha,\beta} \\ \vdots & & \vdots \\ (\ddot{\theta}_{II})^{(n_\alpha+2)}_{\alpha,\beta} & \cdots & (\ddot{\theta}_{II})^{(\|n\|+n_\alpha-n_\beta+1)}_{\alpha,\beta} \end{bmatrix}.$$

Proof. The normalized NPHS is obtained from an unnormalized one by multiplying it on the right by the diagonal matrix $\operatorname{diag}[\gamma_0^{-1},\ldots,\gamma_k^{-1}]$. Similarly, the normalized NSPS is obtained from an unnormalized one by multiplying it on the left by the diagonal matrix $\operatorname{diag}[\gamma_0^{*-1},\ldots,\gamma_k^{*-1}]$. The result now follows directly from (37). □

COROLLARY 9. *If the conditions of Corollary* 6 *are satisfied, then*[3]

$$\|\mathcal{M}_n^{*-1}\|_\infty \le 2\kappa \cdot |a_0^{(0)}| \cdot \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\|. \tag{43}$$

Proof. From (36),

$$\begin{aligned}\|\ddot{\theta}\|_\infty &\le (k+1)\|\ddot{\theta}_{II}(z)\| \\ &\le 2\kappa^2(k+1)^2 \cdot |a_0^{(0)}| \cdot \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\| \cdot \left\{(k+1)\|\delta T^t(z)\| + \|\delta T^*(z)\|\right\}.\end{aligned}$$

Thus,

$$\begin{aligned}\|\ddot{\theta}^*_{IV}\|_\infty &= \left\| \ddot{\theta} - (\gamma_0\gamma_0^*)^{-1}\delta R^t \left[a_0^{(i-j)}\right]^{-1} \mathcal{V}^* - \sum_{\beta=1}^{k} (\gamma_\beta\gamma_\beta^*)^{-1}\delta W_\beta^t \left[a_0^{(i-j)}\right]^{-1} \mathcal{Q}_\beta^* \right\|_\infty \\ &\le \|\ddot{\theta}\|_\infty + (\gamma_0\gamma_0^*)^{-1}(k+1) \cdot \|\delta T^t(z)\| \cdot \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\| \cdot \|S^*(z)\| \\ &\quad + \sum_{\beta=1}^{k} (\gamma_\beta\gamma_\beta^*)^{-1}(k+1) \cdot \|\delta T^t(z)\| \cdot \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\| \cdot \|S^*(z)\| \\ &\le \kappa(k+1)^2 \cdot \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\| \\ &\quad \cdot \left\{ \|\delta T^t(z)\| + 2\kappa|a_0^{(0)}| \cdot \left[(k+1)\|\delta T^t(z)\| + \|\delta T^*(z)\|\right] \right\} \\ &\le 4|a_0^{(0)}| \cdot \left[(\kappa+1)(k+2)(\|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\| + 1)\right]^2 \\ &\quad \cdot \left[(k+2)\|\delta T^t(z)\| + \|\delta T^*(z)\|\right].\end{aligned}$$

[3] The ∞-norm, rather than the 1-norm, is used here because it is more suitable for purposes in [5].

Therefore, using the assumption (34),

$$\left\| \left\{ (a_0^{(0)})^{-1} I_{k\|n\|} + \ddot{\theta}_{IV}^* \right\}^{-1} \right\|_\infty \leq 2|a_0^{(0)}|$$

and so

$$\begin{aligned} \|\mathcal{M}_n^{*-1}\|_\infty \leq & \left\| \left\{ (a_0^{(0)})^{-1} I_{k\|n\|} + \ddot{\theta}_{IV}^* \right\}^{-1} \right\|_\infty \cdot \left\| (\gamma_0 \gamma_0^*)^{-1} \mathcal{Q}^t \left[a_0^{(i-j)} \right]^{-1} \mathcal{V}^* \right. \\ & \left. + \sum_{\beta=1}^{k} (\gamma_\beta \gamma_\beta^*)^{-1} \mathcal{V}_\beta^t \left[a_0^{(i-j)} \right]^{-1} \mathcal{Q}_\beta^* \right\|_\infty \\ \leq & \, 2\kappa |a_0^{(0)}| \cdot \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\|. \end{aligned}$$

□

6. Conclusions. In this paper we have presented new formulae for the "near" inverses of striped and mosaic Sylvester matrices. The formulae are given in terms of numerical Padé–Hermite and simultaneous Padé systems. They are important for numerical computation because they incorporate errors caused by floating-point arithmetic. In particular, the formulae can be used to determine good estimates for the condition numbers of these matrices.

Our primary motivation for obtaining these formulae is the numerically stable computation of Padé–Hermite and simultaneous Padé approximants, the subject of the companion paper [5]. As such we have restricted our attention to a striped and a specific mosaic Sylvester matrices. We conjecture that a similar approach can also be used for determining near inverse formulae of other structured matrices, for example, of mosaic Hankel, Toeplitz, or Sylvester matrices [12, 15]. Some preliminary work on this topic has already been done in [7].

Together with the results of [5], we believe that the formulae given in this paper can be used to stably invert striped and mosaic Sylvester matrices and to stably solve systems of linear equations with these as coefficient matrices. This matter requires formal verification, such as that reported in [10] for the case $k = 1$ and $a_0(z) = 1$.

Acknowledgment. We are very grateful to a referee who contributed much in terms of the correctness of results and the clarity of presentation.

REFERENCES

[1] B. BECKERMANN, *Zur Interpolation mit polynomialen Linearkombinationen beliebiger Funktionen*, Ph.D. thesis, Institut für Angewandte Mathematik, Universität Hannover, 1988.

[2] ———, *A reliable method for computing M-Padé approximants on arbitrary staircases*, J. Comput. Appl. Math., 40 (1992), pp. 19–42.

[3] B. BECKERMANN AND G. LABAHN, *A uniform approach for the fast computation of matrix-type Padé approximants*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 804–823.

[4] S. CABAY, A. JONES, AND G. LABAHN, *A stable algorithm for multi-dimensional Padé systems and the inversion of generalized Sylvester matrices*, Tech. Report TR 94-07, Dept. Comp. Sci., Univ. Alberta, 1994.

[5] ———, *Computation of numerical Padé–Hermite and simultaneous Padé systems* II: *A weakly stable algorithm*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 268–297.

[6] S. CABAY AND G. LABAHN, *A superfast algorithm for multi-dimensional Padé systems*, Numerical Algorithms, 2 (1992), pp. 201–224.

[7] ———, *Fast, stable inversion of mosaic Hankel matrices*, Systems and Networks: Mathematical Theory and Applications, Proceedings of MTNS 93, (1994), pp. 625–630.

[8] S. CABAY, G. LABAHN, AND B. BECKERMANN, *On the theory and computation of non-perfect Padé–Hermite approximants*, J. Comput. Appl. Math., 39 (1992), pp. 295–313.

[9] S. CABAY AND R. MELESHKO, *A weakly stable algorithm for Padé approximants and inversion of Hankel matrices*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 735–765.
[10] M. H. GUTKNECHT AND M. HOCHBRUCK, *The stability of inversion formulas for Toeplitz matrices*, Tech. Report IPS 93–13, IPS-Zürich, 1993.
[11] G. HEINIG AND K. ROST, *Algebraic Methods for Toeplitz-like Matrices and Operators*, Birkhauser Verlag, Basel, 1984.
[12] G. HEINIG AND A. TEWODROS, *On the inverses of Hankel and Toeplitz mosaic matrices*, Seminar Analysis Operator Equation and Numerical Analysis 1987/1988, (1988), pp. 53–65.
[13] T. JONES, *The numerical computation of Padé-Hermite systems*, master's thesis, Dept. Comp. Sci., Univ. Alberta, 1992.
[14] G. LABAHN, *Inversion components for block Hankel-like matrices*, Linear Algebra Appl., 177 (1992), pp. 7–48.
[15] G. LABAHN, B. BECKERMANN, AND S. CABAY, *Inversion of mosaic Hankel matrices via matrix polynomial systems*, Linear Algebra Appl , 221 (1995). pp. 253–280.
[16] G. LABAHN, D. K. CHOI, AND S. CABAY, *The inverses of block Hankel and block Toeplitz matrices*, SIAM J. Comput., 19 (1990), pp. 98–123.
[17] G. LABAHN AND T. SHALOM, *Inversion of Toeplitz structured matrices using only standard equations*, Linear Algebra Appl., (1994), pp. 49–70.
[18] L. LERER AND M. TISMENETSKY, *Generalized Bezoutians and the inversion problem for block matrices*, Integral Equations Operator Theory, 9 (1986), pp. 790–819.
[19] ———, *Toeplitz classification of matrices and inversion formulas II: Block-Toeplitz and perturbed block-Toeplitz matrices*, Tech. Report 88.197, IBM-Israel Scientific Center, Haifa, 1986.
[20] K. MAHLER, *Perfect systems*, Composition Math., 19 (1968), pp. 95–166.
[21] G. W. STEWART, *Introduction to Matrix Computations*, Academic Press, 1973.
[22] M. VAN BAREL AND A. BULTHEEL, *The computation of non-perfect Padé–Hermite approximants*, Numerical Algorithms, 1 (1991), pp. 285–304.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 2, pp. 268–297, April 1996

003

COMPUTATION OF NUMERICAL PADÉ–HERMITE AND SIMULTANEOUS PADÉ SYSTEMS II: A WEAKLY STABLE ALGORITHM*

STAN CABAY†, ANTHONY R. JONES‡, AND GEORGE LABAHN§

Abstract. For $k+1$ power series $a_0(z),\ldots,a_k(z)$, we present a new iterative, look-ahead algorithm for numerically computing Padé–Hermite systems and simultaneous Padé systems along a diagonal of the associated Padé tables. The algorithm computes the systems at all those points along the diagonal at which the associated striped Sylvester and mosaic Sylvester matrices are well conditioned. The operation and the stability of the algorithm is controlled by a single parameter τ which serves as a threshold in deciding if the Sylvester matrices at a point are sufficiently well conditioned. We show that the algorithm is weakly stable and provide bounds for the error in the computed solutions as a function of τ. Experimental results are given which show that the bounds reflect the actual behavior of the error.

The algorithm requires $\mathcal{O}(\|n\|^2+s^3\|n\|)$ operations to compute Padé–Hermite and simultaneous Padé systems of type $n=[n_0,\ldots,n_k]$, where $\|n\|=n_0+\cdots+n_k$ and s is the largest step-size taken along the diagonal. An additional application of the algorithm is the stable inversion of striped and mosaic Sylvester matrices.

Key words. Padé–Hermite approximants, simultaneous Padé approximants, striped Sylvester inverses, mosaic Sylvester inverses, numerical algorithm, numerical stability

AMS subject classifications. 41A21, 65F05, 65G05

1. Introduction.

Let $A^t(z)=[a_0(z),\ldots,a_k(z)]$, $k\geq 1$, be a vector of formal power series over the real numbers[1] with $a_0(0)\neq 0$ and let $n=[n_0,\ldots,n_k]$ be a vector of integers with $n_\beta\geq -1, 0\leq\beta\leq k$, and with at least one $n_\beta\geq 0$. A *Padé–Hermite approximant* of type n for $A(z)$ is a nontrivial vector $[q_0(z),\ldots,q_k(z)]$ of polynomials $q_\beta(z)$ over the real numbers having degrees[2] at most $n_\beta, 0\leq\beta\leq k$, such that

$$a_0(z)q_0(z)+\cdots+a_k(z)q_k(z)=c_{\|n\|+k}z^{\|n\|+k}+c_{\|n\|+k+1}z^{\|n\|+k+1}+\cdots, \tag{1}$$

with $\|n\|=n_0+\cdots+n_k$.

The *Padé–Hermite approximation problem* was introduced in 1873 by Hermite and has been widely studied by several authors (for a bibliography, see, for example, [27, 2, 4, 5, 23]). Note that for $A^t(z)=[-1,a(z)]$, (1) becomes

$$a(z)q_1(z)-q_0(z)=O(z^{n_0+n_1+1}).$$

Thus, as a special case we have the classical Padé approximation problem for the power series $a(z)$. The Padé–Hermite approximation problem also includes other

* Received by the editors May 31, 1994; accepted for publication (in revised form) by M. Gutknecht May 2, 1995.

† Department of Computing Science, University of Alberta, Edmonton, Alberta T6G 2H1, Canada (cabay@cs.ualberta.ca). The research of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant A8035.

‡ Bell Northern Research, P. O. Box 3511, Station C, Ottawa, Ontario K1Y 4H7, Canada (anthonyj@bnr.ca).

§ Department of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada (glabahn@daisy.uwaterloo.ca). The research of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant FS1525C.

[1] The restriction to real numbers is made in order to simplify floating-point analysis. All of the results given in this paper also hold with minor modifications for the field of complex numbers.

[2] By convention, a polynomial of degree -1 is the zero polynomial.

classical approximation problems such as the algebraic approximants where $A^t(z) = [1, a(z), a(z)^2, \ldots, a(z)^k]$ (see [25] for the special case $k = 2$) and G^3J approximants where $A^t(z) = [1, a(z), a'(z)]$. Additional examples can be found in [1].

Closely related to Padé–Hermite approximants are *simultaneous Padé approximants.* A simultaneous Padé approximant of type n for $A(z)$ is a nontrivial vector $[q_0^*(z), \ldots, q_k^*(z)]$ of polynomials $q_\beta^*(z)$ over the real numbers having degrees of at most $\|n\| - n_\beta, 0 \leq \beta \leq k$, such that

$$q_0^*(z) \cdot a_\beta(z) + q_\beta^*(z) \cdot a_0(z) = c_{\|n\|+1}^{(\beta)} z^{\|n\|+1} + c_{\|n\|+2}^{(\beta)} z^{\|n\|+2} + \cdots \tag{2}$$

for $\beta = 1, \ldots, k$. Simultaneous Padé approximants were also defined by Hermite and were used in his famous proof of the transcendence of e. Again, for $A^t(z) = [-1, a(z)]$, the simultaneous Padé approximation problem becomes the classical Padé approximation problem for $a(z)$.

By equating coefficients in (1), the Padé–Hermite approximation problem can be viewed as solving a system of linear equations of size $\|n\| \times \|n\|$. Thus, one can use Gaussian elimination to solve this problem with a complexity of $\mathcal{O}(\|n\|^3)$ operations. However, the coefficient matrix of the corresponding linear system has a structured form, so it is not surprising that there are a number of fast [27, 14] $\mathcal{O}(\|n\|^2)$ and superfast [5, 12] $\mathcal{O}(\|n\| \log^2 \|n\|)$ algorithms for determining Padé–Hermite approximants. All these algorithms have the property that they work for any input vector of power series. In addition, these algorithms all make important use of exact arithmetic; in particular, they all depend on knowing that certain quantities are known to be zero or not. A similar statement also applies for the fast and superfast computation of simultaneous Padé approximants.

In the special case of Padé approximants it has long been known that most fast and superfast algorithms have problems with numerical stability for their computation. The first known numerically stable algorithm for fast Padé approximation was presented by Cabay and Meleshko [15]. Alternate algorithms for fast Padé computation that also consider the issue of numerical stability include [6, 13, 18], and for superfast computation [19]. Algorithms dealing with the closely associated problem of stably computing fast rational interpolation include [8].

In this paper, we present a new algorithm for the computation of Padé–Hermite and simultaneous Padé *systems.* These systems are matrix polynomials that contain the desired multidimensional Padé approximant along with quantities that can be used to recursively or iteratively compute the next approximant along a diagonal path in the associated Padé tables. The algorithm works for all vectors of power series and is fast in the sense that it computes a system in $\mathcal{O}(\|n\|^2)$ operations in the generic case. In addition, we show that this algorithm is *weakly stable* in the sense that it provides good answers to well-conditioned problems. The algorithm is a look-ahead procedure that computes the systems of type n by computing all the Padé systems at the well-conditioned locations along the diagonal path passing through the point n. In the case of Padé approximation ($k = 1$), the algorithm reduces to the Cabay and Meleshko algorithm.

It is known (cf. [12] or [23]) that in exact arithmetic a Padé–Hermite system (PHS) exists uniquely if and only if the striped Sylvester coefficient matrix of the corresponding associated linear system is nonsingular. This is also true for a simultaneous Padé system (SPS) where the coefficient matrix of the associated linear system is now a mosaic Sylvester rather than a striped Sylvester matrix. However, in the case of floating-point arithmetic, determining that such coefficient matrices are nonsingu-

lar is not good enough. Instead one must know, at least in a reasonably computable way, that the linear systems are also well conditioned. Central to the stable operation of our algorithm is the ability to estimate the condition numbers of the associated striped Sylvester and mosaic Sylvester matrices. The estimates follow from some "near" inverse formulae for these matrices that are derived in the companion paper [11] and which are expressed in terms of both Padé–Hermite and simultaneous Padé systems. This is the reason why our algorithm computes Padé–Hermite and simultaneous Padé systems in tandem; the inverse formulae, and consequently the estimates for the condition numbers, require that both the Padé–Hermite and the simultaneous Padé systems be available. The striped Sylvester and mosaic Sylvester matrices are deemed to be well conditioned if the computed estimates of the condition numbers are bounded by some specified "stability" tolerance τ.

As a corollary to the results [11], there is a formula which gives the inverse of a striped Sylvester matrix expressed in terms of the associated Padé–Hermite system only. One attempt to use this formula to develop a stable algorithm for computing Padé–Hermite systems (independent of simultaneous Padé systems) was only partly successful [22]; bounds for the inverse of the associated striped Sylvester matrix (and consequently bounds for its condition number) using the formula were often too pessimistic and impractical.

This paper is organized as follows. Preliminary definitions and basic facts about Padé–Hermite and simultaneous Padé systems are given in the next two sections, and the algorithm for computing these systems is given in §4. The remainder of the paper is devoted to showing that the algorithm is weakly stable for the computation of either system. To this end, §5 discusses the errors that result from the iterative steps of the algorithm, while §6 gives the proof of stability. Section 7 provides results of some numerical experiments that reflect the theoretic results of the previous sections. The final section gives some conclusions and a discussion of further areas of research.

We conclude this section by defining some norms that are used in the analysis of the errors made by the algorithm. Let

$$a(z) = \sum_{\ell=0}^{\infty} a^{(\ell)}\, z^{\ell} \in \mathcal{R}\,[[z]]\,,$$

where $\mathcal{R}[[z]]$ is the domain of power series with coefficients from $\mathcal{R}$, and define the bounded power series

$$\mathcal{R}^{B}[[z]] = \left\{ a(z) \;\middle|\; a(z) \in \mathcal{R}\,[[z]]\,,\ \sum_{\ell=0}^{\infty} |a^{(\ell)}| < \infty \right\}.$$

A norm for $a(z) \in \mathcal{R}^{B}\,[[z]]$ is

$$\|a(z)\| = \sum_{\ell=0}^{\infty} |a^{(\ell)}|.$$

$\mathcal{R}^{B}[[z]]$ includes the domain of polynomials $\mathcal{R}[z]$. So, for

$$s(z) = \sum_{\ell=0}^{\partial} s^{(\ell)}\, z^{\ell} \ \in \mathcal{R}\,[z]\,,$$

we use the norm

$$\|s(z)\| = \sum_{\ell=0}^{\partial} | s^{(\ell)} | .$$

For vectors and matrices over $\mathcal{R}^B[[z]]$, we use the 1-norm unless otherwise specified. So, for example, the norm for $A^t(z)$ is

$$\|A^t(z)\| = \max_{0\le\beta\le k} \{\|a_\beta(z)\|\}$$

and the norm for $S(z) \in \mathcal{R}_{(k+1)\times(k+1)}[z]$ is

$$\|S(z)\| = \max_{0\le\beta\le k} \left\{ \sum_{\alpha=0}^{k} \|S_{\alpha,\beta}(z)\| \right\}.$$

It is easy to verify that various compatibility conditions are satisfied. For example,

$$\|A^t(z) \cdot S(z)\| \le \|A^t(z)\| \cdot \|S(z)\|$$

and

$$\|a(z) \cdot b(z)\| \le \|a(z)\| \cdot \|b(z)\|,$$

where $b(z)$ is also a bounded power series. In addition, for $S^*(z) \in \mathcal{R}_{(k+1)\times(k+1)}[z]$ and $A^*(z) \in \mathcal{R}^B_{(k+1)\times k}[[z]]$,

$$\|S^*(z) \cdot A^*(z)\| \le \|S^*(z)\| \cdot \|A^*(z)\|,$$

$$\|S(z) \cdot S^*(z)\| \le \|S(z)\| \cdot \|S^*(z)\|.$$

In the subsequent development, we also make use of the inequality

$$\|a(z) \pmod{z^{\|n\|+1}}\| \le \|a(z)\|,$$

where

$$a(z) \pmod{z^{\|n\|+1}} = \sum_{\ell=0}^{\|n\|} a^{(\ell)} z^\ell + \sum_{\ell=\|n\|+1}^{\infty} 0 \cdot z^\ell \in \mathcal{R}^B[[z]].$$

2. Padé–Hermite systems (PHS). In this section, we give the definition of a PHS for a vector of formal power series. Let $n = [n_0, \dots, n_k]$ and define $\|n\| = n_0 + \cdots + n_k$. Let

$$A^t(z) = [a_0(z), \dots, a_k(z)], \tag{3}$$

where

$$a_\beta(z) = \sum_{\ell=0}^{\infty} a_\beta^{(\ell)} z^\ell, \qquad \beta = 0, \dots, k,$$

with $a_\beta^{(\ell)} \in \mathcal{R}$, the field of real numbers. Assume that $a_0^{(0)} \neq 0$, which means that $a_0^{-1}(z)$ exists. Assume also that $A^t(z)$ is scaled so that $\|a_\beta(z) \pmod{z^{\|n\|+1}}\| = 1$, $0 \leq \beta \leq k$.

The $(k+1) \times (k+1)$ matrix of polynomials

$$S(z) = \left[\begin{array}{c|c} z^2p(z) & U^t(z) \\ \hline z^2Q(z) & V(z) \end{array}\right] = \left[\begin{array}{c|ccc} z^2p(z) & u_1(z) & \cdots & u_k(z) \\ \hline z^2q_1(z) & v_{1,1}(z) & \cdots & v_{1,k}(z) \\ \vdots & \vdots & & \vdots \\ z^2q_k(z) & v_{k,1}(z) & \cdots & v_{k,k}(z) \end{array}\right] \tag{4}$$

is a PHS [14] of type n for $A(z)$ if the following conditions are satisfied.

I (*Degree conditions*). For $1 \leq \alpha, \beta \leq k$,

$$\begin{aligned} p(z) &= \sum_{\ell=0}^{n_0-1} p^{(\ell)} z^\ell, & u_\beta(z) &= \sum_{\ell=0}^{n_0} u_\beta^{(\ell)} z^\ell, \\ q_\alpha(z) &= \sum_{\ell=0}^{n_\alpha-1} q_\alpha^{(\ell)} z^\ell, & v_{\alpha,\beta}(z) &= \sum_{\ell=0}^{n_\alpha} v_{\alpha,\beta}^{(\ell)} z^\ell. \end{aligned} \tag{5}$$

II (*Order condition*).

$$A^t(z)S(z) = z^{\|n\|+1}T^t(z), \tag{6}$$

where $T^t(z) = [r(z), W^t(z)]$ with $W^t(z) = [w_1(z), \ldots, w_k(z)]$ is the residual.

III (*Nonsingularity condition*). The constant term of $V(z)$ is a diagonal matrix

$$V(0) = \text{diag } [\gamma_1, \ldots, \gamma_k] \tag{7}$$

and

$$\gamma \equiv (a_0^{(0)})^{-1} \prod_{\alpha=0}^{k} \gamma_\alpha \neq 0, \tag{8}$$

where $\gamma_0 = r(0)$.

Remark 1. Only the first column of $S(z)$ is a Padé–Hermite approximant as defined in §1, this being of type $[n_0 - 1, \ldots, n_k - 1]$. The remaining columns of $S(z)$ do not quite satisfy the order condition (1) and are therefore not Padé–Hermite approximants; these columns serve primarily to facilitate the computation of the first column using the algorithm given later in §4. But there are other uses for these columns of $S(z)$, such as that of expressing the inverse of a striped Sylvester matrix [9, 11].

Remark 2. The nonsingularity condition III is equivalent to the condition that $r(0) \neq 0$ and that $V(0)$ be a nonsingular diagonal matrix.

Remark 3. The PHS is said to be *normalized* [14] if the nonsingularity condition III is replaced by $r(0) = 1$ and $V(0) = I_k$. This can be achieved by multiplying $S(z)$ on the right by Γ^{-1}, where

$$\Gamma = \text{diag}\,[\gamma_0, \ldots, \gamma_k]\,. \tag{9}$$

The PHS is said to be *scaled* [22] if each column of $S(z)$ has norm equal to 1 for some norms and if, in addition, $\gamma_\beta > 0$, $0 \leq \beta \leq k$. Here, also, scaling a PHS is accomplished by multiplying it on the right by an appropriate diagonal matrix.

Remark 4. The nonsingularity condition III, namely $\gamma \neq 0$, refers to the nonsingularity of $S(z)$; that is, $S(z)$ is nonsingular iff $\gamma \neq 0$ (in [9], it is shown that $\det S(z) = \gamma \cdot z^{\|n\|+1}$). Equivalently, the nonsingularity condition refers to the nonsingularity of the associated striped Sylvester matrix $\mathcal{M}_n$ defined in (11) below; in [14] it is shown that a PHS (with $\gamma \neq 0$) exists iff $\mathcal{M}_n$ is nonsingular.

If the order condition (6) is not satisfied exactly, but rather

$$A^t(z)S(z) = z^{\|n\|+1}T^t(z) + \delta T^t(z), \tag{10}$$

where $\delta T^t(z) = \left[z^2\, \delta r(z), \delta W^t(z)\right]$ with $\delta W^t(z) = [\delta w_1(z), \ldots, \delta w_k(z)]$ is a relatively "small" residual error, then $S(z)$ is called a numerical Padé–Hermite system (NPHS). In (10), for $1 \leq \beta \leq k$,

$$\delta r(z) = \sum_{\ell=0}^{\|n\|-2} \delta r^{(\ell)}\, z^\ell,$$

$$\delta w_\beta(z) = \sum_{\ell=0}^{\|n\|} \delta w_\beta^{(\ell)}\, z^\ell.$$

If $\delta T^t(z) = 0$, then $S(z)$ is an exact (rather than a numerical) PHS. To distinguish it from an NPHS $S(z)$, an exact system is denoted by $S_E(z)$.

Associated with $A(z)$, let $\mathcal{M}_n$ be the striped Sylvester matrix of order $\|n\|$,

$$\mathcal{M}_n = \left[\begin{array}{ccc|c|ccc} a_0^{(0)} & & & & a_k^{(0)} & & \\ & \ddots & & & & \ddots & \\ \vdots & & a_0^{(0)} & \cdots & \vdots & & a_k^{(0)} \\ & & \vdots & & & & \vdots \\ a_0^{(\|n\|-1)} & \cdots & a_0^{(\|n\|-n_0)} & & a_k^{(\|n\|-1)} & \cdots & a_k^{(\|n\|-n_k)} \end{array}\right]. \tag{11}$$

Then $S(z)$ can be obtained by solving two sets of linear equations with $\mathcal{M}_n$ as the coefficient matrix [14]. From (10),

$$a_0(z)\, p(z) + \sum_{\alpha=1}^{k} a_\alpha(z)\, q_\alpha(z) = z^{\|n\|-1} r(z) + \delta r(z), \tag{12}$$

which gives rise to

$$\mathcal{M}_n \cdot \mathcal{X} = [0, \ldots, 0, \gamma_0]^t, \tag{13}$$

where

$$\mathcal{X} = \left[p^{(0)}, \ldots, p^{(n_0-1)} | q_1^{(0)}, \ldots, q_1^{(n_1-1)} | \cdots | q_k^{(0)}, \ldots, q_k^{(n_k-1)}\right]^t.$$

The solution $\mathcal{X}$ yields the first column $S_{0,0}(z), S_{1,0}(z), \ldots, S_{k,0}(z)$ of $S(z)$. In (13), we require that $\gamma_0 = r(0) \neq 0$; $\gamma_0 = 1$ for a normalized NPHS. The existence of a solution to (13) is assured if $\mathcal{M}_n$ is nonsingular. The term $\delta r(z)$ in (12) represents the residual error made in solving (13).

Next, to compute $U^t(z)$ and $V(z)$ (i.e., the remaining columns of $S(z)$), again we use (6); namely,

$$a_0(z)\, u_\beta(z) + \sum_{\alpha=1}^{k} a_\alpha(z)\, v_{\alpha,\beta}(z) = z^{\|n\|+1} w_\beta(z) + \delta w_\beta(z),\ 1 \le \beta \le k. \tag{14}$$

For $\alpha, \beta = 1, \ldots, k$, set

$$u_\beta^{(0)} = -\frac{a_\beta^{(0)}}{a_0^{(0)}}\gamma_\beta, \tag{15}$$

$$v_{\alpha,\beta}^{(0)} = \begin{cases} \gamma_\beta, & \alpha = \beta, \\ 0, & \alpha \neq \beta. \end{cases}$$

This yields the constant terms $U^t(0)$ and $V(0)$ of $U^t(z)$ and $V(z)$, respectively. The remaining components

$$\mathcal{Y} = \left[\begin{array}{ccc|ccc|c|ccc} u_1^{(1)} & \cdots & u_1^{(n_0)} & v_{1,1}^{(1)} & \cdots & v_{1,1}^{(n_1)} & \cdots & v_{k,1}^{(1)} & \cdots & v_{k,1}^{(n_k)} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ u_k^{(1)} & \cdots & u_k^{(n_0)} & v_{1,k}^{(1)} & \cdots & v_{1,k}^{(n_1)} & & v_{k,k}^{(1)} & \cdots & v_{k,k}^{(n_k)} \end{array}\right]^t \tag{16}$$

can be obtained by solving

$$\mathcal{M}_n \cdot \mathcal{Y} = -\begin{bmatrix} a_0^{(1)} & \cdots & a_k^{(1)} \\ \vdots & & \vdots \\ a_0^{(\|n\|)} & \cdots & a_k^{(\|n\|)} \end{bmatrix} \left[\frac{U^t(0)}{V(0)}\right]. \tag{17}$$

In (17), we require that $\gamma_\beta \neq 0, 1 \le \beta \le k$; $\gamma_\beta = 1$ for a normalized NPHS. Again, the existence of a solution to (17) is assured if $\mathcal{M}_n$ is nonsingular. The terms $\delta w_\beta(z)$, $1 \le \beta \le k$, in (14) represent the residual errors made when solving (15) and (17).

For the special case when $n = [n_0, 0, \ldots, 0]$, the NPHS becomes

$$S(z) = \left[\begin{array}{c|c} [a_0^{(0)}]^{-1} z^{n_0+1} & U^t(z) \\ \hline 0 & I_k \end{array}\right] \cdot \operatorname{diag}[\gamma_0, \ldots, \gamma_k], \tag{18}$$

where $U^t(z) = -[a_0(z)]^{-1} \cdot [a_1(z), \ldots, a_k(z)] \pmod{z^{n_0+1}}$. For initialization purposes in the algorithm given later in §4, we adopt (18) even in the cases $n_0 = 0$ and $n_0 = -1$, despite the fact that it no longer strictly meets all the requirements of an NPHS.

3. Simultaneous Padé systems (SPS). A PHS gives an approximation of a vector of formal power series using matrix multiplication on the right. In this section we give the definition of an SPS which corresponds to a similar approximation but with matrix multiplication on the left and with degree constraints that can be thought of as being "dual" to the degree constraints of a PHS. As in the previous section, an SPS exists if and only if a particular matrix of Sylvester type is nonsingular; in this case it is a mosaic Sylvester matrix.

Let

$$A^*(z) = \left[\begin{array}{ccc} a_{0,1}^*(z) & \cdots & a_{0,k}^*(z) \\ \hline a_{1,1}^*(z) & \cdots & a_{1,k}^*(z) \\ \vdots & & \vdots \\ a_{k,1}^*(z) & \cdots & a_{k,k}^*(z) \end{array}\right] = \left[\frac{B^{*t}(z)}{C^*(z)}\right] \tag{19}$$

be a $(k+1) \times k$ matrix of power series with $\det(C^*(0)) \neq 0$. The $(k+1) \times (k+1)$ matrix of polynomials

$$(20)\quad S^*(z) = \left[\begin{array}{c|c} v^*(z) & U^{*t}(z) \\ \hline z^2 Q^*(z) & z^2 P^*(z) \end{array}\right] = \left[\begin{array}{c|ccc} v^*(z) & u_1^*(z) & \cdots & u_k^*(z) \\ \hline z^2 q_1^*(z) & z^2 p_{1,1}^*(z) & \cdots & z^2 p_{1,k}^*(z) \\ \vdots & \vdots & & \vdots \\ z^2 q_k^*(z) & z^2 p_{k,1}^*(z) & \cdots & z^2 p_{k,k}^*(z) \end{array}\right]$$

is an SPS [12, 14] of type n for $A^*(z)$ if the following conditions are satisfied.

I (*Degree conditions*). For $1 \leq \alpha, \beta \leq k$,

$$(21)\quad v^*(z) = \sum_{\ell=0}^{\|n\|-n_0} v^{*(\ell)} z^\ell, \qquad u_\beta^*(z) = \sum_{\ell=0}^{\|n\|-n_\beta} u_\beta^{*(\ell)} z^\ell,$$

$$q_\alpha^*(z) = \sum_{\ell=0}^{\|n\|-n_0-1} q_\alpha^{*(\ell)} z^\ell, \qquad p_{\alpha,\beta}^*(z) = \sum_{\ell=0}^{\|n\|-n_\beta-1} p_{\alpha,\beta}^{*(\ell)} z^\ell.$$

II (*Order condition*).

$$(22)\quad S^*(z)A^*(z) = z^{\|n\|+1} T^*(z),$$

where $T^{*t}(z) = [W^*(z) | R^{*t}(z)]$ with $R^*(z)$ a $k \times k$ matrix.

III (*Nonsingularity condition*). The constant term of $R^*(z)$ is a diagonal matrix

$$(23)\quad R^*(0) = \operatorname{diag}\,[\gamma_1^*, \ldots, \gamma_k^*]$$

and

$$(24)\quad \gamma^* \equiv (a_0^{(0)})^{-1} \prod_{\alpha=0}^{k} \gamma_\alpha^* \neq 0,$$

where $\gamma_0^* = v^*(0)$.

Remark 5. The SPS is said to be *normalized* [12] if the nonsingularity condition III is replaced by $v^*(0) = 1$ and $R^*(0) = I_k$. This can be achieved by multiplying $S^*(z)$ on the left by Γ^{*-1}, where

$$(25)\quad \Gamma^* = \operatorname{diag}\,[\gamma_0^*, \ldots, \gamma_k^*].$$

The SPS is said to be *scaled* when each row of $S^*(z)$ has a norm equal to 1 for some norm and if, in addition, $\gamma_\alpha^* > 0$, $0 \leq \alpha \leq k$. Here, also, scaling an SPS is accomplished by multiplying it on the left by an appropriate diagonal matrix.

Remark 6. The nonsingularity condition III, namely $\gamma^* \neq 0$, refers to the nonsingularity of $S^*(z)$; that is, $S^*(z)$ is nonsingular iff $\gamma^* \neq 0$ (this follows from Theorem 1 and from an observation about $\det(S(z))$ made in Remark 4). Equivalently, the nonsingularity condition refers to the nonsingularity of the associated mosaic Sylvester matrix $\mathcal{M}_n^*$ defined in (27); in [12] it is shown that an SPS exists iff $\mathcal{M}_n^*$ is nonsingular.

As for the PHS, if the order condition (22) is not satisfied exactly, but rather

$$(26)\quad S^*(z)A^*(z) = z^{\|n\|+1} T^*(z) + \delta T^*(z),$$

where $\delta T^{*t}(z) = \left[\delta W^*(z)|z^2\ \delta R^{*t}(z)\right]$ (with $\delta R^*(z)$ a $k \times k$ matrix) is a relatively "small" residual error, then $S^*(z)$ is called a numerical simultaneous Padé system (NSPS). In (26), for $1 \le \alpha, \beta \le k$,

$$\delta w^*_\beta(z) = \sum_{\ell=0}^{\|n\|} \delta w^{*(\ell)}_\beta\ z^\ell,$$

$$\delta r^*_{\alpha,\beta}(z) = \sum_{\ell=0}^{\|n\|-2} \delta r^{*(\ell)}_{\alpha,\beta}\ z^\ell.$$

As with the NPHS $S(z)$, an NSPS for which $\delta T^*(z) = 0$ is denoted by $S^*_E(z)$.

Associated with $A^*(z)$, let $\mathcal{M}^*_n$ be the mosaic Sylvester matrix of order $k\|n\|$,

$$\mathcal{M}^*_n = \begin{bmatrix} \mathcal{S}^*_{0,1} & \cdots & \mathcal{S}^*_{0,k} \\ \vdots & & \vdots \\ \mathcal{S}^*_{k,1} & \cdots & \mathcal{S}^*_{k,k} \end{bmatrix}, \tag{27}$$

where, for $0 \le \alpha \le k$ and $1 \le \beta \le k$,

$$\mathcal{S}^*_{\alpha,\beta} = \begin{bmatrix} a^{*(0)}_{\alpha,\beta} & & \cdots & a^{*(\|n\|-1)}_{\alpha,\beta} \\ & \ddots & & \vdots \\ & & a^{*(0)}_{\alpha,\beta} & a^{*(n_\alpha)}_{\alpha,\beta} \end{bmatrix}.$$

Also define the order $k(\|n\| + 1)$ matrix

$$\mathcal{N}^*_n = \left[\begin{array}{c|ccc|c|ccc} & a^{*(1)}_{1,1} & \cdots & a^{*(\|n\|)}_{1,1} & & a^{*(1)}_{1,k} & \cdots & a^{*(\|n\|)}_{1,k} \\ C^*(0) & \vdots & & \vdots & \cdots & \vdots & & \vdots \\ & a^{*(1)}_{k,1} & \cdots & a^{*(\|n\|)}_{k,1} & & a^{*(1)}_{k,k} & \cdots & a^{*(\|n\|)}_{k,k} \\ \hline \mathbf{0} & \multicolumn{7}{c}{\mathcal{M}^*_n} \end{array}\right]. \tag{28}$$

Then, as for the NPHS, $S^*(z)$ can be obtained by solving two sets of linear equations with $\mathcal{M}^*_n$ and $\mathcal{N}^*_n$ as the coefficient matrices (also see [12]).

To obtain $S^*_{0,1}(z), \ldots, S^*_{0,k}(z)$ of $S^*(z)$, we use

$$v^*(z)\ a^*_{0,\beta}(z) + \sum_{\alpha=1}^{k} u^*_\alpha(z) a^*_{\alpha,\beta}(z)\ = z^{\|n\|+1} w^*_\beta(z) + \delta w^*_\beta(z),\ \ 1 \le \beta \le k, \tag{29}$$

which is the first row of (26). Matching coefficients of $1, z, \ldots, z^{\|n\|}$ in (29) gives

$$\mathcal{X}^{*t} \cdot \mathcal{N}^*_n = -v^{*(0)} \left[B^{*t}(0)\middle|a^{*(1)}_{0,1}, \ldots, a^{*(\|n\|)}_{0,1}\middle| \cdots \middle|a^{*(1)}_{0,k}, \ldots, a^{*(\|n\|)}_{0,k}\right], \tag{30}$$

where

$$\mathcal{X}^{*t} = [u^{*(0)}_1, \ldots, u^{*(0)}_k | v^{*(1)}, \ldots, v^{*(\|n\|-n_0)} | u^{*(1)}_1, \ldots, u^{*(\|n\|-n_1)}_1 | \cdots | u^{*(1)}_k, \ldots, u^{*(\|n\|-n_k)}_k].$$

With $v^{*(0)} = \gamma_0^* \neq 0$ specified ($\gamma_0^* = 1$ for a normalized NSPS), a unique solution to (30) is assured if $\mathcal{M}_n^*$ is nonsingular because by assumption $\det[C^*(0)] \neq 0$. The terms $\delta w_\beta^*(z)$ in (29) represent the residual errors made in solving (30).

Next, to compute $P^*(z)$ and $Q^*(z)$ (i.e., the remaining rows of $S^*(z)$), again we use (26); namely,

$$q_\alpha^*(z)\, a_{0,\beta}^*(z) + \sum_{\rho=1}^{k} p_{\alpha,\rho}^*(z)\, a_{\rho,\beta}^*(z) = z^{\|n\|-1} r_{\alpha,\beta}^*(z) + \delta r_{\alpha,\beta}^*(z), \quad 1 \le \alpha, \beta \le k. \tag{31}$$

Let

$$\mathcal{Y}_\alpha^{*t} = \left[q_\alpha^{*(0)}, \ldots, q_\alpha^{*(\|n\|-n_0-1)} | p_{\alpha,1}^{*(0)}, \ldots, p_{\alpha,1}^{*(\|n\|-n_1-1)} | \cdots | p_{\alpha,k}^{*(0)}, \ldots, p_{\alpha,k}^{*(\|n\|-n_k-1)} \right].$$

Then, (31) and the requirement that $R^*(0) = \operatorname{diag}[\gamma_1^*, \ldots, \gamma_k^*]$ yield

$$\mathcal{Y}_\alpha^{*t} \cdot \mathcal{M}_n^* = \gamma_\alpha^* E_{\alpha\|n\|}^t, \quad 1 \le \alpha \le k, \tag{32}$$

where $E_{\alpha\|n\|}^t$ is the unit row vector of length $k\|n\|$ with a single 1 in position $\alpha\|n\|$. With $\operatorname{diag}[\gamma_1^*, \ldots, \gamma_k^*]$ specified ($\gamma_\alpha^* = 1$ for a normalized NSPS), a solution of (32) exists uniquely if $\mathcal{M}_n^*$ is nonsingular. The solution $\mathcal{Y}_\alpha^*$ provides the αth row of $S^*(z)$; namely, $S_{\alpha,0}^*(z) = z^2 \cdot q_\alpha^*(z)$ and $S_{\alpha,\beta}^*(z) = z^2 \cdot p_{\alpha,\beta}^*(z)$, $1 \le \beta \le k$. The terms $\delta r_{\alpha,\beta}^*(z)$ in (31) represent the residual errors made in solving (32).

In the remainder of the paper, without loss of generality, we make the simplifying assumption that

$$A^*(z) = \left[\begin{array}{ccc} -a_1(z) & \cdots & -a_k(z) \\ \hline a_0(z) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & a_0(z) \end{array}\right]. \tag{33}$$

With $A^*(z)$ defined by (33), for the special case when $n = [n_0, 0, \ldots, 0]$, the NSPS becomes

$$S^*(z) = \operatorname{diag}[\gamma_0^*, \ldots, \gamma_k^*] \left[\begin{array}{c|c} 1 & U^{*t}(z) \\ \hline 0 & [a_0^{(0)}]^{-1} z^{n_0+1} I_k \end{array}\right], \tag{34}$$

where $U^{*t}(z) = [a_0(z)]^{-1} \cdot [a_1(z), \ldots, a_k(z)] \pmod{z^{n_0+1}}$. For initialization purposes in the algorithm given in §4, we adopt (34) even in the case when $n_0 = 0$ and $n_0 = -1$, despite the fact that it no longer strictly meets all the requirements of an NSPS.

In addition, with $A^*(z)$ defined by (33), there is an important commutativity relationship between PHS and SPS, given in Theorem 1 below. This relationship is used later in §5. But, in our presentation, the residual $T^*(z)$ continues to take the more general form (19) rather than (33), because, for the computation of the NSPS for $T^*(z)$, which is required by the algorithm given in §4, the conversion of $T^*(z)$ from the form (19) to the form (33) by means of multiplication on the right by $R^{*-1}(z)$ introduces undesirable instabilities.

THEOREM 1. *If $S(z)$ is an NPHS of type n for $A(z)$ and $S^*(z)$ is an NSPS of type n for $A^*(z)$, then*

$$S^*(z) \cdot S(z) = z^{\|n\|+1} (a_0^{(0)})^{-1} \Gamma^* \Gamma + \theta_I(z), \tag{35}$$

where

$$\theta_I(z) = a_0^{-1}(z)\left\{\begin{bmatrix} v^*(z) \\ z^2Q^*(z) \end{bmatrix}\delta T^t(z) + \delta T^*(z)\left[\; z^2Q(z) \mid V(z) \;\right]\right\} \pmod{z^{D+1}}$$

with

$$D = \left[\begin{array}{c|ccc} \|n\|+1 & \|n\| & \cdots & \|n\| \\ \hline \|n\|+2 & \|n\|+1 & \cdots & \|n\|+1 \\ \vdots & \vdots & & \vdots \\ \|n\|+2 & \|n\|+1 & \cdots & \|n\|+1 \end{array}\right]$$

and with the modulo operation applied componentwise.

Proof. See [9]. □

Thus, given an NPHS, an NSPS can be computed using (35). However, the stability of such a computation is not known, and we choose instead to compute NPHS and NSPS systems in tandem by the algorithm described in the next section.

4. The algorithm. To compute an NPHS of type n for $A(z)$ and an NSPS of type n for $A^*(z)$, the systems (13), (17), (30), and (32) can be solved using a method such as Gaussian elimination. This method, while not restricting the input power series, does not take advantage of the inherent structure of the coefficient matrices $\mathcal{M}_n$ and $\mathcal{M}_n^*$. Alternatively, a variety of recurrence relations which do take advantage of this structure have been described in the literature; see, e.g., [27, 4, 12, 14]. These recurrence relations usually lead to much more efficient algorithms for algebraically computing PHS and SPS. The recurrence relations given in [12] and [14] appear to be the most easily adaptable to numerical computation and it is the detailed study of the numerical behavior of these recurrences to which we devote the remainder of this paper. We begin by briefly describing these recurrences in the algebraic case.

Let $e_0 = [1, 0, \ldots, 0]$ be a $1 \times k+1$ vector, set

$$M = \min\left\{n_0, \max_{1\le\beta\le k}\{n_\beta\}\right\} + 1,$$

and define integer vectors $n^{(i)} = (n_0^{(i)}, \ldots, n_k^{(i)})$ for $0 \le i \le M$ by $n^{(0)} = -e_0$ and, for $i > 0$,

$$n_\beta^{(i)} = \max\{0, n_\beta - M + i\}, \qquad \beta = 0, \ldots, k.$$

Then the sequence $\{n^{(i)}\}_{i=0,1,\ldots}$ lies on a piecewise linear path with $n_\beta^{(i+1)} \ge n_\beta^{(i)}$ for each i, β and[3] $n^{(M)} = n$. The sequence $\{n^{(i)}\}$ contains a subsequence $\{m^{(\sigma)}\}$ called the *sequence of nonsingular points* for $A(z)$ and $A^*(z)$. This sequence is defined by $m^{(\sigma)} = n^{(i_\sigma)}$, where

$$i_\sigma = \begin{cases} 0, & \sigma = 0, \\ \min\{i > i_{\sigma-1} : \det(\mathcal{M}_{n^{(i)}}) \ne 0\}, & \sigma \ge 1, \end{cases}$$

where $\det(\mathcal{M}_{n^{(i)}})$ is the determinant[4] of $\mathcal{M}_{n^{(i)}}$. Corresponding to the sequence of nonsingular points $\{m^{(\sigma)}\}$ is the sequence $\{S_E^{(\sigma)}(z)\}$ of PHS with residuals $\{T_E^{(\sigma)^t}(z)\}$

[3] We assume here with loss of generality that $n_\beta \ge 0, 0 \le \beta \le k$, because if $n_\beta = -1$ for some β, we can simply remove n_β from n and $a_\beta(z)$ from $A^t(z)$ and decrease k by 1.

[4] By convention, the determinant of a null matrix is defined to be equal to 1.

and the sequence $\{S_E^{*(\sigma)}(z)\}$ of SPS with residuals $\{T_E^{*(\sigma)}(z)\}$. For $\sigma = 0$, set $\{S_E^{(0)}(z)\} = \{S_E^{*(0)}(z)\} = I_{k+1}$. We have that

$$A^t(z) \cdot S_E^{(\sigma)}(z) = z^{\|m^{(\sigma)}\|+1} T_E^{(\sigma)t}(z)$$

and

$$S_E^{*(\sigma)}(z) \cdot A^*(z) = z^{\|m^{(\sigma)}\|+1} T_E^{*(\sigma)}(z).$$

The following theorem provides a relation of the $(\sigma+1)$th exact systems in terms of the σth exact systems.

THEOREM 2. *For $\sigma \geq 0$ and $i > i_\sigma$, let $\nu = n^{(i)} - m^{(\sigma)} - e_0$. Then, the following statements are equivalent.*

1. *$n^{(i)}$ is a nonsingular point for $A(z)$ and $A^*(z)$.*
2. *ν is a nonsingular point for $T_E^{(\sigma)}(z)$.*
3. *ν is a nonsingular point for $T_E^{*(\sigma)}(z)$.*

Furthermore, we have the recurrence relations

$$S_E^{(\sigma+1)}(z) = S_E^{(\sigma)}(z) \cdot \widehat{S}_E(z), \quad T_E^{(\sigma+1)}(z) = \widehat{T}_E(z), \tag{36}$$

and

$$S_E^{*(\sigma+1)}(z) = \widehat{S}_E^*(z) \cdot S_E^{*(\sigma)}(z), \quad T_E^{*(\sigma+1)}(z) = \widehat{T}_E^*(z), \tag{37}$$

where $\widehat{S}_E(z)$ is the PHS of type $(m^{(\sigma+1)} - m^{(\sigma)} - e_0)$ for $T_E^{(\sigma)}(z)$ with residual $\widehat{T}_E(z)$ and $\widehat{S}_E^*(z)$ is the SPS of type $(m^{(\sigma+1)} - m^{(\sigma)} - e_0)$ for $T_E^{*(\sigma)}(z)$ with residual $\widehat{T}_E^*(z)$.

Proof. The proof for the NPHS is given in [14] and for the NSPS in [12]. □

Theorem 2 reduces the problem of determining a PHS and an SPS of types $m^{(\sigma+1)}$ to two smaller problems: determine systems of type $m^{(\sigma)}$ for the original power series and then determine systems of type $\nu = m^{(\sigma+1)} - m^{(\sigma)} - e_0$ for the residual power series. For the residual power series, the system $\widehat{S}_E(z)$ is obtained by solving the linear equations (13) and (17), where in the following the associated matrix is now denoted by $\widehat{\mathcal{M}}_\nu$ rather than by $\mathcal{M}_\nu$; and, the system $\widehat{S}_E^*(z)$ is obtained by solving the linear equations (30) and (32), where in the following the associated matrix is now denoted by $\widehat{\mathcal{M}}_\nu^*$ rather than by $\mathcal{M}_\nu^*$. The overhead cost of each step of this iterative scheme is the cost of determining the residual power series and the cost of combining the solutions, i.e., the cost of computing $S_E^{(\sigma+1)}(z)$ and $S_E^{*(\sigma+1)}(z)$ in (36) and (37). This overhead cost summed over all the steps, in general, is an order of magnitude less than the cost of solving the linear systems (13), (17), (30), and (32) directly.

Numerically, the recurrences (36) and (37) perform badly if $\mathcal{M}_{m^{(\sigma)}}$ and $\mathcal{M}_{m^{(\sigma)}}^*$ are ill conditioned at any point $m^{(\sigma)}$. Rather than moving from nonsingular point to nonsingular point along the diagonal, what we would like to do is move from a well-conditioned point to the next well-conditioned point. This is the motivation for the algorithm PHS_SPS, where the points $m^{(\sigma)}$, $\sigma = 0, 1, \ldots$, correspond to stable points rather than to nonsingular points and we step over unstable blocks.

A quantitative measure of the stability of a point $m^{(\sigma)}$ is provided by the stability parameter

$$\kappa^{(\sigma)} = \sum_{\beta=0}^{k} (\gamma_\beta^{(\sigma)} \gamma_\beta^{*(\sigma)})^{-1}. \tag{38}$$

It is shown in [9, 11] that $2\kappa^{(\sigma)}|a_0^{(0)}| \cdot \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\|$ is an upper bound for the condition numbers $\|\mathcal{M}_{m^{(\sigma)}}\| \cdot \|\mathcal{M}_{m^{(\sigma)}}^{-1}\|$ of $\mathcal{M}_{m^{(\sigma)}}$ and $\|\mathcal{M}_{m^{(\sigma)}}^{*}\|_\infty \cdot \|\mathcal{M}_{m^{(\sigma)}}^{*-1}\|_\infty$ of $\mathcal{M}_{m^{(\sigma)}}^{*}$. For the parameter (38), as well as considerations of §§5 and 6 it is assumed that $S^{(\sigma)}(z)$ and $S^{*(\sigma)}(z)$ are both scaled and that $\|a_\beta(z)\| \le 1, \quad 0 \le \beta \le k$. The norms used for the various scaling are defined in §1. In (38), it is also assumed that the residual errors $\delta T^{(\sigma)}(z)$ and $\delta T^{*(\sigma)}(z)$ in the order equations

$$A^t(z) \cdot S^{(\sigma)}(z) = z^{\|m^{(\sigma)}\|+1} T^{(\sigma)t}(z) + \delta T^{(\sigma)t}(z) \tag{39}$$

and

$$S^{*(\sigma)}(z) \cdot A^*(z) = z^{\|m^{(\sigma)}\|+1} T^{*(\sigma)}(z) + \delta T^{*(\sigma)}(z) \tag{40}$$

at the point $m^{(\sigma)}$ are relatively insignificant. We say that $m^{(\sigma)}$ is a *stable point* (or a well-conditioned point) if for some preassigned tolerance τ, $\kappa^{(\sigma)} \le \tau$. In the algorithm below, the user supplies the tolerance value τ.

PHS_SPS($A(z)$, n, k, τ)
$\sigma \leftarrow 0$; $m^{(0)} \leftarrow -e_0$; $S^{(0)} \leftarrow I_{k+1}$; $S^{*(0)} \leftarrow I_{k+1}$;
$M \leftarrow \min\{n_0, \max_{1\le\beta\le k}\{n_\beta\}\} + 1$
$i \leftarrow 0$; *stable* $\leftarrow$ true
While (($i < M$) and *stable*) do
  $\nu \leftarrow n - m^{(\sigma)} - e_0$
  $s \leftarrow 0$; *stable* $\leftarrow$ false
  While ($s < M - i$) and (not *stable*) do
    $s \leftarrow s + 1$
    $\nu_\beta^{(s)} \leftarrow \max\{0, \nu_\beta + i - M + s\}, \qquad \beta = 0, \ldots, k$
    Compute the residuals $T^{(\sigma)}(z)$ and $T^{*(\sigma)}(z)$ in (39) and (40)
    Construct the matrices $\mathcal{M}_{\nu^{(s)}}$ for $T^{(\sigma)}(z)$ and $\mathcal{M}_{\nu^{(s)}}^{*}$ for $T^{*(\sigma)}(z)$
    If $\mathcal{M}_{\nu^{(s)}}$ is numerically nonsingular then
      $m^{(\sigma+1)} \leftarrow m^{(\sigma)} + \nu^{(s)} + e_0$
      Obtain $\widehat{S}(z)$ by solving (13) and (17) by Gaussian elimination
      $S^{(\sigma+1)}(z) \leftarrow S^{(\sigma)}(z)\, \widehat{S}(z)$
      Scale $S^{(\sigma+1)}(z)$ and compute $\Gamma^{(\sigma+1)}$
      Obtain $\widehat{S}^*(z)$ by solving (30) and (32) by Gaussian elimination
      $S^{*(\sigma+1)}(z) \leftarrow \widehat{S}^*(z)\, S^{*(\sigma)}(z)$
      Scale $S^{*(\sigma+1)}(z)$ and compute $\Gamma^{*(\sigma+1)}$
      Using (38), compute $\kappa^{(\sigma+1)}$
      stable $\leftarrow \kappa^{(\sigma+1)} \le \tau$
    end If
  end While
  If *stable* then $\sigma \leftarrow \sigma + 1$; $i \leftarrow i + s$
end While
If *stable* then return $(S^{(\sigma)}(z), S^{*(\sigma)}(z), \kappa^{(\sigma)})$ else return $(S^{(\sigma+1)}(z), S^{*(\sigma+1)}(z), \kappa^{(\sigma+1)})$

In the algorithm above, by the numerical nonsingularity of a matrix, we mean that no zero pivot elements are encountered during the triangularization of the matrix by the Gaussian elimination method with partial pivoting.

5. Bounds on errors in the order conditions. In this section, we give bounds for the errors in the order conditions for the NPHS and the NSPS computed by the algorithm PHS_SPS. Some of the details of the derivations are omitted and can be found in [9]; in particular, for the NSPS the final result only (without proof) is given.

We begin by giving some standard results from the field of floating-point error analysis. Let μ denote the unit floating-point error and assume that the degrees of all polynomials and the orders of all matrices are bounded by some N, where $N\mu \leq 0.01$ (this restriction comes from Forsythe and Moler [16]). Indeed, as an assumption for all the lemmas and theorems below, we require that $(\|n\| + k + 1)\mu \leq 0.01$. After Wilkinson [28], we denote a floating-point operation by $fl[\cdot]$. In the following results, it is assumed that the operands consist of floating-point numbers.

LEMMA 3. *If* $\partial\mu \leq 0.01$, *then*

$$fl\left[\sum_{k=1}^{\partial} u_k v_k\right] = \sum_{k=1}^{\partial} u_k v_k(1+\delta_k),$$

where $|\delta_k| \leq 1.01\partial\mu$.

LEMMA 4. *If* $S(z)$ *is an NPHS of type* n *for* $A(z)$, *then*

$$fl[A^t(z)\cdot S(z)] = A^t(z)\cdot S(z) + \Psi^t(z),$$

where

$$\|\Psi^t(z)\| \leq 1.01\mu(\|n\| + k + 1)\|A^t(z)\|\cdot\|S(z)\|.$$

Proof. Using Lemma 3, for $0 \leq \beta \leq k$,

$$\begin{aligned} fl\left[\sum_{\alpha=0}^{k} a_\alpha(z) S_{\alpha,\beta}(z)\right] &= \sum_{\ell=0}^{\infty} z^\ell fl\left[\sum_{\alpha=0}^{k}\sum_{j=0}^{n_\alpha} a_\alpha^{(\ell-j)} S_{\alpha,\beta}^{(j)}\right] \\ &= \sum_{\ell=0}^{\infty} z^\ell \sum_{\alpha=0}^{k}\sum_{j=0}^{n_\alpha} a_\alpha^{(\ell-j)} S_{\alpha,\beta}^{(j)}\left(1+\delta_{\alpha,\beta,j,\ell}\right), \end{aligned}$$

where $|\delta_{\alpha,\beta,j,\ell}| \leq 1.01(n_\alpha + k + 1)\mu$. So,

$$\Psi_\beta(z) = \sum_{\ell=0}^{\infty} z^\ell \sum_{\alpha=0}^{k}\sum_{j=0}^{n_\alpha} a_\alpha^{(\ell-j)} S_{\alpha,\beta}^{(j)}\delta_{\alpha,\beta,j,\ell},$$

and

$$\begin{aligned} \|\Psi^t(z)\| &= \max_{0\leq\beta\leq k}\{\|\Psi_\beta(z)\|\} \\ &\leq \max_{0\leq\beta\leq k}\left\{\sum_{\ell=0}^{\infty}\sum_{\alpha=0}^{k}\sum_{j=0}^{n_\alpha} |a_\alpha^{(\ell-j)}|\cdot|S_{\alpha,\beta}^{(j)}|\cdot|\delta_{\alpha,\beta,j,\ell}|\right\} \\ &\leq 1.01\mu \max_{0\leq\beta\leq k}\left\{\sum_{\alpha=0}^{k}(n_\alpha+k+1)\sum_{j=0}^{n_\alpha}|S_{\alpha,\beta}^{(j)}|\sum_{\ell=0}^{\infty}|a_\alpha^{(\ell-j)}|\right\} \\ &\leq 1.01\mu \max_{0\leq\alpha\leq k}\{n_\alpha+k+1\}\|A^t(z)\| \max_{0\leq\beta\leq k}\left\{\sum_{\alpha=0}^{k}\|S_{\alpha,\beta}(z)\|\right\} \\ &\leq 1.01\mu(\|n\|+k+1)\|A^t(z)\|\cdot\|S(z)\|. \qquad \square \end{aligned}$$

We begin the analysis of the error in the order condition in the NSPS by first examining the floating-point errors introduced by one iteration of the algorithm. At the σth iteration, the NPHS $S^{(\sigma)}(z)$ of type $m^{(\sigma)}$ for $A^t(z)$ is available and satisfies

$$A^t(z) \cdot S^{(\sigma)}(z) = \delta T^{(\sigma)^t}(z) + \mathcal{O}(z^{\|m^{(\sigma)}\|+1}).$$

The algorithm proceeds to compute $S^{(\sigma+1)}(z)$ of type $m^{(\sigma+1)}$.

An iterative step consists of three parts. In the first part, the first $\|\nu^{(\sigma)}\| + 1$ terms of $T^{(\sigma)}(z)$ are computed; a bound for the floating-point errors introduced in this part is given in Lemma 5 below. In the second part, the NPHS $\widehat{S}^{(\sigma)}(z)$ of type $\nu^{(\sigma)}$ for $T^{(\sigma)}(z)$ is computed; an error analysis is given in Lemma 6. In the third part, Lemma 7 provides bounds for the floating-point errors introduced in computing $S^{(\sigma+1)}(z) = S^{(\sigma)}(z) \cdot \widehat{S}^{(\sigma)}(z)$. At this point in the algorithm, $S^{(\sigma+1)}(z)$ is scaled so that the norm of each column is 1. We assume for the sake of simplicity that this scaling introduces no additional errors. This is a reasonable assumption because errors due to scaling are comparatively insignificant.[5]

LEMMA 5. *The computed residual* $T^{(\sigma)}(z)$ *satisfies*

$$z^{\|m^{(\sigma)}\|+1} T^{(\sigma)^t}(z) = A^t(z) \cdot S^{(\sigma)}(z) - \delta T^{(\sigma)^t}(z) + z^{\|m^{(\sigma)}\|+1} \theta_{II}^{(\sigma)^t}(z),$$

where

$$\|\theta_{II}^{(\sigma)^t}(z)\| \leq 1.01(\|m^{(\sigma)}\| + k + 1) \cdot \mu.$$

Proof. The result is an immediate consequence of Lemma 4 because $A^t(z)$ and $S^{(\sigma)}(z)$ are both scaled. For details, see [9]. □

LEMMA 6. *If* $\widehat{\mathcal{M}}_{\nu^{(\sigma)}}$ *is nonsingular and* $\widehat{S}^{(\sigma)}(z)$ *is obtained by solving* (13) *and* (17), *then*

$$T^{(\sigma)^t}(z) \cdot \widehat{S}^{(\sigma)}(z) = \theta_{III}^{(\sigma)^t}(z) + O(z^{\|\nu^{(\sigma)}\|+1}),$$

where

$$\|\theta_{III}^{(\sigma)^t}(z)\| \leq (8\|\nu^{(\sigma)}\|^3 \cdot \rho_\sigma \cdot \mu + O(\mu^2)) \cdot \|\widehat{S}^{(\sigma)}(z)\|.$$

Proof. First we obtain bounds for the first component of $\theta_{III}^{(\sigma)^t}(z)$. The first column of $\widehat{S}^{(\sigma)}(z)$ corresponds to the solution $\widehat{\mathcal{X}}$ of (13) obtained by Gaussian elimination. The vector $\widehat{\mathcal{X}}$ is the exact solution of

$$(\widehat{\mathcal{M}}_{\nu^{(\sigma)}} + \mathcal{E}) \cdot \widehat{\mathcal{X}} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix},$$

where[6]

$$\|\mathcal{E}\|_1 \leq 8\|\nu^{(\sigma)}\|^3 \cdot \rho_\sigma \cdot \|\widehat{\mathcal{M}}_{\nu^{(\sigma)}}\|_1 \cdot \mu + O(\mu^2)$$

[5] Note also that $\widehat{S}^{(\sigma)}(z)$ can be determined a posteriori with appropriate values of $\hat{\gamma}^{(\sigma)}$ so that $S^{(\sigma+1)}(z)$ is already scaled. None of the subsequent error bounds would change, and so in reality this assumption is made without loss of generality.

[6] The results in [17] use the ∞-norm, but it is easy to show that they are also valid using the 1-norm. With partial pivoting, ρ_σ is of order unity in practice. Examples can be constructed, however, where the growth factor ρ grows exponentially if partial pivoting is used, but in practice ρ_σ is usually comparable to the modest growth that results when complete pivoting is used (which is approximately 10 in practice) [17, p. 69]. Further discussion and new results regarding the growth factor ρ_σ can be found in [21] and [26].

and ρ_σ is the growth factor associated with the LU-decomposition of $\widehat{\mathcal{M}}_{\nu^{(\sigma)}}$ [17, p. 67]. But, from Lemma 4,

$$\|T^{(\sigma)^t}(z)\| \le 1 + 1.01 \cdot (\|m^{(\sigma)}\| + k + 1) \cdot \mu,$$

because $A(z)$ and $S^{(\sigma)}(z)$ are both scaled. So,

$$\|\widehat{\mathcal{M}}_{\nu^{(\sigma)}}\|_1 \le \|T^{(\sigma)^t}(z)\| \le 1 + O(\mu).$$

Thus,

$$\widehat{\mathcal{M}}_{\nu^{(\sigma)}} \cdot \widehat{\mathcal{X}} - \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} = -\mathcal{E} \cdot \widehat{\mathcal{X}},$$

where

$$\|\mathcal{E} \cdot \widehat{\mathcal{X}}\|_1 \le \{8\|\nu^{(\sigma)}\|^3 \cdot \rho_\sigma \cdot \mu + O(\mu^2)\} \cdot \|\widehat{\mathcal{X}}\|_1.$$

A similar analysis can be done for solving (17) to obtain $\widehat{\mathcal{Y}}$. But $\widehat{\mathcal{X}}$ yields the first column of $\widehat{S}^{(\sigma)}(z)$ with residual error $\mathcal{E} \cdot \widehat{\mathcal{X}}$ and $\widehat{\mathcal{Y}}$ yields the remaining columns of $\widehat{S}^{(\sigma)}(z)$ with a corresponding residual error. Thus,

$$T^{(\sigma)^t}(z) \cdot \widehat{S}^{(\sigma)}(z) = \theta_{III}^{(\sigma)^t}(z) + O(z^{\|\nu^{(\sigma)}\|+1}),$$

where

$$\|\theta_{III}^{(\sigma)^t}(z)\| \le \left\{8\|\nu^{(\sigma)}\|^3 \cdot \rho_\sigma \cdot \mu + O(\mu^2)\right\} \cdot \|\widehat{S}^{(\sigma)}(z)\|. \qquad \square$$

LEMMA 7. *If* $S^{(\sigma+1)}(z) = fl(S^{(\sigma)}(z) \cdot \widehat{S}^{(\sigma)}(z))$, *then*

$$S^{(\sigma+1)}(z) = S^{(\sigma)}(z) \cdot \widehat{S}^{(\sigma)}(z) + \theta_{IV}^{(\sigma)}(z),$$

where

$$\|\theta_{IV}^{(\sigma)}(z)\| \le 1.01(\|\nu^{(\sigma)}\| + k + 1) \cdot \|S^{(\sigma)}(z)\| \cdot \|\widehat{S}^{(\sigma)}(z)\|\mu.$$

Proof. For $1 \le \alpha, \beta \le k$, the (α, β)-component of $S^{(\sigma+1)}(z)$ is

$$\begin{aligned}
&fl\left[z^2 q_\alpha(z) \cdot \widehat{u}_\beta(z) + \sum_{\rho=1}^{k} v_{\alpha,\rho}(z) \cdot \widehat{v}_{\rho,\beta}(z)\right] \\
&= fl\left[z^2 \sum_{\ell=0}^{m_\alpha^{(\sigma)}+\nu_0^{(\sigma)}-1} z^\ell \sum_{j=0}^{\nu_0^{(\sigma)}} q_\alpha^{(\ell-j)} \widehat{u}_\beta^{(j)} + \sum_{\rho=1}^{k} \sum_{\ell=0}^{m_\alpha^{(\sigma)}+\nu_\rho^{(\sigma)}} z^\ell \sum_{j=0}^{\nu_\rho^{(\sigma)}} v_{\alpha,\rho}^{(\ell-j)} \widehat{v}_{\rho,\beta}^{(j)}\right] \\
&= \sum_{\ell=0}^{m_\alpha^{(\sigma)}+\nu_0^{(\sigma)}-1} z^{\ell+2} \sum_{j=0}^{\nu_0^{(\sigma)}} q_\alpha^{(\ell-j)} \widehat{u}_\beta^{(j)} \cdot (1 + \delta_{\alpha,\beta,j,\ell,0}) \\
&\qquad + \sum_{\ell=0}^{m_\alpha^{(\sigma)}+\nu_\rho^{(\sigma)}} z^\ell \sum_{\rho=1}^{k} \sum_{j=0}^{\nu_\rho^{(\sigma)}} v_{\alpha,\rho}^{(\ell-j)} \widehat{v}_{\rho,\beta}^{(j)} \cdot (1 + \delta_{\alpha,\beta,j,\ell,\rho}),
\end{aligned}$$

where $|\delta_{\alpha,\beta,j,\ell,\rho}| \leq 1.01 \cdot (\nu_\rho^{(\sigma)} + k + 1) \cdot \mu$. Here, we have used Lemma 3 with the assumption that $(\|\nu^{(\sigma)}\| + k + 1)\mu \leq 0.01$. So,

$$\left(\theta_{IV}^{(\sigma)}(z)\right)_{\alpha,\beta} = z^2 \sum_{\ell=0}^{m_\alpha^{(\sigma)}+\nu_0^{(\sigma)}-1} z^\ell \sum_{j=0}^{\nu_0^{(\sigma)}} q_\alpha^{(\ell-j)} \cdot \widehat{u}_\beta^{(j)} \cdot \delta_{\alpha,\beta,j,\ell,0} + \sum_{\rho=1}^{k} \sum_{\ell=0}^{m_\alpha^{(\sigma)}+\nu_\rho^{(\sigma)}} z^\ell \sum_{j=0}^{\nu_\rho^{(\sigma)}} v_{\alpha,\rho}^{(\ell-j)} \cdot \widehat{v}_{\rho,\beta}^{(j)} \cdot \delta_{\alpha,\beta,j,\ell,\rho}.$$

Thus,

$$\left\| \left(\theta_{IV}^{(\sigma)}(z)\right)_{\alpha,\beta} \right\| \leq 1.01 \cdot (\|\nu^{(\sigma)}\| + k + 1) \cdot \left\{ \|q_\alpha(z)\| \cdot \|\widehat{u}_\beta(z)\| + \sum_{\rho=1}^{k} \|v_{\alpha,\rho}(z)\| \cdot \|\widehat{v}_{\rho,\beta}(z)\| \right\} \mu.$$

An equivalent result holds for $\alpha = \beta = 0$. The lemma now follows. □

The use of the results of the three lemmas above enables us to express the residual error $\delta T^{(\sigma+1)^t}(z)$ in the order condition at the $(\sigma+1)$th iteration in terms of the residual error $\delta T^{(\sigma)^t}(z)$ at the σth iteration plus the floating-point errors introduced "locally" by the σth iteration.

LEMMA 8.

$$\delta T^{(\sigma+1)^t}(z) = \delta T^{(\sigma)^t}(z) \cdot \widehat{S}^{(\sigma)}(z) + \mathcal{L}^{(\sigma)^t}(z), \tag{41}$$

where

$$\mathcal{L}^{(\sigma)^t}(z) = \left\{ A^t(z) \cdot \theta_{IV}^{(\sigma)}(z) + z^{\|m^{(\sigma)}\|+1} \left[\theta_{III}^{(\sigma)^t}(z) - \theta_{II}^{(\sigma)^t}(z) \cdot \widehat{S}^{(\sigma)}(z) \right] \right\} \pmod{z^{\|m^{(\sigma+1)}\|+1}}.$$

Proof. The result is an immediate consequence of Lemmas 5, 6, and 7. □

Thus, the residual error $\delta T^{(\sigma+1)^t}(z)$ is composed of the local error $\mathcal{L}^{(\sigma)^t}(z)$ introduced by the σth iteration plus the residual error $\delta T^{(\sigma)^t}(z)$ from the previous iteration propagated by $\widehat{S}^{(\sigma)}(z)$. Applying (41) recursively, we obtain the following.

THEOREM 9. *The residual error satisfies*

$$\delta T^{(\sigma+1)^t}(z) = \sum_{j=0}^{\sigma} \mathcal{L}^{(j)^t}(z) \cdot \mathcal{G}_j^{(\sigma)}(z), \tag{42}$$

where

$$\mathcal{G}_j^{(\sigma)}(z) = \begin{cases} \widehat{S}^{(j+1)}(z) \cdot \widehat{S}^{(j+2)}(z) \cdots \widehat{S}^{(\sigma)}(z), & 0 \leq j < \sigma, \\ I_{k+1}, & j = \sigma. \end{cases} \tag{43}$$

Proof. The result follows by induction from Lemma 8. □

From (42), we see that the residual error $\delta T^{(\sigma+1)^t}(z)$ is composed of the local errors $\mathcal{L}^{(j)^t}(z)$ propagated by $\mathcal{G}_j^{(\sigma)}$. Lemmas 5, 6, and 7 provide bounds for $\mathcal{L}^{(j)^t}(z)$. To obtain a bound for $\delta T^{(\sigma+1)^t}(z)$, it remains to determine bounds for the propagation matrices $\mathcal{G}_j^{(\sigma)}$. The concern is that the $\widehat{S}^{(j)}(z)$ making up $\mathcal{G}_j^{(\sigma)}$ will cause $\mathcal{G}_j^{(\sigma)}$ to grow

exponentially with σ. The next lemma and theorem show that this is not the case; a bound is obtained for $\mathcal{G}_j^{(\sigma)}$ which is independent of σ. Hence, the local error $\mathcal{L}^{(j)^t}(z)$ introduced at iteration j and propagated to iteration $\sigma+1$ by $\mathcal{G}_j^{(\sigma)}$ does not grow with σ. Thus, in this sense, the error grows additively; that is, $\delta T^{(\sigma+1)^t}(z)$ is bounded by the sum of the bounds of the local errors at each iteration j.

LEMMA 10. *If μ is so small and $\delta T^{(\sigma)^t}(z)$ and $\delta T^{*(\sigma)}(z)$ are not too large so that*

$$\kappa^{(\sigma)} \cdot |a_0^{(0)}| \cdot \Big\{ \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\| \Big[(k+1)\|\delta T^{(\sigma)^t}(z)\| + \|\delta T^{*(\sigma)}(z)\| \Big] + 1.01(k+1)(\|\nu^{(\sigma)}\| + k + 1) \cdot \mu \Big\} \le \frac{1}{2},$$

then

$$\|\widehat{S}^{(\sigma)}(z)\| \le 2\kappa^{(\sigma)} \cdot (k+1) \cdot |a_0^{(0)}|.$$

Proof. From (38),

$$\|(\Gamma^{*(\sigma)} \cdot \Gamma^{(\sigma)})^{-1} \cdot S^{*(\sigma)}(z) \cdot S^{(\sigma+1)}(z)\| \le \|(\Gamma^{*(\sigma)} \cdot \Gamma^{(\sigma)})^{-1}\| \cdot \|S^{*(\sigma)}(z)\| \cdot \|S^{(\sigma+1)}(z)\| \le \kappa^{(\sigma)} \cdot (k+1).$$

But, using Lemma 7 and Theorem 1 (adjusted to apply at the point $m^{(\sigma)}$ rather than at n),

$$\begin{aligned}
&\|(\Gamma^{*(\sigma)} \cdot \Gamma^{(\sigma)})^{-1} \cdot S^{*(\sigma)}(z) \cdot S^{(\sigma+1)}(z)\| \\
&= \Big\| (\Gamma^{*(\sigma)} \cdot \Gamma^{(\sigma)})^{-1} \cdot S^{*(\sigma)}(z) \cdot \Big\{ S^{(\sigma)}(z) \cdot \widehat{S}^{(\sigma)}(z) + \theta_{IV}^{(\sigma)}(z) \Big\} \Big\| \\
&= \Big\| (\Gamma^{*(\sigma)} \cdot \Gamma^{(\sigma)})^{-1} \cdot \Big\{ z^{\|m^{(\sigma)}\|+1} \cdot (a_0^{(0)})^{-1} \cdot \Gamma^{*(\sigma)} \cdot \Gamma^{(\sigma)} + \theta_I^{(\sigma)}(z) \Big\} \cdot \widehat{S}^{(\sigma)}(z) \\
&\quad + (\Gamma^{*(\sigma)} \cdot \Gamma^{(\sigma)})^{-1} \cdot S^{*(\sigma)}(z) \cdot \theta_{IV}^{(\sigma)}(z) \Big\| \\
&\ge |a_0^{(0)}|^{-1} \cdot \|\widehat{S}^{(\sigma)}(z)\| \\
&\quad - \|(\Gamma^{*(\sigma)} \cdot \Gamma^{(\sigma)})^{-1}\| \cdot \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\| \\
&\quad \cdot \Big[(k+1)\|\delta T^{(\sigma)^t}(z)\| + \|\delta T^{*(\sigma)}(z)\| \Big] \cdot \|\widehat{S}^{(\sigma)}(z)\| \\
&\quad - \|(\Gamma^{*(\sigma)} \cdot \Gamma^{(\sigma)})^{-1}\| \cdot \Big\{ 1.01 \cdot (\|\nu^{(\sigma)}\| + k + 1) \Big\} \cdot \|S^{(\sigma)}(z)\| \cdot \|\widehat{S}^{(\sigma)}(z)\| \cdot \|S^{*(\sigma)}(z)\| \cdot \mu \\
&\ge \|\widehat{S}^{(\sigma)}(z)\| \cdot \Big\{ |a_0^{(0)}|^{-1} - \kappa^{(\sigma)} \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\| \\
&\quad \cdot \Big[(k+1)\|\delta T^{(\sigma)^t}(z)\| + \|\delta T^{*(\sigma)}(z)\| \Big] - 1.01\, \kappa^{(\sigma)} \cdot (\|\nu^{(\sigma)}\| + k + 1) \cdot (k+1) \cdot \mu \Big\} \\
&\ge |a_0^{(0)}|^{-1} \cdot \|\widehat{S}^{(\sigma)}(z)\|/2.
\end{aligned}$$

The result now follows. □

THEOREM 11. *If μ is so small and $\delta T^{(j)^t}(z)$ and $\delta T^{*(j)}(z)$ are not too large so that*

$$\kappa^{(\sigma)} \cdot |a_0^{(0)}| \cdot \Big\{ \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\| \Big[(k+1)\|\delta T^{(j)^t}(z)\| + \|\delta T^{*(j)}(z)\| \Big] + 1.01(k+1)(\|\nu^{(\sigma)}\| + k + 1) \cdot \mu \Big\} \le \frac{1}{2},$$

then

$$\|\mathcal{G}_{j-1}^{(\sigma)}(z)\| \leq 2\kappa^{(j)} \cdot (k+1) \cdot |a_0^{(0)}| + O(\mu), \quad j \leq \sigma.$$

Proof. From (43) and Lemma 7

$$S^{(\sigma+1)}(z) = S^{(j)}(z) \cdot \mathcal{G}_{j-1}^{(\sigma)}(z) + \sum_{\ell=j}^{\sigma} \theta_{IV}^{(\ell)}(z) \cdot \mathcal{G}_{\ell}^{(\sigma)}(z).$$

We proceed by induction. Assume the theorem is true for $\mathcal{G}_{\sigma-1}^{(\sigma)}(z)$, $\mathcal{G}_{\sigma-2}^{(\sigma)}(z), \ldots, \mathcal{G}_{j}^{(\sigma)}(z)$ (the initial case, $j = \sigma - 1$, is proved in Lemma 10 because $\mathcal{G}_{\sigma-1}^{(\sigma)}(z) = \widehat{S}^{(\sigma)}(z)$). From (38),

$$\|(\Gamma^{*(j)} \cdot \Gamma^{(j)})^{-1} S^{*(j)}(z) \cdot S^{(\sigma+1)}(z)\| \leq \kappa^{(j)}(k+1).$$

But, using Lemma 7, Theorem 1, and the inductive hypothesis,

$$\begin{aligned}
&\|(\Gamma^{*(j)} \cdot \Gamma^{(j)})^{-1} \cdot S^{*(j)}(z) \cdot S^{(\sigma+1)}(z)\| \\
&\geq \Bigg\| (\Gamma^{*(j)} \cdot \Gamma^{(j)})^{-1} \cdot S^{*(j)}(z) \cdot S^{(j)}(z) \cdot \mathcal{G}_{j-1}^{(\sigma)}(z) \\
&\quad + \sum_{\ell=j}^{\sigma} (\Gamma^{*(j)} \cdot \Gamma^{(j)})^{-1} \cdot S^{*(j)}(z) \cdot \theta_{IV}^{(\ell)}(z) \cdot \mathcal{G}_{\ell}^{(\sigma)}(z) \Bigg\| \\
&\geq \left\| (\Gamma^{*(j)} \cdot \Gamma^{(j)})^{-1} \cdot \left\{ z^{\|m^{(j)}\|+1} (a_0^{(0)})^{-1} \Gamma^{*(j)} \cdot \Gamma^{(j)} + \theta_I^{(j)}(z) \right\} \cdot \mathcal{G}_{j-1}^{(\sigma)}(z) \right\| \\
&\quad - \kappa^{(j)} \sum_{\ell=j}^{\sigma} \{k+1\} \cdot \left\{ 2.02\kappa^{(\ell)} \cdot (k+1) \cdot (\|\nu^{(\ell)}\| + k + 1) \cdot |a_0^{(0)}| \cdot \mu \right\} \\
&\qquad\qquad \cdot \left\{ 2\kappa^{(\ell+1)} \cdot (k+1) \cdot |a_0^{(0)}| + O(\mu) \right\} \\
&\geq \|\mathcal{G}_{j-1}^{(\sigma)}(z)\| \left\{ |a_0^{(0)}|^{-1} - \kappa^{(j)} \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\| \right. \\
&\qquad\qquad \left. \cdot \left[(k+1)\|\delta T^{(j)^t}(z)\| + \|\delta T^{*(j)}(z)\| \right] \right\} - O(\mu) \\
&\geq |a_0^{(0)}|^{-1} \|\mathcal{G}_{j-1}^{(\sigma)}(z)\|/2 - O(\mu). \qquad \square
\end{aligned}$$

In the above theorem, we have taken the liberty of replacing a summation involving terms linear in μ with an $O(\mu)$ expression. We could have left the summation in explicitly, but, as we shall see, this summation becomes quadratic in μ when it is used to obtain a bound on $\delta T^{(\sigma)^t}(z)$.

Finally, we can give the bound on the residual error.

THEOREM 12. *If μ is so small and $\delta T^{(j)^t}(z)$ and $\delta T^{*(j)}(z)$ are not too large so that*

$$(\|n\| + k + 1)\mu \leq 0.01$$

and

$$\begin{aligned}
\kappa^{(j)} \cdot |a_0^{(0)}| \cdot \Big\{ &\|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\| \left[(k+1)\|\delta T^{(j)^t}(z)\| + \|\delta T^{*(j)}(z)\| \right] \\
&+ 1.01(k+1)(\|\nu^{(j)}\| + k + 1) \cdot \mu \Big\} \leq \frac{1}{2}, \quad j \leq \sigma,
\end{aligned}$$

then

$$\|\delta T^{(\sigma+1)^t}(z)\| \le F_\sigma + 2(k+1)\cdot|a_0^{(0)}|\sum_{j=0}^{\sigma-1}\kappa^{(j+1)}F_j, \tag{44}$$

where

$$F_j = 4\kappa^{(j)}(k+1)\cdot|a_0^{(0)}|\cdot\mu \cdot\Big\{(\|m^{(j)}\|+k+1)+4\rho_j\|\nu^{(j)}\|^3+(\|\nu^{(j)}\|+k+1)\Big\} \tag{45}$$

and ρ_j is the growth factor associated with the LU-decomposition of $\widehat{\mathcal{M}}_{\nu^{(j)}}$ by Gaussian elimination.

Proof. To simplify the analysis, we now split the local error $\mathcal{L}^{(\sigma)^t}(z)$ into three parts and analyze the propagation of each part separately. Let

$$\mathcal{L}_1^{(\sigma)^t}(z) = \begin{cases} 0, & \sigma = 0, \\ -z^{\|m^{(\sigma)}\|+1}\,\theta_{II}^{(\sigma)^t}(z)\,\widehat{S}^{(\sigma)}(z)\ (\bmod z^{\|m^{(\sigma+1)}\|+1}), & \sigma \ge 1, \end{cases} \tag{46}$$

$$\mathcal{L}_2^{(\sigma)^t}(z) = z^{\|m^{(\sigma)}\|+1}\theta_{III}^{(\sigma)^t}(z)\ (\bmod z^{\|m^{(\sigma+1)}\|+1}), \qquad \sigma \ge 0, \tag{47}$$

$$\mathcal{L}_3^{(\sigma)^t}(z) = \begin{cases} 0, & \sigma = 0, \\ A^t(z)\theta_{IV}^{(\sigma)}(z)\ (\bmod z^{\|m^{(\sigma+1)}\|+1}), & \sigma \ge 1, \end{cases} \tag{48}$$

and define

$$\mathcal{E}_i^{(\sigma+1)}(z) \;=\; \sum_{j=0}^{\sigma}\mathcal{L}_i^{(j)^t}(z)\cdot\mathcal{G}_j^{(\sigma)}(z), \qquad i=1,2,3. \tag{49}$$

Then, according to Lemma 8 and Theorem 9,

$$\delta T^{(\sigma+1)^t}(z) \;=\; \sum_{i=1}^{3}\mathcal{E}_i^{(\sigma+1)^t}(z)\,.$$

We now bound $\mathcal{E}_i^{(\sigma+1)^t}(z)$, $1 \le i \le 3$.

From (46), (49), Lemmas 5 and 10, and Theorem 11,

$$\begin{aligned}
\|\mathcal{E}_1^{(\sigma+1)^t}(z)\| &= \left\|\sum_{j=0}^{\sigma}\mathcal{L}_1^{(j)^t}(z)\cdot\mathcal{G}_j^{(\sigma)}(z)\right\| \\
&\le \|\theta_{II}^{(\sigma)^t}(z)\|\cdot\|\widehat{S}^{(\sigma)}(z)\| + \sum_{j=0}^{\sigma-1}\|\theta_{II}^{(j)^t}(z)\|\cdot\|\widehat{S}^{(j)}(z)\|\cdot\|\mathcal{G}_j^{(\sigma)}(z)\| \\
&\le \Big\{1.01(\|m^{(\sigma)}\|+k+1)\cdot\mu\Big\}\Big\{2\kappa^{(\sigma)}(k+1)|a_0^{(0)}|\Big\} \\
&\quad + \sum_{j=0}^{\sigma-1}\Big\{1.01(\|m^{(j)}\|+k+1)\mu\Big\}\cdot\Big\{2\kappa^{(j)}(k+1)|a_0^{(0)}|\Big\} \\
&\qquad\qquad \cdot\Big\{2\kappa^{(j+1)}(k+1)|a_0^{(0)}|+O(\mu)\Big\} \\
&\le 4\kappa^{(\sigma)}\cdot(k+1)\cdot(\|m^{(\sigma)}\|+k+1)\cdot|a_0^{(0)}|\cdot\mu
\end{aligned} \tag{50}$$

$$+ \ 8(k+1)^2 \cdot |a_0^{(0)}|^2 \cdot \mu \sum_{j=0}^{\sigma-1} \kappa^{(j)} \cdot \kappa^{(j+1)} \cdot (\|m^{(j)}\| + k + 1)$$
$$+ \ O(\mu^2).$$

From (47), (49), Lemmas 6 and 10, and Theorem 11,

$$\|\mathcal{E}_2^{(\sigma+1)^t}(z)\| = \left\| \sum_{j=0}^{\sigma} \mathcal{L}_2^{(j)^t}(z) \mathcal{G}_j^{(\sigma)}(z) \right\|$$

$$\leq \|\theta_{III}^{(\sigma)^t}(z)\| + \sum_{j=0}^{\sigma-1} \|\theta_{III}^{(j)^t}(z)\| \cdot \|\mathcal{G}_j^{(\sigma)}(z)\| \tag{51}$$

$$\leq \left\{8\|\nu^{(\sigma)}\|^3 \cdot \rho_\sigma \cdot \mu + O(\mu^2)\right\} \cdot \|\widehat{S}^{(\sigma)}(z)\|$$
$$+ \sum_{j=0}^{\sigma-1} \left\{8\|\nu^{(j)}\|^3 \cdot \rho_j \cdot \mu + O(\mu^2)\right\} \cdot \|\widehat{S}^{(j)}(z)\| \cdot \|\mathcal{G}_j^{(\sigma)}(z)\|$$

$$\leq \left\{8\|\nu^{(\sigma)}\|^3 \cdot \rho_\sigma \cdot \mu + O(\mu^2)\right\} \cdot \left\{2\kappa^{(\sigma)}(k+1)|a_0^{(0)}|\right\}$$
$$+ \sum_{j=0}^{\sigma-1} \left\{8\|\nu^{(j)}\|^3 \cdot \rho_j \cdot \mu + O(\mu^2)\right\} \cdot \left\{2\kappa^{(j)}(k+1)|a_0^{(0)}|\right\}$$
$$\cdot \left\{2\kappa^{(j+1)}(k+1)|a_0^{(0)}| + O(\mu)\right\}$$

$$\leq 16 \cdot \kappa^{(\sigma)} \cdot (k+1)\|\nu^{(\sigma)}\|^3 \cdot \rho_\sigma \cdot |a_0^{(0)}| \cdot \mu$$
$$+32(k+1)^2 \cdot |a_0^{(0)}|^2 \sum_{j=0}^{\sigma-1} \kappa^{(j)} \cdot \kappa^{(j+1)} \cdot \rho_j \cdot \|\nu^{(j)}\|^3 \cdot \mu$$
$$+O(\mu^2).$$

From (48), (49), Lemmas 7 and 10, and Theorem 11,

$$\|\mathcal{E}_3^{(\sigma+1)^t}(z)\| = \left\| \sum_{j=0}^{\sigma} \mathcal{L}_3^{(j)^t}(z) \cdot \mathcal{G}_j^{(\sigma)}(z) \right\|$$

$$\leq \|A^t(z) \cdot \theta_{IV}^{(\sigma)}(z)\| + \sum_{j=0}^{\sigma-1} \|A^t(z) \cdot \theta_{IV}^{(j)}(z)\| \cdot \|\mathcal{G}_j^{(\sigma)}(z)\| \tag{52}$$

$$\leq 1.01(\|\nu^{(\sigma)}\| + k + 1) \cdot \|\widehat{S}^{(\sigma)}(z)\| \cdot \mu$$
$$+ \sum_{j=0}^{\sigma-1} \left\{1.01(\|\nu^{(j)}\| + k + 1) \cdot \mu\right\} \cdot \|\widehat{S}^{(j)}(z)\| \cdot \|\mathcal{G}_j^{(\sigma)}(z)\|$$

$$\leq \left\{1.01 \cdot (\|\nu^{(\sigma)}\| + k + 1) \cdot \mu\right\} \cdot \left\{2\kappa^{(\sigma)}(k+1)|a_0^{(0)}|\right\}$$
$$+ \sum_{j=0}^{\sigma-1} \left\{1.01(\|\nu^{(j)}\| + k + 1) \cdot \mu\right\} \cdot \left\{2\kappa^{(j)}(k+1)|a_0^{(0)}|\right\}$$
$$\cdot \left\{2\kappa^{(j+1)}(k+1)|a_0^{(0)}| + O(\mu)\right\}$$

$$\leq 4\kappa^{(\sigma)} \cdot (k+1) \cdot (\|\nu^{(\sigma)}\| + k + 1) \cdot |a_0^{(0)}| \cdot \mu$$
$$+8(k+1)^2 \cdot |a_0^{(0)}|^2 \cdot \mu \sum_{j=0}^{\sigma-1} \kappa^{(j)}\kappa^{(j+1)}(\|\nu^{(j)}\| + k + 1)$$
$$+O(\mu^2).$$

The result follows by summing (50), (51), and (52). □

In Theorem 12, the bound for $\delta T^{(\sigma+1)^t}(z)$ involves the products $\kappa^{(j)}\kappa^{(j+1)}$. These result from inequalities involving the expression $\|\widehat{S}^{(j)}(z)\| \cdot \|\mathcal{G}_j^{(\sigma)}(z)\|$. However, it is seen that $\widehat{S}^{(j)}(z) \cdot \mathcal{G}_j^{(\sigma)}(z) = \mathcal{G}_{j-1}^{(\sigma)}(z)$, so it is felt that the inequalities are crude and the bounds should just involve a single $\kappa^{(j)}$. Experimental results [10] support this conjecture.

This completes the analysis of the error in the order condition for computing an NPHS. Proceeding in an analogous manner we can obtain the following theorem which gives bounds for the error in the order condition for the NSPS computed by PHS_SPS.

THEOREM 13. *If the conditions of Theorem* 12 *are satisfied, then*

$$\|\delta T^{*(\sigma+1)}(z)\| \leq F_\sigma^* + 2(k+1) \cdot |a_0^{(0)}| \sum_{j=0}^{\sigma-1} \kappa^{(j+1)} F_j^*, \tag{53}$$

where

$$F_j^* = 8\kappa^{(j)}(k+1)^2 \cdot |a_0^{(0)}| \cdot \mu$$
$$\cdot \left\{ (\|m^{(j)}\| + 1) + 4(k+1)^5 \rho_j^* \|\nu^{(j)}\|^3 + (\|\nu^{(j)}\| + k + 1) \right\} \tag{54}$$

and ρ_j^* is the growth factor associated with the LU-decomposition of $\widehat{\mathcal{M}}^*_{\nu^{(j)}}$ by Gaussian elimination.

Proof. See [9]. □

Theorems 12 and 13 assure us that if $\|\delta T^{(\sigma)^t}(z)\|$ and $\delta T^{*(\sigma)}(z)$ are small and $\kappa^{(\sigma)}$ is not too large, then $\|\delta T^{(\sigma+1)^t}(z)\|$ and $\delta T^{*(\sigma+1)}(z)$ will also be small. Thus, $\|\delta T^{(\sigma)^t}(z)\|$ and $\delta T^{*(\sigma)}(z)$ will remain small for all σ as long as, at every iteration j, a step $\nu^{(j)}$ is chosen (stepping over unstable blocks) so that $\kappa^{(j)}$ is not too large. Consequently, the assumptions of Theorems 12 and 13 are satisfied in practice.

6. Stability. In this section, bounds for the errors $\delta S(z) = S(z) - S_E(z)$ and $\delta S^*(z) = S^*(z) - S_E^*(z)$ are obtained. Because $S(z)$ and $S^*(z)$ are scaled, these same bounds serve also as bounds for the relative errors in $S(z)$ and $S^*(z)$. To make the comparisons meaningful in the above, we insist that $S_E(z)$ and $S_E^*(z)$ are such that

$$V_E(0) = V(0) = \text{diag}[\gamma_1, \ldots, \gamma_k],$$
$$r_E(0) = r(0) = \gamma_0,$$

and

$$v_E^*(0) = v^*(0) = \gamma_0^*,$$
$$R_E^*(0) = R^*(0) = \text{diag}[\gamma_1^*, \ldots, \gamma_k^*].$$

We begin by first finding bounds for $\delta S(z)$. From (6) and (10)

$$A^t(z) \cdot \delta S(z) = \delta T^t(z) + \mathcal{O}(z^{\|n\|+1}).$$

So, the constant terms[7] $\delta u_\beta^{(0)}$ and $\delta v_{\alpha,\beta}^{(0)}$ for $0 \le \alpha, \beta \le k$ of $S(z)$ are zero. It then follows that the remaining components of $\delta S(z)$ satisfy

$$\mathcal{M}_n \cdot \delta\mathcal{X} = [\delta r^{(0)}, \ldots, \delta r^{(\|n\|-1)}]^t, \tag{55}$$

where

$$\delta\mathcal{X} = \left[\delta p^{(0)}, \ldots, \delta p^{(n_0-1)} | \delta q_1^{(0)}, \ldots, \delta q_1^{(n_1-1)} | \cdots | \delta q_k^{(0)}, \ldots, \delta q_k^{(n_k-1)}\right]^t,$$

and

$$\mathcal{M}_n \cdot \delta\mathcal{Y} = \begin{bmatrix} \delta w_1^{(1)} & \cdots & \delta w_k^{(1)} \\ \vdots & & \vdots \\ \delta w_1^{(\|n\|)} & \cdots & \delta w_k^{(\|n\|)} \end{bmatrix}, \tag{56}$$

where

$$\delta\mathcal{Y} = \left[\begin{array}{ccc|ccc|c|ccc} \delta u_1^{(1)} & \cdots & \delta u_1^{(n_0)} & \delta v_{1,1}^{(1)} & \cdots & \delta v_{1,1}^{(n_1)} & \cdots & \delta v_{k,1}^{(1)} & \cdots & \delta v_{k,1}^{(n_k)} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ \delta u_k^{(1)} & \cdots & \delta u_k^{(n_0)} & \delta v_{1,k}^{(1)} & \cdots & \delta v_{1,k}^{(n_1)} & & \delta v_{k,k}^{(1)} & \cdots & \delta v_{k,k}^{(n_k)} \end{array}\right]^t.$$

From (55) and (56), it follows that

$$\begin{aligned} \|\delta S(z)\| &\le \max\{\|\delta\mathcal{X}\|_1, \|\delta\mathcal{Y}\|_1\} \\ &\le \|\mathcal{M}_n^{-1}\|_1 \cdot \max\{\|\delta r(z)\|, \|\delta W^t(z)\|\} \\ &\le \|\mathcal{M}_n^{-1}\|_1 \cdot \|\delta T^t(z)\|. \end{aligned} \tag{57}$$

Thus, to obtain a bound for $\delta S(z)$, we need only to obtain bounds for $\mathcal{M}_n^{-1}$ and $\delta T^t(z)$. This is done formally in Theorem 15. But first, in a similar fashion, we show that bounds for $\delta S^*(z)$ can be expressed in terms of bounds for $\mathcal{M}_n^{*-1}$ and $\delta T^*(z)$.

From (22) and (26),

$$S^*(z)A^*(z) = \delta T^*(z) + \mathcal{O}(z^{\|n\|+1}).$$

As for the NSPS, for the sake of simplicity, here again we ignore that the constant term errors $\delta w_\beta^{*(0)}$ for $1 \le \beta \le k$. This is done with no great loss of generality because these are the comparatively small errors made in computing $\delta u_\beta^{*(0)}(z)$ from

$$u_\beta^{*(0)} a_0^{(0)} + v^{*(0)} a_\beta^{(0)} = 0$$

with $v^{*(0)} = \gamma_0^*$. It then follows, in a fashion similar to solving (30) and (32), that the remaining components of $\delta S^*(z)$ satisfy

$$\delta\mathcal{X}^{*t} \cdot \mathcal{M}_n^* = [\delta w_1^{*(1)}, \ldots, \delta w_1^{*(\|n\|)} | \cdots | \delta w_k^{*(1)}, \ldots, \delta w_k^{*(\|n\|)}], \tag{58}$$

[7] In fact, the computations in (15) may yield errors resulting in nonzero values of $\delta u_\beta^{(0)}$ for $1 \le \beta \le k$. But, these errors, each resulting from two floating-point operations, are comparatively small and are ignored in order to simplify the analysis.

where

$$\delta\mathcal{X}^{*t} = \left[\delta v^{*(1)}, \ldots, \delta v^{*(\|n\|-n_0)} | \delta u_1^{*(1)}, \ldots, \delta u_1^{*(\|n\|-n_1)} | \cdots | \delta u_k^{*(1)}, \ldots, \delta u_k^{*(\|n\|-n_k)}\right],$$

and, for $1 \le \alpha \le k$,

$$\delta\mathcal{Y}_\alpha^{*t} \cdot \mathcal{M}_n^* = [\delta r_{\alpha,1}^{*(0)}, \ldots, \delta r_{\alpha,1}^{*(\|n\|-1)} | \cdots | \delta r_{\alpha,k}^{*(0)}, \ldots, \delta r_{\alpha,k}^{*(\|n\|-1)}], \tag{59}$$

where

$$\delta\mathcal{Y}_\alpha^{*t} = \left[\delta q_\alpha^{*(0)}, \ldots, \delta q_\alpha^{*(\|n\|-n_0-1)} | \delta p_{\alpha,1}^{*(0)}, \ldots, \delta p_{\alpha,1}^{*(\|n\|-n_1-1)} | \cdots | \delta p_{\alpha,k}^{*(0)}, \ldots, \delta p_{\alpha,k}^{*(\|n\|-n_k-1)}\right].$$

From (58) and (59), we get

$$\begin{aligned} \|\delta S^*(z)\| &\le (k+1) \max_{1\le\alpha\le k} \{\|\delta\mathcal{X}^*\|_1, \|\delta\mathcal{Y}_\alpha^*\|_1\} \\ &\le (k+1)^2 \|\mathcal{M}_n^{*-1}\|_\infty \cdot \|\delta T^*(z)\|. \end{aligned} \tag{60}$$

We are now ready to give the main results of this paper in the two theorems to follow; the first theorem shows that the algorithm PHS_SPS is weakly stable, whereas the second provides bounds for the errors $\delta S(z)$ and $\delta S^*(z)$. But, first note some notational details. Let $\delta T^t(z)$ and $\delta T^*(z)$ denote the residual errors corresponding, respectively, to the NPHS and NSPS computed by the algorithm PHS_SPS in $\sigma+1$ steps. So, $n = m^{(\sigma+1)}$ and a bound for $\|\delta T^t(z)\|$ is given by Theorem 12 in which $\delta T^{(\sigma+1)^t}(z) = \delta T^t(z)$ and a bound for $\|\delta T^*(z)\|$ is given by Theorem 13 in which $\delta T^{*(\sigma+1)}(z) = \delta T^*(z)$. At the point $m^{(\sigma+1)}$, we drop the superscript $\sigma+1$ so that $\kappa = \kappa^{(\sigma+1)}$, $S(z) = S^{(\sigma+1)}(z)$, $S^*(z) = S^{*(\sigma+1)}(z)$, and so on. The point $m^{(\sigma)}$ is the last stable point (i.e., $\kappa^{(\sigma)} \le \tau$) prior to the point n along the diagonal passing through n. The point n itself need not be stable.

THEOREM 14. *The algorithm PHS_SPS for computing $S(z)$ and $S^*(z)$ is weakly stable.*

Proof. From (44), (53), (57), and (60), it follows that, if the problem is well conditioned (i.e., if the condition number κ associated with the matrices $\mathcal{M}_n$ and $\mathcal{M}_n^*$ is not too large), then the computed solution $S(z)$ is close to the exact solution $S_E(z)$ and $S^*(z)$ is close to the exact solution $S_E^*(z)$. The algorithm is therefore weakly stable [7]. □

Note that the bounds (44) and (53) for the residual errors $\delta T^t(z)$ and $\delta T^*(z)$ (and therefore also the weak stability of PHS_SPS) do not depend on $a_0^{-1}(z)$. So $\kappa^{(j)}$ defined by (38) (i.e., excluding the term $\|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\|$ that appears in the bounds for $\mathcal{M}_n^{-1}$ and $\mathcal{M}_n^{*-1}$ [11]) is an appropriate choice for a stability parameter. Bounds for the errors $\delta S(z)$ and $\delta S^*(z)$ in the solutions, given in Theorem 15 do, however, depend on $a_0^{-1}(z)$.

THEOREM 15. *If κ is not too large and $\delta T^t(z)$ and $\delta T^*(z)$ are sufficiently small,*[8]

[8] In addition to satisfying the assumptions of Theorem 12 at all the stable points $m^{(j)}$, $1 \le j \le \sigma$, at the final point $n = m^{(\sigma+1)}$, we require $\delta T^t(z)$ and $\delta T^*(z)$ to be sufficiently small so that

$$\left[(\kappa+1)(k+2)|a_0^{(0)}|(\|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\|+1)\right]^2 \left[(k+2)\|\delta T^t(z)\| + \|\delta T^*(z)\|\right] \le 1/8.$$

This assumption at the last point n is used in [11] in obtaining bounds for $\mathcal{M}_n^{-1}$ and $\mathcal{M}_n^{*-1}$. All these assumptions are easily satisfied if all the points, including the last one, are reasonably stable.

then

$$\|\delta S(z)\| \leq 2\kappa \cdot |a_0^{(0)}| \cdot \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\| \left\{ \bar{F}_\sigma + 2\tau(k+1) \cdot |a_0^{(0)}| \cdot \sum_{j=0}^{\sigma-1} \bar{F}_j \right\},$$

where

$$\bar{F}_j = 4\tau(k+1) \cdot |a_0^{(0)}| \cdot \mu \left\{ (\|m^{(j)}\| + k + 1) + 4\rho_j \|\nu^{(j)}\|^3 + (\|\nu^{(j)}\| + k + 1) \right\}$$

and

$$\|\delta S^*(z)\| \leq 2\kappa(k+1)^2 \cdot |a_0^{(0)}| \cdot \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\| \cdot \left\{ \bar{F}_\sigma^* + 2\tau(k+1) \cdot |a_0^{(0)}| \sum_{j=0}^{\sigma-1} \bar{F}_j^* \right\},$$

where

$$\bar{F}_j^* = 8\tau(k+1)^2 \cdot |a_0^{(0)}| \cdot \mu \cdot \left\{ (\|m^{(j)}\| + 1) + 4(k+1)^5 \rho_j^* \|\nu^{(j)}\|^3 + (\|\nu^{(j)}\| + k + 1) \right\}.$$

Proof. For κ not too large and $\delta T^t(z)$ and $\delta T^*(z)$ sufficiently small, bounds for $\mathcal{M}_n^{-1}$ and $\mathcal{M}_n^{*-1}$ are derived in [11] to be

$$\|\mathcal{M}_n^{-1}\|_1,\ \|\mathcal{M}_n^{*-1}\|_\infty \leq 2\kappa \cdot |a_0^{(0)}| \cdot \|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\|.$$

The results of the theorem follow from (57) and (60) using (44) and (53). □

7. Experimental results. Numerical experiments have been performed to compare the analysis of the algorithm with its practice. A summary of the conclusions is presented here; details appear in [10].

The algorithm PHS_SPS was implemented using Sun Fortran 1.3.1. All calculations were performed in double precision. The linear systems (13), (17), (30), and (32) arising at intermediate steps of the algorithm were solved using the LINPACK routines SGEFA and SGESL. The results were then compared with the exact answers, obtained via the Maple computer algebra system.

Tables A1 and A2 give the results of a small but typical experiment for which $n = (18, 19, 19)$ and $A^t(z) = [a_0(z), a_1(z), a_2(z)]$ with $a_0(z) = 1$ and with coefficients of $a_1(z)$, $a_2(z)$ randomly and uniformly distributed between -1 and 1 and then scaled. The tables give results at all intermediate points along the diagonal through n. In these tables, the errors (represented in scientific notation with two digits of accuracy and the exponent enclosed in parentheses) in the computed $S^{(j)}(z)$ and $S^{*(j)}$ and in the order conditions are given for two values of the stability parameter τ. The value $\tau = 10^4$ in Table A1 indicates a willingness to accept only those striped Sylvester matrices $\mathcal{M}_{m^{(j)}}$ and mosaic Sylvester matrices $\mathcal{M}_{m^{(j)}}^*$ with condition numbers less than 10^4, approximately (i.e., those for which $\kappa^{(j)} \leq 10^4$). Striped and mosaic Sylvester matrices not satisfying this criterion are assumed to lie in an unstable block and are skipped over. An unstable point is identified by the value "–" in the column labeled "j". In Table A2, the value $\tau = 10^9$ permits a much greater tolerance for ill conditioning and results in an expected deterioration in the accuracy.

TABLE A1
$a_0(z) = 1$. *Errors at intermediate steps:* $\tau = 10^4$.

j	$\kappa^{(j)}$	$\|\delta T^{(j)^t}(z)\|$	$\frac{\|\delta S^{(j)}(z)\|}{\|S_E^{(j)}(z)\|}$	$\|\delta T^{*(j)}(z)\|$	$\frac{\|\delta S^{*(j)}(z)\|}{\|S_E^{*(j)}(z)\|}$
1	1.2(2)	1.7(−18)	1.4(−16)	2.2(−18)	7.0(−17)
2	1.8(2)	1.0(−17)	6.5(−16)	2.0(−17)	6.5(−16)
3	1.6(2)	1.7(−17)	9.8(−16)	3.3(−17)	1.8(−15)
4	9.5(2)	1.6(−17)	8.3(−16)	6.6(−17)	2.0(−15)
5	6.6(2)	2.0(−17)	1.7(−15)	9.0(−17)	2.9(−15)
–	4.1(7)	2.3(−17)	1.2(−15)	8.7(−17)	2.1(−15)
6	1.1(3)	3.3(−17)	2.3(−15)	1.3(−16)	4.8(−15)
7	1.5(3)	3.6(−17)	1.2(−15)	1.2(−16)	6.6(−15)
8	9.1(3)	5.6(−17)	1.8(−15)	1.9(−16)	4.4(−15)
9	3.7(3)	8.2(−17)	4.9(−15)	2.2(−16)	1.3(−14)
10	2.9(3)	1.2(−16)	3.3(−15)	3.9(−16)	1.2(−14)
–	3.2(6)	7.7(−17)	5.6(−15)	5.7(−16)	4.6(−14)
11	2.0(3)	2.8(−16)	8.1(−15)	5.6(−16)	1.4(−14)
–	1.6(4)	2.8(−16)	7.4(−15)	4.5(−16)	1.8(−14)
12	2.9(3)	2.9(−16)	9.5(−15)	6.8(−16)	2.2(−14)
–	4.1(4)	2.5(−16)	1.0(−14)	7.5(−16)	2.3(−14)
–	6.3(4)	2.7(−15)	2.4(−14)	8.2(−16)	2.9(−14)
–	1.1(4)	2.3(−16)	1.7(−14)	8.9(−16)	3.3(−14)
–	1.1(5)	2.5(−16)	1.3(−13)	8.0(−16)	1.4(−13)

TABLE A2
$a_0(z) = 1$. *Errors at intermediate steps:* $\tau = 10^9$.

j	$\kappa^{(j)}$	$\|\delta T^{(j)^t}(z)\|$	$\frac{\|\delta S^{(j)}(z)\|}{\|S_E^{(j)}(z)\|}$	$\|\delta T^{*(j)}(z)\|$	$\frac{\|\delta S^{*(j)}(z)\|}{\|S_E^{*(j)}(z)\|}$
1	1.2(2)	1.7(−18)	1.4(−16)	2.2(−18)	7.0(−17)
2	1.8(2)	1.0(−17)	6.5(−16)	2.0(−17)	6.5(−16)
3	1.6(2)	1.7(−17)	9.8(−16)	3.3(−17)	1.8(−15)
4	9.5(2)	1.6(−17)	8.3(−16)	6.6(−17)	2.0(−15)
5	6.6(2)	2.0(−17)	1.7(−15)	9.0(−17)	2.9(−15)
6	4.1(7)	2.3(−17)	1.2(−15)	8.7(−17)	2.1(−15)
7	1.1(3)	6.9(−13)	3.4(−11)	3.2(−12)	1.2(−10)
8	1.5(3)	6.8(−13)	1.9(−11)	3.7(−12)	1.5(−10)
9	9.1(3)	1.1(−12)	3.7(−11)	6.6(−12)	5.6(−10)
10	3.7(3)	1.6(−12)	9.5(−11)	6.4(−12)	3.7(−10)
11	2.9(3)	1.1(−12)	7.3(−11)	9.5(−12)	3.5(−10)
12	3.2(6)	1.2(−12)	1.7(−10)	1.2(−11)	1.9(−9)
13	2.0(3)	5.0(−12)	1.3(−10)	8.6(−12)	2.2(−10)
14	1.6(4)	4.9(−12)	1.4(−10)	8.3(−12)	1.9(−10)
15	2.9(3)	3.3(−12)	1.1(−10)	1.5(−11)	3.5(−10)
16	4.1(4)	3.6(−12)	1.1(−10)	9.8(−12)	3.5(−10)
17	6.3(4)	2.4(−12)	1.5(−10)	1.3(−11)	6.5(−10)
18	1.1(4)	2.8(−12)	1.8(−10)	1.1(−11)	4.4(−10)
19	1.1(5)	3.7(−12)	2.2(−10)	1.3(−11)	8.1(−10)

Tables B1 and B2 give the results of a similar experiment but for which $a_0(z)$, $a_1(z)$, and $a_2(z)$ were all first randomly generated (except that $a_0^{(0)}$ is initially set to 1) and then modified so as to introduce some pronounced instabilities. To introduce an instability at $m^{(j+1)}$, the coefficients of $a_1(z)$ and $a_2(z)$ were changed to make almost dependent the columns of coefficient matrix $\widehat{\mathcal{M}}_\nu$ corresponding to the residual $T^{(j)^t}(z)$ at the point $m^{(j)}$. The power series were then scaled. For this particular experiment, $\|a_0^{-1}(z) \pmod{z^{\|n\|+1}}\| = 2.3 \times 10^2$, approximately.

TABLE B1
Random $a_0(z)$. Errors at intermediate steps: $\tau = 10^5$.

j	$\kappa^{(j)}$	$\|\delta T^{(j)^t}(z)\|$	$\frac{\|\delta S^{(j)}(z)\|}{\|S_E^{(j)}(z)\|}$	$\|\delta T^{*(j)}(z)\|$	$\frac{\|\delta S^{*(j)}(z)\|}{\|S_E^{*(j)}(z)\|}$
1	3.2(0)	0.0	9.8(−17)	6.9(−18)	7.6(−17)
2	3.9(3)	1.5(−17)	7.1(−17)	1.7(−17)	4.7(−16)
3	3.7(3)	3.6(−17)	6.6(−16)	2.5(−17)	2.9(−15)
4	7.7(3)	1.0(−16)	5.7(−15)	3.6(−17)	2.7(−15)
–	6.4(14)	1.1(−16)	1.0(−14)	4.5(−17)	3.6(−10)
5	1.1(4)	9.3(−17)	1.5(−14)	5.7(−17)	8.4(−15)
–	3.8(5)	9.2(−17)	1.3(−14)	4.1(−16)	2.0(−14)
6	1.1(4)	1.1(−16)	8.5(−15)	4.2(−16)	2.2(−14)
–	1.3(14)	1.1(−16)	2.1(−14)	2.1(−16)	8.2(−10)
7	3.9(4)	1.2(−16)	7.7(−15)	4.2(−16)	3.5(−14)
–	3.8(8)	9.4(−17)	3.2(−11)	4.3(−16)	5.1(−10)
–	1.9(9)	8.9(−17)	1.7(−10)	4.1(−16)	7.1(−10)
–	1.1(15)	9.0(−17)	2.7(−10)	4.0(−16)	2.8(−9)
–	1.3(9)	9.2(−17)	1.9(−10)	4.5(−16)	1.0(−9)
–	2.1(5)	3.3(−16)	6.2(−14)	4.4(−16)	3.5(−14)
8	3.0(4)	3.2(−16)	6.9(−14)	4.2(−16)	4.9(−14)
–	1.4(13)	3.2(−16)	2.3(−13)	5.1(−16)	6.9(−10)
9	6.4(4)	5.4(−16)	7.6(−13)	6.0(−16)	2.1(−13)
–	2.3(5)	5.5(−16)	5.3(−13)	2.3(−15)	4.6(−13)

TABLE B2
Random $a_0(z)$. Errors at intermediate steps: $\tau = 10^9$.

j	$\kappa^{(j)}$	$\|\delta T^{(j)^t}(z)\|$	$\frac{\|\delta S^{(j)}(z)\|}{\|S_E^{(j)}(z)\|}$	$\|\delta T^{*(j)}(z)\|$	$\frac{\|\delta S^{*(j)}(z)\|}{\|S_E^{*(j)}(z)\|}$
1	3.2(0)	0.0	9.8(−17)	6.9(−18)	7.6(−17)
2	3.9(3)	1.5(−17)	7.1(−17)	1.7(−17)	4.7(−16)
3	3.7(3)	3.6(−17)	6.6(−15)	2.5(−17)	2.9(−15)
4	7.7(3)	1.0(−16)	5.7(−15)	3.6(−17)	2.7(−15)
–	6.4(14)	1.1(−16)	1.0(−14)	4.5(−17)	3.6(−10)
5	1.1(4)	9.3(−17)	1.5(−14)	5.7(−17)	8.4(−15)
6	3.8(5)	9.2(−17)	1.3(−14)	4.1(−16)	2.0(−14)
7	1.1(4)	2.2(−16)	1.1(−14)	1.6(−15)	1.1(−13)
–	1.3(14)	1.1(−16)	1.1(−14)	6.7(−15)	7.7(−9)
8	3.9(4)	2.5(−16)	1.1(−14)	4.8(−15)	2.3(−13)
9	3.8(8)	1.7(−16)	1.6(−10)	6.0(−15)	4.1(−9)
–	1.9(9)	1.6(−16)	2.9(−10)	8.9(−15)	1.6(−8)
–	1.1(15)	1.1(−16)	1.0(−9)	8.2(−15)	4.1(−8)
–	1.3(9)	1.3(−16)	1.6(−10)	6.9(−15)	1.3(−8)
10	2.1(5)	1.3(−12)	1.9(−10)	2.2(−13)	2.1(−10)
11	3.0(4)	1.9(−11)	2.3(−9)	8.3(−13)	2.8(−10)
–	1.4(13)	7.2(−12)	1.1(−9)	1.6(−12)	1.4(−6)
12	6.4(4)	1.7(−11)	1.3(−9)	3.8(−12)	1.0(−9)
13	2.3(5)	3.4(−11)	1.1(−9)	2.1(−11)	3.7(−9)

It was observed that the large powers of k that occur in the bounds derived above are not manifested in the experiments. Also, $\|\delta T^t(z)\|$ and $\|\delta T^*(z)\|$ appear to depend on $\kappa^{(j)}$ and not $\kappa^{(j)}\kappa^{(j+1)}$ and the overall error is proportional to the largest $\kappa^{(j)}$ encountered. Thus, the bounds are crude, but they do appear to reflect the behavior of the error. As Wilkinson points out [29, p. 567], "The main object of such an analysis is to expose the potential instabilities, if any, of an algorithm so that hopefully from the insight thus obtained one might be led to improved algorithms. Usually the bound itself is weaker than it might have been because of the necessity

of restricting the mass of detail to a reasonable level and because of the limitations imposed by expressing the errors in terms of matrix norms."

From these and other experiments [10], operational bounds on the errors in the order conditions (as for the case $k = 1$ reported in [15]) appear to be

$$\|\delta T^t(z)\| \leq C(k+1)\mu \left(\sum_{j=0}^{\sigma} \kappa^{(j)} \rho_j \|m^{(j)}\| \right) + O(\mu^2)$$

and

$$\|\delta T^*(z)\| \leq C(k+1)^2\mu \left(\sum_{j=0}^{\sigma} \kappa^{(j)} \rho_j \|m^{(j)}\|^2 \right) + O(\mu^2),$$

where C is a moderate constant. In addition, for the errors in the solutions, operational bounds appear to be

$$\|\delta S(z)\| \leq C\kappa(k+1)\mu \left(\sum_{j=0}^{\sigma} \kappa^{(j)} \rho_j \|m^{(j)}\| \right) + O(\mu^2)$$

and

$$\|\delta S^*(z)\| \leq C\kappa(k+1)^3\mu \left(\sum_{j=0}^{\sigma} \kappa^{(j)} \rho_j \|m^{(j)}\|^2 \right) + O(\mu^2).$$

8. Conclusions. In this paper we have presented a new, fast, weakly stable algorithm for the computation of PHS and SPS. The algorithm requires $\mathcal{O}(\|n\|^2 + s^3\|n\|)$ operations to compute a PHS and an SPS of type $n = [n_0, \ldots, n_k]$, where $\|n\| = n_0 + \cdots + n_k$ and s is the largest distance from one well-conditioned subproblem to the next. The algorithm can also be used for fast stable inversion of striped or mosaic Sylvester matrices (see [20] for the case $k = 1$ and $a_0(z) = 1$). The algorithm relies on the ability to specify when a given subproblem is well conditioned. The stability estimates come as a result of "near" inversion formulae for striped and mosaic Sylvester matrices given in [11]. In addition to a complete stability analysis, we have also provided some numerical experiments that verify that the algorithm performs as the theoretic results imply.

There is a number of open research problems that result from this work. The algorithm that has been presented is fast rather than superfast as is possible in the case of exact arithmetic [12]. It is possible to modify the algorithm so that it takes steps in a quadratic fashion as done in [12]. However, while this approach will work in the generic case, it is possible to find examples where not all the required subproblems are well conditioned. In these cases the algorithm might not be numerically stable. It would be of interest to find a superfast algorithm that works in all cases and in addition is numerically stable.

In cases where the largest step-size is small the algorithm has complexity $\mathcal{O}(\|n\|^2)$. However, there are cases where the algorithm may require a very large step-size and then have a higher cost than Gaussian elimination. This will happen if there is a very large unstable block, or if the stability parameter τ is chosen to be too low. It would be of interest to find a fast, stable algorithm that has complexity $\mathcal{O}(\|n\|^2)$ in all cases.

Our algorithm proceeds along a diagonal path in the corresponding Padé tables of our approximants. It would be of interest to find fast, stable algorithms that proceed along alternate paths in the Padé tables. An example of this in the Padé case is found in [18] where the computation proceeds along straight-line paths. In the context of matrix solvers this is the difference between giving a Toeplitz solver instead of a Hankel solver as is done in [15].

The *M-Padé approximation problem* is a generalization of the Padé–Hermite approximation problem which requires that the residual in (1) vanishes at a given set of knots $z_0, z_1, \ldots, z_{N-1}$, counting multiplicities [2, 3, 4, 24]. The case where all the z_i are equal to 0 is just the Padé–Hermite problem. In this case the coefficient matrix for the associated linear system is the matrix of divided differences. It would be of interest to determine stability parameters for such matrices, with a view to developing fast, stable algorithms for computing this approximation problem. Along these lines, some experiments for the case $k = 1$ are reported in [8].

Acknowledgment. We are very grateful to a referee who contributed much in terms of the correctness of results and the clarity of presentation.

REFERENCES

[1] G. Baker and P. Graves-Morris, *Padé Approximants, Part II*, Addison-Wesley, Reading, MA, 1981.

[2] B. Beckermann, *Zur Interpolation mit polynomialen Linearkombinationen beliebiger Funktionen*, Ph.D. thesis, Institut für Angewandte Mathematik, Universität Hannover, 1988.

[3] ———, *The structure of the singular solution table of the M-Padé approximation problem*, J. Comput. Appl. Math., 32 (1990), pp. 3–15.

[4] ———, *A reliable method for computing M-Padé approximants on arbitrary staircases*, J. Comput. Appl. Math., 40 (1992), pp. 19–42.

[5] B. Beckermann and G. Labahn, *A uniform approach for the fast computation of matrix-type Padé approximants*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 804–823.

[6] A. W. Bojanczyk, R. P. Brent, F. D. De Hoog, and D. R. Sweet, *On the stability of the Bariss and related Toeplitz factorization algorithms*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 40–57.

[7] J. R. Bunch, *The weak and strong stability of algorithms in numerical linear algebra*, Linear Algebra Appl., 88/89 (1987), pp. 49–66.

[8] S. Cabay, M. Gutknecht, and R. Meleshko, *Stable rational interpolation?*, Systems and Networks: Mathematical Theory and Applications, Proceedings of MTNS 93, (1994), pp. 631–633.

[9] S. Cabay, A. Jones, and G. Labahn, *A stable algorithm for multi-dimensional Padé systems and the inversion of generalized Sylvester matrices*, Tech. Report TR 94-07, Dept. Comp. Sci., Univ. Alberta, 1994.

[10] ———, *Experiments with a weakly stable algorithm for computing Padé-Hermite and simultaneous Padé approximants*, submitted to ACM Trans. Math. Software.

[11] ———, *Computation of numerical Padé-Hermite and simultaneous Padé systems I: Near inversion of generalized Sylvester matrices*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 248–267.

[12] S. Cabay and G. Labahn, *A superfast algorithm for multi-dimensional Padé systems*, Numerical Algorithms, 2 (1992), pp. 201–224.

[13] ———, *Fast, stable inversion of mosaic Hankel matrices*, Systems and Networks: Mathematical Theory and Applications, Proceedings of MTNS 93, (1994), pp. 625–630.

[14] S. Cabay, G. Labahn, and B. Beckermann, *On the theory and computation of non-perfect Padé-Hermite approximants*, J. Comput. Appl. Math., 39 (1992), pp. 295–313.

[15] S. Cabay and R. Meleshko, *A weakly stable algorithm for Padé approximants and inversion of Hankel matrices*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 735–765.

[16] G. E. Forsythe and C. B. Moler, *Computer Solution of Linear Algebraic Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[17] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[18] M. GUTKNECHT, *Stable row recurrences in the Padé table and generically superfast lookahead solvers for non-hermitian Toeplitz solvers*, Linear Algebra Appl., 188/189 (1993), pp. 351–421.

[19] M. H. GUTKNECHT AND M. HOCHBRUCK, *Look-ahead Levinson and Schur algorithms for non-Hermitian Toeplitz systems*, Tech. Report IPS 93–11, IPS-Zürich, 1993.

[20] ———, *The stability of inversion formulas for Toeplitz matrices*, Tech. Report IPS 93–13, IPS-Zürich, 1993.

[21] N. J. HIGHAM AND D. J. HIGHAM, *Large growth factors in Gaussian elimination with pivoting*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 155–164.

[22] T. JONES, *The numerical computation of Padé-Hermite systems*, master's thesis, Dept. Comp. Sci., Univ. Alberta, 1992.

[23] G. LABAHN, *Inversion components for block Hankel-like matrices*, Linear Algebra Appl., 177 (1992), pp. 7–48.

[24] K. MAHLER, *Perfect systems*, Compositio Math., 19 (1968), pp. 95–166.

[25] R. SHAFER, *On quadratic approximation*, SIAM J. Numer. Anal., 11 (1974), pp. 447–460.

[26] L. N. TREFETHEN AND R. S. SCHREIBER, *Average-case stability of Gaussian elimination*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 335–360.

[27] M. VAN BAREL AND A. BULTHEEL, *The computation of non-perfect Padé-Hermite approximants*, Numerical Algorithms, 1 (1991), pp. 285–304.

[28] J. WILKINSON, *Rounding Errors in Algebraic Processes*, Prentice-Hall, Englewood Cliffs, NJ, 1963.

[29] J. H. WILKINSON, *Modern error analysis*, SIAM Rev., 13 (1971), pp. 548–568.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 2, pp. 298–312, April 1996

004

ON TWO-SIDED BOUNDS RELATED TO WEAKLY DIAGONALLY DOMINANT M-MATRICES WITH APPLICATION TO DIGITAL CIRCUIT DYNAMICS*

P. N. SHIVAKUMAR†, JOSEPH J. WILLIAMS‡, QIANG YE§, AND CORNELIU A. MARINOV¶

Abstract. Let A be a real weakly diagonally dominant M-matrix. We establish upper and lower bounds for the *minimal* eigenvalue of A, for its corresponding eigenvector, and for the entries of the inverse of A. Our results are applied to find meaningful two-sided bounds for both the ℓ_1-norm and the weighted *Perron-norm* of the solution $x(t)$ to the linear differential system $\dot{x} = -Ax$, $x(0) = x_0 > 0$. These systems occur in a number of applications, including compartmental analysis and RC electrical circuits. A detailed analysis of a model for the transient behaviour of digital circuits is given to illustrate the theory.

Key words. weakly diagonally dominant matrix, M-matrix, bounds, digital circuit dynamics

AMS subject classifications. 15A42, 15A45, 15A48, 94C30

1. Introduction.

A strictly diagonally dominant matrix is invertible and moreover its inverse can be bounded. Results of this type are well known with the bounds depending on the minimal diagonal dominance [11], [16], and they have applications in problems such as estimating the condition number of a matrix. If the matrix is weakly diagonally dominant with at least one row being strictly diagonally dominant, there are conditions that guarantee invertibility, e.g., its irreducibility [17, p. 23] or, more generally, a chain condition [14]. However, the known results do not give a finite bound for the inverse. In this paper, we derive an upper bound for the infinity norm of the inverse of a weakly diagonally dominant M-matrix and a lower bound for the entries of its inverse. We also apply these results to bound the Perron root of the inverse of A and the components of the corresponding normalized eigenvector.

Our interest in these bounds is motivated by a problem related to a system of ordinary differential equations that arises from the study of the dynamics of digital circuits. The system is given by

$$\frac{dx}{dt} = -A\,x(t), \qquad x(0) = x_0 > 0, \tag{1}$$

where $x(t), x_0 \in \mathbb{R}^n$ and A is a constant real $n \times n$ weakly diagonally dominant M-matrix. Using the Perron–Frobenius theorem, we establish upper and lower bounds on the ℓ_1-norm of the solution, $x(t)$. In fact, we prove in §5 that

* Received by the editors October 28, 1994; accepted for publication (in revised form) by T. Ando May 7, 1995.

† Institute of Industrial Mathematical Sciences (IIMS) and Department of Applied Mathematics, University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada (insmath@cc.umanitoba.ca). The research of this author was supported in part by Natural Sciences and Engineering Research Council of Canada grant OGP0007899.

‡ Institute of Industrial Mathematical Sciences (IIMS) and Department of Applied Mathematics, University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada.

§ Institute of Industrial Mathematical Sciences (IIMS) and Department of Applied Mathematics, University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada. The research of this author was supported in part by Natural Sciences and Engineering Research Council of Canada grant OGPG0137369.

¶ Faculty of Electrotechnics, Polytechnical University of Bucharest, Bucharest 77206, Romania.

$$\frac{1}{R}e^{-qt} \leq \frac{\|x(t)\|_1}{\|x_0\|_1} \leq R\, e^{-qt}, \tag{2}$$

where $q = q(A) = 1/\rho(A^{-1})$, $\rho(A^{-1})$ is the spectral radius of A^{-1}, and $R = \max\{z_i/z_j : i,j = 1,2,\ldots,n\}$, where $z = (z_i)^T$ is the positive eigenvector of A^T, the transpose of A, that corresponds to $q(A) = q(A^T)$. When the matrix A is strongly diagonally dominant, there are earlier results giving upper bounds on $\|x(t)\|_1$ that exhibit exponential decay in t (see [7], for example). However, if at least one row is weakly diagonally dominant, then these bounds do not exhibit decay.

The paper is organized as follows. In §2, we present notation and some preliminary results. In §3, we derive upper and lower bounds on A^{-1}. In §4, we obtain upper and lower bounds on $q(A)$ and $\{z_i : i = 1,\ldots,n\}$ for irreducible A, for use in (2). In §5, we will establish (2) and then in §6 apply all of our results to study the dynamics and design of some digital circuits.

2. Preliminaries and notation. We begin by listing all conditions on the matrix A that will be assumed at some point of the paper. In particular, we will always assume (A_1), (A_2), and either (A_3) or (A_6). Let $N = \{1,2,\ldots,n\}$, and let A be an $n \times n$ matrix.

- (A_1) For all $i,j \in N$ with $i \neq j$, $a_{ij} \leq 0$ and $a_{ii} > 0$.
- (A_2) For all $i \in N$, $|a_{ii}| \geq \sum_{j\neq i} |a_{ij}|$ and $J(A) = \{i \in N : |a_{ii}| > \sum_{j\neq i} |a_{ij}|\} \neq \emptyset$.
- (A_3) A is irreducible.
- (A_4) For all $i \in N$ with $i \geq 2$, there exists a $j \in N$ such that $j < i$ and $a_{ij} \neq 0$.
- (A_5) For all i, $j \in N$, $a_{ij} \neq 0$ implies $a_{ji} \neq 0$.
- (A_6) Definition: A is w.c.d.d.(weakly chained diagonally dominant) if A satisfies (A_2), and for all $i \in N$, $i \notin J(A)$, there exist indices $i_1, i_2, \ldots, i_k$ in N with $a_{i_r,i_{r+1}} \neq 0$, $0 \leq r \leq k-1$, where $i_0 = i$ and $i_k \in J(A)$. (We call the above sequence of nonzero entries a chain from i to i_k.)

Remarks. A matrix A satisfying (A_1) is called an L-matrix. (A_2) is the definition of a weakly diagonally dominant matrix. If (A_2) holds with $J(A) = N$ then we say that A is strictly diagonally dominant. A is irreducible (A_3) if and only if for all $i,j \in N$ there is a chain (as in (A_6)) starting with i and ending with j [17, p. 20].

(A_2) and (A_3) state that A is irreducibly diagonally dominant, which implies that A is invertible [17, p. 23]. (A_3) implies that (A_4) is true with a suitable permutation of the indices (see Lemma 2.5 below). (A_5) states that the pattern of nonzero entries of A is symmetric. (A_6) w.c.d.d. has been called J-d.d. or Λ-d.d. in [3] and [15].

A matrix satisfying (A_1) and (A_6) is said to be of generalized positive type with respect to $\zeta = (1,1,\ldots,1)^T$ (see [1] or [18]). We note that if A is strictly diagonally dominant or if A is irreducibly diagonally dominant, then clearly A is w.c.d.d.

LEMMA 2.1. *A w.c.d.d. matrix (A_6) is nonsingular* [14].

For example, the matrix $\left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]$ is w.c.d.d. but is neither strictly diagonally dominant nor irreducible.

DEFINITION. *An $n \times n$ matrix A is an M-matrix if A is an L-matrix (A_1), nonsingular, and $A^{-1} \geq 0$ (i.e., each entry of A^{-1} is nonnegative)* [17].

LEMMA 2.2. *A w.c.d.d. L-matrix $((A_1),(A_6))$ is an M-matrix.*

See [14, Cor. 4] or [18].

We will denote $A^{(n_1,n_2)}$ as the principal submatrix of A formed from all rows and all columns with indices between n_1 and n_2 inclusively; e.g., $A^{(2,n)}$ is the submatrix of A obtained by deleting the first row and the first column of A.

For convenience, we will let the entries of any submatrix of A retain the same indices as they had in A (e.g., the top row of $A^{(2,n)}$ will be indexed as row two).

LEMMA 2.3. *Let A be an $n \times n$ w.c.d.d. M-matrix $((A_1),(A_6))$. Then $B = A^{(2,n)}$ is an $(n-1) \times (n-1)$ w.c.d.d. M-matrix (i.e., $B^{-1} = (\beta_{ij})$ exists and $\beta_{ij} \geq 0$, $i,j = 2,3,\ldots,n$).*

Proof. Let $J(B)$ be the J-set for B relating to (A_2). We show $J(B) \neq \emptyset$. The only difficulty is the case where $J(A) = \{1\}$. In this case, using (A_6), there is some k with $2 \leq k \leq n$ and $a_{k1} \neq 0$; then, using (A_2) and $a_{k1} \neq 0$, $|a_{kk}| > \sum_{j=2,j\neq k}^{n} |a_{kj}|$ and $k \in J(B) \neq \emptyset$. Thus, conditions (A_1) and (A_2) hold for B. To show (A_6) for B, let $i \in \{2,\ldots,n\}, i \notin J(B)$. Then $a_{ii} = \sum_{j=2,j\neq i}^{n} |a_{ij}|$ and $a_{i1} = 0$ (using (A_2) for A). Thus, $i \notin J(A)$, and there exists a chain of nonzero entries $a_{ii_1}, a_{i_1,i_2}, \ldots, a_{i_{k-1},i_k}$ with $i_k \in J(A)$. If all $i_r \neq 1$, then the above gives a chain in B from i to $i_k \in J(B)$. If some $i_r = 1$, then $a_{i_{r-1},1} \neq 0$, and using (A_2) for A, row number $m = i_{r-1}$ in B is strictly diagonally dominant and $a_{ii_1}, a_{i_1,i_2}, \ldots, a_{i_{r-2},i_{r-1}}$ gives a chain from i to $m \in J(B)$. So, B is w.c.d.d. From Lemma 2.2, B is nonsingular and $\beta_{ij} \geq 0$. □

Remarks. The above lemma is false if we replace the w.c.d.d. condition with the irreducible condition (A_3). For example,

$$A = \begin{bmatrix} 2 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix} \text{ satisfies } (A_1)\text{–}(A_3);$$

$$B = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \text{ is reducible } ((A_2) \text{ and } (A_3) \text{ fail});$$

however, B is w.c.d.d. It is primarily for this reason as well as for greater generality that we prefer to use w.c.d.d. rather than irreducible matrices (as much as possible). We next give some properties of irreducible matrices.

LEMMA 2.4. *An invertible matrix A is irreducible if and only if A^{-1} is irreducible.*

Proof. A is reducible if and only if there exists a permutation matrix P such that

$$B = PAP^T = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}, \tag{3}$$

where $B_{11} \in \mathbb{R}^{k\times k}$ and $B_{22} \in \mathbb{R}^{(n-k)\times(n-k)}$. Since B^{-1} has the same structure as B and $B^{-1} = PA^{-1}P^T$, A is reducible if and only if A^{-1} is reducible. □

LEMMA 2.5. (a) *Let A be an irreducible matrix. Then there exists a permutation matrix P such that $B = PAP^T$ satisfies condition (A_4) and $1 \in J(B)$.*

(b) *If A is irreducible and satisfies (A_5), then so does B, and for each $k \in N$, $B^{(1,k)}$ is irreducible.*

Proof. (a) is equivalent to finding a permutation $(i_1, i_2, \ldots, i_n)$ of $(1,2,\ldots,n)$ such that for each $j = 2,3,\ldots,n$, there exists a k with $1 \leq k < j$ and $a_{i_j,i_k} \neq 0$. Then P is formed by applying the above permutation to the rows of I.

Let $i_1 \in J(A)$. Let $i_2 \in N - \{i_1\}$ such that $a_{i_2,i_1} \neq 0$ (this exists since A is irreducible). For the same reason, there exist $i_3 \in N - \{i_1,i_2\}$ and $j \in \{i_1,i_2\}$ such that $a_{i_3,j} \neq 0$. Note that $j = i_k$ with $k < 3$. We repeat this process to find $i_4, i_5, \ldots, i_{n-1}$. The remaining index is i_n. By irreducibility, there exists a j with $1 \leq j \leq n$, $j \neq i_n$ such that $a_{i_n,j} \neq 0$; then $j = i_k$ for some $k < n$.

(b) follows easily by induction on k. $B^{(1,1)}$ and $B^{(1,2)}$ have all nonzero entries and thus are irreducible. The third row and column of $B^{(1,3)}$ each has a nonzero entry besides b_{33}, and thus, there exists a connection from 3 to 1 or 2 and back. Thus $B^{(1,3)}$ is irreducible (similarly for $B^{(1,4)}, \ldots, B^{(1,n)}$). (A connection from i to j means $b_{ij} \neq 0$.) □

DESCRIPTION OF $q(A)$ AND z. Assume that A is a w.c.d.d. M-matrix $((A_1), (A_6))$. Since $A^{-1} \geq 0$, by an extension of the *Perron–Frobenius* theorem [17, Thm. 2.7], the spectral radius of $A^{-1}, \rho(A^{-1}) = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } A^{-1}\}$ is an eigenvalue of A^{-1}. Since $Au = \lambda u$ if and only if $A^{-1}u = \lambda^{-1}u$, it follows that

$$q(A) = \frac{1}{\rho(A^{-1})}$$

is the *minimal* eigenvalue of A: $q(A)$ is an eigenvalue of A, and for any eigenvalue λ of A, $|\lambda| \geq q(A)$. If, in addition, A is irreducible, then by Lemma 2.4, so is A^{-1}. So the *Perron–Frobenius* theorem [17] tells us that $q(A^T)$ is a simple eigenvalue of A^T corresponding to a positive eigenvector, $z = (z_1, z_2, \ldots, z_n)^T > 0$, which is unique if we assume that $\|z\|_1 = \sum_{i \in N} |z_i| = 1$. Note that $z > 0$ means that $z_i > 0$ for all $i \in N$.

By the Gerschgorin theorem, $q(A) \leq \lambda$ for any real eigenvalue λ. Our results in §4 include finding a positive lower bound for $q(A)$, which will be an improvement over the Gerschgorin theorem which gives only $q(A) \geq 0$. Since A^T and A have the same eigenvalues, $q(A^T) = q(A)$.

We now introduce some notation related to the diagonal dominant condition. We define the relative row and column sums:

$$\rho_i = \frac{1}{|a_{ii}|} \sum_{j=1, j \neq i}^{n} |a_{ij}|, \qquad \chi_j = \frac{1}{|a_{jj}|} \sum_{i=1, i \neq j}^{n} |a_{ij}|,$$

the partial left and right row sums:

$$l_i = \frac{1}{|a_{ii}|} \sum_{j=1}^{i-1} |a_{ij}|, \qquad r_i = \frac{1}{|a_{ii}|} \sum_{j=i+1}^{n} |a_{ij}|,$$

and the partial upper and lower column sums:

$$u_j = \frac{1}{|a_{jj}|} \sum_{i=1}^{j-1} |a_{ij}|, \qquad d_j = \frac{1}{|a_{jj}|} \sum_{i=j+1}^{n} |a_{ij}|.$$

Clearly, $\rho_i = (l_i + r_i)$ and $\chi_j = (u_j + d_j)$.

If A is a w.c.d.d. M-matrix, then for each $i \in N$, $0 \leq \rho_i \leq 1$, $J(A) = \{i \in N : \rho_i < 1\} \neq \emptyset$, $\sum_{j \in N, j \neq i} a_{ij} = -a_{ii}\rho_i$, $\sum_{j \in N} a_{ij} = a_{ii}(1 - \rho_i)$, and $\sum_{j \in N} |a_{ij}| = a_{ii}(1 + \rho_i)$.

LEMMA 2.6. *If A is a w.c.d.d. matrix (A_6) and $A^{-1} = (\alpha_{ij})$, then for $i \neq j$,*

$$|\alpha_{ij}| \leq \rho_i |\alpha_{jj}| \leq |\alpha_{jj}|,$$

and if $i \in J(A)$,

$$\frac{1}{|a_{ii}|(1 + \rho_i)} \leq |\alpha_{ii}| \leq \frac{1}{|a_{ii}|(1 - \rho_i)}.$$

This result was proved by Ostrowski [11, Eqs. (13) and (14)] for a strictly diagonally dominant matrix (i.e., $\rho_i < 1$), but the same proof is valid for the case here. This can also be obtained through a perturbation argument.

3. Upper and lower bounds for A^{-1}. In this section, we derive upper and lower bounds for the inverse of a w.c.d.d. M-matrix (i.e., a matrix satisfying (A_1) and (A_6); see Lemma 2.2). Let $B = A^{(2,n)}$. We require expressions for the entries of A^{-1} in terms of those of B^{-1}.

LEMMA 3.1. *Let A be a w.c.d.d. M-matrix $((A_1),(A_6))$, $B = A^{(2,n)}$, $A^{-1} = (\alpha_{ij})_{i,j=1}^n$, and $B^{-1} = (\beta_{ij})_{i,j=2}^n$. Then, for $i,j = 2,\dots,n$,*

$$\alpha_{11} = \frac{1}{\Delta}, \tag{4}$$

$$\alpha_{i1} = \frac{1}{\Delta}\sum_{k=2}^{n}\beta_{ik}(-a_{k1}), \tag{5}$$

$$\alpha_{1j} = \frac{1}{\Delta}\sum_{k=2}^{n}\beta_{kj}(-a_{1k}), \tag{6}$$

and

$$\alpha_{ij} = \beta_{ij} + \alpha_{1j}\sum_{k=2}^{n}\beta_{ik}(-a_{k1}), \tag{7}$$

where

$$\Delta = a_{11} - \sum_{k=2}^{n} a_{1k}\left[\sum_{i=2}^{n}\beta_{ki}a_{i1}\right] \geq a_{11}(1-\rho_1) \tag{8}$$

and $\Delta > 0$.

Proof. By Lemmas 2.2 and 2.3, A and B are nonsingular and $\beta_{ij} \geq 0$. We partition

$$A = \begin{bmatrix} a_{11} & x^T \\ y & B \end{bmatrix}$$

and let

$$A^{-1} = (\alpha_{ij}) = \begin{bmatrix} \alpha_{11} & \xi^T \\ \eta & \Gamma \end{bmatrix},$$

where we have split off the first row and first column of A and A^{-1}. By expanding $AA^{-1} = I$, it can be verified that $\alpha_{11}\Delta = 1$, where

$$\Delta = a_{11} - x^T B^{-1} y. \tag{9}$$

Thus, $\Delta \neq 0$ and $\alpha_{11} = \Delta^{-1}$. Furthermore, $\eta = -\Delta^{-1}B^{-1}y$, which gives (5). Similarly, we have $\Gamma = B^{-1}(I - y\xi^T)$ and $\xi^T = -\Delta^{-1}x^TB^{-1}$, which give (6) and (7).

Now, expanding (9), we obtain

$$\begin{aligned}
\Delta &= a_{11} - \sum_{k=2}^{n} a_{1k}\left\{\sum_{i=2}^{n}\beta_{ki}a_{i1}\right\} \\
&= a_{11} - \sum_{k=2}^{n} a_{1k}\left\{\sum_{i=2}^{n}\beta_{ki}\left[\sum_{j=1}^{n}a_{ij} - \sum_{j=2}^{n}a_{ij}\right]\right\} \\
&= a_{11} - \sum_{k=2}^{n} a_{1k}\left[\sum_{i=2}^{n}\beta_{ki}a_{ii}(1-\rho_i) - \sum_{j=2}^{n}\sum_{i=2}^{n}\beta_{ki}a_{ij}\right].
\end{aligned}$$

ρ_i was defined in §2. Hence, for $k,\ j \geq 2$,

$$\sum_{i=2}^{n} \beta_{ki} a_{ij} = (B^{-1}B)_{kj} = \begin{cases} 1 & \text{if } k = j, \\ 0 & \text{if } k \neq j. \end{cases}$$

Thus,

$$\begin{aligned} \Delta &= a_{11} - \sum_{k=2}^{n} a_{1k} \left[\sum_{i=2}^{n} \beta_{ki} a_{ii} (1 - \rho_i) - 1 \right] \\ &= \sum_{k=1}^{n} a_{1k} + \sum_{k=2}^{n} (-a_{1k}) \sum_{i=2}^{n} \beta_{ki} a_{ii} (1 - \rho_i) \\ &\geq \sum_{k=1}^{n} a_{1k} = a_{11}(1 - \rho_1) \geq 0, \end{aligned}$$

where the last inequality follows from Lemma 2.3 and the conditions on A. Since $\Delta \neq 0$, $\Delta > 0$. □

By interchanging rows and corresponding columns (i.e., a permutation of the indices), we may assume that the first row of A is strictly diagonally dominant, i.e., $1 \in J(A)$.

LEMMA 3.2. *Let A and $B = A^{(2,n)}$ be as in Lemma* 3.1, *and assume that* $1 \in J(A)$, *i.e.,* $\rho_1 < 1$. *Then*

$$\|A^{-1}\|_\infty \leq \frac{1}{a_{11}(1 - \rho_1)} + \frac{\|B^{-1}\|_\infty}{1 - \rho_1}.$$

Proof. Let $s_i = \sum_{k=1}^{n} \alpha_{ik}$, $M_1 = \|A^{-1}\|_\infty$, and $M_2 = \|B^{-1}\|_\infty$. Then $M_1 = \max\{s_i : 1 \leq i \leq n\}$ and $M_2 = \max\{\sum_{k=2}^{n} \beta_{ik} : 2 \leq i \leq n\}$. Using Lemma 3.1,

$$\begin{aligned} s_1 &= \alpha_{11} + \sum_{k=2}^{n} \alpha_{1k} \\ &= \frac{1}{\Delta} + \sum_{k=2}^{n} \frac{1}{\Delta} \sum_{p=2}^{n} \beta_{pk}(-a_{1p}) \\ &= \frac{1}{\Delta} + \frac{1}{\Delta} \sum_{p=2}^{n} (-a_{1p}) \sum_{k=2}^{n} \beta_{pk} \\ &\leq \frac{1}{\Delta} + \frac{1}{\Delta} \sum_{p=2}^{n} (-a_{1p}) M_2 \\ &\leq \frac{1}{a_{11}(1 - \rho_1)} (1 + a_{11} \rho_1 M_2). \end{aligned} \tag{10}$$

Let $2 \leq i \leq n$. Then, using Lemma 2.6,

$$\alpha_{i1} \leq \alpha_{11}. \tag{11}$$

Using this in (4) and (5) gives $\sum_{k=2}^{n} \beta_{ik}(-a_{k1}) \leq 1$, and thus, from (7), with $2 \leq j \leq n$,

$$\alpha_{ij} \leq \beta_{ij} + \alpha_{1j}. \tag{12}$$

Thus, for $i \geq 2$,

$$\begin{aligned} s_i &= \alpha_{i1} + \sum_{k=2}^{n} \alpha_{ik} \\ &\leq \alpha_{11} + \sum_{k=2}^{n} (\beta_{ik} + \alpha_{1k}) \\ &\leq s_1 + M_2, \end{aligned}$$

where we have used (11) and (12). Thus, $M_1 \leq s_1 + M_2$. Using (10), the result now follows. □

If the matrix $B = A^{(2,n)}$ is strictly diagonally dominant, then the known results [11], [16] can be used to give a bound on $\|B^{-1}\|_\infty$. Thus, the lemma immediately gives a bound on $\|A^{-1}\|_\infty$.

Note that $\rho_1 = r_1$. It follows from Lemma 2.3 and the comments just prior to Lemma 3.2 that we can permute the indices so that the resulting matrix, which we still call A, has the property that for $k = 1, \ldots, n-1$, the matrix $A^{(k,n)}$ has its first row strictly diagonally dominant (within $A^{(k,n)}$), i.e., $r_k < 1$. Note that $r_n = 0$.

THEOREM 3.3. *Let A be an $n \times n$ w.c.d.d. M-matrix $((A_1), (A_6))$ such that $r_k < 1$ for all $k \in N$. Then, $\|A^{-1}\|_\infty \leq \sum_{i=1}^{n} [a_{ii} \prod_{j=1}^{i} (1 - r_j)]^{-1}$, i.e.,*

$$\|A^{-1}\|_\infty \leq \frac{1}{a_{11}(1-r_1)} + \frac{1}{a_{22}(1-r_1)(1-r_2)} + \cdots + \frac{1}{a_{nn}(1-r_1)\cdots(1-r_n)}.$$

Proof. Apply induction with respect to k to $A^{(k,n)}$, using Lemma 3.2. □

LEMMA 3.4. *Let A and B be as in Lemma 3.1. Let $m_1 = \min\{\sum_{j=1}^{n} \alpha_{ij}, i = 1, \ldots, n\}$, $m_2 = \min\{\sum_{j=2}^{n} \beta_{ij}, i = 2, \ldots, n\}$, and let $p_1 = a_{11}^{-1} + m_2 \rho_1$. Then*

$$m_1 \geq \min\left\{p_1, m_2\left(1 + p_1 \min_{2 \leq k \leq n} |a_{k1}|\right)\right\}.$$

Also,

$$m_1 \geq \left(\max_{1 \leq i \leq n} a_{ii}(1 + \rho_i)\right)^{-1}. \tag{13}$$

Proof. Let $s_i = \sum_{j=1}^{n} \alpha_{ij}$. Then, as in the proof of Lemma 3.2,

$$\begin{aligned} s_1 &= \frac{1}{\Delta} + \frac{1}{\Delta} \sum_{k=2}^{n} \sum_{p=2}^{n} \beta_{pk}(-a_{1p}) \\ &\geq \frac{1}{\Delta} + \frac{1}{\Delta} m_2 \sum_{k=2}^{n} (-a_{1k}) \\ &= \frac{1}{\Delta} + \frac{1}{\Delta} m_2 a_{11} \rho_1. \end{aligned}$$

From Lemma 3.1, $\Delta \leq a_{11}$. Thus,

$$s_1 \geq \frac{1}{a_{11}} + m_2 \rho_1 = p_1. \tag{14}$$

Let $i = 2, \ldots, n$. Then

$$\sum_{k=2}^{n} \beta_{ik}(-a_{k1}) \geq \left(\min_{2 \leq k \leq n} |a_{k1}| \right) m_2. \tag{15}$$

Using Lemma 3.1 and (15),

$$\begin{aligned} s_i &\geq \frac{1}{\Delta} \left(\min_{2 \leq k \leq n} |a_{k1}| \right) m_2 + \sum_{j=2}^{n} \left[\beta_{ij} + \alpha_{1j} \left(\min_{2 \leq k \leq n} |a_{k1}| \right) m_2 \right] \\ &\geq m_2 + \left(\frac{1}{\Delta} + \sum_{j=2}^{n} \alpha_{1j} \right) \left(\min_{2 \leq k \leq n} |a_{k1}| \right) m_2 \\ &= m_2 + s_1 \left(\min_{2 \leq k \leq n} |a_{k1}| \right) m_2 \qquad \text{(from (4))} \\ &\geq m_2 \left[1 + p_1 \min_{2 \leq k \leq n} |a_{k1}| \right] \qquad \text{(from (14))}. \end{aligned}$$

Since $m_1 = \min\{s_i : i = 1, \ldots, n\}$, the first result now follows from (14) and the above. The second inequality, (13), follows from Lemma 2.6 since

$$\sum_{j=1}^{n} \alpha_{ij} \geq \alpha_{ii} \geq \frac{1}{a_{ii}(1 + \rho_i)}. \qquad \square$$

Remark. Let A be an $n \times n$ w.c.d.d. M-matrix such that $r_k < 1$ for all $k \in N$. Then the result of Lemma 3.4 can be used inductively, starting with $m_n = {a_{nn}}^{-1}$, then using Lemma 3.4 applied to $A^{(n-1,n)}$ to find a lower bound for m_{n-1}, and so forth. The final result would have to be compared with (13).

We now consider lower bounds on the elements of A^{-1}. Recall that d_i, r_i were defined in §2.

THEOREM 3.5. *Let A be a w.c.d.d. M-matrix $((A_1), (A_6))$ and let $A^{-1} = (\alpha_{ij})$. Then*

$$\min_{j,k} \alpha_{jk} \geq \frac{1}{a_{nn}} \prod_{i=1}^{n-1} \min\{d_i, r_i\}.$$

Proof. Let $\gamma_1 = \min_{j,k} \alpha_{jk}$ and $\gamma_2 = \min_{j,k} \beta_{jk}$, where $B = A^{(2,n)}$ and $B^{-1} = (\beta_{ij})_{i,j=2}^{n}$. From Lemma 2.6, $\alpha_{11} \geq \alpha_{i1}$, and for $i \geq 2$, by Lemma 3.1,

$$\begin{aligned} \alpha_{i1} &= \frac{1}{\Delta} \sum_{k=2}^{n} \beta_{ik} |a_{k1}| \\ &\geq \frac{\gamma_2}{\Delta} \sum_{k=2}^{n} |a_{k1}| \geq \gamma_2 d_1, \end{aligned}$$

where $\Delta \leq a_{11}$ by Lemma 3.1. On the other hand, for $j \geq 2$,

$$\alpha_{1j} = \frac{1}{\Delta} \sum_{k=2}^{n} \beta_{kj} |a_{1k}| \geq \gamma_2 r_1$$

and for $2 \le i, j \le n$,

$$\begin{aligned}\alpha_{ij} &= \beta_{ij} + \alpha_{1j} \sum_{k=2}^{n} \beta_{ik} |a_{k1}| \\ &\ge \beta_{ij} \ge \gamma_2.\end{aligned}$$

Thus, we obtain

$$\gamma_1 \ge \gamma_2 \min\{d_1, r_1\}, \tag{16}$$

where $d_1, r_1 \le 1$. Note that B satisfies (A_1) and (A_6), by Lemma 2.3; then Theorem 3.5 follows by repeatedly applying (16) until we reach $\gamma_n = a_{nn}^{-1}$. □

Note that if the matrix A is reducible, then A^{-1} necessarily contains a zero entry and thus, $\min_{j,k} \alpha_{jk} = 0$. In this case, the above bound is trivial. We have, however, stated the result in terms of w.c.d.d. matrices (including reducible matrices) for convenience, since each principal submatrix $A^{(i,k)}$ is a w.c.d.d. matrix in the proof.

If some d_i or r_i is zero but the matrix is irreducible, our lower bound is zero. Note that the lower bound for the row sums obtained from Lemma 3.4 is still nontrivial. For the application problem in which we are interested, assumptions (A_4) and (A_5) hold. Then the above theorem yields a nontrivial lower bound in the following corollary.

COROLLARY 3.6. *Let A be a matrix that satisfies assumptions $(A_1) - (A_5)$. Then*

$$\min_{j,k} \alpha_{jk} \ge \frac{1}{a_{11}} \prod_{i=2}^{n} \min\{u_i, l_i\} > 0.$$

Proof. The first part follows from applying Theorem 3.5 to the permutation of A given by $(n, n-1, \ldots, 1)$. Now, from (A_4), $l_i > 0$ and from (A_5), $u_i > 0$. Thus, the lower bound is strictly positive. □

4. Upper and lower bounds for $q(A)$ and z. We continue to consider a w.c.d.d. M-matrix. We recall that $q(A) = 1/\rho(A^{-1})$ is the *minimal* eigenvalue of A.

THEOREM 4.1. *Let A be an $n \times n$ w.c.d.d. M-matrix, let $A^{-1} = (\alpha_{ij})$, and let $q = q(A)$. Then*

$$q \le \min\{a_{ii} \ : \ i \in N\}, \tag{17}$$

$$q \le \max\{a_{ii}(1 - \rho_i) : i \in N\} = \max\left\{ \sum_{j \in N} a_{ij} \ : \ i \in N \right\}, \tag{18}$$

$$q \ge \min\{a_{ii}(1 - \rho_i) : i \in N\} = \min\left\{ \sum_{j \in N} a_{ij} \ : \ i \in N \right\}, \tag{19}$$

and

$$\frac{1}{M} \le q \le \frac{1}{m}, \tag{20}$$

where

$$M = \max_{i \in N} \sum_{j \in N} \alpha_{ij} = \|A^{-1}\|_\infty \quad \text{and} \quad m = \min_{i \in N} \sum_{j \in N} \alpha_{ij}.$$

Moreover, since $q(A) = q(A^T)$, each of the above bounds remains valid when A and A^{-1} are replaced with their respective transposes; thus, row sums become column sums.

Proof. By replacing a_{ij} with $a_{ij} - \varepsilon$ for $i \neq j$ and a_{ii} with $a_{ii} + n\varepsilon$, $\varepsilon > 0$, and then taking $\varepsilon \to 0^+$, we may assume that A is irreducible. Then, let $z > 0$ be an eigenvector of A corresponding to q, i.e., $Az = qz$. For each $i \in N$, $\sum_{j \in N} a_{ij} z_j = q z_i$ and

$$(a_{ii} - q)\, z_i = \sum_{j \neq i} (-a_{ij}) z_j \geq 0. \tag{21}$$

Since $z_i > 0$, $q \leq a_{ii}$, and (17) follows. Let $z_m = \min\{z_j : j \in N\}$; then, from (21),

$$(a_{mm} - q) z_m \geq \sum_{j \neq m} (-a_{mj}) z_m.$$

Since $z_m > 0$, $q \leq a_{mm} + \sum_{j \neq m} a_{mj} = \sum_{j \in N} a_{mj}$, and (18) follows. (19) follows similarly from considering $z_M = \max\{z_j : j \in N\}$ (or the Gershgorin theorem [17, p. 16]). (20) can be proved similarly using $A^{-1} z = \rho(A^{-1}) z$ and is due to Frobenius [10, Thm. 1.1, p. 24]. □

Remark. If A is not strictly diagonally dominant, then (19) gives only $q \geq 0$; however, (20) together with Theorem 3.3 always gives a *positive* lower bound for q. In (20), Theorems 3.3 and 3.5 and Lemma 3.4 can be used to give bounds for q in terms of the (a_{ij}); when using Theorem 3.3, we first must permute the indices of A (or A^T) so that $r_k < 1$ for all $k \in N$. On the other hand, since $\|A^{-1}\|_\infty \geq 1/q$ from (20) the above upper bounds for q will give lower bounds for $\|A^{-1}\|_\infty$.

Finding bounds on $\rho(A)$ or $q(A)$ is a subject of interest on its own and various refined bounds can be found in Chapter 2 of [10]. However, (20) is the only one that is applicable here.

We now give upper and lower bounds for the components of the eigenvector z for an irreducible matrix, which will be used in the next section in our application to systems of linear ordinary differential equations.

THEOREM 4.2. *Let A satisfy (A_1)–(A_3), $A^{-1} = (\alpha_{ij})$, and let $z = (z_1, z_2, \ldots, z_n)^T$ be the positive eigenvector of A corresponding to $q(A)$ with $\|z\|_1 = 1$. Then*

$$q(A) \min_{j,k} \alpha_{jk} \leq z_i \leq q(A) \max_{j,k} \alpha_{jk}.$$

Furthermore,

$$\max_{i,j} \frac{z_i}{z_j} \leq \max_{k,j} \frac{\alpha_{kk}}{\alpha_{jk}}.$$

Proof. From the assumptions, A^{-1} exists and is strictly positive. From $A^{-1} z = q(A)^{-1} z$ and $z > 0$, we obtain

$$z_i = q(A) \sum_{k=1}^{n} \alpha_{ik} z_k \leq q(A) \max_{j,k} \alpha_{jk},$$

where $\sum_{k=1}^{n} z_k = 1$. The lower bound for z_i is proved similarly.

Also, by [10, Thm. 3.1, p. 41],

$$\max_{i,j} \frac{z_i}{z_j} \leq \max_{i,j,k} \frac{\alpha_{ik}}{\alpha_{jk}}.$$

By Lemma 2.6, $\alpha_{ik} \leq \alpha_{kk}$. Thus,

$$\max_{i,j,k} \frac{\alpha_{ik}}{\alpha_{jk}} = \max_{k,j} \frac{\alpha_{kk}}{\alpha_{jk}}. \qquad \square$$

Note that we also have $z_i \leq 1$. The bounds given here are in terms of the inverse elements of A. Note that from Theorem 3.3 and Corollary 3.6, we can bound z_i in terms of the elements of A without computing the inverse.

5. Systems of differential equations. In this section we deal with a system of linear differential equations on $\mathbb{R}^n$ governed by a matrix A.

THEOREM 5.1. *Let A satisfy (A_1)–(A_3) and let*

$$\frac{dx}{dt} = -A\, x(t), \qquad x(0) = x_0. \tag{22}$$

Then, for all $t \geq 0$,

$$\sum_{i=1}^{n} z_i x_i(t) = e^{-qt} \sum_{i=1}^{n} z_i x_{0,i}, \tag{23}$$

where $q = q(A)$ and $z = (z_1, \ldots, z_n)^T$ is the positive eigenvector of A^T corresponding to q.

Proof. After multiplication by z^T, (22) becomes

$$\frac{d}{dt} z^T x = -z^T A x = -q z^T x.$$

Integrating gives $z^T x = Ce^{-qt}$, from which we obtain the result. $\square$

Let us now assume $x_0 \geq 0$. Then (A_1) implies that $x(t) \geq 0$ for all $t \geq 0$ [19] and $\sum z_i x_i(t) = \|x\|_z$ is a weighted norm for $x(t)$, which we call the z-norm. In particular, (23) means that

$$\|x(t)\|_z = e^{-qt} \|x_0\|_z, \qquad t \geq 0. \tag{24}$$

This implies a global stability property and asymptotic convergence of the solution with a rate determined by q, so that

$$\lim_{t \to \infty} x(t) = 0. \tag{25}$$

Of course, this is valid for those initial vectors with $x_0 > 0$ only.

Furthermore, if we have upper and lower bounds for q, namely $0 < q_m \leq q \leq q_M$, then we have bounds for the z-norm of the solution:

$$\|x_0\|_z e^{-q_M t} \leq \|x(t)\|_z \leq \|x_0\|_z e^{-q_m t} \tag{26}$$

or, for the ℓ_1-norm:

$$\frac{z_{\min}}{z_{\max}} \|x_0\|_1 e^{-q_M t} \leq \|x(t)\|_1 \leq \frac{z_{\max}}{z_{\min}} \|x_0\|_1 e^{-q_m t}, \tag{27}$$

where $z_{\min} \leq z_i \leq z_{\max}$ for all $i = 1, \ldots, n$.

These results are of the same type as the classcial results of Wazewski for the ℓ_2-norm [5] and of [4], [13] for the ℓ_1- and ℓ_∞-norms. See also [3], [7], [9].

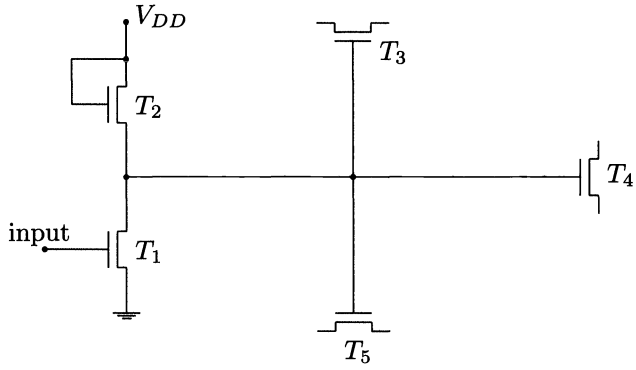


FIG. 1. *A MOS inverter with T_3, T_4, and T_5 loads.*

6. Application to electrical circuits. Computational digital systems are made of interconnected transistors which are switched between two stages in accordance with specific tasks. Figure 6.1 represents a simple MOS inverter, a basic and specific component of a digital circuit. It consists of transistors T_1 and T_2 and it drives the "gates" T_3, T_4, and T_5 through interconnection lines.

The structure is implemented on a semiconductor wafer. In Fig. 6.2 we show a linear RC model for this circuit after transistor T_1 was switched off by the input signal.

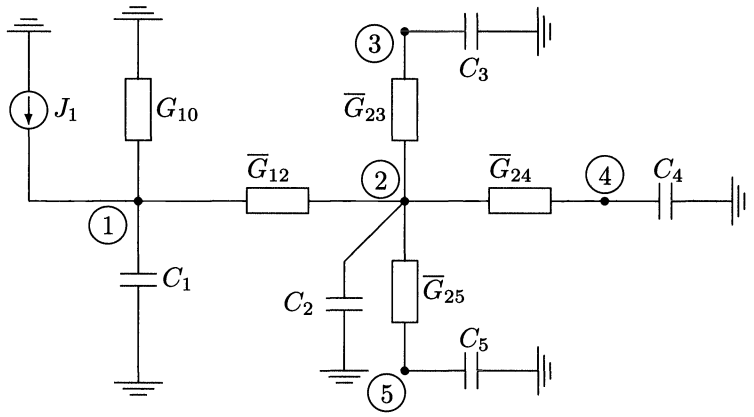


FIG. 2. *A model for the circuit in Fig.* 6.1.

Keeping in mind this example, we can consider a general RC network with n nodes, in which the ith node is capacitively grounded by $C_i > 0$ and eventually resistively grounded by a conductance $G_{i_0} \geq 0$. Also, each node i is connected by a conductance $\overline{G}_{ij} \geq 0$ with the node j. The circuit is a "conex" one, in the sense that for a node i there exists at least one node j connected to it by $\overline{G}_{ij} > 0$. Note that $\overline{G}_{ij} = \overline{G}_{ji}$. A possible nonzero constant source between node i and "ground" is denoted by J_i. Then, if we denote by $v = [v_1(t), v_2(t), \ldots, v_n(t)]^T$ the vector of node voltages, it can be easily shown that the transient evolution of this circuit is described

by the equation

$$C\frac{dv}{dt} = -G\,v + J, \tag{28}$$

where $C = \mathrm{diag}(C_1, C_2, \ldots, C_n)$, $J = [J_1, J_2, \ldots, J_n]^T$, and G is a matrix of conductances with the elements

$$G_{ii} = G_{i0} + \sum_{j=1, j\neq i}^{n} \overline{G}_{ij} \quad \text{and} \quad G_{ij} = -\overline{G}_{ij}, \quad i \neq j.$$

If $v^\infty \in \mathbb{R}^n$ is the stationary regime voltage vector, i.e., $0 = -G\,v^\infty + J$, and if we denote $x(t) = v(t) - v^\infty$, we can see from (28) that

$$\frac{dx}{dt} = -C^{-1}G\,x. \tag{29}$$

It can be easily verified that the matrix $A = C^{-1}G$ satisfies (A_1)–(A_5), which we listed in §2. In fact, (A_1) and the first half of (A_2) are clear, while the second half of (A_2) means that at least one node is resistively grounded. This is the case for at least the source node (node 1 in Fig. 6.2). The irreducibility (A_3) of A is assured by the connectivity of the circuit. The fact that $G_{ij} = G_{ji}$ implies (A_5). Finally, assumption (A_4) is fulfilled by the connectivity of the circuit, which always allows a suitable node labeling (see Lemma 2.5).

The crucial performance of the above digital circuit is the high operating speed [6] that is measured by the delay time for which the signal will reach a point which is ϵ times its initial value after constant inputs are applied, where $0 < \epsilon < 1$; i.e.,

$$T(\epsilon) = \sup\left\{t \;:\; \frac{\|x(t)\|}{\|x_0\|} = \epsilon\right\}. \tag{30}$$

It is standard in engineering practice to take $\epsilon = 0.1$.

Providing an appropriate value of $T(\epsilon)$ by an optimal choice of circuit parameters is one of the primary goals of the VLSI design process. In order to speed up the initial stages of this process (where repeated simulations are done), it appeared ideal to include in CAD tools simple computable formulae for T (or its bounds). There have been a large number of contributions in this area during the past decade (see [2], [6], [8], [12], for example) and [9] contains a fairly complete list up to 1991.

Coming back to the bound of T defined in (30), we observe that the norm chosen must be independent of circuit parameters in the set where we search for their optimal value. The largest possible set is, of course, $\mathbb{R}^+$ for all entries of C and G. Then, in (30) we can choose the ℓ_1-norm, $\|x\|_1 = \sum |x_i|$, and denote this delay by $T_1(\epsilon)$. From (27) we obtain

$$\frac{z_{\min}}{z_{\max}} e^{-q_M T_1} \leq \epsilon = \frac{\|x(T_1(\epsilon))\|_1}{\|x_0\|_1} \leq \frac{z_{\max}}{z_{\min}} e^{-q_m T_1},$$

i.e.,

$$\frac{1}{q_M} \ln \frac{z_{\min}}{\epsilon z_{\max}} \leq T_1(\epsilon) \leq \frac{1}{q_m} \ln \frac{z_{\max}}{\epsilon z_{\min}}. \tag{31}$$

Note that we always have $T_1(\epsilon) > 0$.

It is also possible and useful to restrict our parameter search to circuits for which $G = \alpha \underline{G}$ and $C = \beta \underline{C}$, i.e., all conductances are the same multiple of some basic conductances, and similarly for capacitances. Obviously, in this class of circuits the eigenvector z of $A^T = GC^{-1}$ is invariant, and $\|\cdot\|_z$ is a suitable norm for (30). For this class of circuits we obtain the following from (26):

$$\frac{\ln(1/\epsilon)}{q_M} \leq T_z(\epsilon) \leq \frac{\ln(1/\epsilon)}{q_m}. \tag{32}$$

Let us consider the circuit example from Fig. 6.2. If we take (for calculation simplicity) all conductances and capacitances with value 1, then we obtain $C = \operatorname{diag}(1,1,1,1,1)$ and the elements of A (and of G at the same time) are $a_{11} = 2$, $a_{12} = a_{21} = -1$, $a_{22} = 4$, $a_{23} = a_{32} = a_{24} = a_{42} = a_{25} = a_{52} = -1$, $a_{33} = a_{44} = a_{55} = 1$, with the others being 0. Now, we can compute $r_1 = \frac{1}{2}$, $r_2 = \frac{3}{4}$, $r_3 = r_4 = r_5 = 0$, and $l_2 = \frac{1}{4}$, $l_3 = l_4 = l_5 = 1$. Thus, Theorem 3.3 gives us

$$M = \|A^{-1}\|_\infty \leq 27,$$

while Corollary 3.6 implies

$$m = \min_{j,k} \alpha_{jk} \geq \frac{1}{8}.$$

Using Theorem 4.1, we obtain

$$\frac{1}{27} \leq \frac{1}{M} \leq q \leq \min a_{ii} = 1,$$

and from Theorem 4.2 we obtain

$$\max_{i,j} \frac{z_i}{z_j} \leq \frac{M}{m} \leq 216.$$

For $\epsilon = 0.1$, from (31) and (32), we obtain $0 \leq T_1 \leq 207.3$ and $2.303 \leq T_z \leq 62.17$; for $\epsilon = 0.001$, we obtain $1.532 \leq T_1 \leq 331.6$ and $6.908 \leq T_z \leq 186.5$.

It is apparent from this example that the T_z delay time bounds provided by (32) are reasonably tight, and because they are simple to calculate, they are useful for large-scale circuit design. If we search for optimal parameters in the largest possible set of values, then (31) will provide useful information.

REFERENCES

[1] J. H. Bramble and B. E. Hubbard, *On a finite difference analogue of an elliptic boundary value problem which is neither diagonally dominant nor of nonnegative type*, J. Math. and Phys., 43 (1964), pp. 117–132.

[2] P. K. Chan and M. D. F. Sehlag, *Bounds on signal delay in RC mesh networks*, IEEE Trans. Circuits Systems I Fund. Theory Appl., CAS 8 (1989), pp. 581–589.

[3] K. H. Chew, *Stability conditions for linear differential systems*, Nanta Math., 9 (1976), pp. 39–48.

[4] W. A. Coppel, *Stability and Asymptotic Behaviour of Differential Equations*, 2nd ed., D. C. Heath, Boston, 1965.

[5] C. Corduneanu, *Principles of Differential and Integral Equations*, Chelsea, New York, 1988.

[6] L. A. Glasser and S. W. Sobberphul, *Design and Analysis of VLSI Circuits*, Addison–Wesley, New York, 1985.

[7] C. Kahane, *Stability of solutions of linear systems with dominant main diagonal*, Proc. Amer. Math. Soc., 33 (1972), pp. 69–71.

[8] C. A. MARINOV, *Dissipativity as a unified approach to Sanberg–Willson type properties of nonlinear transistor networks*, Internat. J. Circ. Theory Appl., 18 (1990), pp. 575–594.
[9] C. A. MARINOV AND P. NEITTAANMAKI, *Mathematical Models in Electrical Circuits: Theory and Applications*, Kluwer–Nijhoff, Dordrecht, Boston, London, 1991.
[10] H. MINC, *Nonnegative Matrices*, John Wiley and Sons, New York, 1987.
[11] A. M. OSTROWSKI, *Note on bounds for determinants with dominant principal diagonal*, Proc. Amer. Math. Soc., 3 (1952), pp. 26–30.
[12] L. T. PILLAGE AND R. A. RAHRER, *Asymptotic waveform evaluation for timing analysis*, IEEE Trans. Comput. Aided Design, CAD 9 (1990), pp. 352–366.
[13] I. W. SANDBERG, *Some theorems on the dynamic response of nonlinear transistor networks*, Bell Syst. Tech. J., 48 (1969), pp. 35–54.
[14] P. N. SHIVAKUMAR AND K. H. CHEW, *A sufficient condition for nonvanishing of determinants*, Proc. Amer. Math. Soc., 43 (1974), pp. 63–66.
[15] ———, *Iterations for diagonally dominant matrices,* Canad. Math. Bull., 19 (1976), pp. 375–377.
[16] J. M. VARAH, *A lower bound for the smallest singular value of a matrix*, Linear Algebra Appl., 11 (1975), pp. 3–5.
[17] R. S. VARGA, *Matrix Iterative Analysis,* Prentice–Hall, Englewood Cliffs, NJ, 1962.
[18] ———, *On recurring theorems on diagonal dominance,* Linear Algebra Appl., 13 (1976), pp. 1–9.
[19] W. WALTER, *Differential and Integral Inequalities*, Springer–Verlag, Berlin, Heidelberg, New York, 1970.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 2, pp. 313–331, April 1996

005

PERTURBATION ANALYSIS OF THE POLE ASSIGNMENT PROBLEM*

JI-GUANG SUN†

Abstract. Condition numbers of the state feedback pole assignment problem having no repeated closed-loop eigenvalues are derived by using the implicit function theorem and its generalization. The absolute condition numbers are then used to derive the first-order perturbation bounds for the solution to the pole assignment problem. Moreover, the conditioning of the state feedback and the conditioning of the resulting closed-loop eigenvalues are discussed.

Key words. controllable system, state feedback, pole assignment, eigenvalues, condition number, perturbation bounds

AMS subject classifications. 15A18, 65F35, 93B55

1. Introduction and basic results.

Let (A, B) denote a system

$$\dot{x} = Ax + Bu, \tag{1.1}$$

where $A \in \mathcal{R}^{n\times n}, B \in \mathcal{R}^{n\times m}, x \in \mathcal{R}^n$, and $u \in \mathcal{R}^m$. The symbol $\mathcal{R}^{m\times n}$ denotes the set of real $m \times n$ matrices and $\mathcal{R}^n = \mathcal{R}^{n\times 1}$.

The state feedback pole assignment problem for the system (1.1), as a special additive inverse eigenvalue problem [4], may be formulated as follows [11], [18], [28].

PROBLEM PA. Given $A \in \mathcal{R}^{n\times n}, B \in \mathcal{R}^{n\times m}$, and a set of n complex numbers, $\mathcal{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, closed under complex conjugation, find an $F \in \mathcal{R}^{n\times m}$ such that the eigenvalues of $A + BF^T$ are λ_j, $j = 1, 2, \ldots, n$.

The following result is known [11], [28].

THEOREM 1.1. *A solution $F \in \mathcal{R}^{n\times m}$ to Problem PA exists for every set $\mathcal{L}$ of self-conjugate complex numbers if and only if (A, B) is controllable, that is, if and only if*

$$\{y^T A = \mu y^T \ \text{ and } \ y^T B = 0\} \Longleftrightarrow y^T = 0.$$

Moreover, in the single-input case (i.e., $m = 1$), if Problem PA has a solution, then the solution is unique.

It is worthwhile to point out that although many approaches have been developed for solving Problem PA (see [1], [3], [5], [6], [15], [16], [18], [19], [25], and the references contained therein), relatively little attention has been paid to the perturbation analysis of the problem [1], [12], [13].

Suppose that a controllable system (A, B) and a set $\mathcal{L}$ of n self-conjugate complex numbers are slightly perturbed to another controllable system and another set of n self-conjugate complex numbers. Then, by Theorem 1.1, there is a solution to Problem PA for the perturbed data. Generally speaking, the solution F changes when the data (A, B) and $\mathcal{L}$ are subject to a perturbation. Hence, there is a question: which quantities can be used to measure the sensitivity of the solution to small changes in the data? The object of this paper is to describe a technique, developed by Hald [9] and the author [23], to give an answer to this question for the case of $\lambda_i \neq \lambda_j, i \neq j$.

* Received by the editors October 25, 1994; accepted for publication (in revised form) by P. Van Dooren May 21, 1995. This work was supported by 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.

Observe that in the single-input case the solution to Problem PA is unique if it exists, but in the multi-input case Problem PA is essentially underdetermined [11]. Hence, in this paper we shall study perturbation analysis for the two different cases separately.

Throughout this paper we shall use the following notational conventions. $\mathcal{C}^{m\times n}$ denotes the set of complex $m\times n$ matrices and $\mathcal{C}^n=\mathcal{C}^{n\times 1}$. A^T, A^H, and $A^\dagger$ denote the transpose, the conjugate transpose, and the Moore–Penrose inverse of a matrix A, respectively. I is the identity matrix, I_n is the identity matrix of order n, and 0 is the null matrix. For $A=(a_1,a_2,\ldots,a_n)=(\alpha_{ij})\in\mathcal{R}^{m\times n}$ (or $\mathcal{C}^{m\times n}$), the symbol $\mathrm{vec}(A)$ denotes an mn-dimensional vector defined by $\mathrm{vec}(A)=(a_1^T,a_2^T,\ldots,a_n^T)^T$. $\|\ \|_2$ denotes the Euclidean vector norm and the spectral norm, and $\|\ \|_F$ the Frobenius norm.

Suppose that the function

$$\phi:\mathcal{D}\subset\mathcal{R}^n\to\mathcal{R}^m\ (\text{or } \mathcal{D}\subset\mathcal{C}^n\to\mathcal{C}^m),$$

with

$$\phi(x)=(\phi_1(x),\ldots,\phi_m(x))^T,\quad x=(x_1,\ldots,x_n)^T,$$

is defined on an open subset $\mathcal{D}$ of $\mathcal{R}^n$ (or $\mathcal{C}^n$), and that its component functions $\phi_i, i=1,\ldots,m$, have continuous first derivatives on $\mathcal{D}$. Then we define the Jacobian matrix ϕ'_x by

$$\phi'_x=\begin{pmatrix}\dfrac{\partial\phi_1(x)}{\partial x_1} & \cdots & \dfrac{\partial\phi_1(x)}{\partial x_n}\\ \vdots & & \vdots\\ \dfrac{\partial\phi_m(x)}{\partial x_1} & \cdots & \dfrac{\partial\phi_m(x)}{\partial x_n}\end{pmatrix},$$

and in the case of $m=n$, we define the Jacobian $\frac{\partial\phi}{\partial x}$ (or more precisely, $\frac{\partial(\phi_1,\ldots,\phi_n)}{\partial(x_1,\ldots,x_n)}$) by

$$\frac{\partial\phi}{\partial x}=\det\phi'_x.$$

The implicit function theorem [2, p. 39] and the following two known results are basic tools in our analysis.

THEOREM 1.2 (a generalization of the implicit function theorem). *Suppose that the complex-valued functions*

$$f_j(\xi_1,\ldots,\xi_k;\eta_1,\ldots,\eta_l),\quad j=1,\ldots,m,$$

are analytic functions of $k+l$ complex variables in some neighborhood $\mathcal{B}$ of the origin of $\mathcal{C}^{k+l}$, where $m<k$. Let

$$f=(f_1,f_2,\ldots,f_m)^T\in\mathcal{C}^m,\quad x=(\xi_1,\xi_2,\ldots,\xi_k)^T,\quad y=(\eta_1,\eta_2,\ldots,\eta_l)^T.$$

If $f_j(0;0)=0$, $j=1,\ldots,m$, and if

$$\mathrm{rank}(f'_x)_{x=0,y=0}=m,$$

then the equations

$$f_j(\xi_1, \ldots, \xi_k; \eta_1, \ldots, \eta_l) = 0, \quad j = 1, \ldots, m,$$

have infinite many solutions

$$\xi_j = g_j(\eta_1, \ldots, \eta_l), \quad j = 1, \ldots, k,$$

vanishing for $\eta_1 = \cdots = \eta_l = 0$ *and analytic in some neighborhood* $\mathcal{B}_y$ *of the origin of* $\mathcal{C}^l$.

Theorem 1.2 can be proved by the implicit function theorem [21].

THEOREM 1.3 (see [22]). *Let* $z = (z_1, \ldots, z_k)^T \in \mathcal{C}^k$, *and let* $A(z) \in \mathcal{C}^{n \times n}$ *be an analytic function of* z *in some neighborhood of the origin of* $\mathcal{C}^k$. *Suppose that* λ_1 *is a simple eigenvalue of* $A(0)$ *and* x_1, y_1 *are associated eigenvectors satisfying*

$$A(0)x_1 = \lambda_1 x_1, \quad y_1^T A(0) = \lambda_1 y_1^T, \quad y_1^T x_1 = 1.$$

Then

(1) *there exists a simple eigenvalue* $\lambda_1(z)$ *of* $A(z)$ *which is an analytic function of* z *in some neighborhood* $\mathcal{B}_0 \subset \mathcal{C}^k$ *of the origin, and* $\lambda_1(0) = \lambda_1$;

(2) *the right and left eigenvectors* $x_1(z)$ *and* $y_1(z)$ *corresponding to* $\lambda_1(z)$ *may be defined to be analytic functions of* $z \in \mathcal{B}_0$, *and* $x_1(0) = x_1, y_1(0) = y_1$;

(3) *there are formulae*

$$\left(\frac{\partial \lambda_1(z)}{\partial z_j}\right)_{z=0} = y_1^T \cdot \left(\frac{\partial A(z)}{\partial z_j}\right)_{z=0} \cdot x_1, \quad j = 1, \ldots, k.$$

The rest of this paper is organized as follows. In §2 we derive condition numbers and the first-order perturbation bounds for the solution to Problem PA for the single-input case, and discuss the conditioning of the state feedback and the conditioning of the resulting closed-loop eigenvalues. In §3 we touch upon the multi-input case. Finally, in §4 we present some results of numerical tests.

Our numerical tests show that there is, presumably, some intrinsic relation between the conditioning of the state feedback (i.e., the conditioning of the pole assignment problem) and the distance of the given controllable system from the nearest uncontrollable system.

2. The single-input case.

2.1. Absolute condition numbers. Given a single-input controllable system (A, b) with $A \in \mathcal{R}^{n \times n}, b \in \mathcal{R}^n$, and given a set of n complex numbers, $\mathcal{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, closed under complex conjugation and $\lambda_i \neq \lambda_j$ for $i \neq j$. Let

$$\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n). \tag{2.1}$$

By Theorem 1.1, there is a unique $f = (f_1, f_2, \ldots, f_n)^T \in \mathcal{R}^n$ with a nonsingular $X = (x_{ij}) \in \mathcal{C}^{n \times n}$, whose columns are the right eigenvectors $x_1, x_2, \ldots, x_n$ of $A + bf^T$ corresponding to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, such that

$$A + bf^T = X \Lambda X^{-1}. \tag{2.2}$$

Let

$$Y = X^{-T} = [y_1, y_2, \ldots, y_n]. \tag{2.3}$$

Then (2.2) is equivalent to

$$Y^T(A+bf^T)=\Lambda Y^T, \tag{2.4}$$

and the columns of Y are the left eigenvectors of $A+bf^T$:

$$y_j^T(A+bf^T)=\lambda_j y_j^T, \quad j=1,2,\dots,n. \tag{2.5}$$

By Theorem 1.1, the relations (2.5) imply that

$$y_j^T b\neq 0, \quad j=1,2,\dots,n. \tag{2.6}$$

Let $(\tilde{A},\tilde{b})$ be a controllable system, and let $\tilde{\mathcal{L}}=\{\tilde{\lambda}_1,\tilde{\lambda}_2,\dots,\tilde{\lambda}_n\}$ be a set of self-conjugate complex numbers. Moreover, let

$$\lambda=(\lambda_1,\lambda_2,\dots,\lambda_n)^T, \quad \tilde{\lambda}=(\tilde{\lambda}_1,\tilde{\lambda}_2,\dots,\tilde{\lambda}_n)^T,$$

and

$$\begin{aligned} &A(t)=A+t(\tilde{A}-A), \quad b(t)=b+t(\tilde{b}-b),\\ &\lambda(t)=\lambda+t(\tilde{\lambda}-\lambda)=(\lambda_1(t),\lambda_2(t),\dots,\lambda_n(t))^T, \quad t\in[-1,1]. \end{aligned} \tag{2.7}$$

We assume that $(\tilde{A},\tilde{b})$ is sufficiently near (A,b) and $\tilde{\lambda}$ is sufficiently near λ, such that the system $(A(t),b(t))$ is controllable and the set $\{\lambda_1(t),\lambda_2(t),\dots,\lambda_n(t)\}$ is closed under complex conjugation for any $t\in[-1,1]$. Then by Theorem 1.1, there is a unique vector $f(t)\in\mathcal{R}^n$ such that the eigenvalues of $A(t)+b(t)f(t)^T$ are $\lambda_j(t)$ for all $t\in[-1,1]$, $j=1,2,\dots,n$.

For investigating the problem of how the vector $f(t)-f$ is dependent on $A(t)-A, b(t)-b$ and $\lambda(t)-\lambda$ when $t\to 0$, we now embed $A, A(t)$ in $\mathcal{C}^{n\times n}$, and embed $b, b(t), \lambda, \lambda(t), f, f(t)$ in $\mathcal{C}^n$. Let

$$\hat{A}=A+E, \quad \hat{b}=b+e, \quad \hat{f}=f+g, \quad \hat{\lambda}=\lambda+\nu,$$

where $E\in\mathcal{C}^{n\times n}$, and $e,g,\nu\in\mathcal{C}^n$. Assume that the elements of E,e,g,ν are sufficiently small in magnitude such that all the eigenvalues of $\hat{A}+\hat{b}\hat{f}^T$ are simple. Then by Theorem 1.3,

(i) the eigenvalues are analytic functions of the elements of $(\hat{A},\hat{b},\hat{f})$ in some neighborhood $\mathcal{B}\subset\mathcal{C}^{n\times n}\oplus\mathcal{C}^n\oplus\mathcal{C}^n$ of the point (A,b,f);

(ii) the associated right and left eigenvectors may be defined to be analytic functions of $(\hat{A},\hat{b},\hat{f})$ in $\mathcal{B}$; and

(iii) the eigenvalues and associated eigenvectors of $\hat{A}+\hat{b}\hat{f}^T$ become λ_j, x_j, y_j when $(\hat{A},\hat{b},\hat{f})=(A,b,f)$.

Let $\mu(\hat{A}+\hat{b}\hat{f}^T)$ denote a vector in $\mathcal{C}^n$ whose jth element is the eigenvalue of $\hat{A}+\hat{b}\hat{f}^T$ nearest to λ_j, $j=1,2,\dots,n$. Then obviously $\mu(A+bf^T)=\lambda$.

Let $a=\mathrm{vec}(A), \hat{a}=\mathrm{vec}(\hat{A})$, and let

$$\phi(\hat{f};\hat{a},\hat{b})=\mu(\hat{A}+\hat{b}\hat{f}^T), \quad \omega(\hat{f};\hat{a},\hat{b},\hat{\lambda})=\phi(\hat{f};\hat{a},\hat{b})-\hat{\lambda}. \tag{2.8}$$

Applying Theorem 1.3 (3) and using simple calculations, from (2.8) we get the expression of the Jacobian matrix $\omega'_{\hat{f}}$ at $\hat{f}=f,\hat{a}=a,\hat{b}=b,\hat{\lambda}=\lambda$:

$$\begin{aligned} (\omega'_{\hat{f}})_{\hat{f}=f,\hat{a}=a,\hat{b}=b,\hat{\lambda}=\lambda} &= (\phi'_{\hat{f}})_{\hat{f}=f,\hat{a}=a,\hat{b}=b}\\ &= \mathrm{diag}(y_1^Tb, y_2^Tb,\dots,y_n^Tb)X^T\equiv W_f. \end{aligned} \tag{2.9}$$

By (2.6), $y_j^T b \neq 0$ for all j. Thus, the relation (2.9) implies that

$$\det(\omega'_{\hat{f}})_{\hat{f}=f,\hat{a}=a,\hat{b}=b,\hat{\lambda}=\lambda} \neq 0. \tag{2.10}$$

Consequently, by the implicit function theorem [2, p. 39], the equation $\omega(\hat{f};\hat{a},\hat{b},\hat{\lambda}) = 0$ has a unique analytic solution $\hat{f} = \hat{f}(\hat{a},\hat{b},\hat{\lambda})$ in some neighborhood $\hat{\mathcal{B}} \subset \mathcal{C}^{n^2} \oplus \mathcal{C}^n \oplus \mathcal{C}^n$ of the point (a,b,λ), and $\hat{f}(a,b,\lambda) = f$.

Now we restrict

$$\hat{a} = a(t) \equiv \mathrm{vec}(A(t)), \quad \hat{b} = b(t), \quad \hat{\lambda} = \lambda(t),$$

where $A(t), b(t), \lambda(t)$ for $t \in [-\epsilon, \epsilon]$ are defined by (2.7) and ϵ is a sufficiently small positive scalar such that $(a(t), b(t), \lambda(t)) \in \hat{\mathcal{B}}$ when $t \in [-\epsilon, \epsilon]$. Then we have proved that the equation

$$\omega(\hat{f}; a(t), b(t), \lambda(t)) = 0$$

has a unique analytic solution

$$\hat{f} = \hat{f}(a(t), b(t), \lambda(t)) \equiv f(t), \quad t \in [-\epsilon, \epsilon],$$

satisfying $f(0) = f$. This means that we have the relation

$$\omega(\hat{f}(a(t), b(t), \lambda(t)); a(t), b(t), \lambda(t)) = 0, \quad t \in [-\epsilon, \epsilon], \tag{2.11}$$

where $a(t), b(t), \lambda(t)$, $\hat{f}(a(t), b(t), \lambda(t))$ and $\omega(\hat{f}(a(t), b(t), \lambda(t)); a(t), b(t), \lambda(t))$ defined by

$$\begin{aligned} &\omega(\hat{f}(a(t), b(t), \lambda(t)); a(t), b(t), \lambda(t)) \\ &\quad = \phi(\hat{f}(a(t), b(t), \lambda(t)); a(t), b(t)) - \lambda(t) \end{aligned} \tag{2.12}$$

are analytic functions of $t \in [-\epsilon, \epsilon]$, and $a(0) = a, b(0) = b, \lambda(0) = \lambda, f(0) = f$.

Differentiating (2.11), we get

$$\omega'_{\hat{f}} df(t) + \omega'_{\hat{a}} da(t) + \omega'_{\hat{b}} db(t) + \omega'_{\hat{\lambda}} d\lambda(t) = 0, \quad t \in [-\epsilon, \epsilon].$$

Notably at $t = 0$, i.e., at $\hat{f} = f, \hat{a} = a, \hat{b} = b, \hat{\lambda} = \lambda$, we have

$$W_f df + W_a da + W_b db + W_\lambda d\lambda = 0, \tag{2.13}$$

where W_f is defined by (2.9), and

$$\begin{aligned} &W_a = (\omega'_{\hat{a}})_{\hat{f}=f,\hat{a}=a,\hat{b}=b,\hat{\lambda}=\lambda}, \quad W_b = (\omega'_{\hat{b}})_{\hat{f}=f,\hat{a}=a,\hat{b}=b,\hat{\lambda}=\lambda}, \\ &W_\lambda = (\omega'_{\hat{\lambda}})_{\hat{f}=f,\hat{a}=a,\hat{b}=b,\hat{\lambda}=\lambda}, \quad df = (df(t))_{t=0}, \\ &da = (da(t))_{t=0}, \quad db = (db(t))_{t=0}, \quad d\lambda = (d\lambda(t))_{t=0}. \end{aligned} \tag{2.14}$$

By (2.10), the matrix W_f is nonsingular. Consequently, from (2.13),

$$df = -W_f^{-1} W_a da - W_f^{-1} W_b db - W_f^{-1} W_\lambda d\lambda. \tag{2.15}$$

Let

$$Z = W_f^{-1}. \tag{2.16}$$

Then from (2.9) and (2.3),

$$Z = Y\mathrm{diag}\left(\frac{1}{y_1^T b}, \frac{1}{y_2^T b}, \ldots, \frac{1}{y_n^T b}\right). \tag{2.17}$$

Applying Theorem 1.3 (3) and using simple calculations, from (2.8) we get the expressions of W_a, W_b, W_λ, the Jacobian matrices $\omega'_{\hat{a}}, \omega'_{\hat{b}}, \omega'_{\hat{\lambda}}$ at $\hat{f} = f, \hat{a} = a, \hat{b} = b, \hat{\lambda} = \lambda$ (i.e., at $t = 0$):

$$\begin{aligned} W_a &= (\phi'_{\hat{a}})_{\hat{f}=f,\hat{a}=a,\hat{b}=b} = \left[D_1(X)X^{-1}, D_2(X)X^{-1}, \ldots, D_n(X)X^{-1}\right], \\ W_b &= (\phi'_{\hat{b}})_{\hat{f}=f,\hat{a}=a,\hat{b}=b} = \mathrm{diag}(f^T x_1, f^T x_2, \ldots, f^T x_n)X^{-1}, \\ W_\lambda &= -I_n, \end{aligned} \tag{2.18}$$

where

$$D_i(X) = \mathrm{diag}(x_{i1}, x_{i2}, \ldots, x_{in}), \quad i = 1, 2, \ldots, n. \tag{2.19}$$

Substituting (2.16)–(2.19) into (2.15), we get the differential relation

$$df = \Phi da + \Psi db + Z d\lambda, \tag{2.20}$$

where

$$\begin{aligned} Z &= -W_f^{-1}W_\lambda = Y\mathrm{diag}\left(\frac{1}{y_1^T b}, \frac{1}{y_2^T b}, \ldots, \frac{1}{y_n^T b}\right) \in \mathcal{C}^{n\times n}, \\ \Phi &= -W_f^{-1}W_a = -Z\left[D_1(X)X^{-1}, D_2(X)X^{-1}, \ldots, D_n(X)X^{-1}\right] \in \mathcal{C}^{n\times n^2}, \\ \Psi &= -W_f^{-1}W_b = -Z\mathrm{diag}(f^T x_1, f^T x_2, \ldots, f^T x_n)X^{-1} \in \mathcal{C}^{n\times n}. \end{aligned} \tag{2.21}$$

Thus, we have proved the following theorem.

THEOREM 2.1. *Let a controllable system (A, b) and a set of self-conjugate complex numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$ be given, where $\lambda_i \neq \lambda_j$, $i \neq j$. Let $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)^T, a = \mathrm{vec}(A)$. Assume that $f \in \mathcal{R}^n$ and $X \in \mathcal{C}^{n\times n}$ satisfy* (2.2). *Then there is a differential relation* (2.20), *where Z, Φ, Ψ are expressed by* (2.21).

Remark 2.2. For any consistent norm $\|\ \|$, the relation (2.20) gives

$$\|df\| \leq \|\Phi\|\|da\| + \|\Psi\|\|db\| + \|Z\|\|d\lambda\|.$$

Consequently, the group of scalars $\kappa_A(f), \kappa_b(f), \kappa_\lambda(f)$ defined by

$$\kappa_A(f) = \|\Phi\|, \quad \kappa_b(f) = \|\Psi\|, \quad \kappa_\lambda(f) = \|Z\| \tag{2.22}$$

is the group of absolute condition numbers of the state feedback f with respect to A, b, and λ. Moreover, the scalar $\kappa(f)$ defined by

$$\kappa(f) = \sqrt{[\kappa_A(f)]^2 + [\kappa_b(f)]^2 + [\kappa_\lambda(f)]^2} \tag{2.23}$$

can be called the absolute condition number of f.

Remark 2.3. Observe that the matrices Φ, Ψ, and Z expressed by (2.21) are independent of the column scaling of X. Hence, there is an important fact: the absolute condition numbers $\kappa_A(f), \kappa_b(f), \kappa_\lambda(f)$ are independent of the norms of the right eigenvectors $x_1, x_2, \ldots, x_n$ of $A + bf^T$.

Remark 2.4. We now define the absolute condition numbers $\kappa_A(f), \kappa_b(f)$, and $\kappa_\lambda(f)$ by using the norm $\| \ \|_2$ in (2.22). If the matrix $X = (x_1, x_2, \ldots, x_n)$ of (2.2) satisfies $\|x_j\|_2 = 1$ for all j, then from (2.21)–(2.22),

$$\kappa_A(f) = \|\Phi\|_2 \leq \|Z\|_2 \|X^{-1}\|_2,$$

$$\kappa_b(f) = \|\Psi\|_2 \leq \|f\|_2 \|Z\|_2 \|X^{-1}\|_2,$$

$$\kappa_\lambda(f) = \|Z\|_2,$$

where the first inequality is deduced from the fact that

$$[D_1(X), D_2(X), \ldots, D_n(X)]\,[D_1(X), D_2(X), \ldots, D_n(X)]^T = I_n.$$

The following result, as a corollary of Theorem 2.1, gives the first-order perturbation bounds for the solution to Problem PA.

COROLLARY 2.5. *Let a controllable system (A, b) and a set $\mathcal{L}$ of self-conjugate complex numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$ be given, where $\lambda_i \neq \lambda_j$, $i \neq j$. Suppose that (A, b) is slightly perturbed to a controllable system $(\tilde{A}, \tilde{b})$ and $\mathcal{L}$ is slightly perturbed to a set $\tilde{\mathcal{L}}$ of self-conjugate complex numbers $\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_n$. Let $f, \tilde{f} \in \mathcal{R}^n$ be the solutions to Problem PA with the data $A, b, \mathcal{L}$ and $\tilde{A}, \tilde{b}, \tilde{\mathcal{L}}$, respectively, and let*

$$a = \operatorname{vec}(A), \quad \tilde{a} = \operatorname{vec}(\tilde{A}), \quad \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)^T, \quad \tilde{\lambda} = (\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_n)^T.$$

Then for any consistent norm $\| \ \|$, we have

$$\begin{aligned} \|\tilde{f} - f\| &\leq \delta_f + O\left(\left\| \begin{pmatrix} \tilde{a} \\ \tilde{b} \\ \tilde{\lambda} \end{pmatrix} - \begin{pmatrix} a \\ b \\ \lambda \end{pmatrix} \right\|^2\right) \\ &\leq \Delta_f + O\left(\left\| \begin{pmatrix} \tilde{a} \\ \tilde{b} \\ \tilde{\lambda} \end{pmatrix} - \begin{pmatrix} a \\ b \\ \lambda \end{pmatrix} \right\|^2\right), \end{aligned} \tag{2.24}$$

where

$$\begin{aligned} \delta_f &= \|\Phi(\tilde{a} - a) + \Psi(\tilde{b} - b) + Z(\tilde{\lambda} - \lambda)\|, \\ \Delta_f &= \|\Phi\|\|\tilde{a} - a\| + \|\Psi\|\|\tilde{b} - b\| + \|Z\|\|\tilde{\lambda} - \lambda\|, \end{aligned} \tag{2.25}$$

and Z, Φ, Ψ are defined in (2.21).

2.2. Relative condition numbers. By Rice [20], there are two kinds of condition numbers: absolute and relative. As above, from the differential relation (2.20) we have derived the absolute condition numbers $\kappa_A(f), \kappa_b(f), \kappa_\lambda(f)$, and $\kappa(f)$ of the state feedback f. In this subsection, we derive relative condition numbers of f.

From the relation (2.20) we get

$$\begin{aligned}\frac{\|f(\tilde{A},\tilde{b},\tilde{\lambda})-f\|_2}{\|f\|_2} &\le \kappa_A(f)\cdot\frac{\|A\|_F}{\|f\|_2}\cdot\frac{\|\tilde{A}-A\|_F}{\|A\|_F}+\kappa_b(f)\cdot\frac{\|b\|_2}{\|f\|_2}\cdot\frac{\|\tilde{b}-b\|_2}{\|b\|_2}\\ &+\kappa_\lambda(f)\cdot\frac{\|\lambda\|_2}{\|f\|_2}\cdot\frac{\|\tilde{\lambda}-\lambda\|_2}{\|\lambda\|_2}\\ &+O\left(\left(\frac{\|\tilde{A}-A\|_F}{\|A\|_F}\right)^2+\left(\frac{\|\tilde{b}-b\|_2}{\|b\|_2}\right)^2+\left(\frac{\|\tilde{\lambda}-\lambda\|_2}{\|\lambda\|_2}\right)^2\right),\end{aligned} \tag{2.26}$$

where $f = f(A,b,\lambda)$; $\kappa_A(f),\kappa_b(f)$, and $\kappa_\lambda(f)$ are defined by (2.22) with the norm $\|\ \|_2$; and $\|\tilde{A}-A\|_F \to 0, \|\tilde{b}-b\|_2 \to 0, \|\tilde{\lambda}-\lambda\|_2 \to 0$. The relation (2.26) shows that the scalar $\kappa_A^{(r)}(f)$ defined by

$$\kappa_A^{(r)}(f) = \kappa_A(f)\cdot\frac{\|A\|_F}{\|f\|_2} \tag{2.27}$$

is the relative condition number of f with respect to A. Similarly, the scalars $\kappa_b^{(r)}(f),\kappa_\lambda^{(r)}(f)$ defined by

$$\kappa_b^{(r)}(f) = \kappa_b(f)\cdot\frac{\|b\|_2}{\|f\|_2},\quad \kappa_\lambda^{(r)}(f) = \kappa_\lambda(f)\cdot\frac{\|\lambda\|_2}{\|f\|_2} \tag{2.28}$$

are the relative condition numbers of f with respect to b and λ, respectively. Moreover, the scalar $\kappa^{(r)}(f)$ defined by

$$\kappa^{(r)}(f) = \sqrt{[\kappa_A^{(r)}(f)]^2+[\kappa_b^{(r)}(f)]^2+[\kappa_\lambda^{(r)}(f)]^2} \tag{2.29}$$

can be called the relative condition number of f. In other words, $\kappa^{(r)}(f)$ can be called the relative condition number of the pole assignment problem.

It is known that a relative condition number can be used to distinguish the conditioning of a problem. Therefore, we will call a pole assignment problem ill conditioned if the relative condition number $\kappa^{(r)}(f)$ is large.

2.3. Conditioning of the closed-loop eigenvalues. Let A, b, and λ be as in Theorem 2.1 and let $f = f(A,b,\lambda)$ be the unique solution to Problem PA with the data A, B, λ. In §2.2 we discussed the conditioning of the solution f to Problem PA, i.e., the conditioning of the pole assignment problem. In this subsection we discuss a related question: how do we distinguish the conditioning of the resulting closed-loop eigenvalues $\lambda_1,\dots,\lambda_n$?

Let $A+bf^T$ be decomposed by (2.2). We now rewrite it as

$$Y^T(A+bf^T)X = \Lambda, \tag{2.30}$$

where $X = [x_1,\dots,x_n], Y = [y_1,\dots,y_n]$, and $\Lambda = \operatorname{diag}(\lambda_1,\dots,\lambda_n)$ with $\lambda_i \ne \lambda_j$ for $i \ne j$. The relation (2.30) shows that x_j and y_j are the right and left eigenvectors of $A+bf^T$ belonging to λ_j for $j = 1,\dots,n$. By Wilkinson [27, Chap. 2], the scalars $c(\lambda_j)$ defined by

$$c(\lambda_j) = \|x_j\|_2\|y_j\|_2 \tag{2.31}$$

are the absolute condition numbers of the closed-loop eigenvalues λ_j. Moreover, by Geurts [8, p. 90], the scalars $c^{(r)}(\lambda_j)$ defined by

$$c^{(r)}(\lambda_j) = c(\lambda_j) \cdot \frac{\|A + bf^T\|_2}{|\lambda_j|} \tag{2.32}$$

are the relative condition numbers of the nonzero closed-loop eigenvalues λ_j.

Assume that $\lambda_j \neq 0$ for all j. Then the scalar $c^{(r)}(\lambda)$ defined by

$$c^{(r)}(\lambda) = \sqrt{\sum_{j=1}^{n} [c^{(r)}(\lambda_j)]^2} \tag{2.33}$$

can be regarded as the relative condition number of the resulting closed-loop eigenvalues $\lambda_1, \ldots, \lambda_n$. Therefore, we will call the resulting closed-loop eigenvalues ill conditioned if the relative condition number $c^{(r)}(\lambda)$ is large.

It is worth pointing out that the conditioning of the feedback and the conditioning of the resulting closed-loop eigenvalues are two entirely different things. This fact will be illustrated by Example 4.3 of §4.

3. The multi-input case. Given a multi-input controllable system (A, B) with $A \in \mathcal{R}^{n\times n}, B \in \mathcal{R}^{n\times m}$ $(m > 1)$, and given a set of n complex numbers, $\mathcal{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, closed under complex conjugation and $\lambda_i \neq \lambda_j$ for $i \neq j$. By Theorem 1.1, there is an $F = (f_1, f_2, \ldots, f_m) \in \mathcal{R}^{m\times m}$ with a nonsingular $X = (x_{ij}) \in \mathcal{C}^{n\times n}$, whose columns are the right eigenvectors $x_1, x_2, \ldots, x_n$ of $A + BF^T$ corresponding to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, such that

$$A + BF^T = X\Lambda X^{-1}, \tag{3.1}$$

where $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. As before, let

$$Y = X^{-T} = [y_1, y_2, \ldots, y_n].$$

Then $y_1, y_2, \ldots, y_n$ are the left eigenvectors of $A + BF^T$ belonging to $\lambda_1, \lambda_2, \ldots, \lambda_n$, respectively, and $y_j^T x_j = 1$ for all j.

Let $(\tilde{A}, \tilde{B})$ be a controllable system and let $\tilde{\mathcal{L}} = \{\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_n\}$ be a set of self-conjugate complex numbers. Moreover, let $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)^T$ and let $\tilde{\lambda} = (\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_n)^T$. After the manner of (2.7) we define $A(t), B(t), \lambda(t)$ for $t \in [-1, 1]$. Assume that $(\tilde{A}, \tilde{B})$ is sufficiently near (A, B) and $\tilde{\lambda}$ is sufficiently near λ, such that the system $(A(t), B(t))$ is controllable and the set $\{\lambda_1(t), \lambda_2(t), \ldots, \lambda_n(t)\}$ is closed under complex conjugation for any $t \in [-1, 1]$. Then by Theorem 1.1, Problem PA with the data $A(t), B(t)$, and $\lambda(t)$ is solvable for each $t \in [-1, 1]$. But observe that there exist extra degrees of freedom in Problem PA for the multi-input case, i.e., Problem PA is essentially underdetermined for the multi-input case [11]. Hence, in general, there are various $F(t) \in \mathcal{R}^{n\times m}$ such that the eigenvalues of $A(t) + B(t)F(t)^T$ are $\lambda_j(t), j = 1, 2, \ldots, n$.

As we have done in §2, we now embed $A, A(t), B, B(t), \lambda, \lambda(t), F, F(t)$ in complex linear spaces. Let

$$\hat{A} = A + E_A, \quad \hat{B} = B + E_B, \quad \hat{F} = F + G, \quad \hat{\lambda} = \lambda + \nu,$$

where $E_A \in \mathcal{C}^{n\times n}$, $E_B, G \in \mathcal{C}^{n\times m}$, and $\nu \in \mathcal{C}^n$, and the elements of E_A, E_B, G, ν are sufficiently small in magnitude such that all the eigenvalues of $\hat{A} + \hat{B}\hat{F}^T$ are simple.

Let $\mu(\hat{A}+\hat{B}\hat{F}^T)$ denote a vector in $\mathcal{C}^n$ whose jth element is the eigenvalue of $\hat{A}+\hat{B}\hat{F}^T$ nearest to λ_j, $j=1,2,\dots,n$. Then obviously $\mu(A+BF^T)=\lambda$.

Define $a=\mathrm{vec}(A), b=\mathrm{vec}(B), f=\mathrm{vec}(F)$, and define $\hat{a},\hat{b},\hat{f}$ similarly. Let

$$\phi(\hat{f};\hat{a},\hat{b})=\mu(\hat{A}+\hat{B}\hat{F}^T),\quad \omega(\hat{f};\hat{a},\hat{b},\hat{\lambda})=\phi(\hat{f};\hat{a},\hat{b})-\hat{\lambda}. \tag{3.2}$$

Applying Theorem 1.3 (3) and using simple calculations, from (3.2) we get the expression of the Jacobian matrix $\omega'_{\hat{f}}$ at $\hat{f}=f,\hat{a}=a,\hat{b}=b,\hat{\lambda}=\lambda$:

$$\begin{aligned}(\omega'_{\hat{f}})_{\hat{f}=f,\hat{a}=a,\hat{b}=b,\hat{\lambda}=\lambda}&=(\phi'_{\hat{f}})_{\hat{f}=f,\hat{a}=a,\hat{b}=b}\\&=\left[S_1X^T,S_2X^T,\dots,S_mX^T\right]\equiv W_f\in\mathcal{C}^{n\times mn},\end{aligned} \tag{3.3}$$

where

$$S_j=\mathrm{diag}(y_1^Tb_j,y_2^Tb_j,\dots,y_n^Tb_j),\quad j=1,2,\dots,m. \tag{3.4}$$

Since $\mathrm{diag}(X^T,X^T,\dots,X^T)$ is nonsingular, the relation (3.3) implies that

$$\mathrm{rank}(W_f)=\mathrm{rank}([S_1,S_2,\dots,S_m]).$$

By Theorem 1.1, $y_j^TB\neq 0$ for all j, i.e., there are indices $1',2',\dots,n'\in\{1,2,\dots,m\}$ such that

$$y_j^Tb_{j'}\neq 0,\quad j=1,2,\dots,n.$$

Therefore, $\mathrm{rank}([S_1,S_2,\dots,S_m])=n$. Thus, we have $\mathrm{rank}(W_f)=n$. Consequently, by Theorem 1.2, the equation $\omega(\hat{f};\hat{a},\hat{b},\hat{\lambda})=0$ has infinite analytic solutions $\hat{f}=\hat{f}(\hat{a},\hat{b},\hat{\lambda})$ in some neighborhood $\hat{\mathcal{B}}\subset\mathcal{C}^{n^2}\oplus\mathcal{C}^{mn}\oplus\mathcal{C}^n$ of the point (a,b,λ), and $\hat{f}(a,b,\lambda)=f$.

Now we restrict

$$\hat{a}=a(t)\equiv\mathrm{vec}(A(t)),\quad \hat{b}=b(t)\equiv\mathrm{vec}(B(t)),\quad \hat{\lambda}=\lambda(t).$$

Then there is a sufficiently small $\epsilon>0$ such that $(a(t),b(t),\lambda(t))\in\hat{\mathcal{B}}$ when $t\in[-\epsilon,\epsilon]$. Thus, we have proved that the equation

$$\omega(\hat{f};a(t),b(t),\lambda(t))=0 \tag{3.5}$$

has various analytic solutions

$$\hat{f}=\hat{f}(a(t),b(t),\lambda(t))\equiv f(t),\quad t\in[-\epsilon,\epsilon] \tag{3.6}$$

satisfying $f(0)=f$.

Let $f(t)$ be any of the solutions to the equation (3.5). Differentiating (3.5) at $t=0$, i.e., at $\hat{f}=f,\hat{a}=a,\hat{b}=b,\hat{\lambda}=\lambda$, we get

$$W_f\,df+W_a\,da+W_b\,db+W_\lambda\,d\lambda=0, \tag{3.7}$$

where W_f is defined by (3.3), and $W_a,W_b,W_\lambda,df,da,db,d\lambda$ have the same definitions as in the single-input case (see (2.14)).

Applying Theorem 1.3 (3) and using simple calculations, we get

$$W_a = \left[D_1(X)X^{-1}, D_2(X)X^{-1}, \ldots, D_n(X)X^{-1}\right] \in \mathcal{C}^{n\times n^2},$$

$$W_b = \operatorname{diag}(T_1X^{-1}, T_2X^{-1}, \ldots, T_mX^{-1}) \in \mathcal{C}^{n\times mn}, \tag{3.8}$$

$$W_\lambda = -I_n,$$

where

$$D_i(X) = \operatorname{diag}(x_{i1}, x_{i2}, \ldots, x_{in}), \quad i = 1, 2, \ldots, n,$$
$$T_j = \operatorname{diag}(f_j^T x_1, f_j^T x_2, \ldots, f_j^T x_n), \quad j = 1, 2, \ldots, m. \tag{3.9}$$

It has been pointed out that the equation (3.5) has various analytic solutions $f(t)$ satisfying $f(0) = 0$. Consequently, there are various df satisfying (3.7). We now take a special solution $f(t)$ such that its differentiation at $t = 0$ is expressed by

$$df = -W_f^\dagger W_a da - W_f^\dagger W_b db + W_f^\dagger d\lambda.$$

Then we have proved the following theorem.

THEOREM 3.1. *Let a controllable system (A, B) and a set of self-conjugate complex numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$ be given, where $A \in \mathcal{R}^{n\times n}, B \in \mathcal{R}^{n\times m}$ $(m > 1)$, and $\lambda_i \neq \lambda_j$, $i \neq j$. Let $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n), \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)^T, a = \operatorname{vec}(A), b = \operatorname{vec}(B)$. Assume that $F \in \mathcal{R}^{n\times m}$ and $X \in \mathcal{C}^{n\times n}$ satisfy* (3.1), *and let $f = \operatorname{vec}(F)$. Then there is a differential relation*

$$df = \Phi da + \Psi db + Z d\lambda, \tag{3.10}$$

where Z, Φ, Ψ are defined by

$$Z = W_f^\dagger \in \mathcal{C}^{mn\times n}, \quad \Phi = -ZW_a \in \mathcal{C}^{mn\times n^2}, \quad \Psi = -ZW_b \in \mathcal{C}^{mn\times mn}, \tag{3.11}$$

and W_f, W_a, W_b are expressed by (3.3)–(3.4) *and* (3.8)–(3.9).

Remark 3.2. Observe that there are various analytic solutions $f(t) \equiv \operatorname{vec}(F(t))$ to the equation (3.5) satisfying $f(0) = f \equiv \operatorname{vec}(F)$, where F is a solution to Problem PA with the data A, B, λ. Hence, the solution F may have various groups of absolute condition numbers that reflect the different sensitivities of F with respect to A, B, λ. From (3.10) we see that the group of scalars κ_A, κ_B, and κ_λ defined by

$$\kappa_A(F) = \|\Phi\|, \quad \kappa_B(F) = \|\Psi\|, \quad \kappa_\lambda(F) = \|Z\| \tag{3.12}$$

is one of the groups of absolute condition numbers of F. Moreover, the scalar $\kappa(F)$ defined by

$$\kappa(F) = \sqrt{[\kappa_A(F)]^2 + [\kappa_B(F)]^2 + [\kappa_\lambda(F)]^2} \tag{3.13}$$

can be regarded as an absolute condition number of F.

Remark 3.3. It is easy to verify that the matrices Φ, Ψ, and Z defined by (3.11) are independent of the column scaling of X. Consequently, the condition numbers $\kappa_A(F), \kappa_B(F), \kappa_\lambda(F)$ are independent of the norms of the right eigenvectors $x_1, x_2, \ldots, x_n$ of $A + BF^T$.

Remark 3.4. We now define the absolute condition numbers $\kappa_A(F), \kappa_B(F)$, and $\kappa_\lambda(F)$ by using the norm $\|\ \|_2$ in (3.12). If the matrix $X = (x_1, x_2, \ldots, x_n)$ of (3.1) satisfies $\|x_j\|_2 = 1$ for all j, then from (3.3)–(3.4), (3.8)–(3.9), and (3.11)–(3.12),

$$\kappa_A(F) = \|\Phi\|_2 \le \|Z\|_2 \|X^{-1}\|_2,$$

$$\kappa_B(F) = \|\Psi\|_2 \le \max_{1\le j\le n} \|f_j\|_2 \|Z\|_2 \|X^{-1}\|_2,$$

$$\kappa_\lambda(F) = \|Z\|_2.$$

From Theorem 3.1 we get the following corollary.

COROLLARY 3.5. *Let a controllable system (A, B) and a set $\mathcal{L}$ of self-conjugate complex numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$ be given, where $\lambda_i \ne \lambda_j$, $i \ne j$. Suppose that (A, B) is slightly perturbed to a controllable system $(\tilde{A}, \tilde{B})$ and $\mathcal{L}$ is slightly perturbed to a set $\tilde{\mathcal{L}}$ of self-conjugate complex numbers $\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_n$. Let $F \in \mathcal{R}^{n\times m}$ be a solution to Problem PA with the data $A, B, \mathcal{L}$. Then there is a solution to Problem PA with the data $\tilde{A}, \tilde{B}, \tilde{\mathcal{L}}$, such that for any consistent norm $\|\ \|$, we have*

$$\begin{aligned} \|\tilde{F} - F\| &\le \delta_F + O\left(\left\|\begin{pmatrix} \tilde{a} \\ \tilde{b} \\ \tilde{\lambda} \end{pmatrix} - \begin{pmatrix} a \\ b \\ \lambda \end{pmatrix}\right\|^2\right) \\ &\le \Delta_F + O\left(\left\|\begin{pmatrix} \tilde{a} \\ \tilde{b} \\ \tilde{\lambda} \end{pmatrix} - \begin{pmatrix} a \\ b \\ \lambda \end{pmatrix}\right\|^2\right), \end{aligned} \tag{3.14}$$

where

$$a = \mathrm{vec}(A), \quad \tilde{a} = \mathrm{vec}(\tilde{A}), \quad b = \mathrm{vec}(B), \quad \tilde{b} = \mathrm{vec}(\tilde{B}),$$

$$\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)^T, \quad \tilde{\lambda} = (\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_n)^T,$$

$$\begin{aligned} \delta_F &= \|\Phi(\tilde{a} - a) + \Psi(\tilde{b} - b) + Z(\tilde{\lambda} - \lambda)\|, \\ \Delta_F &= \|\Phi\|\|\tilde{a} - a\| + \|\Psi\|\|\tilde{b} - b\| + \|Z\|\|\tilde{\lambda} - \lambda\|, \end{aligned} \tag{3.15}$$

and Z, Φ, Ψ are defined by (3.11).

Remark 3.6. Let $(A, B), \lambda$ be as in Theorem 3.1, and let F be a solution to Problem PA with the data A, B, λ. In a similar manner as described in §2.2 we can define the relative condition numbers $\kappa_A^{(r)}(F), \kappa_B^{(r)}(F), \kappa_\lambda^{(r)}(F)$, and $\kappa^{(r)}(F)$ of F by

$$\kappa_A^{(r)}(F) = \kappa_A(F)\cdot\frac{\|A\|_F}{\|F\|_F}, \quad \kappa_B^{(r)}(F) = \kappa_B(F)\cdot\frac{\|B\|_F}{\|F\|_F}, \quad \kappa_\lambda^{(r)}(F) = \kappa_\lambda(F)\cdot\frac{\|\lambda\|_2}{\|F\|_F} \tag{3.16}$$

and

$$\kappa^{(r)}(F) = \sqrt{[\kappa_A^{(r)}(F)]^2 + [\kappa_B^{(r)}(F)]^2 + [\kappa_\lambda^{(r)}(F)]^2}, \tag{3.17}$$

where $\kappa_A(F), \kappa_B(F)$, and $\kappa_\lambda(F)$ are defined by (3.12) with the norm $\|\ \|_2$. Moreover, rewrite (3.1) as

$$Y^T(A + BF^T)X = \Lambda,$$

where $X = [x_1, \ldots, x_n], Y = [y_1, \ldots, y_n]$, and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ with $\lambda_i \neq \lambda_j$ for $i \neq j$. Then in the same way as described in §2.3 we can define the relative condition numbers $c^{(r)}(\lambda_j)$ and $c^{(r)}(\lambda)$ of the closed-loop eigenvalues by

$$c^{(r)}(\lambda_j) = c(\lambda_j) \cdot \frac{\|A + BF^T\|_2}{|\lambda_j|} \quad \text{if } \lambda_j \neq 0 \tag{3.18}$$

and

$$c^{(r)}(\lambda) = \sqrt{\sum_{j=1}^{n} [c^{(r)}(\lambda_j)]^2} \quad \text{if } \lambda_j \neq 0 \quad \forall j, \tag{3.19}$$

where $c(\lambda_j) = \|x_j\|_2 \|y_j\|_2$ for $j = 1, \ldots, n$.

Remark 3.7. Let (A, B), $\lambda_1, \ldots, \lambda_n$, and F be as in Theorem 3.1, and let

$$A + BF^T = UTU^H$$

be the Schur decomposition of $A + BF^T$, where U is a unitary matrix and $T = \Lambda + M$ is an upper triangular matrix with a diagonal matrix $\Lambda = (\lambda_i)$ and a strictly upper triangular matrix M. Moreover, let A, B, and Λ be perturbed to $A + \Delta A, B + \Delta B$, and $\Lambda + \Delta\Lambda$, and let U and F be perturbed to $U + \Delta U$ and $F + \Delta F$. By using a first-order perturbation equation for U and F (the unknowns of the equation are ΔU and ΔF), Konstantinov and Petkov [13] derived upper bounds for $\|\Delta U\|_F$ and $\|\Delta F\|_F$, as well as the absolute condition numbers of the pole assignment problem with respect to A, B, and Λ. Note that there are several differences between [13] and this paper:

(i) the paper [13] does not restrict all the eigenvalues $\lambda_1, \ldots, \lambda_n$ to be simple;

(ii) the techniques for deriving condition numbers and perturbation bounds are different; and

(iii) the coefficient matrix W_f of the equation (3.7) is an $n \times mn$ matrix, but the coefficient matrix of an analogous equation in [13] is an $\frac{n(n+1)}{2} \times (\frac{n(n-1)}{2} + mn)$ matrix.

Consequently, the amount of work in computing the condition numbers and perturbation bounds by (3.12) and (3.14) is less than that of [13]. Besides, the condition numbers $\kappa_A(F)$ and $\kappa_\lambda(F)$ defined by (3.12) are, in general, different, but in [13], the condition numbers of the problem with respect to A and Λ are equal. Note that the paper [13] studies the condition numbers not only of the pole assignment problem, but also of the general feedback synthesis problem.

4. Numerical examples. Now we present some results of numerical tests.

The first three examples are for the single-input case. It is known that (i) for each controllable system (A, b) there exists a unique system Hessenberg decomposition

$$A = QHQ^T, \quad b = Qh,$$

where Q is orthogonal and

$$H = \begin{pmatrix} h_{11} & \cdots & \cdots & h_{1n} \\ h_{21} & \ddots & & \vdots \\ 0 & \ddots & \ddots & \vdots \\ 0 & 0 & h_{n,n-1} & h_{nn} \end{pmatrix}, \quad h = \begin{pmatrix} h_{10} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{4.1}$$

with $h_{j,j-1} > 0$ for all j [14]; (ii) a preliminary stage of the current pole assignment algorithms is to transform the pair (A, b) into the system Hessenberg form (4.1) via an orthogonal similarity transformation [15], [19]. Therefore, for simplicity, we assume that the system (A, b) of Examples 4.1–4.2 is in system Hessenberg form.

Example 4.1 (see [14]). Consider the system (A, b) with

$$A = \begin{pmatrix} -4 & 0 & 0 & 0 & 0 \\ \alpha & -3 & 0 & 0 & 0 \\ 0 & \alpha & -2 & 0 & 0 \\ 0 & 0 & \alpha & -1 & 0 \\ 0 & 0 & 0 & \alpha & 0 \end{pmatrix}, \quad b = \begin{pmatrix} \beta \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \tag{4.2}$$

where α and β are small positive numbers. Let

$$f = (3.12, -1.67, 7.45, -2.98, 0.37)^T \tag{4.3}$$

be the solution to Problem PA with data A, b, and $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)^T \in \mathcal{C}^5$, where the set $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$ is closed under complex conjugation.

Taking the spectral norm $\| \ \|_2$, by (2.27)–(2.29) we can compute the relative condition numbers $\kappa_A^{(r)}(f), \kappa_b^{(r)}(f), \kappa_\lambda^{(r)}(f)$, and $\kappa^{(r)}(f)$. Some numerical results obtained by using MATLAB are listed in Table 1, where

$$\lambda^{(1)} = \begin{pmatrix} -2.980127\text{e}+00 + i1.796999\text{e}+00 \\ -2.980127\text{e}+00 - i1.796999\text{e}+00 \\ -4.955320\text{e}-01 + i4.276059\text{e}-01 \\ -4.955320\text{e}-01 - i4.276059\text{e}-01 \\ 7.131756e-02 \end{pmatrix},$$

$$\lambda^{(2)} = \begin{pmatrix} -2.872401\text{e}+00 \\ -2.082527\text{e}+00 \\ -9.625392\text{e}-01 + i9.227306\text{e}-03 \\ -0.625392\text{e}-01 - i9.227306\text{e}-03 \\ 6.675582\text{e}-06 \end{pmatrix}, \tag{4.4}$$

$$\lambda^{(3)} = \begin{pmatrix} -2.991735\text{e}+00 \\ -2.000678\text{e}+00 \\ -8.876000\text{e}-01 \\ -9.999867\text{e}-01 \\ 6.964499\text{e}-10 \end{pmatrix}, \quad \lambda^{(4)} = \begin{pmatrix} -2.999208\text{e}+00 \\ -8.807849\text{e}-01 \\ -2.000007\text{e}+00 \\ -1.000000\text{e}+00 \\ 7.003156\text{e}-14 \end{pmatrix},$$

and $c(A, b)$, the condition number of the system Hessenberg form of the system (A, b), is defined as follows [24]. Let $\mathcal{L}_{\mathrm{s}}^{n\times n}$ denote the set of real $n \times n$ strictly lower triangular matrices. For $Y = (y_{ij}) \in \mathcal{R}^{n\times n}$, define the operator low : $\mathcal{R}^{n\times n} \to \mathcal{L}_{\mathrm{s}}^{n\times n}$ by

$$\mathrm{low}(Y) = (\eta_{ij}), \quad \eta_{ij} = \begin{cases} y_{ij} & \text{if } i > j, \\ 0 & \text{if } i \le j. \end{cases}$$

Further, define the operator $\mathbf{L} : \mathcal{L}_{\mathrm{s}}^{n\times n} \to \mathcal{L}_{\mathrm{s}}^{n\times n}$ by

$$\mathbf{L}X_{\mathrm{L}} = \mathrm{low}(X_{\mathrm{L}}(h, H(1:n-1)) - H(0, X_{\mathrm{L}}(1:n-1))), \quad X_{\mathrm{L}} \in \mathcal{L}_{\mathrm{s}}^{n\times n}, \tag{4.5}$$

TABLE 1
(take $\beta = 1.00e+00$).

α	1.00e+00	1.00e-01	1.00e-02	1.00e-03
λ	$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$
$\kappa_A^{(r)}(f)$	1.1109e+01	3.4957e+04	3.3155e+08	3.2969e+12
$\kappa_b^{(r)}(f)$	7.1558e+00	1.0247e+04	1.0118e+08	1.0117e+12
$\kappa_\lambda^{(r)}(f)$	5.1808e+00	2.4116e+04	2.3240e+08	2.3134e+12
$\kappa^{(r)}(f)$	1.4194e+01	4.3686e+04	4.1734e+08	4.1527e+12
$c(A,b)$	3.4910e+01	2.4839e+04	2.4740e+07	2.4739e+10

where H, h are expressed by (4.1), and $H(1:n-1)$ denotes the matrix that consists of the first $n-1$ columns of H. Then $c(A,b)$ is defined by

$$c(A,b) = \|\mathbf{L}^{-1}\|, \tag{4.6}$$

in which $\|\ \|$ is the operator norm induced from the Frobenius matrix norm.

The author's numerical tests show that the relative sensitivity of the solution f to Problem PA is increasing along with the increases of $c(A,b)$, the condition number of the system Hessenberg form of the controllable system (A,b) [24]. Observe the fact that the system Hessenberg form (H,h) of the system (A,b) is ill conditioned if the system (A,b) is very near to an uncontrollable one [14], [24]. Hence, presumably there is some intrinsic relation between the conditioning of the solution f to Problem PA and the distance of (A,b) from the nearest uncontrollable system [17]. For instance, perhaps the condition number $\kappa^{(r)}(f)$ is inversely proportional to the distance of (A,b) from the nearest uncontrollable system [7].

Example 4.2. Let A, b be expressed by (4.2), where $\alpha = 0.1, \beta = 1$, and let $\lambda = \lambda^{(2)}$ be the vector expressed by (4.4). It is known that the vector f expressed by (4.3) is the solution to Problem PA with the data A, b, λ. Let $\tilde{f}$ be the solution to Problem PA with the data $\tilde{A}, \tilde{b}, \tilde{\lambda}$:

$$\tilde{A} = A + E, \quad \tilde{b} = b + e, \quad \tilde{\lambda} = \lambda + \nu,$$

where

$$E = \epsilon E_0, \quad e = \epsilon e_0, \quad \nu = \epsilon \nu_0,$$

$$E_0 = \begin{pmatrix} 0.182 & -0.378 & 0.394 & 0.223 & -0.556 \\ -0.481 & 0.274 & 0.683 & -0.911 & 0.025 \\ 0.284 & -0.179 & 0.932 & 0.573 & -0.447 \\ 0.116 & -0.523 & 0.379 & -0.432 & 0.332 \\ -0.765 & 0.337 & -0.184 & -0.367 & 0.386 \end{pmatrix}, \tag{4.7}$$

$$e_0 = (1,\ -2,\ 0.3,\ -4,\ -1)^T, \quad \nu_0 = (1+i,\ 1-i,\ 3,\ -2,\ 1)^T,$$

and ϵ is a very small positive scalar.

Taking the norm $\|\ \|_2$ and different values of ϵ, by (2.25) we can compute approximate upper bounds δ_f and Δ_f for $\|\tilde{f} - f\|_2$. Some numerical results obtained by using MATLAB and the file SEVAS are listed in Table 2. Note that SEVAS and MEVAS (used in Example 4.4) are computer programs written by G. S. Miminis, Department

TABLE 2

ϵ	1.00e-04	1.00e-06	1.00e-08	1.00e-10
$\|\tilde{f}-f\|_2$	5.2352e+00	2.8292e-01	1.8205e-03	1.8144e-05
δ_f	1.9661e+01	1.9661e-01	1.9661e-03	1.9661e-05
Δ_f	8.0117e+01	8.0117e-01	8.0117e-03	8.0117e-05

of Computer Science, Memorial University of Newfoundland, Canada. The programs are implementations of an algorithm for pole assignment by Miminis and Paige [16].

The results of Table 2 show that, by using the estimates (2.25), the computed approximate upper bounds δ_f and Δ_f of $\|\tilde{f}-f\|_2$ are satisfactory.

Example 4.3 (see [10, Ex. 1]). Let

$$A = \operatorname{diag}(0.1, 0.2, 0.3, 0.4, 0.5, 0.6), \quad b = (1, 2, 3, 4, 5, 6)^T. \tag{4.8}$$

Suppose that we wish to assign the eigenvalues -6, -5, -4, -3, -1.1, -1. Varga [26] has pointed out that the computed feedback by any of the stable algorithms has a very high relative accuracy (about 10^{-9}), but the resulting closed-loop eigenvalues have only about two correct digits. This means that the solution to the pole assignment problem is well conditioned but the closed-loop eigenvalues are extremely ill conditioned. Note that this fact can be clarified by our analysis in §2. Let

$$\lambda = (-6, -5, -4, -3, -1.1, -1)^T. \tag{4.9}$$

By using MATLAB and the file SEVAS, from A, b, and λ we get the computed feedback and then by the formulae (2.27)–(2.29) and (2.31)–(2.33) we get the relative condition numbers of the feedback and those of the resulting closed-loop eigenvalues as follows:

(4.10)
$$\kappa_A^{(r)}(f) = 2.0439\text{e}+01, \quad \kappa_b^{(r)}(f) = 1.9456\text{e}+00, \quad \kappa_\lambda^{(r)}(f) = 1.0535\text{e}+01,$$
$$\kappa^{(r)}(f) = 2.3077\text{e}+01,$$

(4.11)
$$c^{(r)}(\lambda_1) = 2.3956\text{e}+15, \quad c^{(r)}(\lambda_2) = 6.6419\text{e}+15, \quad c^{(r)}(\lambda_3) = 6.4221\text{e}+15,$$
$$c^{(r)}(\lambda_4) = 2.3411\text{e}+15, \quad c^{(r)}(\lambda_5) = 4.5093\text{e}+14, \quad c^{(r)}(\lambda_6) = 3.2512\text{e}+14,$$
$$c^{(r)}(\lambda) = 9.8431\text{e}+15.$$

From (4.10) and (4.11) we can understand why the computed feedback by any of the stable algorithms has a very high relative accuracy but the resulting closed-loop eigenvalues have only about two correct digits.

Let A, b, λ be expressed by (4.8)–(4.9), and let $A_\alpha = \alpha A$, where α is a positive number. Moreover, let f_α be the solution to Problem PA with the data A_α, b, λ, and let $c(A_\alpha, b)$ be the condition number of the system Hessenberg form of the controllable system (A_α, b) defined by (4.5)–(4.6). From the results listed in Table 3 we see a similar phenomenon as shown by the results in Table 1: the relative sensitivity of f_α is increasing along with the increases of $c(A_\alpha, b)$.

TABLE 3

α	1.00e+00	1.00e-01	1.00e-02	1.00e-3
$\kappa^{(r)}(f_\alpha)$	2.3077e+01	1.6510e+03	1.6308e+04	1.6313e+05
$c(A_\alpha, b)$	2.1956e+01	2.1948e+02	2.1948e+03	2.1948e+04

Example 4.4 (see [19]). Consider the system (A, B) with

$$A = \begin{pmatrix} -0.10940 & 0.06280 & 0.00000 & 0.00000 & 0.00000 \\ 1.30600 & -2.13200 & 0.98070 & 0.00000 & 0.00000 \\ 0.00000 & 1.59500 & -3.14900 & 1.54700 & 0.00000 \\ 0.00000 & 0.03550 & 2.63200 & -4.25700 & 1.85500 \\ 0.00000 & 0.00227 & 0.00000 & 0.16360 & -0.16250 \end{pmatrix},$$

$$B = \begin{pmatrix} 0.0000 & 0.0632 & 0.0838 & 0.1004 & 0.0063 \\ 0.0000 & 0.0000 & -0.1396 & -0.2060 & -0.0128 \end{pmatrix}^T,$$

and let

$$\lambda = (-1+i,\ -1-i,\ -0.5,\ -0.2,\ -1)^T.$$

By using MATLAB and the file MEVAS, we obtain a solution F to Problem PA with the data A, B, λ:

$$F = \begin{pmatrix} 2.269923501498446\text{e}+01 & 4.328501909623649\text{e}+00 \\ -5.312109137216190\text{e}+01 & 1.066982012066820\text{e}+01 \\ 5.597202515142414\text{e}+01 & -7.995243974279445\text{e}+01 \\ -6.697515863715235\text{e}+01 & 6.098543501544316\text{e}+01 \\ -6.435235435440497\text{e}+01 & -8.502047279072825\text{e}+01 \end{pmatrix}.$$

By using MATLAB and by (3.16)–(3.17) we get a group of relative condition numbers of the feedback F as follows:

$$\kappa_A^{(r)}(F) = 3.9150\text{e}+00, \quad \kappa_B^{(r)}(F) = 4.7480\text{e}+00, \quad \kappa_\lambda^{(r)}(F) = 3.9134\text{e}-01,$$

$$\kappa^{(r)}(F) = 9.5921\text{e}+00.$$

Moreover, by (3.18)–(3.19) we get the relative condition number $c^{(r)}(\lambda)$ of λ : $c^{(r)}(\lambda) = 2.1926\text{e}+04$.

Let the data A, B, λ be perturbed to $\tilde{A}, \tilde{B}, \tilde{\lambda}$:

$$\tilde{A} = A + E_A, \quad \tilde{B} = B + E_B, \quad \tilde{\lambda} = \lambda + \nu,$$

where $E_A = \epsilon E_0, E_B = \epsilon E_1, \nu = \epsilon\nu_0$, in which E_0 and ν_0 are expressed by (4.7), ϵ is a very small positive scalar, and

$$E_1 = \begin{pmatrix} 1.0 & -2.0 & 0.3 & -4.0 & -1.0 \\ -1.0 & 3.0 & 2.0 & 1.0 & 0.6 \end{pmatrix}^T.$$

For each group of data $\tilde{A}, \tilde{B}, \tilde{\lambda}$, we obtain a solution $\tilde{F}$ to Problem PA by using MATLAB and the file MEVAS. On the other hand, by (3.15) we can compute approximate upper bounds δ_F and Δ_F. Some numerical results are listed in Table 4.

The results of Table 4 show that, by using the estimates (3.15), the computed approximate upper bounds δ_F and Δ_F are satisfactory.

TABLE 4

ϵ	1.00e-04	1.00e-06	1.00e-08	1.00e-10
$\|\tilde{F} - F\|_F$	1.6593e+00	1.6432e-02	1.6431e-04	1.6431e-06
δ_F	1.6810e+00	1.6810e-02	1.6810e-04	1.6810e-06
Δ_F	3.3854e+00	3.3854e-02	3.3854e-04	3.3854e-06

Acknowledgments. I am grateful to Dr. A. Varga for very helpful discussions and suggestions. I am also grateful to Professor Bo Kågström, Dr. Chunyang He, and the referees for helpful comments.

REFERENCES

[1] M. ARNOLD, *Conditioning and Algorithm for the Eigenvalue Assignment Problem*, Ph.D. thesis, Department of Mathematical Sciences, Northern Illinois University, Dekalb, IL, 1993.

[2] S. BOCHNER AND W. T. MARTIN, *Several Complex Variables*, Princeton University Press, Princeton, NJ, 1948.

[3] R. BRU, J. MAS, AND A. M. URBANO, *An algorithm for the single-input pole assignment problem*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 393–407.

[4] C. I. BYRNES AND X. WANG, *The additive inverse eigenvalue problem for Lie perturbations*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 113–117.

[5] C. L. COX AND W. F. MOSS, *Backward error analysis for a pole assignment algorithm*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 446–456.

[6] ———, *Backward error analysis for a pole assignment algorithm* II: *The complex case*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1159–1171.

[7] J. DEMMEL, *On condition numbers and the distance to the nearest ill-posed problem*, Numer. Math., 51 (1987), pp. 251–289.

[8] A. J. GEURTS, *A contribution to theory of condition*, Numer. Math., 39 (1982), pp. 85–96.

[9] OLE H. HALD, *Inverse eigenvalue problems for Jacobi matrices*, Linear Algebra Appl., 14 (1976), pp. 63–85.

[10] C. HE AND V. MEHRMANN, *Stabilization of linear systems*, Report spc 94-21, Fakultät für Mathematik, TU Chemnitz-Zwickau, Chemnitz, Germany, 1994.

[11] J. KAUTSKY, N. K. NICHOLS, AND P. VAN DOOREN, *Robust pole assignment in linear state feedback*, Internat. J. Control, 41 (1986), pp. 1129–1155.

[12] M. M. KONSTANTINOV, N. D. CHRISTOV, AND P. HR. PETKOV, *Perturbation analysis of linear control problems*, IFAC 10th Congress, Vol. 9, pp. 16–21, Munich, 1987.

[13] M. M. KONSTANTINOV AND P. HR. PETKOV, *Conditioning of linear state feedback*, Report 93-61, Department of Engineering, Leicester University, Leicester, United Kingdom, November 1993.

[14] A. J. LAUB AND A. LINNEMANN, *Hessenberg and Hessenberg/triangular forms in linear system theory*, Internat. J. Control, 44 (1986), pp. 1523–1547.

[15] G. S. MIMINIS AND C. C. PAIGE, *An algorithm for pole assignment of time invariant linear systems*, Internat. J. Control, 35 (1982), pp. 341–354.

[16] ———, *A QR-like approach for the eigenvalue assignment problem*, to appear in the "2nd Hellenic Conference on Mathematics and Informatics", Athens, Greece, September 1994.

[17] C. C. PAIGE, *Properties of numerical algorithms related to computing controllability*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 130–138.

[18] R. V. PATEL, A. J. LAUB, AND P. M. VAN DOOREN, *Introduction and survey*, Numerical Linear Algebra Techniques for Systems and Control, A Selected Reprint Volume, IEEE Control Systems Society, Sponsor, The Institute of Electrical and Electronics Engineers, Inc., New York, 1994.

[19] P. HR. PETKOV, N. D. CHRISTOV, AND M. M. KONSTANTINOV, *A computational algorithm for pole assignment of linear single-input systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 1045–1048.

[20] J. R. RICE, *A theory of condition*, SIAM J. Numer. Anal., 3 (1966), pp. 287–310.

[21] M. SHUB, *The implicit function theorem revisited*, IBM J. Res. Develop., 38 (1994), pp. 259–264.

[22] J.-G. SUN, *Eigenvalues and eigenvectors of a matrix dependent on several parameters*, J. Comput. Math., 3 (1985), pp. 351–364.

[23] J.-G. SUN, *The stability analysis of the solutions of inverse eigenvalue problems*, J. Comput. Math., 4 (1986), pp. 345–353.
[24] ———, *Perturbation analysis of system Hessenberg and Hessenberg-triangular forms*, to appear in Linear Algebra Appl.
[25] A. VARGA, *A Schur method for pole assignment*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 517–519.
[26] ———, private communication, December 1994.
[27] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon, Oxford, UK, 1965.
[28] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Springer–Verlag, 2nd ed., New York, 1979.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 2, pp. 332–347, April 1996

006

A SUBSPACE MODEL IDENTIFICATION SOLUTION TO THE IDENTIFICATION OF MIXED CAUSAL, ANTI-CAUSAL LTI SYSTEMS*

MICHEL VERHAEGEN†

Abstract. This paper describes the modification of the family of MOESP[1] subspace algorithms when identifying mixed causal and anti-causal systems. It is assumed that these class of systems have a regular pencil $zE - A$, where E is possibly singular. The key numerical problem in solving this identification problem is the separation of the extended observability matrix of the causal part from that of the anti-causal part when a mixture of both is determined from the input–output data. For the general mixed causal, anti-causal case, this requires a partial calculation of the Kronecker canonical form of the pencil $zE - A$, where the pair $[A\ E]$ has been determined from the recorded input–output data. For the descriptor case, that is, when E is nilpotent, this problem is solved without computing the Kronecker canonical form.

All existing members of the MOESP family applicable to causal, linear, time-invariant systems are generalized. This allows a broad scope of identification problems for mixed causal, anti-causal systems to be addressed.

Key words. linear systems, descriptor systems, subspace identification, causal, anti-causal systems

AMS subject classifications. 15A18, 65F20

1. Introduction.

Let us consider the discrete time-generalized state–space model [1], [2], [3], [6].

$$\bar{E}\chi_{k+1} = \bar{A}\chi_k + \bar{B}u_k, \tag{1}$$

$$y_k = \bar{C}\chi_k + \bar{D}u_k, \tag{2}$$

where $u_k \in \mathbb{R}^m$, $y_k \in \mathbb{R}^\ell$, and $\chi_k \in \mathbb{R}^n$ and $\bar{E}, \bar{A}, \bar{B}, \bar{C}$, and $\bar{D}$ are constant matrices of appropriate dimensions.

If $\bar{E}$ is invertible, the system is a causal (strictly proper) system. In that case we can write (1) as

$$\chi_{k+1} = \bar{E}^{-1}\bar{A}\chi_k + \bar{E}^{-1}\bar{B}u_k.$$

When the pencil $z\bar{E} - \bar{A}$ is regular, i.e., the $\det(z\bar{E} - \bar{A}) \neq 0$, the so-called Kronecker canonical form has the following specific structure:

$$z\begin{pmatrix} I & 0 \\ 0 & E \end{pmatrix} - \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix}.$$

Correspondingly, the generalized state–space description is transformed into the following mixed causal, anti-causal form:

$$\begin{aligned} x^c_{k+1} &= Ax^c_k + B^c u_k \qquad \text{(causal part)}, \\ Ex^{ac}_{k+1} &= x^{ac}_k - B^{ac}u_k \qquad \text{(anti-causal part)}, \\ y_k &= \begin{pmatrix} C^c & C^{ac} \end{pmatrix}\begin{pmatrix} x^c_k \\ x^{ac}_k \end{pmatrix} + Du_k. \end{aligned} \tag{3}$$

* Received by the editors January 26, 1994; accepted for publication (in revised form) by P. Van Dooren May 21, 1995. This research was supported by a senior research fellowship from the Royal Dutch Academy of Arts and Sciences.

† Department of Electrical Engineering, Network Theory Section, P. O. Box 5031, NL-2600 GA Delft, The Netherlands (m.verhaegen@et.tudelft.nl).

[1] The acronym MOESP stands for multivariable output error state–space model identification schemes and was introduced in [8].

Here $x_k^c \in \mathbb{R}^{n_c}, x_k^{ac} \in \mathbb{R}^{n_{ac}}$, $n = n_c + n_{ac}$, and the eigenvalues of A, denoted by $\lambda(A)$, satisfy $|\lambda(A)| \leq 1$, while those of E satisfy $|\lambda(E)| < 1$. The particular class of systems that have the above generalized state–space form in combination with the fact that the matrix E is nilpotent, i.e., $E^{n_{ac}} \equiv 0$, is sometimes indicated by the class of descriptor systems [6].

Mixed causal, anti-causal, or descriptor systems frequently occur in a system analysis of practical problems. Examples are the biomedical application of identifying the human joint dynamics [7], the inversion of causal systems [6], and the description of electrical networks and economical systems [5]. Despite their importance, only a very limited number of solutions are available to directly identify mixed causal, anti-causal systems in the state–space form given in (3) from recorded input–output sequences.

One class of algorithms first estimates the impulse response of the system (3) from input–output data and then realizes a state–space representation for this impulse response. Examples in this class are [7], [3], and [14]. As outlined for causal systems in [8] and for descriptor systems in [4], a better alternative is to identify the state–space model directly from input–output data without having to estimate the impulse response first.

Within this second class of so-called subspace model identification algorithms we present a number of algorithms belonging to the MOESP family that allow one to tackle very general identification problems for mixed causal, anti-causal systems. These algorithms differ from that presented in [4] in the following ways.

(1) They do not introduce unobservable modes as done in [4] when the matrix E in (3) is singular. This is because in [4], the observable part of the following modified system

$$\begin{bmatrix} x_{k+1}^c \\ x_{k-1}^{ac} \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & E \end{bmatrix} \begin{bmatrix} x_k^c \\ x_k^{ac} \end{bmatrix} + \begin{bmatrix} B^c \\ -B^{ac} \end{bmatrix} u_k,$$

$$y_k = \begin{bmatrix} C^c & C^{ac}E \end{bmatrix} \begin{bmatrix} x_k^c \\ x_k^{ac} \end{bmatrix} + (D - C^{ac}B^{ac})u_k$$

is identified instead of identifying (3) directly. As remarked in [4], the anti-causal part of the above modified system is unobservable when the matrix E is singular. This is so even when the original anti-causal part was observable.

(2) The MOESP approach allows one to identify descriptor systems without the need to calculate the Kronecker canonical form. This is explicitly outlined in Theorem 3 in §3.3.

(3) We extend all members of the MOESP family that are applicable to time-invariant linear and finite-dimensional systems to the generalized class of systems treated in this paper. This leads to a number of algorithms that allow the solution of a wide variety of identification problems in a statistically consistent manner.

In this paper, we subsequently treat the following topics. In §2, we state the identification problems analyzed throughout the paper and in §3 we extend the basic member of the MOESP family to mixed causal, anti-causal systems, namely, the ordinary MOESP scheme. Here we also give special attention to the solvability of the set of equations that arise in calculating the different system matrices of the state–space representation in (3) (up to a similarity transformation). In the final section, we present the extension to the other members of the MOESP family.

2. Formulation of the identification problem for mixed causal, anti-causal systems. First, we state the deterministic form, that is, when the input–output records are noise free.

The deterministic identification problem for mixed causal, anti-causal systems. Let the data sequences $[u_j, u_{j+1}, \ldots, u_{N+j-1}]$ and $[y_j, y_{j+1}, \ldots, y_{N+j-1}]$ denote an input–output pair for the system (3). Then the problem is to determine a similarly equivalent generalized state–space model having the following explicit mixed causal, anti-causal form:

$$
\begin{aligned}
\eta_{k+1}^{c} &= (T^{c})^{-1} A T^{c} \eta_{k}^{c} + (T^{c})^{-1} u_{l} && := A_T \eta_{k}^{c} + B_T^{c} u_k, \\
\eta_{k}^{ac} &= (T^{ac})^{-1} E T^{ac} \eta_{k+1}^{ac} + (T^{ac})^{-1} B^{ac} u_k && := E_T \eta_{k+1}^{ac} + B_T^{ac} u_k, \\
y_k &= \begin{pmatrix} C^{c}T^{c} & C^{ac}T^{ac} \end{pmatrix} \begin{pmatrix} \eta_{k}^{c} \\ \eta_{k}^{ac} \end{pmatrix} + D u_k := \begin{pmatrix} C_T^{c} & C_T^{ac} \end{pmatrix} \begin{pmatrix} \eta_{k}^{c} \\ \eta_{k}^{ac} \end{pmatrix} + D u_k
\end{aligned} \tag{4}
$$

from the given input–output data sequences. Here := denotes the definition of the quantities on the right of the assignment.

The real identification problem that will be considered in this paper is the determination of statistically consistent estimates of the similarly equivalent matrices $A_T, E_T, \ldots$ when the output sequence is perturbed by additive, unmeasurable errors. Three types of errors will be considered: (1) zero-mean white noise errors, (2) arbitrary colored errors including a deterministic bias which is independent from the input sequence u_k, and (3) zero-mean errors in the following innovation type of model structure:

$$
\begin{aligned}
x_{k+1}^{c} &= A x_{k}^{c} + B^{c} u_k + K^{c} w_k, \\
E x_{k+1}^{ac} &= x_{k}^{ac} - B^{ac} u_k + K^{ac} w_k, \\
y_k &= \begin{pmatrix} C^{c} & C^{ac} \end{pmatrix} \begin{pmatrix} x_{k}^{c} \\ x_{k}^{ac} \end{pmatrix} + D u_k + v_k,
\end{aligned} \tag{5}
$$

where w_k and v_k are zero-mean, discrete white noise sequences.

3. Description of the extension of the ordinary MOESP algorithm.

3.1. Data equation and definitions of observability and persistency of excitation. From the input sequence $\{u_k\}$ construct the following Hankel matrix:

$$
U_{j,s} = \begin{bmatrix} u_j & u_{j+1} & \cdots & u_{N+j-1} \\ u_{j+1} & u_{j+2} & \cdots & u_{N+j} \\ \vdots & & \ddots & \vdots \\ u_{j+s-1} & u_{j+s} & \cdots & u_{N+j+s-2} \end{bmatrix}.
$$

Similarly construct the Hankel matrix $Y_{j,s}$ from the output data. Let the state vector sequences x_k^c and x_k^{ac} be stored in the matrices X_j^c and X_j^{ac}, respectively, as

$$
X_j^{c} = \begin{bmatrix} x_j^{c} & x_{j+1}^{c} & \cdots & x_{N+j-1}^{c} \end{bmatrix}, \quad X_j^{ac} = \begin{bmatrix} x_j^{ac} & x_{j+1}^{ac} & \cdots & x_{N+j-1}^{ac} \end{bmatrix}.
$$

Then, when defining the extended observability matrix Γ_s and the near Toeplitz matrix H_s as

$$
\Gamma_s = \left[\begin{array}{c|c} C^{c} & C^{ac}E^{s-1} \\ C^{c}A & C^{ac}E^{s-2} \\ \vdots & \vdots \\ C^{c}A^{s-1} & C^{ac} \end{array}\right] := \left[\begin{array}{c|c} \Gamma_s^{c} & \Gamma_s^{ac} \end{array}\right],
$$

$$H_s = \begin{bmatrix} D + C^{ac}B^{ac} & C^{ac}EB^{ac} & \cdots & C^{ac}E^{s-2}B^{ac} & 0 \\ C^cB^c & D + C^{ac}B^{ac} & \cdots & C^{ac}E^{s-3}B^{ac} & 0 \\ \vdots & & \ddots & & \vdots \\ C^cA^{s-2}B^c & C^cA^{s-3}B^c & \cdots & C^cB^c & D \end{bmatrix},$$

we can relate the data Hankel matrices $U_{1,s}$ and $Y_{1,s}$ as

$$Y_{1,s} = \Gamma_s \begin{bmatrix} X_1^c \\ X_s^{ac} \end{bmatrix} + H_s U_{1,s}. \tag{6}$$

From [16] we state the following definition of observability of the generalized state–space system (3).

DEFINITION 1. *The linear time-invariant generalized state–space system* (3) *is observable if and only if*

$$\rho\begin{bmatrix} \Gamma_{s_1}^c \end{bmatrix} = n_c, \quad \rho\begin{bmatrix} \Gamma_{s2}^{ac} \end{bmatrix} = n_{ac}, \quad \text{and} \quad \rho[\Gamma_s] = n,$$

where $\rho(.)$ *denotes the rank of the matrix* $(.)$ *and* $s_1 \geq n_c$, $s_2 \geq n_{ac}$, *and* $s \geq n$. □

From [8] we recall the following definition of persistency of excitation.

DEFINITION 2. *The input* u_k *is persistently exciting to the linear time-invariant generalized state–space system* (3) *if*

$$\rho\begin{bmatrix} U_{1,s} \\ \hline X_1^c \\ X_s^{ac} \end{bmatrix} = ms + n. \qquad \square$$

It should be remarked that this definition requires the underlying system to be controllable. For a proper definition of controllability for the generalized systems (3) we refer to [16].

3.2. Calculating the extended observability matrix Γ_s. The key step in the algorithms belonging to the MOESP family, described in [8], [9], [10], [11], [12], and [13], is the calculation of the column space of the extended observability matrix Γ_s.

Because the data equation (6) is exactly the same as that obtained for causal time-invariant systems analyzed in the above series of papers, we state without proof the following theorem.

THEOREM 1. *Let the linear time-invariant descriptor system* (3) *be observable and let the input be persistently exciting over the time interval* $[1, N]$. *In addition, let*

1. $s \geq n$ *and* $N \geq s\ell + (s-1)m + (s-1)$,
2. *the following RQ factorization be given:*

$$\begin{bmatrix} U_{1,s} \\ \hline Y_{1,s} \end{bmatrix} = \begin{bmatrix} R_{11} & 0 \\ \hline R_{21} & R_{22} \end{bmatrix} \begin{bmatrix} Q_1 \\ \hline Q_2 \end{bmatrix}, \tag{7}$$

3. *and the following SVD of the matrix* R_{22} *be given:*

$$R_{22} = \begin{bmatrix} U_n & U_n^{\perp} \end{bmatrix} \begin{bmatrix} S_n & 0 \\ \hline 0 & S_2 \end{bmatrix} \begin{bmatrix} V_n^T \\ \hline V_2^T \end{bmatrix} \tag{8}$$

with $S_n \in R^{n \times n}$.

Then $S_n > 0$, $S_2 = 0$, *and* $\exists T \in R^{n \times n}$ *with* T *nonsingular, such that*

$$U_n = \Gamma_s T. \tag{9}$$

3.3. Calculating the matrices A_T, E_T, and C_T^c, C_T^{ac}. Because of the special structure of Γ_s and the relationship (9), the following holds:

$$U_n(1:(s-1)\ell,:)T^{-1}\begin{bmatrix} A & 0 \\ 0 & I \end{bmatrix} = U_n(\ell+1:s\ell,:)T^{-1}\begin{bmatrix} I & 0 \\ 0 & E \end{bmatrix}.$$

Therefore, by solving the following overdetermined set of equations for $\overline{A}_T$ and $\overline{E}_T$,

$$\left[\; U_n(1:(s-1)\ell,:) \mid -U_n(\ell+1:s\ell,:) \;\right]\left[\begin{array}{c} \overline{A}_T \\ \hline \overline{E}_T \end{array}\right] = 0, \tag{10}$$

the pencil $z\overline{E}_T - \overline{A}_T$ is similarly equivalent to the pencil $z\left[\begin{smallmatrix} I & 0 \\ 0 & E \end{smallmatrix}\right] - \left[\begin{smallmatrix} A & 0 \\ 0 & I \end{smallmatrix}\right]$. As a consequence, we can compute A_T and E_T by reducing the pencil $z\overline{E}_T - \overline{A}_T$ into the Kronecker canonical form.

To determine the additional system matrices C_T^c and C_T^{ac} and later on also B_T^c, B_T^{ac}, and D, we look for an invertible $n \times n$ transformation matrix P such that

$$U_n P = \left[\; \Gamma_s^c \mid \Gamma_s^{ac} \;\right]\left[\begin{array}{c|c} T_c & 0 \\ \hline 0 & T_{ac} \end{array}\right], \tag{11}$$

where both T_c and T_{ac} are square, nonsingular matrices. For that purpose, we first reduce the matrix pencil $z\overline{E}_T - \overline{A}_T$ into the following similarly equivalent form:

$$(P')^{-1}\left(z\overline{E}_T - \overline{A}_T\right)Q' = z\begin{bmatrix} I & 0 \\ 0 & E_T \end{bmatrix} - \begin{bmatrix} A_T & 0 \\ 0 & I \end{bmatrix}, \tag{12}$$

where the eigenvalues of A_T and E_T coincide with those of A and E, respectively. For computational details we refer to §3.5.

In the next theorem, we show that for $P = P'$, (11) indeed holds.

THEOREM 2. *Let the matrix pencil $z\overline{E}_T - \overline{A}_T$, with $\overline{A}_T$ and $\overline{E}_T$ computed by solving* (10) *for $s > n$, be reduced by the pair of invertible matrices (P', Q') into the structured form as indicated in* (12)*, with the spectrum of the matrices A_T and E_T equal to that of the matrices A and E in* (3)*, respectively. Then the relationship* (11) *holds for $P = P'$.*

Proof. Suppose that the transformation P' computed in (12) yields

$$U_n P' = \left[\; \Gamma_s^c \mid \Gamma_s^{ac} \;\right]\left[\begin{array}{c|c} T_c' & T_2 \\ \hline T_1 & T_{ac}' \end{array}\right].$$

Using the special form of the transformed pencil on the right of (12), (10) can be denoted as

$$U_n(1:(s-1)\ell,:)P'\begin{bmatrix} A_T & 0 \\ 0 & I \end{bmatrix} = U_n(\ell+1:s\ell,:)P'\begin{bmatrix} I & 0 \\ 0 & E_T \end{bmatrix}.$$

Hence,

$$\left[\; \Gamma_s^c(1:(s-1)\ell,:) \mid \Gamma_s^{ac}(1:(s-1)\ell,:) \;\right]\left[\begin{array}{c|c} T_c' & T_2 \\ \hline T_1 & T_{ac}' \end{array}\right]\begin{bmatrix} A_T & 0 \\ 0 & I \end{bmatrix}$$

$$= \left[\; \Gamma_s^c(\ell+1:s\ell,:) \mid \Gamma_s^{ac}(\ell+1:s\ell,:) \;\right]\left[\begin{array}{c|c} T_c' & T_2 \\ \hline T_1 & T_{ac}' \end{array}\right]\begin{bmatrix} I & 0 \\ 0 & E_T \end{bmatrix},$$

which is denoted more compactly as

$$\left[\begin{array}{c|c} \Gamma_c^{(1)} & \Gamma_{ac}^{(1)} \end{array}\right] \left[\begin{array}{c|c} T_c' A_T & T_2 \\ \hline T_1 A_T & T_{ac}' \end{array}\right] = \left[\begin{array}{c|c} \Gamma_c^{(2)} & \Gamma_{ac}^{(2)} \end{array}\right] \left[\begin{array}{c|c} T_c' & T_2 E_T \\ \hline T_1 & T_{ac}' E_T \end{array}\right]. \tag{13}$$

The matrices A and E satisfy the equations

$$\Gamma_c^{(1)} A = \Gamma_c^{(2)}, \quad \Gamma_{ac}^{(2)} E = \Gamma_{ac}^{(1)}.$$

Using these two relationships, (13) leads to the following two equations:

$$\begin{aligned} \Gamma_c^{(1)} (T_c' A_T - A T_c') &= -\Gamma_{ac}^{(2)} (E T_1 A_T - T_1), \\ \Gamma_c^{(1)} (T_2 - A T_2 E_T) &= \Gamma_{ac}^{(2)} (T_{ac}' E_T - E T_{ac}'). \end{aligned}$$

Since $s > n$, we have that $\rho(\ \Gamma_c^{(1)} \quad \Gamma_{ac}^{(2)}\ ) = \rho(\Gamma_{s-1}) = n$. As a consequence, the above equations only hold if

$$\begin{aligned} T_c' A_T - A T_c' &= 0, \\ T_{ac}' E_T - E T_{ac}' &= 0, \\ E T_1 A_T - T_1 &= 0, \\ T_2 - A T_2 E_T &= 0. \end{aligned}$$

In the last two Sylvestor equations, the matrices A_T and A (respectively, E and E_T) have a spectrum located inside or on the unit circle (respectively, inside the unit circle). Therefore, both Sylvestor equations have the zero solution as a unique solution. Hence,

$$T_1 \equiv 0, \quad T_2 \equiv 0$$

and the proof is completed. □

If we define the quantity

$$U_n' = U_n P', \tag{14}$$

then the matrices A_T and E_T satisfy

$$U_n'(1 : (s-1)\ell, 1 : n_c) A_T = U_n'(\ell + 1 : s\ell, 1 : n_c), \tag{15}$$

$$U_n'(1 : (s-1)\ell, n_c + 1 : n) = U_n'(\ell + 1 : s\ell, n_c + 1 : n) E_T, \tag{16}$$

and C_T^c, C_T^{ac},

$$C_T^c = U_n'(1 : \ell, 1 : n_c), \quad C_T^{ac} = U_n'((s-1)\ell : s\ell, n_c + 1 : n). \tag{17}$$

When the underlying system is a descriptor system, it is possible to "split" the column space of U_n into one part related only to the causal part and one related only to the anti-causal part of the underlying system without needing to compute the Kronecker canonical form. This is outlined in the following theorem.

THEOREM 3. *When the system given in* (3) *is a descriptor system such that* $E^{n_{ac}} \equiv 0$ *and the conditions of Theorem* 1 *are satisfied for* $s \geq \max(2n_{ac}, 2n_c)$ *and* s

even, then

$$\exists P \in \mathbb{R}^{n\times n}, T_{11} \in \mathbb{R}^{n_c\times n_c}, \text{ and } T_{22} \in \mathbb{R}^{n_{ac}\times n_{ac}} : U_n P$$
$$= \left[\begin{array}{c|c} \begin{bmatrix} C^c \\ C^cA \\ \vdots \\ C^cA^{\frac{s}{2}-1} \end{bmatrix} T_{11} & 0 \\ \hline \star & \begin{bmatrix} C^{ac}E^{\frac{s}{2}-1} \\ \vdots \\ C^{ac}E \\ C^{ac} \end{bmatrix} T_{22} \end{array}\right]$$

for P, T_{11}, T_{22} nonsingular and $\star$ is a matrix in $\mathbb{R}^{\frac{s}{2}\ell\times n_c}$.

Proof. First note that since $\frac{s}{2} \geq n_{ac}$, Γ_s has the following form:

$$\Gamma_s = \left[\begin{array}{c|c} C^c & \\ C^cA & \\ \vdots & 0 \\ C^cA^{\frac{s}{2}-1} & \\ \hline C^cA^{\frac{s}{2}} & C^{ac}E^{\frac{s}{2}-1} \\ \vdots & \vdots \\ C^cA^{s-1} & C^{ac} \end{array}\right] = \left[\begin{array}{c|c} \Gamma_{11} & 0 \\ \hline \Gamma_{21} & \Gamma_{22} \end{array}\right].$$

By Theorem 1, we have that

$$U_n = \Gamma_s T = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} \quad \text{with } \gamma_1 \in \mathbb{R}^{\frac{s}{2}\ell\times n}.$$

Since T is nonsingular and $\frac{s}{2} \geq n_c$, this shows that $\rho(\gamma_1) = n_c$. Therefore,

$$\exists P \in \mathbb{R}^{n\times n}, P \text{ nonsingular} : \gamma_1 P = \begin{bmatrix} \overline{\gamma}_1 & 0 \end{bmatrix}$$

with $\overline{\gamma}_1 \in \mathbb{R}^{\frac{s}{2}\ell\times n_c}$ having the same column space as γ_1. This transformation P yields the desired result. For

$$U_n P = \begin{bmatrix} \overline{\gamma}_1 & 0 \\ \overline{\gamma}_2 & \overline{\gamma}_3 \end{bmatrix} = \begin{bmatrix} \Gamma_{11} & 0 \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix},$$

which shows that

$$\overline{\gamma}_1 = \Gamma_{11} T_{11}, \tag{18}$$
$$0 = \Gamma_{11} T_{12}, \tag{19}$$
$$\overline{\gamma}_3 = \Gamma_{21} T_{12} + \Gamma_{22} T_{22}. \tag{20}$$

Since $\frac{s}{2} \geq n_c$ we have that $\rho(\Gamma_{11}) = n_c$, and (19) yields $T_{12} = 0$ and the matrices T_{11} and T_{22} have to be nonsingular. Furthermore, (20) reduces to

$$\overline{\gamma}_3 = \Gamma_{22} T_{22}$$

and the proof is completed. □

The result of this theorem enables the calculation of A_T, E_T, C_T^c, and C_T^{ac} and from those matrices we construct the matrix U_n' as

$$U_n' = \left[\begin{array}{c|c} \begin{bmatrix} C^c \\ C^cA \\ \vdots \\ C^cA^{\frac{s}{2}-1} \\ \hline C^cA^{\frac{s}{2}} \\ \vdots \\ C^cA^{s-1} \end{bmatrix} T_{11} & \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \hline C^{ac}E^{\frac{s}{2}-1} \\ \vdots \\ C^{ac} \end{bmatrix} T_{22} \end{array} \right].$$

3.4. Calculating the matrices B_T^c, B_T^{ac}, and D. Substituting relationships obtained in the RQ factorization (7) into the data equation (6) yields, under the conditions of Theorem 1,

$$(U_n^{\perp})^T R_{21} R_{11}^{-1} = (U_n^{\perp})^T H_s. \tag{21}$$

When we know the column space of the causal part and the anti-causal part of the extended observability matrix Γ_s, as given by those of the matrices $U_n'(:, 1 : n_c)$ and $U_n'(:, n_c + 1 : n)$, respectively, we can rewrite (21) into a set of equations which is linear in the unknowns B_T^c, B_T^{ac}, and D. If we denote the left-hand side of (21) by the matrix Ξ, we can write the first m columns of (21) as

$$\Xi(:, 1 : m) = (U_n^{\perp})^T \begin{bmatrix} I_\ell & 0 \\ 0 & U_n'(1 : (s-1)\ell, 1 : n_c) \end{bmatrix} \begin{bmatrix} D + C_T^{ac} B_T^{ac} \\ B_T^c \end{bmatrix},$$

where I_ℓ denotes the identity matrix of order ℓ. The second to last block of m-columns can be denoted as

$$\begin{aligned} &\Xi(:, (s-2)m + 1 : (s-1)m) \\ &= (U_n^{\perp})^T \begin{bmatrix} U_n'(\ell + 1 : (s-1)\ell, n_c + 1 : n) & 0 & 0 \\ 0 & I_\ell & 0 \\ 0 & 0 & U_n'(1 : \ell, 1 : n_c) \end{bmatrix} \begin{bmatrix} B_T^{ac} \\ D + C_T^{ac} B_T^{ac} \\ B_T^c \end{bmatrix}. \end{aligned}$$

By now padding the matrix $(U_n^{\perp})^T$ with zeros on the left or on the right, we can denote the above equations and those obtained from intermediate columns of the matrix Ξ in the following unified way:

$$\begin{matrix} \\ \ell s-n \\ \ell s-n \\ \vdots \\ \ell s-n \end{matrix} \begin{matrix} \overset{m}{} \\ \left(\begin{array}{c} \Xi(:, 1 : m) \\ \Xi(:, m + 1 : 2m) \\ \vdots \\ \Xi(:, m(s-2) + 1 : m(s-1)) \end{array} \right. \end{matrix}$$

$$= \underbrace{\begin{bmatrix} 0_{\ell s-n \times (s-2)\ell} & (U_n^{\perp})^T \\ 0_{\ell s-n \times (s-3)\ell} & (U_n^{\perp})^T \quad 0_{\ell s-n \times \ell} \\ \vdots & \\ (U_n^{\perp})^T & 0_{\ell s-n \times (s-2)\ell} \end{bmatrix} \begin{bmatrix} U_n'(\ell + 1 : (s-1)\ell, n_c + 1 : n) & 0 & 0 \\ 0 & I_\ell & 0 \\ 0 & 0 & U_n'(1 : (s-1)\ell, 1 : n_c) \end{bmatrix}}$$

(22)

$$\cdot \begin{bmatrix} B_T^{ac} \\ D + C_T^{ac} B_T^{ac} \\ B_T^c \end{bmatrix}.$$

Here $0_{p\times q}$ denotes the zero matrix of dimensions $p \times q$.

The last m-columns of (21) become

$$\underbrace{U_n^{\perp}(\ell(s-1)+1:\ell s,:)^T}\, D = \Xi(:, m(s-1)+1:ms). \tag{23}$$

The set of equations (22) and (23) can be solved in least-squares sense if the under-braced matrices have full column rank. We now investigate when this is satisfied and start with the matrix in (23). The key here is summarized in the following theorem.

THEOREM 4. *Let the quantities $U_n^{\perp}$, Γ_s be defined as in Theorem* 1, *let $\rho(U_n^{\perp}(\ell(s-1)+1:\ell s,:)^T) := \rho(U_-) < \ell$, and let $\ell_- = \ell - \rho(U_-)$. Then*

$$\exists C_2 \in \mathbb{R}^{\ell\times\ell_-} : \begin{bmatrix} 0_{\ell(s-1)\times\ell_-} \\ C_2 \end{bmatrix} \subset \operatorname{span}_{\mathrm{col}}(\Gamma_s^{ac}) \quad \textit{and} \quad \rho(\begin{bmatrix} C_2 & U_- \end{bmatrix}) = \ell.$$

Proof. Denote the matrix $\begin{bmatrix} U_n \mid U_n^{\perp} \end{bmatrix}$ in (7) as

$$\begin{bmatrix} U_n \mid U_n^{\perp} \end{bmatrix} = \left[\begin{array}{c|c} U_{11} & U_{12} \\ U_{21} & U_- \end{array}\right].$$

Then since $\rho(U_-) < \ell$, $\exists Q \in \mathbb{R}^{(s\ell-n)\times(s\ell-n)}, \pi \in \mathbb{R}^{\ell\times\ell}$, $QQ^T = I, \pi\pi^T = I$, and $\pi U_- Q^T = R = \begin{bmatrix} R_- \mid 0 \end{bmatrix}$ with $R_- \in \mathbb{R}^{\ell\times\rho(U_-)}$. Let us define $\overline{\pi} := \operatorname{diag}(\pi, \ldots, \pi) \in \mathbb{R}^{\ell(s-1)\times\ell(s-1)}$. Then

$$\begin{bmatrix} \overline{\pi}U_{11} & \overline{\pi}U_{12}Q^T \\ \pi U_{21} & \pi U_- Q^T \end{bmatrix} = \begin{bmatrix} \overline{\pi}U_{11} & \overline{\pi}U_{12}Q^T \\ \pi U_{21} & R_- \;\; 0 \end{bmatrix}.$$

Since $\rho(U_-) < \ell$ and $R_- \in \mathbb{R}^{\ell\times\rho(U_-)}$, $\exists Q_2 \in \mathbb{R}^{\ell\times\ell} : Q_2 R_- = \begin{bmatrix} R_+ \\ 0 \end{bmatrix}$ with $R_+ \in \mathbb{R}^{\rho(U_-)\times\rho(U_-)}$. Hence, defining $\overline{Q}_2$ from Q_2 as $\overline{\pi}$ was defined from π yields

$$\begin{bmatrix} \overline{Q}_2\overline{\pi}U_{11} & \overline{Q}_2\overline{\pi}U_{12}Q^T \\ Q_2\pi U_{21} \;\; Q_2 \;\; R_- \;\; 0 \end{bmatrix} = \begin{bmatrix} \overline{Q}_2\overline{\pi}U_{11} & \overline{Q}_2\overline{\pi}U_{12}Q^T \\ Q_2\pi U_{21} & \begin{matrix} R_+ & 0 \\ 0 & 0 \end{matrix} \end{bmatrix}.$$

The left and right transformations performed to the matrix $\begin{bmatrix} U_n \, U_n^{\perp} \end{bmatrix}$ preserve the orthonormality of that matrix. Hence, $\exists Q_3 \in \mathbb{R}^{n\times n}$:

$$\left[\begin{array}{c|cc} \overline{Q}_2\overline{\pi}U_{11}Q_3^T & \overline{Q}_2\overline{\pi}U_{12}Q^T \\ \hline Q_2\pi U_{21}Q_3^T & \begin{matrix} R_+ & 0 \\ 0 & 0 \end{matrix} \end{array}\right] = \left[\begin{array}{cc|cc} \overline{Q}_2\overline{\pi}U'_{11} & 0_{(s-1)\ell\times\ell_-} & \overline{Q}_2\overline{\pi}U_{12}Q^T \\ \hline * & 0 & \begin{matrix} R_+ & 0 \end{matrix} \\ 0 & I_{\ell_-} & \begin{matrix} 0 & 0 \end{matrix} \end{array}\right].$$

Now we restore the fact that the first n-columns of this matrix span the column space of Γ_s. This yields the matrix

$$\left[\begin{array}{cc|c} U'_{11} & 0_{(s-1)\ell\times\ell_-} & U_{12}Q^T \\ \hline * & C_2 & U_- Q^T \end{array}\right] = \begin{bmatrix} U_n \mid U_n^{\perp} \end{bmatrix} \begin{bmatrix} Q_3^T & \\ & Q^T \end{bmatrix}.$$

This relationship concludes the proof of the theorem, since

$$\begin{aligned} \rho(\begin{bmatrix} C_2 \mid U_- \end{bmatrix}) &= \rho\left(\left[\pi^T Q_2^T \begin{pmatrix} 0 \\ I_{\ell_-} \end{pmatrix} \mid U_- Q^T \right]\right) \\ &= \rho\left(\left[\begin{array}{c|cc} 0 & R_+ & 0 \\ I_{\ell_-} & 0 & 0 \end{array}\right]\right) = \ell. \end{aligned}$$

□

The modes that correspond to the column space of the matrix $\left[\begin{smallmatrix} 0 \\ C_2 \end{smallmatrix}\right]$ can be considered as pure D-action. In order to see this, such modes can be represented by the following descriptor state–space model:

$$\begin{aligned} 0x'^{ac}_{k+1} &= x'^{ac}_k - B'^{ac}u_k, \\ y'_k &= C_2 x'^{ac}_k + D'u_k, \end{aligned}$$

which is indeed equivalent to

$$y'_k = \underbrace{(D' + C_2 B'^{ac})}_{D} u_k.$$

The above analysis shows that when $\rho\left(U_n^{\perp}(\ell(s-1)+1 : s\ell, :)^T\right) < \ell$, we can always reduce the descriptor part of the system such that $\rho\left(U_{n'}^{\perp}(\ell(s-1)+1 : s\ell, :)^T\right) = \ell$, where $n' < n$. Therefore, to calculate the matrix D, (23) can always be made solvable in least-squares sense.

Now we analyze the full rank condition of the underbraced term in (22). A necessary condition here is that the second matrix in this product have full column rank. When the matrix E in (3) is singular, this necessary condition is not satisfied. In that case, we first solve the matrix D from (23) and then we use this solution to transform (22) into

$$\begin{pmatrix} \Xi(:, 1:m) \\ \Xi(:, m+1:2m) \\ \vdots \\ \Xi(:, m(s-2)+1 : m(s-1)) \end{pmatrix} - \begin{bmatrix} 0_{\ell s - n \times (s-2)\ell} & (U_n^{\perp})^T \\ 0_{\ell s - n \times (s-3)\ell} & (U_n^{\perp})^T \quad 0_{\ell s - n \times \ell} \\ & \vdots \\ (U_n^{\perp})^T & 0_{\ell s - n \times (s-2)\ell} \end{bmatrix} \begin{bmatrix} 0 \\ D \\ 0 \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} 0_{\ell s - n \times (s-2)\ell} & (U_n^{\perp})^T \\ 0_{\ell s - n \times (s-3)\ell} & (U_n^{\perp})^T \quad 0_{\ell s - n \times \ell} \\ & \vdots \\ (U_n^{\perp})^T & 0_{\ell s - n \times (s-2)\ell} \end{bmatrix} \begin{bmatrix} U'_n(\ell+1 : s\ell, n_c+1 : n) & 0 \\ 0 & U'_n(1 : (s-1)\ell, 1 : n_c) \end{bmatrix}} \begin{bmatrix} B_T^{ac} \\ B_T^{c} \end{bmatrix}. \tag{24}$$

Here the second matrix in the underbraced term has full column rank when $s > n$, even for the case when the matrix E is singular.

Suppose that we can solve (22) for the matrices B_T^{ac}, B_T^{c}, and $(D + C_T^{ac} B_T^{ac})$. Then the matrix D is computed by solving the equation

$$\begin{bmatrix} U_n^{\perp}(\ell(s-1)+1 : \ell s, :)^T \\ I_\ell \end{bmatrix} D = \begin{bmatrix} \Xi(:, m(s-1)+1 : ms) \\ (D + C_T^{ac} B_T^{ac}) - C_T^{ac} B_T^{ac} \end{bmatrix}. \tag{25}$$

Therefore, in order to show the solvability of (22), we may assume that the second matrix in the underbraced term of this equation has full column rank. When this is not the case, we treat the solvability of (24). Here the second matrix in the underbraced term fulfills the full rank condition when $s > n$. Similar to the analysis of the related question for identifying causal systems within the MOESP framework, as given in [8], [9], we state the following result.

THEOREM 5. *Let there exist an even s such that for $k \geq \frac{s}{2} - 1$, $A^k \equiv 0$ and $E^k \equiv 0$, and the second matrix in the underbraced term in either* (22) *or* (24) *has full column rank. Then we can solve both equations in least-squares sense.*

Proof. With the s defined in the theorem, the matrix $\left[\begin{array}{c|c} U_n & U_n^{\perp} \end{array}\right]$ has the following structure:

$$\left[\begin{array}{c|c} U_n & U_n^{\perp} \end{array}\right] = \left[\begin{array}{cc|cc|c} \Gamma^{c}_{\frac{s}{2}-1} & 0 & 0 & 0 & * \\ \hline 0 & 0 & I_\ell & 0 & * \\ 0 & 0 & 0 & I_\ell & * \\ 0 & \Gamma^{ac}_{\frac{s}{2}-1} & 0 & 0 & * \end{array}\right],$$

where $*$ are irrelevant matrices and where it is implicitly assumed that the columns of $\Gamma^{c}_{\frac{s}{2}-1}$ and $\Gamma^{ac}_{\frac{s}{2}-1}$ are orthogonal. With this structure, the underbraced terms in (22) and (24) can be denoted as

$$\left[\begin{array}{cccccc} 0_{2\ell\times(s-2)\ell} & 0_{2\ell\times(\frac{s}{2}-1)\ell} & & I_\ell & 0 & \\ & & & 0 & I_\ell & 0_{2\ell\times(\frac{s}{2}-1)\ell} \\ 0_{2\ell\times(s-3)\ell} & 0_{2\ell\times(\frac{s}{2}-1)\ell} & I_\ell & 0 & & \\ & & 0 & I_\ell & & 0_{2\ell\times\frac{s}{2}\ell} \\ & & & \vdots & & \\ & I_\ell & 0 & & & \\ 0_{2\ell\times(\frac{s}{2}-1)\ell} & 0 & I_\ell & & 0_{2\ell\times(s-2)\ell} & 0_{2\ell\times(\frac{s}{2}-1)\ell} \\ & & & * & & \end{array}\right]$$

$$\cdot \left[\begin{array}{c|c} 0_{(\frac{s}{2}-2)\ell\times n_{ac}} & \\ 0_{2\ell\times n_{ac}} & 0_{(s-1)\ell\times n_c} \\ \Gamma^{ac}_{\frac{s}{2}-1} & \\ & \Gamma^{c}_{\frac{s}{2}-1} \\ 0_{(s-1)\ell\times n_{ac}} & 0_{2\ell\times n_c} \\ & 0_{(\frac{s}{2}-2)\ell\times n_c} \end{array}\right] = \left[\begin{array}{cc|cc|c} & & & I_\ell & \\ & & & 0 & \\ \hline & & I_\ell & 0 & \\ & & 0 & I_\ell & \\ \hline & I_\ell & 0 & & \\ & 0 & I_\ell & & \\ \hline & \vdots & 0 & & \\ & & I_\ell & & \\ \hline & \vdots & & & \\ \hline I_\ell & & & & \\ 0 & & & & \\ \hline 0 & & & & \\ I_\ell & & & & \\ \hline & * & & & \end{array}\right] \cdot \left[\begin{array}{c|c} \Gamma^{ac}_{\frac{s}{2}-1} & \\ & \Gamma^{c}_{\frac{s}{2}-1} \end{array}\right].$$

Clearly this product has full column rank. □

The above theorem shows that the set of equations (22) or (24) can be made solvable by increasing the s parameter when both the causal and anti-causal parts are asymptotically stable. This conclusion is completely similar to that drawn in the same context for pure causal systems analyzed in [8] and [9].

3.5. Summary of the algorithm (the algorithm ACC_OM (ordinary MOESP scheme for mixed anti-causal, causal systems)).

Given:

- An estimate of the underlying system order $n = n_c + n_{ac}$. The detection of the order proceeds in exactly the same way as outlined for the "causal" variants of the MOESP family of algorithms. It is based on partitioning the singular values of the matrix R_{22} in (8) into "noise" and "signal" singular

values. For a more elaborate discussion on order detection we refer to [12], [10].

- A dimension parameter s satisfying

$$s > n.$$

- The input and output data sequences

$$[u_1, u_2, \ldots, u_{N+s-1}] \quad \text{and} \quad [y_1, y_2, \ldots, y_{N+s-1}]$$

with $N >> m.s$.

Do the following:

- Step 1: Construct the Hankel matrices $U_{1,s}$ and $Y_{1,s}$ defined in (6).
- Step 2: Achieve a data compression via an RQ factorization, of which the R-factor is partitioned as in (7) of Theorem 1.
- Step 3: Compute the SVD of the matrix R_{22} as given in (8) of Theorem 1. From this SVD we can read off the column span of the extended observability matrix Γ_s.
- Step 3′: *Split the column span of* Γ_s *into causal and anit-causal parts.* This is done by the following sequence of computations.
  - Solve the set of equations (10); for example, by using the singular value decomposition we can do that in total least-squares sense.
  - Determine orthogonal matrices P_1, Q_1, such that

$$P_1 \overline{A}_T Q_1 = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \quad \text{and} \quad P_1 \overline{E}_T Q_1 = \begin{bmatrix} E_{11} & E_{12} \\ 0 & E_{22} \end{bmatrix}$$

    with $A_{11}, E_{11} \in \mathbb{R}^{n_c \times n_c}$ and $A_{22}, E_{22} \in \mathbb{R}^{n_{ac} \times n_{ac}}$, and such that $|\lambda(A_{11}E_{11}^{-1})| \leq 1$ and $|\lambda(A_{22}^{-1}E_{22})| < 1$.
  - Solve the set of Sylvestor equations [15]:

$$\begin{aligned} A_{11}R - LA_{22} &= -A_{12}, \\ E_{11}R - LE_{22} &= -E_{12}. \end{aligned}$$

  - Then P' of Theorem 2 becomes

$$P' = P_1^T \cdot \begin{bmatrix} I & L \\ 0 & I \end{bmatrix}$$

    and defines the matrix U_n' as in (14).
  - Calculate A_T, E_T from the set of equations 15–16 and read off the matrices C_T^c, C_T^{ac} as indicated in (17).
- Step 4: *Calculate the system matrices* A_T, E_T via solving (15) and (16), respectively, reading off the system matrices C_T^c, C_T^{ac} as indicated in (17), and calculating the matrices B_T^c, B_T^{ac}, and D by solving (22), (25) or (24), (23).

We remark that Step 3′ can be considered as an extra step in the ordinary MOESP algorithm that makes the scheme applicable to the identification of mixed causal, anti-causal systems. In fact, in the next section we will highlight that when squeezing this additional step in all the members of the MOESP family, we make them applicable to the class of generalized state–space systems analyzed in this paper. When the underlying system is a descriptor system, we can use the results of Theorem 3 to split

the column space of Γ_s into causal and anti-causal parts without (partially) computing the Kronecker canonical form. The matrix P in this theorem can simply be computed by using a QR factorization with column pivoting of the matrix γ_1 defined in the proof of Theorem 3.

4. Tackling various identification problems.

4.1. Statistical consistency of the ACC_OM algorithm. The reason that the main part of the algorithmic structure of the ordinary MOESP scheme is preserved by the ACC_OM algorithm, namely Steps 1–3, is that the data equation (6) has exactly the same structure as that obtained for pure causal systems. More precisely, Steps 1–3 remain valid when the matrices Γ_s and H_s in (6) are arbitrary. The only assumption made on these matrices is that $\rho(\Gamma_s) = n$.

The same is true in the analysis of the statistical consistency of Steps 1–3 of the ordinary MOESP scheme; see [10] for causal systems. Therefore, based on Theorem 1 of [10], we can state that under the assumptions in Theorem 1 of this paper and for the case where the output y_k in (3) is perturbed by *zero-mean white noise errors*,

$$\exists T \in \mathbb{R}^{n \times n}, T \text{ nonsingular} : \lim_{N \to \infty} U_{n,N} = \Gamma_s T,$$

$$\lim_{N \to \infty} (U_n^{\perp})_N^T R_{21}^N (R_{11}^N)^{-1} - (U_n^{\perp})_N^T H_s = 0,$$

where the additional index N of the different quantities now represents the dependency of these quantities on the number of columns of the Hankel matrices $U_{1,s}, Y_{1,s}$ processed by the algorithm ACC_OM.

4.2. Extension of the PI scheme. In this subsection, we consider the case where the output y_k is perturbed by an additive error ν_k. The latter is assumed to be zero-mean, having arbitrary coloring and possibly containing a deterministic term which is independent from the input u_k. For this perturbation, the data equation (6) transforms into

$$Y_{j,s} = \Gamma_s \begin{bmatrix} X_j^c \\ X_{j+s-1}^{ac} \end{bmatrix} + H_s U_{j,s} + V_{j,s},$$

where $V_{j,s}$ is a Hankel matrix constructed from the perturbations ν_k. Again, because of the fact that the structure of the matrices Γ_s and H_s is of no importance in Steps 1–3 of the PI scheme, derived in [10], the latter three steps will produce a consistent estimate of the column space of Γ_s and will yield a consistent estimate of the equation similar to (21), denoted as

$$\lim_{N \to \infty} \Xi_N - (U_n^{\perp})_N^T H_s = 0$$

when the underlying system is mixed causal, anti-causal. As a consequence, we can determine consistent estimates of the system matrices in (4) when using the calculated column space of Γ_s and the above equation in Step 3′. Therefore, the PI scheme for mixed causal, anti-causal systems, denoted by ACC_PI, is the same as the original PI scheme of [10] with Step 4 replaced by Step 3′ and Step 4 of the ACC_OM algorithm.

4.3. Extension of the PO scheme. In this subsection, we consider the case where the output y_k is perturbed by filtered white noise sequences w_k and v_k as denoted in (5). For this case, the data equation (6) becomes

$$Y_{j,s} = \Gamma_s \begin{bmatrix} X_j^c \\ X_{j+s-1}^{ac} \end{bmatrix} + H_s U_{j,s} + H_s^w W_{j,s} + V_{j,s},$$

where H_s^w is equal to H_s with D, B^{ac}, and B^c replaced by 0, K^{ac}, and K^c, respectively. Furthermore, the matrices $W_{j,s}$ and $V_{j,s}$ are Hankel matrices constructed from w_k and v_k, respectively.

Again, because the structure of the data equation is preserved, we only need to substitute Step 4 of the original PO scheme by Steps 3′ and 4 of the ACC_OM algorithm in order to make this variant of the MOESP family applicable to the class of systems described by (5). The latter scheme will be indicated by the ACC_PO scheme.

4.4. Incorporating a bilinear transformation of the shift operator in the ACC_PI scheme. A last variant of the MOESP family of algorithms for LTI systems is the incorporation of a bilinear transformation of the shift operator in the PI scheme as discussed in [13]. In the context of causal systems, it is shown in [13] that the latter transformation drastically improves the accuracy of the estimated state–space matrices when the eigenvalues of the transition matrix of the system to be identified are in the vicinity of point 1 in the complex plane. Extending this use of bilinear transformation of the shift operator to the class of mixed causal, anti-causal systems will lead to improved estimates when the eigenvalues of A and E are both close to 1.

In terms of complex variables the following bilinear transformation is considered:

$$w = \frac{z-a}{1-az} \quad \Leftrightarrow \quad z = \frac{w+a}{1+aw}, \quad -1 < a < 1, \quad w, z \in \mathbb{C}.$$

This transformation was also considered in [17] in the context of using Laguerre series in parametric model identification.

In order to compactly denote the relationships between the state–space models related to the complex z parameter and those related to the w parameter we introduce the shift operator Z. Let the entry x_0 at time instant 0 of the double infinite time sequence $x = \begin{bmatrix} \cdots & x_{-1} & x_0 & x_1 & \cdots \end{bmatrix}$ be indicated by a square box. Then the operation of Z is represented as

$$\begin{bmatrix} \cdots & x_{-1} & \boxed{x_0} & x_1 & \cdots \end{bmatrix} Z = \begin{bmatrix} \cdots & x_0 & \boxed{x_1} & x_2 & \cdots \end{bmatrix}. \tag{26}$$

This operator defines the operator W and vice versa as

$$W = (Z - aI)(I - aZ)^{-1} \quad \Leftrightarrow \quad Z = (W + aI)(I + aW)^{-1}. \tag{27}$$

In this subsection, we demonstrate that incorporating this transformation in mixed causal, anti-causal state–space models of the type (3) again preserves the structure of the data equation compared to that obtained for causal systems [13].

For the sake of brevity, we only highlight that the structure of the data equation is preserved when the latter is given in operator form. Let x^c, x^{ac}, u, and y denote ℓ_2-sequences with components in $\mathbb{R}^{n_c}$, $\mathbb{R}^{n_{ac}}$, $\mathbb{R}^m$, and $\mathbb{R}^\ell$, respectively. Then the definition of the Z-operator in (26) allows one to write the system (3) in operator format as

$$\begin{aligned} x^c Z &= Ax^c + B^c u, \\ x^{ac} &= Ex^{ac} Z + B^{ac} u, \\ y &= C^c x + (C^{ac} E) x^{ac} Z + (C^{ac} B^{ac} + D) u, \end{aligned} \tag{28}$$

where the matrices A, B, etc. are finite size matrices defined in (3). Substituting the expression for the Z-operator in terms of the W-operator of (27) transforms this state–space model into

$$
\begin{aligned}
x^c W &= F^c x^c + G^c u(I + aW), \\
x^{ac} &= F^{ac} x^{ac} W + G^{ac} u(I + aW), \\
y(I + aW) &= H^c x + (H^{ac} F^{ac}) x^{ac} W + (H^{ac} G^{ac} + J) u(I + aW),
\end{aligned} \tag{29}
$$

where

$$
\begin{array}{lllllll}
F^c & = & (I - Aa)^{-1}(A - aI), & F^{ac} & = & (I - Ea)^{-1}(E - aI), \\
G^c & = & (I - Aa)^{-1} B^c, & G^{ac} & = & (I - Ea)^{-1} B^{ac}, \\
H^c & = & (1 - a^2) C^c (I - Aa)^{-1}, & (H^{ac} F^{ac}) & = & (1 - a^2)\,(C^{ac} E)\,(I - Ea)^{-1},
\end{array}
$$

$$
(J + H^{ac} G^{ac}) = (C^{ac} B^{ac} + D) + a C^c (I - Aa)^{-1} B^c + a\,(C^{ac} E)\,(I - Ea)^{-1} B^{ac}.
$$

The inverse of the matrices $(I - Aa)$ and $(I - Ea)$ exists under the conditions assumed in this paper, that is, $|\lambda(A)| \leq 1$ and $|\lambda(E)| < 1$ and $|a| < 1$. For a formal proof of this assertion we refer to Theorem 1 of [13]. Conversely, when the state–space model in (29) is given, we can derive the system matrices in (28) as follows:

$$
\begin{array}{lllllll}
A & = & (I + F^c a)^{-1}(F^c + aI), & E & = & (I + Ea)^{-1}(E + aI), \\
B^c & = & (1 - a^2)(I + F^c a)^{-1} G^c, & B^{ac} & = & (1 - a^2)(I + F^{ac} a)^{-1} G^{ac}, \\
C^c & = & H^c (I + F^c a)^{-1}, & (C^{ac} E) & = & (H^{ac} F^{ac})\,(I + F^{ac} a)^{-1},
\end{array}
$$

$$
(C^{ac} B^{ac} + D) = (J + H^{ac} G^{ac}) - a H^c (I + F^c a)^{-1} G^c - a\,(H^{ac} F^{ac})\,(I + F^{ac} a)^{-1} G^{ac},
$$

where the inverses again exist under the above stated conditions; see [13].

For the state–space model (29), the data equation in operator form is as follows:

$$
\left[\begin{array}{c} y(I + aW) \\ y(I + aW)W \\ \cdots \\ y(I + aW)W^{s-1} \end{array}\right]
= \left[\begin{array}{c|c} H^c & H^{ac}(F^{ac})^{s-1} \\ H^c F^c & H^{ac}(F^{ac})^{s-2} \\ \vdots & \vdots \\ H^c (F^c)^{s-1} & H^{ac} \end{array}\right]
\left[\begin{array}{c} x^c \\ x^{ac} W^{s-1} \end{array}\right]
$$

$$
+ \left[\begin{array}{cccc} J + H^{ac} G^{ac} & H^{ac}(F^{ac}) G^{ac} & \cdots & H^{ac}(F^{ac})^{s-2} G^{ac} \quad 0 \\ H^c G^c & J + H^{ac} G^{ac} & \cdots & H^{ac}(F^{ac})^{s-3} G^{ac} \quad 0 \\ \vdots & & \ddots & \vdots \\ H^c (F^c)^{s-2} G^c & H^c (F^c)^{s-3} G^c & \cdots & H^c G^c \quad J \end{array}\right]
\left[\begin{array}{c} u(I + aW) \\ u(I + aW)W \\ \cdots \\ u(I + aW)W^{s-1} \end{array}\right].
$$

Here the operations such as $y(I + aW)W^k$ for $s - 1 \geq k \geq 0$ represent an anti-causal filtering operation of the output data sequence. The algorithmic details for generating these sequences when only a finite number of output (and input) samples are available are discussed in [13].

When denoting the data equation (6) in operator format, namely,

$$
\left[\begin{array}{c} y \\ yZ \\ \cdots \\ yZ^{s-1} \end{array}\right]
= \Gamma_s \left[\begin{array}{c} x^c \\ x^{ac} Z^{s-1} \end{array}\right] + H_s \left[\begin{array}{c} u \\ uZ \\ \cdots \\ uZ^{s-1} \end{array}\right],
$$

we clearly observe that the structure of the data equation is preserved under a bilinear transformation of the shift operator. It is again this property that leads to a straightforward extension of the PI_BTZ (the PI scheme using a bilinear transformation of the Z-operator [13]) scheme toward mixed causal, anti-causal systems.

5. Summarizing remarks. In this paper, a number of subspace algorithms are presented that allow one to solve a broad class of identification problems for mixed causal, anti-causal systems that have a regular pencil.

The algorithms are complete in the sense that it becomes tempting to apply them to realistic problems.

Acknowledgement. The author wants to acknowledge Mrs. X. Yu for fruitful discussions related to §§3.1 and 3.2 of this paper.

REFERENCES

[1] D. G. LUENBERGER, *Dynamic equations in descriptor form*, IEEE Trans. Automat. Control, AC–22 (1977), pp. 312–321.

[2] G. C. VERGHESE, B. C. LÉVY, AND T. KAILATH, *A generalized state space for singular systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 811–831.

[3] M. A. CHRISTODOULOU AND B. G. MERTZIOS, *Realization of singular systems via Markov parameters*, Internat. J. Control, 42 (1985), pp. 1433–1441.

[4] M. MOONEN, B. D. MOOR, J. RAMOS, AND S. TAN, *A subspace identification algorithm for descriptor systems*, Systems Control Lett., 19 (1992), pp. 47–52.

[5] F. L. LEWIS, *A survey of linear singular systems*, Circuits Systems Signal Process, 5 (1986), pp. 3–36.

[6] J. D. APLEVICH, *Implicit Linear Systems Lecture Notes in Control and Information Sciences*, M. Thoma and A. Wyner, eds., Springer–Verlag, Berlin, New York, 1991

[7] J. B. MACNEIL, R. E. KEARNEY, AND I. W. HUNTER, *Identification of time-varying biological systems from ensemble data*, IEEE Trans. on Biomedical Engineering, 39 (1992), pp. 1213–1225.

[8] M. VERHAEGEN AND P. DEWILDE, *Subspace model identification. Part* I: *The output-error state space model identification class of algorithms*, Internat. J. Control, 56 (1992), pp. 1187–1210.

[9] ———, *Subspace model identification. Part* II: *Analysis of the elementary output-error state space model identification algorithm*, Internat. J. Control, 56 (1992), pp. 1211–1241.

[10] M. VERHAEGEN, *Subspace model identification. Part* III: *Analysis of the ordinary output-error state space model identification algorithm*, Internat. J. Control, 58 (1993), pp. 555–586.

[11] ———, *Application of a subspace model identification technique to identify LTI systems operating in closed-loop*, Automatica, 29 (1993), pp. 1027–1040.

[12] ———, *Identification of the deterministic part of MIMO state space models given in innovation form from input–output data*, Automatica, Special Issue on Statistical Signal Processing and Control, 30 (1994), pp. 61–74.

[13] M. VERHAEGEN, D. WESTWICK, AND R. E. KEARNEY, *The use of a bilinear transformation of the shift operator in subspace model identification*, IEEE Trans. Automat. Control, 40 (1995), pp. 1422–1428.

[14] W. S. GRAY, E. I. VERRIEST, AND F. L. LEWIS, *A Hankel approach to singular system realization theory*, Proc. 29th CDC, Honolulu, Hawaii, 1990, pp. 73–78.

[15] B. KÅGSTRÖM AND L. WESTIN, *Generalized Schur methods with condition estimators for solving generalized Sylvestor equations*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 745–751.

[16] L. DAI, *Singular Control Systems, Lecture Notes in Control and Information Sciences*, Springer–Verlag, Berlin, New York, 1989.

[17] B. WAHLBERG, *System identification using Laguerre models*, IEEE Trans. on Automat. Control, 36 (1991), pp. 551–562.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 2, pp. 348–354, April 1996

007

IS THE POLAR DECOMPOSITION FINITELY COMPUTABLE?*

ALAN GEORGE† AND Kh. IKRAMOV‡

Abstract. The polar decomposition of a square matrix is a major step toward the singular value decomposition, and is an important device in its own right. Singular values of a matrix A, being the eigenvalues for a matrix closely related to A, generally cannot be computed by a finite process if only arithmetic operations and radicals are allowed. However, this consideration alone does not prove that the polar decomposition is not finitely computable.

The problem of finite computability of the polar decomposition is not settled in this paper, but we do show it to be equivalent to the following simpler-looking problem. Suppose f is a real polynomial of degree $n > 4$, and all the roots of f are distinct positive numbers. Denote by g a polynomial of the same degree whose zeros are the positive square roots of the zeros of f. Can this polynomial g always be computed finitely for a given polynomial f? In the Appendix we discuss one nontrivial situation where the polar decomposition can indeed be computed finitely.

Key words. polar decomposition, finite computation

AMS subject classification. 65F10

1. Introduction.

In this paper, a computational problem $\mathcal{P}$ over $\mathbb{R}$ or $\mathbb{C}$ is called finitely solvable, or solvable by radicals, if the solution(s) of $\mathcal{P}$ can be obtained by a finite algorithm using only operations from the following list:

(a) arithmetic operations $+, -, \times, \div$;

(b) extraction of radicals of arbitrary integer degree;

(c) comparison with zero.

Checking to see if a general real or complex expression is identically zero is known to be a hard problem (even an algorithmically intractable one!). Nevertheless, we include operation (c) in our basic set, because practically all computational algorithms rely on such comparisons.

The question we address is whether the problem of computing the polar decomposition of a general real or complex matrix A is solvable by radicals. For background material on the polar decomposition, see Gantmacher [2]. To be more definite, by the polar decomposition of the square matrix A we mean its factorization of the form

$$A = HV, \quad H = H^* \geq 0, \quad VV^* = I. \tag{1}$$

We remind the reader of the existence of another type of polar decomposition, with the order of Hermitian and unitary factors reversed. The considerations in our paper are applicable to this type of decomposition as well.

We believe that the answer to the question posed in the title is in general "no." Indeed, the polar decomposition of A is very closely related to its singular value decomposition. Moreover, the singular values of A, being the eigenvalues for the matrix H in (1), cannot be, in general, computed finitely for $n > 4$.

This argument, although persuasive, is not a rigorous proof, and we do not settle the question in this paper. Our contribution is to present four problems, including the one in the title, and to show that they are equivalent. This may provide others with

* Received by the editors September 2, 1994; accepted for publication (in revised form) by G. Cybenko May 26, 1995. This research was supported by Natural Sciences and Engineering Research Council of Canada grant OGP 000811.

† Department of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada.

‡ Moscow State University, Faculty of Computational Mathematics and Cybernetics, Moscow 119899, Russia.

new avenues through which to settle the question of if, or under what restrictions, the polar decomposition of a general matrix can be finitely computed. Of course this work is largely of theoretical interest; there are a number of powerful and robust (nonfinite) methods for computing the polar decomposition. A nice description can be found in Higham [3].

In the Appendix we discuss one nontrivial situation where the polar decomposition can indeed be computed finitely.

2. The polar decomposition and the extraction of the square root of a polynomial. Below we state the four problems (P1)–(P4), including the one contained in the title of the paper. We then show that these problems are equivalent. We do that by proving the implications

$$P_2 \Longrightarrow P_1, \quad P_3 \Longrightarrow P_2, \quad P_4 \Longrightarrow P_3, \quad P_1 \Longrightarrow P_4.$$

The relation $P_i \Longrightarrow P_j$ means that the solvability by radicals of the problem P_i implies that the problem P_j is solvable by radicals as well.

(P1) For a given matrix $A \in \mathbb{C}^{n,n}$, find a polar decomposition of A;

(P2) for a given nonsingular matrix $A \in \mathbb{R}^{n,n}$, find the polar decomposition of A;

(P3) for a given symmetric positive definite matrix $A \in \mathbb{R}^{n,n}$, find the unique positive definite square root of A;

(P4) for a given real polynomial f of degree n with distinct positive zeros $\lambda_1, \ldots, \lambda_n$, find the polynomial g of the same degree with the positive zeros $\lambda_1^{1/2}, \ldots, \lambda_n^{1/2}$,

$$P_2 \Longrightarrow P_1.$$

Suppose A is a given complex $n \times n$ matrix. Applying to A the well-known bidiagonalization procedure (for a description of the procedure in the real case see [5, Chap. 18, §3]), we obtain

$$B = PAQ = \begin{bmatrix} q_1 & e_2 & & \\ & q_2 & \ddots & \\ & & \ddots & e_n \\ & & & q_n \end{bmatrix}. \tag{2}$$

The matrices P and Q here are computed as finite products of elementary Hermitian matrices (Householder matrices, in the real case) and, possibly, diagonal unitary matrices. We can always make the numbers q_i and e_i in (2) real and nonnegative.

Now if B is nonsingular and $B = H_B V_B$ is the polar decomposition of B, then

$$A = P^* B Q = (P^* H_B P)(P^* V_B Q) = HV$$

is the polar decomposition of A. Therefore, from the finite solvability of P_2 it follows that P_1 is also finitely solvable.

If B is singular, i.e., some numbers q_i in (2) are zero, then applying to B a finite number of rotations, as described in [5, p. 111], we can decompose B into a direct sum of nonsingular bidiagonal matrices and a zero matrix:

$$C = T_k \ldots T_1 B R_1 \ldots R_\ell = \begin{bmatrix} C_1 & & & \\ & \ddots & & \\ & & C_m & \\ & & & 0 \end{bmatrix}. \tag{3}$$

The polar decomposition of C can now be obtained by computing the polar decompositions of each of the blocks $C_i (1 \leq i \leq m)$. Then, retracing transformations (3) and (2), we finally get the polar decomposition of A:

$$P_3 \Longrightarrow P_2.$$

Suppose A is a given real nonsingular matrix. Then, applying to AA^* a finite algorithm which solves the problem P_3, we obtain the Hermitian factor H in (1). This matrix H is also nonsingular, and the unitary factor V for the polar decomposition of A can now be computed as

$$V = H^{-1}A.$$

The inversion of a nonsingular matrix is, of course, a finite (even rational!) procedure:

$$P_4 \Longrightarrow P_3.$$

Suppose A is a given real symmetric positive definite $n \times n$ matrix. Denote by $f(\lambda)$ the characteristic polynomial of A. This polynomial can obviously be computed finitely. Then, dividing f by the greatest common divisor of f and its derivative f', we obtain the polynomial $\hat{f}$ with distinct positive zeros, which is the minimal polynomial of A. Next, applying to $\hat{f}$ a finite algorithm which solves the problem P_4, we obtain the polynomial g whose positive zeros are the square roots of the distinct eigenvalues $\lambda_1, \ldots, \lambda_m$ of A. This polynomial is therefore the minimal polynomial of the matrix X, the square root of A.

If we let

$$g(\lambda) = \lambda^m + a_1\lambda^{m-1} + \cdots + a_{m-1}\lambda + a_m,$$

then

$$g(X) = X^m + a_1X^{m-1} + \cdots + a_{m-1}X + a_mI_n = 0. \tag{4}$$

Replacing X^2 by A, we rewrite (4) as

$$\varphi(A) + \psi(A)X = 0, \tag{5}$$

where φ and ψ are (known) polynomials of degree $\leq \lfloor m/2 \rfloor$.

If $\psi(A)$ in (5) is nonsingular, then we immediately find X as

$$X = -[\psi(A)]^{-1}\, \phi(A).$$

Suppose now $Y = \psi(A)$ is singular. On the other hand, Y cannot be the zero matrix. Otherwise, the polynomial of ψ of degree $\leq \lfloor m/2 \rfloor$ annihilates A, which is impossible for a matrix with m distinct eigenvalues.

The image $\mathcal{L}$ and the null space $\mathcal{N}$ of the nonzero matrix Y are two nontrivial invariant subspaces of A. Moreover,

$$\mathbb{R}^n = \mathcal{L} \oplus \mathcal{N}.$$

One can find orthonormal bases of $\mathcal{L}$ and $\mathcal{N}$ by computing the QR decomposition of Y. In the orthonormal basis of $\mathbb{R}^n$, comprised of the orthonormal bases of $\mathcal{L}$ and $\mathcal{N}$, the matrix A decomposes into the direct sum of two matrices of lower order:

$$Q^{-1}AQ = \begin{bmatrix} A_{\mathcal{L}} & 0 \\ 0 & A_{\mathcal{N}} \end{bmatrix}.$$

The extraction of the square root of A is now reduced to the same problem with the smaller matrices $A_{\mathcal{L}}$ and $A_{\mathcal{N}}$. Applying the reasoning above to these matrices, we either find the matrices $A_{\mathcal{L}}^{1/2}$ and/or $A_{\mathcal{N}}^{1/2}$, or reduce the problem of extracting the square root even further. Continuing in this way, we can construct $A^{1/2}$.

Remark. If $\psi(A)$ in (5) is singular then $\varphi(A)$ is singular as well. Moreover, if $x \in \ker \psi(A)$ then $x \in \ker \varphi(A)$ as well. This means an eigenvalue λ_0 of A exists such that

$$\varphi(\lambda_0) = \psi(\lambda_0) = 0.$$

Because

$$g(\lambda) = \varphi(\lambda) + \lambda\varphi(\lambda),$$

we have $g(\lambda_0) = 0$. Therefore,

$$\lambda_0^2 = \lambda_i$$

for some eigenvalue λ_i of A. We conclude that the matrix $\psi(A)$ in (5) can only be singular if the spectrum of A contains a pair of the form (λ_0, λ_0^2). We have

$$P_1 \Longrightarrow P_4.$$

Suppose f is a given real polynomial of degree n with distinct positive zeros. Finite procedures have recently been devised [1, 7] which allow one to construct a symmetric matrix S_f with the characteristic polynomial f prescribed. If f is a real polynomial with all the roots real then S_f is real as well. For our polynomial f, the real symmetric matrix S_f should be positive definite. Hence, the Cholesky procedure can be applied to S_f giving

$$S_f = LL^T$$

with the lower triangular matrix L. Now, using for L a finite algorithm which solves the problem P_1, we obtain, in particular, the matrix

$$H = (LL^T)^{1/2} = S_f^{1/2}.$$

The characteristic polynomial of H is the polynomial g desired.

3. Appendix. Since problems (P1)–(P4) are equivalent, we can consider the extraction of the positive definite (p.d.) square root of a real symmetric p.d. matrix A, instead of the polar decomposition.

Suppose B is an unreduced tridiagonal matrix:

$$B = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \alpha_{n-1} & \beta_{n-1} \\ & & & \beta_{n-1} & \alpha_n \end{bmatrix}. \tag{6}$$

We may assume the numbers $\beta_1, \ldots, \beta_{n-1}$ to be positive.

It is obvious that the matrix $A = B^2$ is pentadiagonal:

$$A = \begin{bmatrix} a_1 & b_1 & c_1 & & & & \\ b_1 & a_2 & b_2 & \ddots & & & \\ c_1 & b_2 & a_3 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & a_{n-2} & b_{n-2} & c_{n-2} \\ & & & \ddots & b_{n-2} & a_{n-1} & b_{n-1} \\ & & & & c_{n-2} & b_{n-1} & a_n \end{bmatrix} \tag{7}$$

with the elements $c_1, \ldots, c_{n-2}$ positive.

Now, suppose we are given a real symmetric p.d. matrix A of the form (7), where the elements $c_1, \ldots, c_{n-2}$ are all positive. If it is known that the p.d. root B of A is a tridiagonal matrix, then B can be computed finitely. Indeed, equating entries on the main diagonal and two superdiagonals in $B^2 = A$, we obtain

$$\alpha_1^2 + \beta_1^2 = a_1, \tag{8}$$

$$\beta_1^2 + \alpha_2^2 + \beta_2^2 = a_2, \tag{9}$$

$$\beta_{i-1}^2 + \alpha_i^2 + \beta_i^2 = a_i, \quad i = 3, \ldots, n-1, \tag{10}$$

$$\beta_{n-1}^2 + \alpha_n^2 = a_n, \tag{11}$$

$$(\alpha_i + \alpha_{i+1})\beta_i = b_i, \quad i = 1, \ldots, n-1, \tag{12}$$

and

$$\beta_i \beta_{i+1} = c_i, \quad i = 1, \ldots, n-2. \tag{13}$$

Equation (13) shows that B is unreduced; i.e., all the β's are nonzero. Now, we deduce from these equations

$$\beta_2 = c_1 \frac{1}{\beta_1}, \tag{14}$$

$$\beta_3 = c_2 \frac{1}{\beta_2} = \frac{c_2}{c_1} \beta_1, \tag{15}$$

$$\beta_4 = c_3 \frac{1}{\beta_3} = \frac{c_1 c_3}{c_2} \frac{1}{\beta_1}, \tag{16}$$

and so on. In general, letting $\beta_1 \equiv \beta$, we have

$$\beta_{2k} = d_{2k} \beta, \quad k = 1, \ldots, \lfloor (n-1)/2 \rfloor \tag{17}$$

and

$$\beta_{2k+1} = d_{2k+1}\frac{1}{\beta}, \quad k = 1, \ldots, \lfloor (n-2)/2 \rfloor. \tag{18}$$

The multipliers d_i can easily be expressed as rational functions of the c's. Rewriting equation (12) in the form

$$\alpha_i + \alpha_{i+1} = \frac{b_i}{\beta_i}, \quad i = 1, \ldots, n-1, \tag{19}$$

we see that $\alpha_2, \ldots, \alpha_n$ can be immediately (and rationally!) computed as long as $\alpha_1 \equiv \alpha$ and β are found. In particular,

$$\alpha_2 = \frac{b_1}{\beta} - \alpha. \tag{20}$$

Using (14), (20), (8), and (9), we have

$$\alpha^2 + \beta^2 = a_1, \tag{21}$$

$$\beta^2 + \left(\frac{b_1}{\beta} - \alpha\right)^2 + \frac{c_1^2}{\beta^2} = a_2, \tag{22}$$

or

$$\beta^4 + b_1^2 - 2b_1\alpha\beta + \alpha^2\beta^2 + c_1^2 = a_2\beta^2. \tag{23}$$

Because $\alpha^2 = a_1 - \beta^2$, we get from (23) the explicit expression for α as a function of β:

$$\alpha = \frac{s\beta^2 + t}{2b_1\beta}, \tag{24}$$

where

$$s = a_1 - a_2, \quad t = b_1^2 + c_1^2. \tag{25}$$

Substituting (24) into (21), we finally obtain

$$\frac{(s\beta^2 + t)^2}{4b_1^2\beta^2} + \beta^2 = a_1, \tag{26}$$

which is the biquadratic equation in β. By assumption, the desired matrix B exists; hence, equation (26) must have real solutions. For any of these solutions, we first find α from (24), then $\beta_2, \ldots, \beta_{n-1}$ from (17)–(18), and $\alpha_2, \ldots, \alpha_n$ from (19). We then check if equations (10) and (11) are satisfied with these values for α's and β's. Completing these calculations, we end up with exactly one set of correct values for the unknown α's and β's.

Unfortunately, the finite computability of B just proved does not help much in the general case. It is true that any symmetric or Hermitian (not necessarily positive definite) matrix can be finitely reduced to a pentadiagonal form. The band Lanczos algorithm by Ruhe (see [6, p. 286]) can be employed for this purpose with (almost) any choice of the initial orthonormal vectors q_1, q_2. The problem is that most

pentadiagonal forms of a given p.d. matrix A do not admit tridiagonal square roots. To obtain the proper pentadiagonal form for A, we must find some way to assure that the chosen vector q_2 belongs, for a given q_1, to the two-dimensional Krylov subspace

$$\mathcal{K}_2(B, q_1) = \text{span}\{q_1, Bq_1\},$$

B being the p.d. square root of A. This is no simpler than our original problem of finding B.

Acknowledgment and some additional observations. We are grateful to one of the referees for raising some interesting points and questions that warrant comment here. First, the referee points out, following [1], that to find $f(A)$ it is sufficient to compute the polynomial that takes on the values $f(\lambda)$ at the eigenvalues of A. This condition appears to be closely related to our condition (P4), but we have been unable to establish whether there is such a relation, or whether it is possible to compute one polynomial from the other via a finite computation.

Additionally, the referee notes that if it is possible to determine the largest and smallest singular value of a matrix, then its polar decomposition is finitely computable via an optimally scaled Newton iteration. (The method can be found in [3] and the finite convergence of the method is provided in [4].) We note that singular symmetric stochastic matrices fall nicely into this category, having extreme singular values of zero and one.

REFERENCES

[1] M. FIEDLER, *Expressing a polynomial as the characteristic polynomial of a symmetric matrix*, Linear Algebra Appl., 141 (1990), pp. 265–270.
[2] F. R. GANTMACHER, *The Theory of Matrices*, Vol. 1, Chelsea, New York, 1974.
[3] N. J. HIGHAM, *Computing the polar decomposition—with applications*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1160–1174.
[4] C. KENNEY AND A. J. LAUB, *On scaling Newton's method for polar decomposition and the matrix sign function*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 688–706.
[5] C. L. LAWSON AND R. J. HANSON, *Solving Least Squares Problems*, Prentice–Hall, Englewood Cliffs, NJ, 1974.
[6] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[7] G. SCHMEISER, *A real symmetric tridiagonal matrix with a given characteristic polynomial*, Linear Algebra Appl., 193 (1993), pp. 11–18.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 2, pp. 355–382, April 1996

008

ON THE PROPERTIES OF A RELATIVE ENTROPY FUNCTIONAL*

J.-P. LE CADRE†

Abstract. The identification of the noise correlations (e.g., between the sensors of an array) is an important problem. It is also ill posed unless some additional conditions are verified. Here, these supplementary conditions are reduced to a low-rank hypothesis and to the knowledge (e.g., an upper bound of its length) of the general structure of the noise correlations. By introducing an original functional (named relative entropy functional), we develop a new approach for solving the above problem. In particular, it is shown that this functional inherits from its definition interesting and useful properties (such as location of the extrema, concavity, etc.). These properties are shown using elementary linear algebra.

Key words. entropy, noise, optimisation, eigenvalues

AMS subject classifications. 94A17, 94A12

1. Introduction and problem statement.

Usually, the signal received on an array of sensors is composed of source and noise parts. The aim of array processing is to estimate source parameters from the array measurements [1]. However, in numerous practical situations, especially in the array processing context, the noise parameters are unknown.

Most of the practical array processing methods are based upon the properties of the covariance matrices (CM) of the various signals impinging on the array. This is particularly true for high-resolution methods for which source and noise parts play symmetric roles [2–7]. For readers unfamiliar with these methods, we note that they are rather similar, in spirit, to principal component analysis methods [8–10].

At a given frequency (after discrete Fourier transform, for example) the problem of separation in source and noise parts is reduced to the following matrix equation:

$$R = S + B$$

$$\text{(1)} \qquad \text{with :} \quad \left| \begin{array}{lcl} R & : & \text{sensor outputs } CM, \\ S & : & \text{source } CM, \\ B & : & \text{noise } CM, \\ R, S, B & & q \times q \text{ matrices.} \end{array} \right.$$

The matrix R is assumed to be known (it is actually estimated from the sensor outputs). The problem we deal with can be stated as follows:

How can we obtain an "estimate" of B from R?

The problem stated above is ill posed and is meaningless without the following additional hypotheses:

$$\text{(2)} \qquad \left| \begin{array}{lcl} H_1 & : & S \text{ is positive semidefinite and rank deficient,} \\ H_2 & : & \text{the general structure of } B \text{ is known,} \\ & & (B \text{ is positive definite}). \end{array} \right.$$

The above hypotheses are generally accepted in the array processing literature [1–7] even if H_2 is frequently replaced by a drastically simpler hypothesis, say, H_2'.

* Received by the editors July 6, 1993; accepted for publication (in revised form) by F. T. Luk May 30, 1995.
† IRISA/CNRS, Campus de Beaulieu, 35042 Rennes Cedex, France.

The matrix B is known except for a scalar multiplier λ (λ is the noise level). The hypothesis H_2 is thus far less restrictive than H_2'. The general structure of B may be simply a banded Toeplitz structure (with the positivity hypothesis) or given by the covariance structure of a moving average (MA) spatial process [11].

The hypothesis H_1 is also quite acceptable since the rank deficiency hypothesis amounts to assuming that the number of sources is strictly inferior to the number of sensors. This assumption is instrumental for high resolution methods.

After the noise matrix B is estimated, the source parameters (defining S) can be estimated [1, 2, 7]. We want to stress that, for this approach, the parameters defining B are estimated independently of the source ones, using only the observation (i.e., the matrix R). For that purpose, an original approach will be derived. It relies on the "separating" properties of a relative entropy functional (REF). Roughly speaking, this functional allows us to "extract" the smooth component (i.e., the noise) of the observations. Another approach consists of using an approximated likelihood functional. This functional involves only the eigenvalues of a whitened matrix. A complete description of this approach is given in [12]. This method presents some (hidden) similarities with the REF method since it too relies on a (hidden) barrier functional. However, it is much more classical in principle and does not present the same possibilities.

The optimal methods [13] (for the statistical meaning) consist in simultaneously estimating the source and noise parameters. These approaches are rather direct although they may involve rather intricate derivations. However, the main criticism comes from the absence of any convergence property for the iterative algorithms that optimize the related functionals. The practical interests of such methods may thus be greatly reduced despite their (theoretical) optimality.

The method that will be presented obeys the following general scheme: we define a barrier functional forbidding the description of sources by the noise model. We shall carefully study the estimation of the noise model (or equivalently of B) as well as iterative optimization of the functional (gradient-like procedure). We stress that this optimization is defined *only with respect to noise parameters*, which constitutes the major novelty of our approach. We recall that the present paper deals with the exact properties of the functional and, thus, that *statistical considerations are not in the paper's scope.*

Actually, the barrier property of the REF appears to be instrumental since it is a means to create a singularity at the boundary of the feasible region. This study can thus be included in the much more general context of barrier methods [14]. According to [14], barrier methods fell from favor during the 1970s partly because of inherent ill-conditioning in the Hessian matrix. We shall see that the proposed method does not suffer from this drawback and enjoys interesting properties.

We use the following notation throughout the text of this paper:

- matrices are represented by capital letters (e.g., R, S, B, U_i, Z_i); all the matrices are $q \times q$ except the matrices Z_i (5) and N,
- vectors are represented by capital bold letters (e.g., $\mathbf{X}, \mathbf{B}, \mathbf{V}, \mathbf{W}$),
- scalars are denoted by small letters (e.g., b_i, l), eigenvalues by small Greek letters (e.g., λ_i),
- R generally represents the observation (or a resume), S the source part, and B the noise part (noise parameters β_i or b_i),
- the symbols det and *tr* denote, respectively, the transpose and the trace (of a $q \times q$ matrix),

- diag denotes a diagonal matrix,
- A^t and A^* denote, respectively, the transpose and the hermitian adjoint of A,
- Id stands for the identity matrix,
- $R(k)$ denotes the spatial density (31) of the observation, and
- $\mathrm{Re}(z)$ denotes the real part of z, $\bar{z}$ the complex conjugate of z.

2. The relative entropy approach. This approach deals with a functional depending only on the matrices R and B. In what follows, this functional will be named the relative entropy functional (REF) and will play the central role in solving the problem (1) under the hypotheses (2). It is defined below as

$$H(B) = \log\det(R - B) + l.\log\det B, \tag{3}$$

where l is a scalar factor and $\det(A)$ denotes the determinant of the matrix A.

A brief statistical motivation of H is presented in Appendix A. The scalar factor l is considered (see Appendix A) as a redundancy factor since it is associated with the number of (statistically) independent noise vectors available along the array. Practically, the choice of the factor l is related to statistical considerations beyond the scope of the present paper.

Now a parameterization of the B matrix is necessary. For that purpose, a banded Toeplitz parameterization is quite convenient, i.e., [15, 16]:

$$B = \sum_{i=1}^{p} \beta_i U_i,$$

$$\text{where :} \quad \left| \begin{array}{l} \beta_i \text{ are scalars (real),} \\ U_i \text{ is a } q \times q \text{ matrix defined as usual by:} \\ U_i(k,\ell) = \left\{ \begin{array}{lll} 1 & \text{if} & |k-\ell| = i-1, \\ 0 & \text{else.} & \end{array} \right. \end{array} \right. \tag{4}$$

The number q represents the sensor number and is consequently the dimension of the square matrices R, S, B. The matrices $\{U_i\}_{i=1}^{q}$ constitute an orthogonal (for the euclidean product) basis of T_q (the vector space of q-dimensional symmetric Toeplitz matrices). For practical applications [7], p is small with respect to q.

Obviously, this parameterization does not ensure the positive definiteness of B, so to avoid such a problem B can be parameterized as the covariance matrix of a MA process:

$$B = \left(\sum_{i=0}^{p-1} b_i \, Z_i \right) \left(\sum_{i=0}^{p-1} b_i \, Z_i \right)^t .$$

(The symbol "t" denotes matrix transposition.)

Here the matrices Z_i are $(p \times p + q)$ rectangular matrices given by

$$Z_i = \begin{pmatrix} 0 & \cdots & 0 & 1 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \end{pmatrix}$$

or

$$Z_i(k,\ell) = \begin{cases} 1 \text{ if } |k-\ell| = i, \\ 0 \text{ else.} \end{cases} \tag{5}$$

(This model is the assumed minimum phase [11].)

This parameterization will be especially useful for the special case study of large arrays (§5), but for the moment our attention will be restricted to the banded Toeplitz parameterization. The REF will thus be defined as follows:

$$H(\beta_1, \ldots, \beta_p) = \log\det(R - B) + l.\log\det B$$

with $B = B(\beta_1, \ldots, \beta_p)$ (R, B are $q \times q$ matrices).

The general optimization problem takes the following form. Find

$$\max H(\beta_1, \ldots, \beta_p)$$

under the constraints

$$\mathcal{C} \left| \begin{array}{lll} R - B & > & 0, \\ B & > & 0. \end{array} \right. \tag{6}$$

($A > B$ means, as usual, that $A - B$ is positive definite.)

Let $\mathbf{B}_*$ be the parameter vector maximizing H under $\mathcal{C}$, i.e.,

$$\mathbf{B}_* = \arg\max H(\beta_1, \ldots, \beta_p) \quad \text{under } \mathcal{C}.$$

As will be seen later, the functional H can be efficiently maximized by iterative methods. But let us first consider the properties of $\mathbf{B}_*$.

3. Properties of $\mathbf{B}_*$. Because the REF depends nonlinearly on the parameters $\{\beta_i\}$, it seems much simpler to formulate the problem in terms of the eigensystems.

The spatially white noise case is presented first. It is not relevant to our problem, but it allows us to obtain a simple result and enlightens the role of the factor l. Then the general case is considered.

3.1. Spatially white noise case. This case is very simple but the reasoning is rather similar to that used in the general case.

The noise is assumed to be uncorrelated (sensor to sensor), so

$$B = \lambda\,\mathrm{Id}$$

(Id : identity matrix, $\lambda > 0, \lambda = \beta_1$).

Consider now an eigendecomposition of R, i.e.,

$$\begin{aligned} &R = \sum_{i=1}^{q} \alpha_i \mathbf{V}_i \mathbf{V}_i^*, \quad \mathbf{V}_i \perp \mathbf{V}_j (i \neq j), \|\mathbf{V}_i\| = 1 \\ &(\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_q > 0). \end{aligned} \tag{7}$$

($*$ denotes transposition and conjugation.)

Then by means of elementary algebra, we have

$$H(\lambda) = \sum_{i=1}^{q} \log(\alpha_i - \lambda) + ql.\log\lambda. \tag{8}$$

The problem now consists of determining the value of λ that maximizes $H(\lambda)$ under the constraints $\mathcal{C}$.

Now

$$\frac{\partial H}{\partial \lambda} = \frac{ql}{\lambda} + \sum_{i=1}^{q} \frac{-1}{(\alpha_q - \lambda) - (\alpha_q - \alpha_i)},$$

and denoting by $\tilde{\lambda}$ the following particular value of λ,

$$\tilde{\lambda} = \alpha_q \left(\frac{l}{l+1} \right),$$

we obtain

$$\left(\frac{\partial H}{\partial \lambda} \right)_{\lambda = \tilde{\lambda}} = (l+1) \frac{q}{\alpha_q} + \sum_{i=1}^{q} \left[-\frac{1}{(l+1)} \alpha_q + (\alpha_q - \alpha_i) \right]^{-1}. \tag{9}$$

Now $\alpha_i - \alpha_q \geq 0$ for $i = 1, 2, \ldots, q$ and therefore

$$\left(\frac{\partial H}{\partial \lambda} \right)_{\lambda = \tilde{\lambda}} \geq 0.$$

Furthermore, if λ tends towards α_q, then $\partial H / \partial \lambda$ tends towards $-\infty$. Since H is a differentiable and concave function of λ in the interval $]0, \alpha_q[$, it follows that the maximum of the REF under the constraints $\mathcal{C}$ is attained for a value λ_* of λ satisfying the following inequalities:

$$\alpha_q \left(\frac{l}{l+1} \right) \leq \lambda_* \leq \alpha_q. \tag{10}$$

3.2. General case. Let B_0 be the exact noise CM and assume that B_0 may be described by the parameterization defining the B matrices. Then the following proposition is valid.

PROPOSITION 3.1. *Let $\{\lambda_i^*\}_{i=1}^q$ be the eigenvalues of the (whitened) matrix $B_0^{-1} B_*$, where B_* denotes the matrix maximizing the REF H under the two constraints (B and $R - B$ positive definite). Then these eigenvalues satisfy the following inequalities:*

$$\frac{l}{l+1} \leq \lambda_i^* \leq 1, \qquad i = 1, 2, \ldots, q.$$

Proof. The decomposition of R in source and noise parts (i.e., $R = S + B_0$) is assumed to be unique (see Proposition 4.3). The rank of S is denoted by s (the source number). Note that s is strictly inferior to q. Then the REF H takes the general following form:

$$H(B) = \log \det (S + (B_0 - B)) + l . \log \det B.$$

Let us now examine the various terms of the functional $H(B)$. For that purpose, let us note the following assumptions:

1. B_0 is positive definite,
2. $(B_0 - B)$ is positive definite.

The first assumption is quite classical in the signal processing context since B_0 is a covariance matrix [1, 3]. The second will be justified later (Comment 1, pp. 361–362).

Since the matrix B_0 is positive definite, it can be factorized in triangular factors (Choleski factorization [17]), i.e.,

$$B_0 = T_0 T_0^*.$$

Thus

$$\begin{aligned}\log\det\left[S+(B_0-B)\right] &= \log\det\left[S(B_0-B)^{-1}+\mathrm{Id}\right] \\ &\quad +\log\det\left(B_0-B\right) \\ &= \log\det\left[ST_0^{-1*}\left(\mathrm{Id}-T_0^{-1}BT_0^{-1*}\right)^{-1}T_0^{-1}+\mathrm{Id}\right] \\ &\quad +\log\det\left(B_0-B\right). \qquad (11)\end{aligned}$$

$$\left((B_0-B)^{-1} = T_0^{-1*}\left(\mathrm{Id}-T_0^{-1}BT_0^{-1*}\right)^{-1}T_0^{-1}\right).$$

Since the matrix $\left(\mathrm{Id}-T_0^{-1}BT_0^{-1*}\right)^{-1}$ is hermitian it is diagonalizable, i.e.,

$$\left(\mathrm{Id}-T_0^{-1}BT_0^{-1*}\right)^{-1} = W\Delta W^*$$

(W : unitary matrix, Δ : diagonal) with

$$\Delta(i,i) = (1-\lambda_i)^{-1}.$$

(λ_i : eigenvalue of $B_0^{-1}B$.)

In what follows, it is necessary to preserve the symmetry of the problem obtained by using elementary algebra as follows:

$$\begin{aligned}\log\det\left[ST_0^{-1*}\left(\mathrm{Id}-T_0BT_0^{-1*}\right)T_0^{-1}+\mathrm{Id}\right] &= \log\det\left[ST_0^{-1*}W\Delta W^*T_0^{-1}+\mathrm{Id}\right] \\ &= \log\det\left[\Delta^{1/2}W^*S'W\Delta^{1/2}+\mathrm{Id}\right] \\ \text{with}: \quad S' &= T_0^{-1}ST_0^{-1*}. \qquad (12)\end{aligned}$$

(This last equality results from intensive use of the classical formula $\det(AB) = \det(BA)$ [17].)

The matrix $\Delta^{1/2}$ in (12) is the diagonal matrix defined by $\Delta^{1/2}(i,i) = (\Delta(i,i))^{1/2}$. Its existence follows from the hypothesis that $(B_0 - B)$ is positive definite. Thus, the following equality holds trivially:

$$\begin{aligned}\log\det\left[\Delta^{1/2}W^*S'W\Delta^{1/2}+\mathrm{Id}\right] &= \log\det\left[\Delta^{1/2}W^*\left(S'+W\Delta^{-1}W^*\right)W\Delta^{1/2}\right] \\ &= \log\det\Delta+\log\det\left(S'+W\Delta^{-1}W^*\right). \qquad (13)\end{aligned}$$

We are now able to calculate the partial derivatives of the REF H with respect to the parameters λ_i. More precisely, using a classical formula for the differentiation of the determinant of a matrix $A(\lambda)$ [18] (i.e., $\partial/\partial\lambda \log\det A(\lambda) = tr\left(A^{-1}(\lambda)\partial/\partial\lambda A(\lambda)\right)$), the partial derivatives $\partial H/\partial\lambda_i$ take the following form:

$$\frac{\partial H}{\partial\lambda_i} = l.\frac{1}{\lambda_i}-\frac{1}{1-\lambda_i}+\left\{\frac{1}{1-\lambda_i}+tr\left[\left(S'+W\Delta^{-1}W^*\right)^{-1}W\left(\frac{\partial}{\partial\lambda_i}\Delta^{-1}\right)W^*\right]\right\} \qquad (14)$$

(tr denotes the trace).

Now the following equality comes from the orthogonality property of the eigenvectors [17]:

$$(15) \quad tr\left[\left(S' + W\Delta^{-1}W^*\right)^{-1} W \frac{\partial \Delta^{-1}}{\partial \lambda_i} W^*\right] = -\mathbf{W}_i^* \left(S' + W\Delta^{-1}W^*\right)^{-1} \mathbf{W}_i^*.$$

($\mathbf{W}_i$ is the ith column of the matrix W.)

Furthermore, one has

$$S' + W\Delta^{-1}W^* \geq W\Delta^{-1}W^*;$$

hence,

$$\left(S' + W\Delta^{-1}W^*\right)^{-1} \leq \left(W\Delta^{-1}W^*\right)^{-1},$$

so that, finally,

$$(16) \quad -\mathbf{W}_i^* \left(S' + W\Delta^{-1}W^*\right)^{-1} \mathbf{W}_i \geq -\mathbf{W}_i^* (W\Delta W^*)^{-1} \mathbf{W}_i = -\frac{1}{1-\lambda_i}.$$

In conclusion, we note that the term between braces in (14) (i.e., $1/(1-\lambda_i) + tr[(S' + W\Delta^{-1}W^*)^{-1} W(\frac{\partial}{\partial \lambda_i}\Delta^{-1})W^*]$) is positive when λ_i runs throughout the open interval $]0, 1[$. Consequently, the partial derivatives $\partial H/\partial \lambda_i$ are positive ($i = 1, \ldots, q$) when λ_i runs throughout the open interval $]\,0, l/l+1\,[$.

Furthermore, the equality

$$B_0 - B = T_0 \left(\mathrm{Id} - T_0^{-1} B T_0^{-1*}\right) T_0^*$$

proves that (under the basic assumptions) the matrix $\left(\mathrm{Id} - T_0^{-1} B T_0^{-1*}\right)$ must be positive definite and that all the eigenvalues (i.e., $1 - \lambda_i$) of the matrix $\mathrm{Id} - B_0^{-1}B$ must be positive. It is thus sufficient to restrict our attention to the parameter values β_i such that all the eigenvalues λ_i are smaller than 1.

Now the following fact is instrumental for the proof of Proposition 3.1: the REF H is a concave functional on the whole domain $\mathcal{C}$ of the constraints (6). This property will be shown in the next section *independently* of Proposition 3.1.

Denote by $\mathcal{C}'$ the following (restricted) constraint domain defined as $\mathcal{C}' = \{$ B such that (s.t.) B and $B_0 - B$ are positive definite $\}$. Then it is directly shown that $\mathcal{C}'$ is a convex subset of $\mathcal{C}$.

When λ_i tends towards 1_- then H tends towards $-\infty$. Since all the partial derivatives $\partial H/\partial \lambda_i$ are positive when λ_i runs through the interval $]0, l/l+1[$ and are continuous on $\mathcal{C}'$, there exists a matrix B_* of $\mathcal{C}'$ such that the maximum of H on $\mathcal{C}'$ is attained for B_*. Note that this maximum is unique on $\mathcal{C}'$ (concavity) and is attained for a matrix B_* such that all the eigenvalues λ_i^* (of $B_0^{-1}B_*$) belong to the interval $]l/l+1, 1[$. So there is a point $(\beta_1^*, \ldots, \beta_p^*)$ of $\mathcal{C}'$ such that all the partial derivatives $\partial H/\partial \lambda_i$ are altogether null.

Because the REF H is concave on the whole domain $\mathcal{C}$, its maximum on $\mathcal{C}$ is unique and is attained for a matrix B_* of $\mathcal{C}'$. This proves Proposition 3.1. □

Comments. The preceding calculations require some comments.

1. In the proof of Proposition 3.1, the positive definite hypothesis $(R - B)$ has been replaced by the positive definite hypothesis $(B_0 - B)$.

Actually, the two subsets $\mathcal{C}$ and $\mathcal{C}'$ ($\mathcal{C}' = \{B$ s.t. B and $B_0 - B$ are positive definite$\}$) are convex and $\mathcal{C}$ contains $\mathcal{C}'$. Because the functional H is concave on $\mathcal{C}$ (Proposition 4.1) and attains its maximum value on $\mathcal{C}'$, this maximum is unique and satisfies Proposition 3.1 on the whole subset $\mathcal{C}$ [19].

2. Consider (16). Then this inequality becomes an equality (for all the values of i) if and only if the source matrix S' is null. In this case, all the partial derivatives $\partial H/\partial\lambda_i$ are null for $\{b_i^*\}$ values s.t.

$$B_0^{-1}B_* = \left(\frac{l}{l+1}\right)\mathrm{Id}$$

or

$$B_* = \left(\frac{l}{l+1}\right)B_0.$$

The matrix B_0 is thus perfectly "estimated" up to a scalar factor. We want to stress that this scalar factor (i.e., $l/l+1$) has no practical importance.

3. As has been seen in the proof of Proposition 3.1, the effect of sources is to move the maximum of H and to cancel the equality of all the λ_i^*. Thus, in the presence of sources, B_0 cannot be perfectly "estimated" by maximizing H. Of course, the "quality" of the estimate increases with the scalar l.
4. Proposition 3.1 is still valid when the noise model is overdetermined (i.e., $p_0 \leq p$); this fact follows clearly from the proof of Proposition 3.1. For practical applications, it is a fundamental point.
5. The following property seems valid (at least for sufficiently great values of q).

Conjecture 1. Consider two distinct values of the parameter l, say l_1 and l_2 ($l_1 > l_2$), and denote $\{\lambda_i^*\}$ (respectively, $\{\mu_i^*\}$) to be the eigenvalues of $B_0^{-1}B_{*l_1}$ (respectively, $B_0^{-1}B_{*l_2}$). Then the following property holds:

$$\lambda_i^* \geq \mu_i^*, \qquad i = 1, \ldots, q.$$

This property has been verified by numerous simulation results (see Figs. 3–6). A very rough "proof" is based on the following fact: two banded Toeplitz matrices commute (approximately).

Actually, the REF method can be considered as a way to tackle the following problem: how to determine the "more random" noise model (i.e., maximizing $\log\det B$) under the positive definiteness constraints (B and $R - B$ positive definite). Clearly, for this sense, the term $\log\det(R - B)$ appears as a barrier functional forbidding interaction between source and noise models. The factor l represents the compromise between the accuracy of estimated parameters (Proposition 3.1) and the statistical variability of the estimated data (i.e., $\hat{R}$). It can thus be considered as a redundancy factor (see Appendix A).

Obviously, this interpretation of the factor l relies upon statistical considerations that are not in the scope of this paper.

4. Maximization of the REF H. The numerical problem now consists of determining the values of the parameters $\{\beta_i\}_{i=1}^p$ that maximize the REF H under the positivity constraints. After a general study of the functional concavity, the problem of practical optimization will be considered.

Actually, the REF enjoys a useful property which has been instrumental in the proof of Proposition 3.1.

PROPOSITION 4.1. *On the constraints domain* $\mathcal{C}$ (6) *the REF is a concave functional with respect to the* $\{\beta_i\}_{i=1}^p$ *parameters.*

Proof. The proof of Proposition 4.1 relies upon classical results of linear algebra. More precisely, we use the following classical lemmas, valid for any differential family of isomorphisms [20]:

$$\left|\begin{array}{lcl} \dfrac{\partial}{\partial\beta}\log\det B(\beta) & = & tr\left(B^{-1}(\beta)\dfrac{\partial B}{\partial\beta}\right), \\ \dfrac{\partial}{\partial\beta}B^{-1}(\beta) & = & -B^{-1}(\beta)\dfrac{\partial B}{\partial\beta}B^{-1}(\beta). \end{array}\right. \tag{17}$$

Then the Hessian matrix (denoted H_2) of H with respect to the $\{\beta_i\}_{i=1}^p$ is easily obtained:

$$H_2(i,j) \triangleq \frac{\partial^2 H}{\partial\beta_i\partial\beta_j} = -tr\left[(R-B)^{-1}U_i(R-B)^{-1}U_j\right] - l\cdot tr\left(B^{-1}U_iB^{-1}U_j\right). \tag{18}$$

Let $\mathbf{X}$ be any vector of $\mathbb{R}^p$ ($\mathbf{X}^t = (x_1,\ldots,x_p)$); then

$$\mathbf{X}^tH_2\mathbf{X} = \sum_{i,j} x_i\frac{\partial^2 H}{\partial\beta_i\partial\beta_j}x_j,$$

and using (18) and the linearity of the trace we get

$$\begin{aligned} \mathbf{X}^tH_2\mathbf{X} = -tr\left[(R-B)^{-1}\left(\sum_{i=1}^p x_iU_i\right)(R-B)^{-1}\left(\sum_{j=1}^p x_jU_j\right)\right] \\ -l.tr\left[B^{-1}\left(\sum_{i=1}^p x_iU_i\right)B^{-1}\left(\sum_{j=1}^p x_jU_j\right)\right]. \end{aligned} \tag{19}$$

The two terms $-tr(--)$ of (19) are of the form $-tr(AC\,AC)$ with

$$A = (R-B)^{-1} \text{ or } B^{-1} \text{ and } C = \sum_{i=1}^p x_iU_i.$$

Since the matrix A is assumed to be positive definite, it admits a Choleski factorization, say $A = TT^*$, so that

$$\begin{aligned} -tr\,[ACAC] &= -tr\,[TT^*CTT^*C] \\ &= -tr\left[(T^*CT)^2\right] = -\|T^*CT\|_F^2. \end{aligned} \tag{20}$$

(The symbol $\|\ \|_F$ denotes the Frobenius norm [17] of a matrix.)

Finally, the quadratic form $\mathbf{X}^tH_2\mathbf{X}$ is negative for any (nonnull) vector $\mathbf{X}$. The matrix H_2 is therefore negative definite and H is therefore a concave functional with respect to $\{\beta_i\}_{i=1}^p$ on $\mathcal{C}$. Consequently, gradient-like methods (with optimal stepsize) will converge on $\mathcal{C}$. □

Actually, this concavity property is very strong and quite dependent on the noise parameterization. Thus Proposition 4.1 holds for a linear parameterization but not for more restrictive (especially nonlinear) parameterizations. Consider, for instance,

the MA parameterization of the noise (5); then the partial derivatives of the functional H (with respect to b_i) are directly calculated, yielding

$$\frac{\partial H}{\partial b_i} = -tr\left[(R-B)^{-1}D_i^1\right] + l.tr\left[B^{-1}D_i^1\right]$$

$$\frac{\partial^2 H}{\partial b_i \partial b_j} = -tr\left[(R-B)^{-1}D_i^1(R-B)^{-1}D_j^1\right] - tr\left[(R-B)^{-1}D_{i,j}^2\right]$$
$$-l.tr\left[B^{-1}D_i^1B^{-1}D_j^1\right] + l.tr\left[B^{-1}D_{i,j}^2\right]$$

with

$$D_i^1 = \frac{\partial B}{\partial b_i}$$
$$= \left(\sum_{i=0}^{p-1} b_i Z_i\right) Z_j^t + Z_j \left(\sum_{i=0}^{p-1} b_i Z_i\right)^t$$

and

$$D_{i,j}^2 = \frac{\partial^2 B}{\partial b_i \partial b_j}$$
$$= Z_i Z_j^t + Z_j Z_i^t.$$

The sign of the quadratic form $\mathbf{X}H_2\mathbf{X}$ is thus not at all evident. So the reasoning previously used for proving Proposition 4.1 cannot be directly extended to this parameterization. The simplicity of the proof of Proposition 4.1 is essentially due to the linear parameterization of the noise matrix B.

The concavity property seems (generally) wrong for the MA parameterization. This is illustrated by Fig. 1, which represents the level curves of the functional $H(b_0, b_1)$. The cross corresponds to the exact values of b_0 and b_1. However, even if Proposition 4.1 is not (generally) valid for the MA parameterization, the following proposition holds.

PROPOSITION 4.2. *The coefficients* $(b_0^*, \ldots, b_p^*)$ *of the MA process maximizing* H *on the constraint domain satisfy the following inequalities:*

$$\left|(b_i^* - b_i^0)\,\frac{1}{b_i^0}\right| \leq 1 - \sqrt{\frac{l}{l+1}}.$$

Furthermore, the gradient of H is null for a unique point of the parameter set; this point verifies the above proposition. This is a direct consequence of Proposition 4.1 and the one-to-one mapping between the coefficients (say $\{b_i\}$) of a min-phase MA model and its covariance set (say $\{\beta_i\}$). Therefore, there is a unique maximum of H for the MA parameterization on the constraint domain $\mathcal{C}$. Locating this point is not at all obvious since the correspondence between the MA parameters and the eigenvalues of the matrix $B_o^{-1}B_*$ is highly nonlinear. So this property will be proved by analytic arguments (see the proof of Proposition 5.1). A direct algebraic proof of Proposition 4.2 seems unfeasible.

Let us now consider practical considerations and, more precisely, iterative methods for maximizing H.

We shall now briefly present the gradient method.

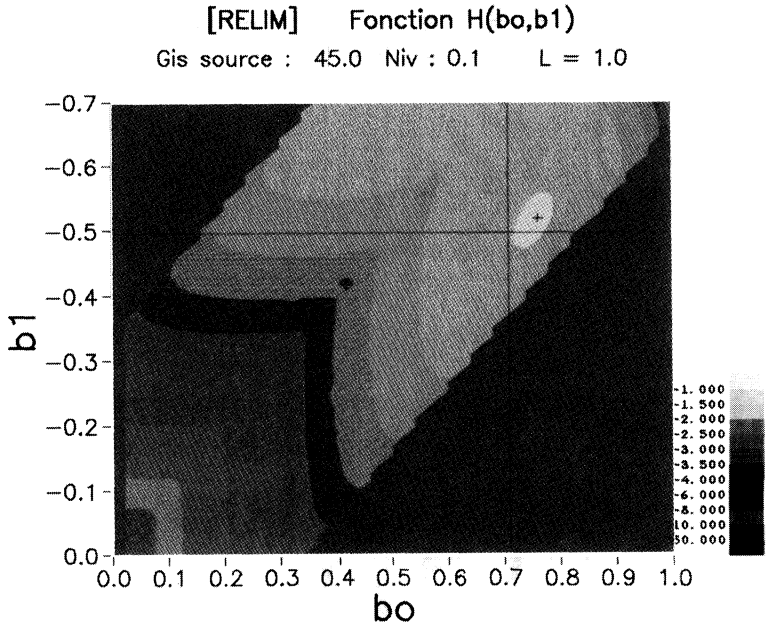


FIG. 1. *Values of the functional* $H(b_0, b_1)$, *one source* (*bearing* : 45 deg., -10dB).

The calculation of the gradient vector is straightforward. The ith component of the gradient vector G_k (at iteration k) is given by

$$G_k(i) = -tr\left[(R - B_k)^{-1} U_i\right] + l.tr\left(B_k^{-1} U_i\right). \tag{21}$$

(B_k is the noise matrix at iteration k.)

The gradient algorithm takes the following general form:

$$\mathbf{X}_{k+1} = \mathbf{X}_k - \rho_k \mathbf{G}_k \tag{22}$$

with

$$\mathbf{X}_k^t \triangleq (\beta_1^k, \ldots, \beta_p^k)$$

and $\mathbf{G}$ defined by (21).

The scalar ρ_k is the stepsize of the algorithm. In order to ensure convergence (on $\mathcal{C}$) of the gradient algorithm, it is worth determining the optimal stepsize.

The matrix translation of (22) is

$$B_{k+1} = B_k - \rho_k D_k$$

with

$$\begin{cases} B_k = \sum_{i=1}^{p} \beta_i^k U_i, \\ D_k = \sum_{i=1}^{p} G_k(i) U_i. \end{cases} \tag{23}$$

The optimal stepsize ρ_k can be easily obtained by using eigendecompositions. The corresponding algorithm is presented below (and detailed in Appendix B).

Step 1. Since B_k and $R - B_k$ are positive definite, decompose them in triangular factors:

$$B_k = T_k T_k^*, \quad R - B_k = S_k S_k^*.$$

Step 2. Compute the eigenvalues $\{\alpha_i^k\}$ and $\{\beta_i^k\}$ of the two hermitian matrices:

$$S_k^{-1} D_k S_k^{-1*} \text{ and } T_k^{-1} D_k T_k^{-1*}.$$

Step 3. Then the REF becomes an explicit function of the parameter (stepsize) ρ, given by

$$H(\rho) = \sum_{i=1}^{q} \log\left(1 + \rho\alpha_i^k\right) + l \sum_{i=1}^{q} \log\left(1 - \rho\beta_i^k\right) + cst. \tag{24}$$

Furthermore (it is perhaps the most important fact), the positivity constraints $\mathcal{C}$ are translated into explicit (relatively to ρ) constraints, i.e.,

$$\mathcal{C} \begin{cases} 1 + \rho\alpha_i^k > 0, & i = 1, \ldots, q, \\ 1 - \rho\beta_i^k > 0, & i = 1, \ldots, q. \end{cases} \tag{25}$$

Step 4. The optimal stepsize ρ is simply obtained by maximizing $H(\rho)$ given by (24) under the constraints (25). This task is easily achieved by means of a unidimensional Newton method initialized at 0.

Obviously, the gradient method may be replaced by more sophisticated iterative methods (Newton, BFGS, etc.). However, this does not appear drastically important since the number of parameters defining the noise model (i.e., p) is generally quite smaller than q and because of Propositions 4.1 and 4.2. A direct extension to the complex case (the noise parameters are complex) is provided in Appendix C.

We now consider the unicity of the decomposition in source and noise parts. This is an important problem of identifiability [21]. The source's CM matrix S is assumed to be Toeplitz. Physically, this corresponds to the plane wave hypothesis and a uniform linear array assumption [1–7]. Since S is rank deficient and semipositive definite, S may be written in the following form (thanks to the theorem of Caratheodory [22]):

$$\left| \begin{array}{l} S = \sum_{j=1}^{s} \sigma_j \mathbf{Z}_j \mathbf{Z}_j^* \\ \text{with } : \sigma_j > 0, \\ \mathbf{Z}_j = \left(1, z_j, \ldots, z_j^{q-1}\right)^t, \ |z_j| = 1, \\ \text{rank } (S) = s. \end{array} \right. \tag{26}$$

Then a sufficient (and very rough) condition ensuring unicity of the decomposition will be obtained as follows. Assume the existence of two such decompositions. Then

$$R = S_1 + B_1 = S_2 + B_2;$$

hence

$$S_1 - S_2 = B_2 - B_1$$

(B_1 and B_2 are p-banded Toeplitz matrices).

In order to annihilate the noise effect, we consider the $\left(\frac{q-p}{2} \times \frac{q-p}{2}\right)$ lower left submatrix L of $S_1 - S_2$ defined by

$$L(i,j) = (S_1 - S_2)(q + i - p, j), \qquad 1 \leq i, j \leq \frac{q-p}{2}.$$

We assume, in the previous equation, that $q-p$ is even. Otherwise, it must be replaced by $q - p - 1$. All the components of the matrix L must be null since B_1 and B_2 are two p-banded Toeplitz matrices. The following equality is then directly obtained from (26):

$$\sum_{j=1}^{s} \sigma_{j,1} \mathbf{Z'}_{j,1} (\mathbf{Z''}_{j,1})^* = \sum_{j=1}^{s} \sigma_{j,2} \mathbf{Z'}_{j,2} (\mathbf{Z''}_{j,2})^*$$

with

$$\left| \begin{array}{l} \mathbf{Z'}_{j,1} \triangleq \left(z_{j,1}^{q-q'}, \ldots, z_{j,1}^{q-1}\right), \\ \mathbf{Z''}_{j,1} \triangleq \left(z_{j,1}^{0}, \ldots, z_{j,1}^{q'-1}\right), \\ q' \triangleq \dfrac{q-p}{2} \qquad (j = 1, \ldots, s). \end{array} \right. \tag{27}$$

Now there exist coefficients $\{a_0, a_1, \ldots, a_s\}$ such that the roots of the polynomial equation

$$A(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_s z^s = 0$$

are $\{z_{1,1}, \ldots, z_{s,1}\}$. Hence the $s \times (q-p)$ matrix N defined by

$$N \triangleq \begin{pmatrix} a_0 & \cdots & a_s & 0 & \cdots & 0 \\ 0 & a_0 & \cdots & a_s & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & a_0 & \cdots & a_s \end{pmatrix}$$

annihilates the columns of Z'_1 (Z'_1 is the rectangular matrix whose colums are the vectors $\mathbf{Z'}_i$; $i = 1, \ldots, s$) when $s \leq q'$. In other words,

$$N Z'_1 = 0,$$

which implies that $N Z'_2 = 0$.

This implication is easily shown by considering the columns of Z'_1 and Z'_2. They span the same space since by (27)

$$Z'_1 \Delta_1 Z''^{,*}_1 = Z'_2 \Delta_2 Z''^{,*}_2$$

with

$$\Delta_1 = \operatorname{diag}(\sigma_{1,1}, \ldots, \sigma_{s,1}), \qquad \Delta_2 = \operatorname{diag}(\sigma_{1,2}, \ldots, \sigma_{s,2}).$$

Consequently, $\{z_{1,2}, \ldots, z_{s,2}\}$ are also the roots of $A(z)$. So there exists a permutation matrix P such that $Z_1 = Z_2 P$.

Now assume that the dimension of $\mathbf{Z}'_j$ (i.e., q') is superior to s. Then using a basic result on Vandermonde determinants [23], the solution to (27) is either

$$\left| \begin{array}{ll} \sigma_{j,1} = 0, & j = 1, \ldots, s, \\ \sigma_{j,2} = 0, & j = 1, \ldots, s, \end{array} \right.$$

or, for each source index j (related to S_1), there exists a source index k (related to S_2) such that

$$\left| \begin{array}{l} \sigma_{j,1} = \sigma_{k,2}, \\ z_{j,1} = z_{k,2} \end{array} \right. \tag{28}$$

or, equivalently, there exists a $s \times s$ permutation matrix P such that

$$Z_1 = Z_2 P. \tag{29}$$

(The matrices Z_1 and Z_2 are the rectangular matrices (27) whose columns are the source vectors.)

Thus, the following proposition holds.

PROPOSITION 4.3. *Assume that S and B are Toeplitz matrices and assume, furthermore, that $q - p$ is greater than $2s$. Then the decomposition in source and noise matrices is unique.*

Note that the plane wave assumption or, equivalently, the Toeplitz hypothesis, has been instrumental for proving Proposition 4.3, which may be only considered as a sufficient and rough identifiability condition. The identifiability problem is greatly complicated by the noise correlation. In this case, the noise subspace has no clear algebraic meaning, as opposed to the white noise case.

We would like to stress that the accuracy of noise parameter estimates (i.e., the $\{\beta_i\}$) is expressed only in terms of the eigenvalues of $B_0^{-1} B_*$ and not directly in terms of the β_i. However, these two subsets are strongly related even if these relations are nonexplicit and highly nonlinear in the general case.

Actually, there is a one to one correspondence between the noise parameter vector β and the vector of the eigenvalues of the matrix $B_0^{-1} B_*$.

Indeed, consider the Jacobian matrix of partial derivatives [17]:

$$J \triangleq \begin{pmatrix} \frac{\partial \lambda_1}{\partial \beta_1} & \cdots & \frac{\partial \lambda_q}{\partial \beta_1} \\ \vdots & & \vdots \\ \frac{\partial \lambda_1}{\partial \beta_p} & \cdots & \frac{\partial \lambda_q}{\partial \beta_p} \end{pmatrix}.$$

Using the classical lemma (simple eigenvalues) [17] we have

$$\frac{\partial \lambda_i}{\partial \beta_j} = \mathbf{V}_i^* \frac{\partial}{\partial \beta_\ell} \left(B_0^{-1} B_* \right) \mathbf{V}_i'$$

and thus

$$J = \begin{pmatrix} tr\left(B_0^{-1} U_1 \mathbf{V}_1 \mathbf{V}_1^*\right) & \cdots & tr\left(B_0^{-1} U_1 \mathbf{V}_q \mathbf{V}_q^*\right) \\ \vdots & & \vdots \\ tr\left(B_0^{-1} U_p \mathbf{V}_1 \mathbf{V}_1^*\right) & \cdots & tr\left(B_0^{-1} U_p \mathbf{V}_q \mathbf{V}_q^*\right). \end{pmatrix}. \tag{30}$$

Now the matrices $\{\mathbf{V}_i\mathbf{V}_i^*\}_{i=1}^q$ are linearly independent in $M_q(\mathbb{C})$ (the space of hermitian $q \times q$ matrices) and, consequently, the rank of the rectangular matrix J is generally equal to p ($q > p$).

Finally, when l becomes great the eigenvalues of $B_0^{-1}B_*$ approach 1 (Proposition 3.1) and the parameter vector $\mathbf{B}_*$ approaches $\mathbf{B}_0$. Using simple algebraic considerations, it seems difficult to go further, but as will be seen in §5, an analytic formulation of the REF will allow us to refine the results of Proposition 3.1.

5. Analytic properties of the REF. The REF properties, previously considered, rely upon algebraic considerations. We shall see now that the REF definition can be translated in terms of analytic functions, allowing us to make the REF properties precise.

Let $R(k)$ be the (spatial) density of the stationary process received by the array. For a uniform array, $R(k)$ is simply the Fourier transform of the covariance matrix R, i.e.,

$$R(k) = \sum_{j=1-q}^{q-1} r_j \exp(2i\pi kjd)$$

with

$$\begin{aligned} &R = \text{Toepl}\ (r_0, r_1, \ldots, r_{q-1}), \\ &d\ :\ \text{intersensor distance}, \\ &k = (d/\lambda)\cdot \sin\theta, \qquad \lambda : \text{wavelength},\ \theta : \text{bearing}. \end{aligned} \tag{31}$$

Even if the scalar d has a physical meaning, this meaning may be forgotten for what follows. Using Szego's theorem [22] one obtains (for q large)

$$\lim_{q\to\infty} \frac{1}{q} \log\det R = \frac{1}{2w}\int_{-w}^{w} \log R(k)dk, \tag{32}$$

where w is the spatial bandwidth.

Once again the physical meaning of w is not at all fundamental for what follows. Usually it is assumed to be $1/2$. Hence for a large value of q, the REF can be expressed as follows:

$$H = \int_{-w}^{w} \log\left(R(k) - B(k)\right) dk + l.\int_{-w}^{w} \log B(k)dk. \tag{33}$$

An MA noise modelling (5) seems to be quite convenient since it avoids the positivity problems while conserving the banded Toeplitz structure of the covariance matrix. For this model, $B(k)$ is given by

$$\left|\begin{array}{l} B(k) = |\ F(z)\ |^2 \\ \text{with} \\ F(z) = b_0 + b_1 z + \cdots + b_{p-1}z^{p-1}, \\ z = \exp(2i\pi kd), \\ i^2 = -1. \end{array}\right. \tag{34}$$

Then the following proposition of the REF holds.

PROPOSITION 5.1. *Assume that the noise may be exactly modelled by an MA process* $(b_0^0, b_1^0, \ldots, b_{p-1}^0)$. *Then for any MA modelling of the same order* (p), *the coefficients* $(b_0^*, b_1^*, \ldots, b_{p-1}^*)$ *of the MA process that maximizes* H (33) *under* C *satisfy the following set of inequalities:*

$$\left|\left(b_i^* - b_i^0\right)\frac{1}{b_i^0}\right| \leq 1 - \sqrt{\frac{l}{l+1}}, \qquad i = 0, 1, \ldots, p-1.^1$$

Proof. Previously, the proofs basically used the tools of linear algebra. From now on they will be replaced by complex analysis arguments. A direct approach will be considered. More precisely, the study of the sign of functionals involving partial derivatives (e.g., $\sum \lambda_i \partial H/\partial b_i$) will be instrumental.

The REF H is given by (33) and its partial derivatives (with respect to the $\{b_i\}$) are straightforwardly obtained:

$$\frac{\partial H}{\partial b_i} = \int_{-w}^{w} \frac{\operatorname{Re}\left(z^i \bar{F}(k)\right) \cdot [lR(k) - (l+1)B(k)]}{(R(k) - B(k))\, B(k)} dk. \tag{35}$$

(Re: real part of a complex number, $\bar{z}$: complex conjugate of z.)

We first consider the noise alone case. Then $R(k) = B_0(k)$ and the partial derivatives $\partial H/\partial b_i$ become

$$\frac{\partial H}{\partial b_i} = \int_{-w}^{w} \frac{\operatorname{Re}\left(z^i \bar{F}(k)\right) [lB_0(k) - (l+1)B(k)]}{(B_0(k) - B(k))\, B(k)} dk.$$

Because of the independence of the functions $\{\operatorname{Re}\left(z^i \bar{F}(k)\right)\}$ in the polynomial space, there exists a set of scalars $\{\lambda_i\}$ such that

$$\sum \lambda_i \operatorname{Re}\left(z^i \bar{F}(k)\right) = lB_0(k) - (l+1)B(k),$$

which implies

$$\sum \lambda_i \frac{\partial H}{\partial b_i} = \int_{-w}^{w} \frac{(lB_0(k) - (l+1)B(k))^2}{(B_0(k) - B(k))\, B(k)} dk. \tag{36}$$

Let us now examine the right member of (36). Clearly, the integrand is positive since the function $B_0(k) - B(k)$ must be positive on $[-w, +w]$ because of the definition of the REF. Therefore, $\sum \lambda_i \partial H/\partial b_i$ is positive and is null if and only if the numerator $(lB_0(k) - (l+1)B(k))$ is almost everywhere (a.e.) null on $[-w, w]$ or

$$B(k) = (l/l+1)B_0(k), \qquad k \in [-w, w],$$

which implies

$$b_i = \sqrt{\frac{l}{l+1}} \cdot b_i^0, \qquad i = 0, 1, \ldots, p-1.$$

Consequently, the gradient of H is null only when $b_i = \sqrt{l/l+1} \cdot b_i^0 (\, i = 0, 1, \ldots, p-1)$. Proposition 5.1 is thus proved for this (special) case.

[1] The coefficients b_i are assumed to be real.

We now assume that at least a source is present and we consider the following functional of the partial derivatives:

$$\sum_{i=0}^{p-1} b_i \frac{\partial H}{\partial b_i} = \int_{-w}^{w} \frac{\operatorname{Re}\left(\left(\sum_{i=1}^{p-1} b_i z^i\right) \bar{F}(k)\right) [lR(k) - (l+1)B(k)]}{(R(k) - B(k))\, B(k)} dk$$
$$= \int_{-w}^{w} \frac{lR(k) - (l+1)B(k)}{R(k) - B(k)}\, dk$$
$$= I_1 + I_2$$

with

$$I_1 = l \int_{-w}^{w} \frac{S(k)}{R(k) - B(k)}\, dk,$$
$$I_2 = l \int_{-w}^{w} \frac{B_0(k) - (l + 1/l)\, B(k)}{R(k) - B(k)}\, dk, \tag{37}$$

and

$$R(k) = S(k) + B_0(k), \qquad B(k) \text{ given by (34).}$$

The scalar product $\sum b_i \partial H / \partial b_i$ is thus written as the sum of the terms I_1 and I_2. We shall now examine them.

First, note that $(R(k) - B(k))$ is positive no matter what k is in the interval $[-w, +w]$. This is due to the definition of the REF and involves the barrier functional $\log(R(k) - B(k))$. Furthermore, $S(k)$ is also positive (it is a power spectral density), so that the term I_1 is always positive.

Second, now assume that a single source is present. Then

$$S(k) = \frac{\sigma^2}{|z - z_0|^2}$$

with z_0 the pole of the source, $z_0 \in D(0, 1)$.

Then the following inequality holds:

$$\int_{-w}^{w} \frac{S(k)}{R(k) - B(k)} dk \geq \frac{1}{\alpha} \int_{-w}^{w} S(k) dk = \frac{1}{\alpha} \frac{\sigma^2}{1 - |z_0|^2}. \tag{38}$$

(α : lower bound of $R(k) - B(k)$ on $[-w, +w]$, σ^2 source power.)

Let us consider the term I_2. For that purpose, it is worth partitioning the parameter domain $(\{b_i\}_i)$ into zones, as depicted below. For the sake of clarity, only the two-dimensional (2-D) case will be presented in Fig. 2.

Let us now prove that the maximum of H cannot be attained on $\mathcal{Z}_1$. More precisely, assume that the coefficients $\{b_i\}$ satisfy the following inequalities (defining $\mathcal{Z}_1$):

$$|b_i| < |b_i^0| \sqrt{\frac{l}{l+1}} \quad \text{for} \quad i = 0, 1, \ldots, p-1.$$

Now Parseval's equality asserts that

$$\int_{-w}^{w} B_0(k) = \sum_{i=0}^{p-1} \left(b_i^0\right)^2$$

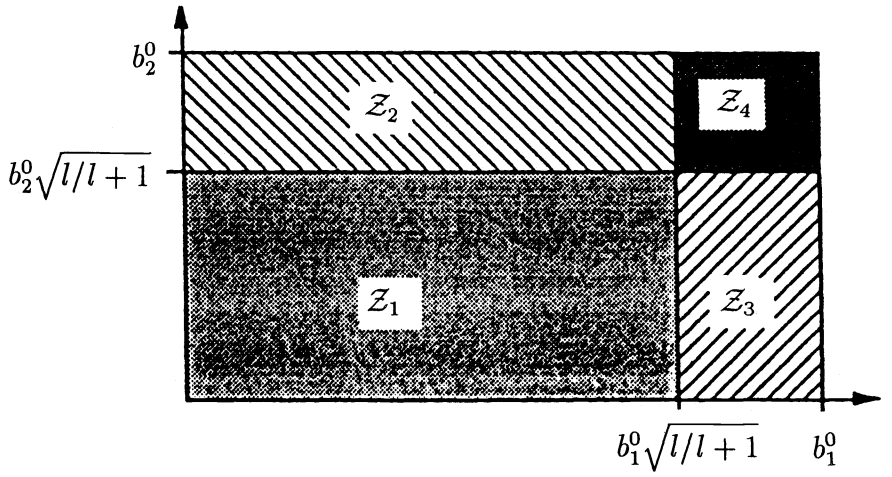


FIG. 2. *Decomposition of the positive orthant in* 4 *zones* $(\mathcal{Z}_1, \mathcal{Z}_2, \mathcal{Z}_3, \mathcal{Z}_4)$ b_1^0, b_2^0, $b_1^0\sqrt{l/l+1}$, $b_2^0\sqrt{l/l+1}$.

and, consequently,

$$\int_{-w}^{w} \left(B_0(k) - \left(\frac{l+1}{l}\right) B(k)\right) dk = \sum_{i=0}^{p-1} \left[(b_i^0)^2 - \left(\frac{l+1}{l}\right)(b_i^2)\right],$$

so that

$$|\, I_2 \,| \leq \left(\frac{1}{\alpha}\right) \cdot \left(\sum_{i=0}^{p-1} \left[(b_i^0)^2 - \left(\frac{l+1}{l}\right)(b_i^2)\right]\right). \tag{39}$$

Therefore, the term I_2 is bounded on $\mathcal{Z}_1$. So when the source pole z_0 approaches the unit circle, then I_1 tends towards $+\infty$. Note that this is equivalent to the rank deficiency hypothesis for source CM. Finally, the following result has been obtained:

$$\sum_{i=0}^{p-1} b_i \frac{\partial H}{\partial b_i} > 0 \quad \text{on} \quad \mathcal{Z}_1. \tag{40}$$

When the source contribution is null (i.e., $S = 0$), then the partial derivatives $\partial H/\partial b_i$ are null for $b_i^* = b_i^0\sqrt{l/l+1}, i = 0, \ldots, p-1$. The effect of the source is thus to displace the location of the maximum of H.

Let us now consider the zones $\mathcal{Z}_2$ and $\mathcal{Z}_3$.

If the coefficients $\{b_i\}$ belong to $\mathcal{Z}_2$ or $\mathcal{Z}_3$, then the term I_2 of (37) is not necessarily positive, but it remains bounded. Therefore when $|z_0|$ tends towards 1 (plane wave hypothesis) then one has once again

$$\sum_{i=1}^{p-1} b_i \frac{\partial H}{\partial b_i} > 0 \quad \text{on} \quad \mathcal{Z}_2, \mathcal{Z}_3. \tag{41}$$

Finally, if the coefficients $\{b_i\}$ approach their exact values $\{b_i^0\}$, then H tends toward $-\infty$. According to (40), (41) the maximum of H is thus attained on $\mathcal{Z}_4$, achieving the proof.

Obviously, the reasoning is strictly similar for the multiple source case. □

Jensen's theorem [24] can be used to calculate the REF. Thus, since $B(k)$ is analytic in $D(0,1)$ we obtain ($w = 1/2$)

$$\int_{-1/2}^{1/2} \log(B(k))dk = 2\left(\log|F(0)| - \sum_{i-1}^{p} \log(|z_i|\right),$$

where $\{z_i\}$ are the zeros of $B(k)$ inside the unit circle.

Thus, the following equality holds:

$$\int_{-1/2}^{1/2} \log(B(k))dk = 2\log(b_0).$$

It is rather surprising that the limit ($q \to \infty$) of the term $(1/q)\log(\det B(b_0,\ldots,b_p))$ is simply $2\log(b_0)$. The first part (noise alone case) of Proposition 5.1 can be proved in this way (Jensen's theorem). However, practically, this proof is restricted to first-order MA models.

Practically, $R(k)$ must be replaced by an estimate $\hat{R}(k)$, generally obtained by Fourier transform of the spatial covariances:

$$\hat{R}(k) = \sum_{j=-q+1}^{j=q-1} \hat{r}(jd)w_j \exp(2i\pi kjd), \tag{42}$$

where $\hat{r}(jd)$ are estimates of the spatial covariances.

Estimates of $\hat{r}(jd)$ are themselves obtained by replacing the exact matrix R by an orthogonal projection of the periodogram matrix [11] on the Toeplitz subspace [15,16]. The scalars w_j represent the array weighting. They are necessary for sidelobe reduction and, overall, to ensure the positivity of $\hat{R}(k)$. For this purpose, the following weighting ensures the positivity constraint of $\hat{R}(k)$:

$$\mathbf{W}^t = (1, 1-1/q, 1-2/q, \ldots, 1/q),$$

since it amounts to a consideration of $\hat{R}(k)$ defined by

$$\hat{R}(k) = \mathbf{D}_k^* \hat{R} \mathbf{D}_k.$$

The REF method can be easily extended to multifrequency analysis. Under the independence assumption, the following formulation of the REF is obtained:

$$H = \sum_{f=f\,\min}^{f\,\max} \left[\int \left[\log\left(\hat{R}(f_i,k) - B(f_i,k)\right) + l \cdot \log B(f_i,k)\right] dk\right]. \tag{43}$$

6. The whitening procedure. The more classical and direct whitening procedure consists of performing a Choleski factorization of the matrix B_* (i.e., $B_* = TT_*$, T triangular factor) and defining the whitened matrix R_w by

$$R_w = T^{-1} R T^{-1*}.$$

However, this approach suffers from some drawbacks, which may become important. Among them are the computation cost (for a large array) and the numerical conditioning if the matrix B_* is near to singularity. So the following procedure is generally preferable.

1. Determine an autoregressive (AR) model "equivalent" to the MA model. Let p' be the AR model order. The term "equivalent" means that the covariance sequence of the AR model is as close as possible to the MA ones. Usually, this is achieved by means of the Yule–Walker equation [11]. Standard procedures exist for this problem [25].

2. Consider the whitened matrix defined as follows:

$$R_w = A_w R A_w^*,$$

where A_w is a rectangular $q - p' \times q$ defined by

$$\begin{pmatrix} a_0 & \cdots & a_{p'} & 0 & \cdots & 0 \\ 0 & a_0 & \cdots & a_{p'} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & a_0 & \cdots & a_{p'} \end{pmatrix}. \tag{44}$$

This whitening method enjoys the following properties.

1. R_w is a Toeplitz matrix.

2. The transform of a source CM matrix (i.e., $\mathbf{D}_\theta \mathbf{D}_\theta^*$) is a rank-one matrix associated with the same bearing and given by

$$A_w \left(\mathbf{D}_\theta \mathbf{D}_\theta^*\right) A_w^* = q(\theta)\, \mathbf{D}_\theta' \mathbf{D}_\theta'^*$$

with

$$q(\theta) = |A(z)|^2.$$

Both the Toeplitz and plane wave structures are preserved by using this whitening procedure. In the case of a very large array, the above formula suggests the following (approximated) whitening:

$$R_w(k) = \left(\frac{1}{q(\theta)}\right) \cdot R(k).$$

7. Computation results and further comments. In this section, the behavior of the REF will be illustrated by computation and simulation results. The covariance matrix of the sensor outputs is given by

$$R = \sum_{i=1}^{s} \sigma_i^2 D_{\theta_i} D_{\theta_i}^* + B_0 \tag{45}$$

(B_0 exact noise CM; D_{θ_i} steering vector [1] associated with a source coming from the bearing θ_i and with spectral density σ_i^2).

The aim of the following results is to illustrate the REF properties.

1. *Effects of signal to noise ratios and of the factor* l. The effects of the signal to noise ratios are illustrated by Figs. 3 and 4. The eigenvalues of $B_0^{-1} B_{l,*}$ are ranked in increasing order. The index of each eigenvalue is plotted in the x-axis and its corresponding value in the y-axis. For these two figures, both the source bearings and the noise parameters are similar. They differ only by the source powers ("level" indicates the source spectral density σ_i^2).

The REF (3) is maximized by using a standard gradient algorithm, initialized on ($\beta_1 = 0.1, \beta_2 = .0, \ldots, \beta_5 = .0$) or, in other words $B_{init} = \lambda \cdot .\mathrm{Id}$ (λ is chosen "small").

Thanks to Propositions 3.1 and 4.1, the convergence of the iterative maximization algorithm is ensured no matter what initialization satisfies the constraints $\mathcal{C}$. The previous choice (for initialization) appears to the simpler. Once the gradient method has converged, a matrix $B_{l,*}$ is obtained for each value of l. For each value of l a horizontal dotted line ($y = l/l+1$) is plotted and the eigenvalues of $B_0^{-1}B_{l,*}$ are compared to this line.

Proposition 3.1 is verified no matter what the value of l and the signal to noise ratios are. The lowest eigenvalue of $B_0^{-1}B_{l,*}$ may be slightly inferior to the theoretical lower bound (i.e., $l/l+1$) because of the stopping rule of the iterative algorithm.

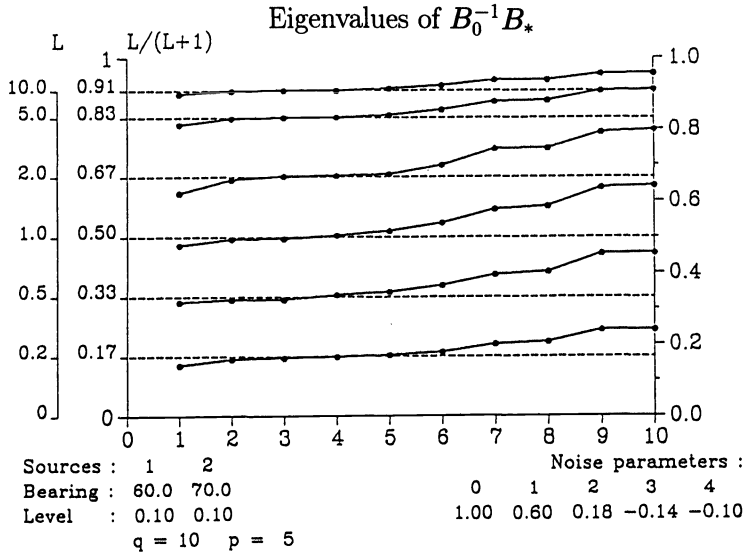


FIG. 3. *Verification of Proposition* 3.1. *Eigenvalues of the matrix* $B_0^{-1}B_{l,*}$ *for various values of* $l, p = 5, q = 10, s = 2$. *Noise parameters :* $(\beta_1 = 1, \beta_2 = 0.6, \beta_3 = 0.18, \beta_4 = -0.14, \beta_5 = -0.10)$, *two sources* (*bearings :* 60 *and* 70 *deg., powers :* 0.1 *and* 0.1).

2. *Noise modelling overdetermination and Proposition* 4.3. The proof of Proposition 3.1 shows that it still holds when the noise model is overdetermined. This fact is illustrated by Fig. 5, where we assumed that the noise model was defined by 7 parameters when the true order was 5. The REF has been maximized with respect to $\beta_1, \ldots, \beta_7$. Note that the true parameters are those of Fig. 4. Proposition 3.1 is still verified in Fig. 4 even if an effect of overdetermination is an increase in the greater eigenvalues of $B_0^{-1}B_{l,*}$, thus enlarging the dispersion of the eigenvalues. Conversely, Proposition 3.1 is not verified if the noise model is underdetermined.

The effects of a "large" source number are presented in Fig. 6. Obviously, the dispersion of the eigenvalues is enlarged, but Proposition 3.1 still holds when the hypotheses of Proposition 4.3 are not satisfied. Proposition 4.3 thus appears to be very pessimistic. Note that Conjecture 1 is valid for all these simulations.

3. *Verification of Proposition* 5.1. Proposition 5.1 is illustrated by Table 1, for which the computation parameters are

$$q = 32,$$
$$B_0 \quad MA(2) : b_0 = 1, \quad b_1 = 0.3, \quad b_2 = -0.3.$$

As can be seen in Table 1, Proposition 5.1 is verified no matter what the value of l is. The parameters $\{b_i^*\}$ have been computed by using a gradient algorithm for

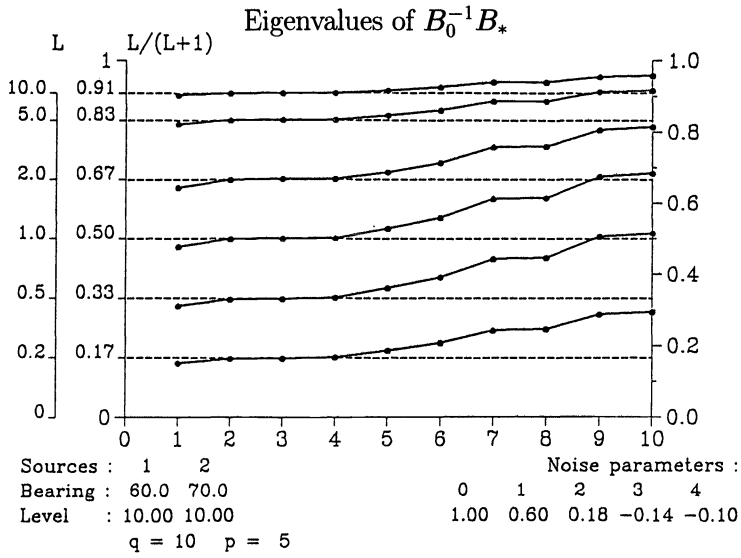


FIG. 4. *Verification of Proposition* 3.1. *Eigenvalues of the matrix* $B_0^{-1}B_{l,*}$ *for various values of* $l, p = 5, q = 10, s = 2$. *Noise parameters :* $(\beta_1 = 1, \beta_2 = 0.6, \beta_3 = 0.18, \beta_4 = -0.14, \beta_5 = -0.10)$, *two sources (bearings :* 60 *and* 70 *deg., powers :* 10 *and* 10).

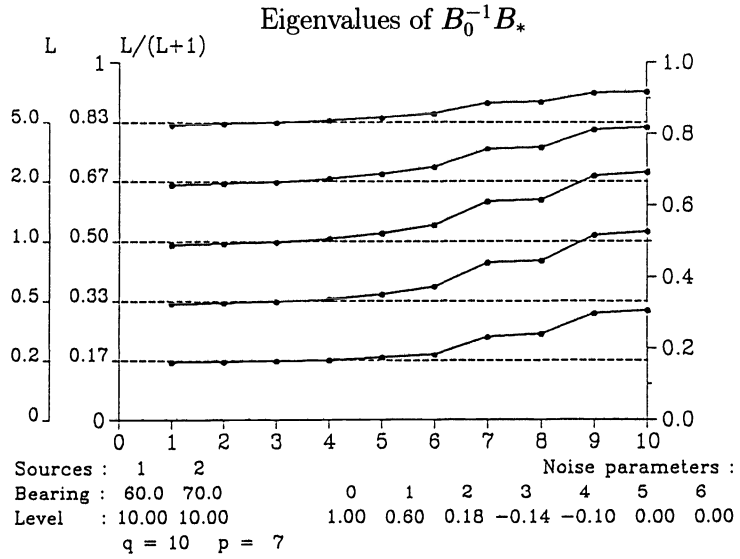


FIG. 5. *Effect of the noise model overdetermination. Eigenvalues of the matrix* $B_0^{-1}B_{l,*}$ *for various values of* $l, p = 7, q = 10, s = 2$. *Noise parameters :* $(\beta_1 = 1, \beta_2 = 0.6, \beta_3 = 0.18, \beta_4 = -0.14, \beta_5 = -0.10, \beta_6 = 0.0, \beta_7 = 0.0)$, *two sources (bearings :* 60 *and* 70 *deg., powers :* 10 *and* 10).

maximizing the REF H (30). Because it is quite direct, the calculation of the gradient is skipped. No convergence problem occurs.

4. *Simulation results.* Practically, the REF H (3) is replaced by the following functional:

$$H = \log\det\left(\hat{R} - B\right) + l\log\det B, \tag{46}$$

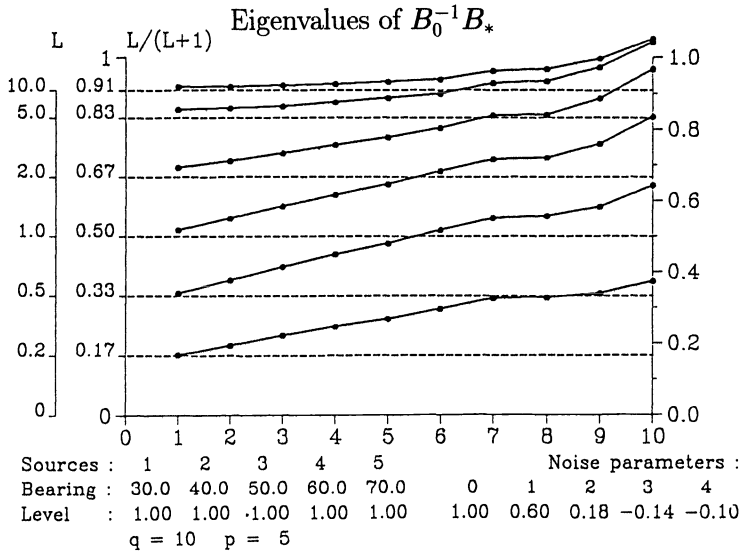


FIG. 6. *Verification of Proposition* 4.3. *Eigenvalues of the matrix* $B_0^{-1}B_{l,*}$ *for various values of* $l, p = 5, q = 10, s = 5$. *Noise parameters :* $(\beta_1 = 1, \beta_2 = 0.6, \beta_3 = 0.18, \beta_4 = -0.14, \beta_5 = -0.10)$, *five sources* (*bearings :* 30, 40, 50, 60, *and* 70 *deg., powers :* 1).

where $\hat{R}$ is an estimated CM of the sensor outputs.

The vectors $\mathbf{X}_i$ of array outputs have then been simulated. The general scheme of the simulation is presented below.

1. Let B_0 be the exact noise matrix, which performs a Choleski factorization of B_0, say,

$$B_0 = TT^*.$$

2. Let $\mathbf{Y}_i$ be a zero-mean gaussian complex vector of dimension q with covariance matrix Id; then a noise vector is $\mathbf{Y}_i' = T\mathbf{Y}_i$.

3. A source vector $\mathbf{S}_i$ is simulated by

$$\mathbf{S}_i = \sum_{j=1}^{s} \alpha_{i,j}\mathbf{D}_{\theta,j} \text{ with : } \alpha_{i,j} \ \mathcal{N}\left(0, \sigma_j^2\right).$$

The covariance matrix $\hat{R}$ is then estimated by the following.

1. $\hat{R}_1 = (1/N)\sum_{i=1}^{n} \mathbf{X}_i\mathbf{X}_i^*, \quad \mathbf{X}_i = \mathbf{S}_i + \mathbf{Y}_i'$.

2. $\hat{R} = \text{proj}(\hat{R}_1)$.

The projection is the orthogonal projection on the Toeplitz subspace, which is simply obtained by averaging along the diagonals [15–16]. The gradient algorithm is once again used for maximizing the REF (cf. §4). Since the initialization is not critical, we simply choose $B_{init} = \lambda$ Id. The value of λ must be inferior to the lowest eigenvalue of $\hat{R}$. After runs of the algorithm, noise estimates are obtained. The eigenvalues of the matrix $B_0^{-1}B_*$ are presented in Figs. 7 and 8 for 10 trials, each corresponding to $N = 300$. In other words, the snapshot number is 300.

Figures 7 and 8 correspond to the same simulated data; they differ only by the value of l. Proposition 3.1 advocates choosing a large l. This is not true for simulated data. If the value of l is 3, then Proposition 3.1 is "almost" valid. The statistical

TABLE 1
Values of $b_i^*(l)$ for various values of l.

Value of l		$\{b_i^*\}$ without source	$\{b_i^*\}$ with source	b_i^0	$b_i^0\sqrt{l/l+1}$
1	b_0	0.71	0.73	1.0	0.71
	b_1	0.21	0.22	0.3	0.21
	b_2	−0.21	−0.23	−0.3	−0.21
2	b_0	0.82	0.84	1.0	0.82
	b_1	0.24	0.25	0.3	0.24
	b_2	−0.24	−0.27	−0.3	−0.24
3	b_0	0.87	0.89	1.0	0.87
	b_1	0.26	0.26	0.3	0.26
	b_2	−0.26	−0.28	−0.3	−0.26
4	b_0	0.89	0.91	1.0	0.89
	b_1	0.27	0.27	0.3	0.27
	b_2	−0.27	−0.28	−0.3	−0.27
5	b_0	0.91	0.93	1.0	0.91
	b_1	0.27	0.28	0.3	0.27
	b_2	−0.27	−0.29	−0.3	−0.27
10	b_0	0.95	0.97	1.0	0.95
	b_1	0.29	0.29	0.3	0.29
	b_2	−0.29	−0.29	−0.3	−0.29

dispersion of the results of the various trials is rather reduced. This is not the case when the value of l is 10. It thus seems that there is an optimal value of l.

The choice of the optimal value of the parameter l results from statistical considerations relating the values of the parameters p, q, l with the statistical properties of the b_i''s estimates. Actually, the quantities defining the statistical behavior (standard deviation bias) of the b_i' estimates can be calculated by using an expansion of the $\{\hat{b}_i^*\}$ around their asymptotic values b_i^*.

This kind of calculation presents no major difficulty, but it is omitted here since it is beyond the scope of this paper. Roughly, there is a compromise between the accuracy of the b_i''s estimates (large values of l) and their variance. Thus l appears as an uncertainty factor describing the redundancy of information relative to the noise structure.

8. Conclusion. The properties of an original functional have been studied. They appear to be quite interesting, proving furthermore that the maximization of the REF is easy and reliable. The REF thus appears to be a promising method for solving an ill-posed problem.

Appendix A: A definition of the REF. We shall consider, for the REF definition, that the *physical* array is constituted of n_s equispaced sensors impinged by s sources, and we shall assume that the correlation length of the noise is null beyond p sensors.

Now consider the vector $\mathcal{B}$ defined by

$$\mathcal{B}^t = \left(\mathbf{B}_1^t, \mathbf{B}_2^t, \ldots, \mathbf{B}_l^t\right),$$

where the vectors $\mathbf{B}_i$ are (statistically) independent sample vectors of the noise im-

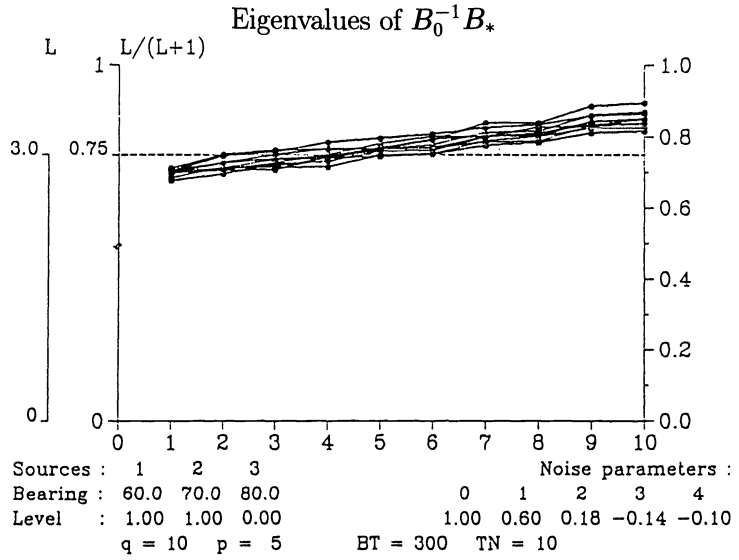


FIG. 7. *Simulated data. Eigenvalues of the matrix* $B_0^{-1}B_{l,*}$ *for* $l = 3$. *Ten trials, number of snapshots* $N = 300$, *two sources (bearings : 60 and 70 deg., powers:1 and 1), noise parameters :*$(\beta_1 = 1, \beta_2 = 0.6, \beta_3 = 0.18, \beta_4 = -0.14, \beta_5 = -0.10)$.

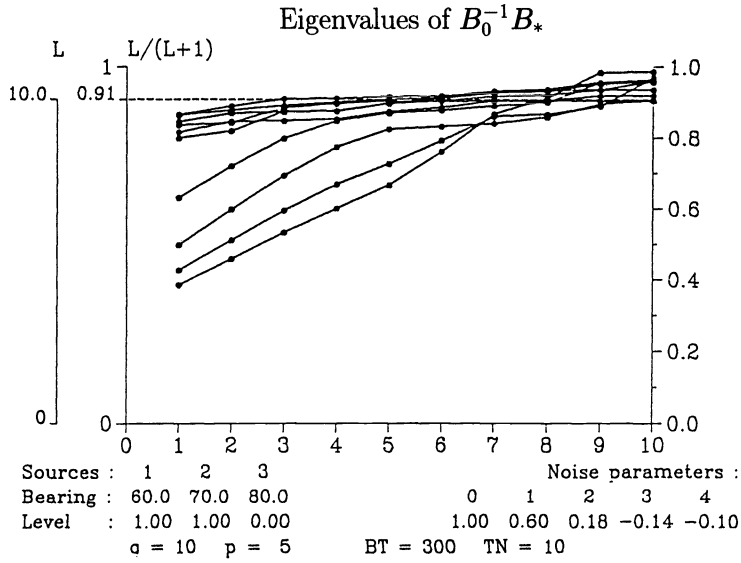


FIG. 8. *Simulated data. Eigenvalues of the matrix* $B_0^{-1}B_{l,*}$ *for* $l = 10$. *Ten trials, number of snapshots* $N = 300$, *two sources (bearings : 60 and 70 deg., powers:1 and 1), noise parameters :*$(\beta_1 = 1, \beta_2 = 0.6, \beta_3 = 0.18, \beta_4 = -0.14, \beta_5 = -0.10)$.

pinging on the array and with

(47) $\mathbf{B}_i$ q – dimensional vector.

Furthermore, let $\mathbf{X}$ be an observation vector (sensor outputs) of the same dimension q. Then denote $\mathbf{B}_i|\mathbf{X}$ to be the linear minimum variance estimate of the zero-mean random vector $\mathbf{X}$ (i.e., the orthogonal projection of the random vector $\mathbf{B}_i$

on the Hilbert space spanned by $\mathbf{X}$). The following results:

$$\mathbf{B}_i|\mathbf{X} = \mathbb{E}\left(\mathbf{B}_i\mathbf{X}^*\right)\left[\mathbb{E}\left(\mathbf{X}\mathbf{X}^*\right)\right]^{-1}\mathbf{X},$$

$\mathbb{E}$ denotes expectation, and by a slight abuse of notation (concatenation of the projections), we have

$$\mathcal{B}|\mathbf{X} = \mathbb{E}\left(\mathcal{B}\mathbf{X}^*\right)\left[\mathbb{E}\left(\mathbf{X}\mathbf{X}^*\right)\right]^{-1}\mathbf{X}. \tag{48}$$

Then a "measure" of the "uncertainty" upon $\mathcal{B}$ which is not "explained" by $\mathbf{X}$ is deduced from the conditional variance of $\mathcal{B}$ and is equal to

$$I(\mathcal{B}, \mathbf{X}) = \log\det\left[\operatorname{covar}\left(\mathcal{B} - \mathcal{B}|\mathbf{X}\right)\right]. \tag{49}$$

Actually, the observation vector $\mathbf{X}$ is the sum of a source part ($\mathbf{X}$) and a noise part ($\mathbf{B}_1$), say,

$$\mathbf{X} = \mathbf{S} + \mathbf{B_1}. \tag{50}$$

Using (48) and (49), the following expression of $I(\mathcal{B}, \mathbf{X})$ is easily derived, yielding

$$\begin{aligned} I(\mathcal{B}, \mathbf{X}) &= \log\det\left\{\begin{pmatrix} B & & & \\ \ddots & \ddots & 0 & \\ & & & \\ 0 & \ddots & & \\ & & & B \end{pmatrix} - \begin{pmatrix} B \\ 0 \\ \vdots \\ 0 \end{pmatrix} R_q^{-1}(B\ 0\ 0)\right\} \\ &= \log\det\left(R_q - B\right) + l\cdot\log\det B - \log\det R_q. \end{aligned} \tag{51}$$

Appendix B: Optimization of the stepsize. This appendix is devoted to the calculation of the optimal stepsize ρ of the gradient's algorithm on $\mathcal{C}$.

The major aim of this appendix is to obtain an explicit formulation of the REF $H\left(R, B_{k+1}\right)$. For that purpose, consider the following factorization:

$$B_k = T_k T_k^* \quad \text{and} \quad R - B_k = S_k S_k^*$$

(by assumption B_k and $R - B_k$ are positive definite) so that

$$\begin{aligned} \log\det\left(R - B_k + \rho D_k\right) &= \log\det\left(S_k S_k^* + \rho D_k\right) \\ &= \log\det\left[S_k\left(\mathrm{Id} + \rho S_k^{-1} D_k S_k^{-1*}\right) S_k^*\right] \\ &= \log\det\left(R - B_k\right) + \log\det\left(\mathrm{Id} + \rho S_k^{-1} D_k S_k^{-1*}\right). \end{aligned} \tag{52}$$

Similarly, one obtains

$$\log\det\left(B_k - \rho D_k\right) = \log\det B_k + \log\det\left(\mathrm{Id} - \rho T_k^{-1} D_k T_k^{-1*}\right). \tag{53}$$

Therefore, using (52) and (53), the following results:

$$H(\rho) = \log\det\left(\mathrm{Id} + \rho S_k^{-1} D_k S_k^{-1*}\right) + l.\log\det\left(\mathrm{Id} - \rho T_k^{-1} D_k T_k^{-1*}\right) + cst. \tag{54}$$

The two matrices $S_k^{-1} D_k S_k^{-1*}$ and $T_k^{-1} D_k T_k^{-1*}$ are hermitian and therefore diagonalizable. Let $\left\{\alpha_i^k\right\}$ and $\left\{\beta_i^k\right\}$ be their respective eigenvalues. The following explicit form of $H(\rho)$ is thus

$$H(\rho) = \sum_{i=1}^{q} \log\left(1 + \rho\alpha_i^k\right) + l.\sum_{i=1}^{q} \log\left(1 - \rho\beta_i^k\right) + cst. \tag{55}$$

The two constraints $\mathcal{C}$ (6) are translated into explicit constraints (with respect to ρ), i.e.,

$$(56) \qquad \left| \begin{array}{lll} 1+\rho\alpha_i^k > 0, & i=1,2,\ldots,q, & R-B_{k+1} \text{ positive definite,} \\ 1-\rho\beta_i^k > 0, & i=1,2,\ldots,q, & B_{k+1} \text{ positive definite.} \end{array} \right.$$

The optimal stepsize ρ_k is obtained by maximizing $H(\rho)$ (55) under the constraints (56). Practically, ρ_k is obtained by means of a unidimensional Newton method initialized at $\rho = 0$. The convergence of Newton's method on $\mathcal{C}$ is ensured since $H(\rho)$ is concave on this domain.

Appendix C: The complex case. The gradient algorithm for REF maximization will now be extended to the complex case. For that purpose, let V_i be the $q \times q$ matrix defined by

$$V_i = \left| \begin{array}{ll} 1 & \text{if} \quad l-k=i-1, \\ 0 & \text{else.} \end{array} \right.$$

The noise matrix B then takes the following form:

$$B = \beta_1 U_1 + \beta_2 V_2 + \bar{\beta}_2 V_2^t + \cdots + \beta_p V_p + \bar{\beta} p V_p^t.$$

A real gradient vector $\mathbf{G}_k$ is then defined by

$$\mathbf{G}_k = \left| \begin{array}{l} g_k^1 = -tr\,(\Delta_k U_1)\,, \\ g_k^2 = tr\,(\Delta_k (V_2 + V_2^t))\,, \\ g_k^{2'} = itr\,(\Delta_k (V_2 - V_2^t))\,, \\ \vdots \\ g_k^p = tr\,\left(\Delta_k (V_p + V_p^t)\right), \\ g_k^{p'} = itr\,\left(\Delta_k (V_p - V_p^t)\right), \end{array} \right.$$

with

$$\Delta_k = l \cdot \left({B_k}^{-1}\right) - (R - B_k)^{-1}.$$

The gradient iteration takes then the following form:

$$B_{k+1} = B_k - \rho_k D_k$$

with

$$D_k = g_k^1 U_1 + g_k^2 (V_2 + V_2^t) + i g_k^{2'} (V_2 - V_2^t) + \cdots + g_k^p (V_p + V_p^t) + i g_k^{p'} (V_p - V_p^t).$$

The rest of the algorithm is strictly similar to the real case.

Acknowledgments. A major part of this research was done while the author worked with GERDSM, a laboratory of the "Direction des Constructions Navales" (DCN). The author wishes to thank DCN for its constant support, his former colleagues of GERDSM for their friendship, and the referees for their comments.

REFERENCES

[1] W. S. BURDIC, *Underwater Acoustic System Analysis*, 2nd ed., Signal Proc. series, Prentice–Hall, Englewood Cliffs, NJ, 1991.

[2] G. BIENVENU AND L. KOPP, *Optimality of high resolution array processing using the eigensystem method*, IEEE Trans. on ASSP, ASSP–31 (1983), pp. 1235–1247.

[3] D. H. JOHNSON, *The application of spectral estimation methods to bearing estimation problems*, Proc. of the IEEE, 70 (1982), pp. 1018–1028.

[4] W. S. LIGGETT, *Passive sonar: Fitting models to multiple time series*, in Signal Processing, J.W.R. Griffiths, P.L. Stocklin, and C. Van Schooneveld, eds., Academic Press, New York, 1979, pp. 327–345.

[5] S. U. PILLAI, *Array Signal Processing*, Springer–Verlag, New York, Berlin, 1989.

[6] S. HAYKIN, ED., *Advances in Spectrum Analysis and Array Processing*, Vols. II and III, Prentice–Hall, Englewood Cliffs, NJ, 1991–1992.

[7] M. BOUVET AND G. BIENVENU, EDS., *High-resolution methods in underwater acoustic*, Lecture Notes in Control and Inform. Sci., 155, Springer–Verlag, New York, Berlin, 1991.

[8] M. S. SRIVASTAVA AND G. G. KHATRI, *An Introduction to Multivariate Statistics*, North–Holland, Amsterdam, 1979.

[9] R. J. MUIRHEAD, *Aspects of Multivariate Statistical Theory*, John Wiley, New York, 1982.

[10] C. R. RAO, *The Use and Interpretation of Principal Component Analysis in Applied Research*, Sankhya Series A, Vol. 26, 1964, pp. 329–358.

[11] S. KAY, *Modern Spectral Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1988.

[12] J. P. LE CADRE, *Parametric methods for spatial signal processing in the presence of unknown colored noise fields*, IEEE Trans. on ASSP, 37, pp. 965–983.

[13] M. WAX, *Detection and localization of multiple sources in noise with unknown covariance*, IEEE Trans. on Signal Processing, 40 (1992), pp. 245–249.

[14] W. MURRAY AND M. H. WRIGHT, *Line search procedures for the logarithmic barrier function*, SIAM J. Optim., 4 (1994), pp. 229–246.

[15] P. A. ROEBUCK AND S. BARNET, *A survey of Toeplitz and related matrices*, Internat. J. Systems Sci., 9 (1978), pp. 921–934.

[16] B. N. MUKHERJEE AND S. S. MAITI, *On some properties of positive definite Toeplitz matrices and their possible applications*, Linear Algebra Appl., 102 (1988), pp. 211–240.

[17] G. H. GOLUB AND C. F. VAN LOAN, *Matrix computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[18] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[19] P. STOICA, personal communication.

[20] G. W. STEWART AND J. G. SUN, *Matrix perturbation theory*, in Computer Science and Scientific Computing. Academic Press, New York, 1990.

[21] L. LJUNG AND T. GLAD, *On global identifiability for arbitrary model parametrizations*, Automatica, 30 (1994), pp. 265–276.

[22] U. GRENANDER AND G. SZEGO, *Toeplitz Forms and their Applications*, University of Caiifornia Press, Berkeley, CA, 1958.

[23] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices*, 2nd ed., Academic Press, New York, 1985.

[24] R. P. BOAS, *Invitation to Complex Analysis*, The Random House Birkhauser Mathematic Series, Random House, New York, 1987.

[25] A. TIKHONOV AND V. ARSENINE, *Méthodes de résolution de problémes mal posés*. Editions MIR, Moscow, 1974.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 2, pp. 383–400, April 1996

009

PERTURBATION ANALYSIS FOR TWO-SIDED (OR COMPLETE) ORTHOGONAL DECOMPOSITIONS*

RICARDO D. FIERRO†

Abstract. Two-sided (or complete) orthogonal decompositions are good alternatives to the singular value decomposition *(SVD)* because they can yield good approximations to the fundamental subspaces associated with a numerically rank-deficient matrix. In this paper we derive perturbation bounds for the subspaces associated with a general two-sided orthogonal decomposition of a numerically rank-deficient matrix. The results imply the subspaces are only *slightly more sensitive* to perturbations than singular subspaces, provided the norm of the off-diagonal blocks of the middle matrices are sufficiently small with respect to the size of the perturbation. We consider regularizing the solution to the ill-conditioned least squares problem by truncating the decomposition and present perturbation theory for the minimum norm solution of the resulting least squares problem. The main results can be specialized to well known *SVD*-based perturbation bounds for singular subspaces as well as the truncated least squares solution.

Key words. orthogonal decompositions, singular value decomposition, rank deficiency, rank revealing, subspaces, perturbation, least squares

AMS subject classifications. 65F25, 65F30

1. Introduction.

A *two-sided (or complete) orthogonal decomposition* of an $m \times n$ matrix A is a product of three matrices: an orthogonal matrix, a middle matrix, and another orthogonal matrix. For practical reasons the middle matrix is usually either triangular or diagonal. The most well known example of such a decomposition is the singular value decomposition *(SVD)*, where the middle matrix is diagonal. The *SVD* has proven to be a valuable and reliable tool in a wide variety of settings. Denote the *SVD* of A (cf. [9, §2.3]) by

$$A = U\Sigma V^T = [U_1\ U_2\ U_\perp]\,\Sigma\,[V_1\ V_2]^T \tag{1}$$

where

$$\Sigma = \begin{array}{c} \begin{array}{cc} k & n-k \end{array} \\ \left[\begin{array}{cc} \Sigma_1 & 0 \\ 0 & \Sigma_2 \\ 0 & 0 \end{array}\right] \begin{array}{l} k \\ n-k \\ m-n. \end{array} \end{array}$$

We assume $(m \geq n)$, otherwise we consider the transposed matrix A^T. The parameter k is the numerical rank of A and the singular values of A, denoted σ_i, are the diagonal elements of Σ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$. We also denote by

$$A_k \equiv U_1\Sigma_1 V_1^T$$

a rank-k matrix approximation to A, where $\Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$. The *SVD* is a very powerful tool because the algorithm is numerically stable, detects near-rank deficiency, and provides a basis for the fundamental subspaces associated with a numerically rank-deficient matrix. The sensitivity of the singular values to perturbations of A can

* Received by the editors March 14, 1994; accepted for publication (in revised form) by C. Van Loan May 26, 1995.

† Department of Mathematics, California State University, San Marcos, CA 92096 (fierro@thunder.csusm.edu).

be found, e.g., in [9, pp. 428–429]. The sensitivity of *SVD*-based subspaces, called singular subspaces, has been analyzed by Wedin [19].

The *SVD* can be used, for example, to analyze the ill-conditioned least squares (LS) problem. It is well known that when A is very ill conditioned, the minimum norm LS solution to

$$\min_{x \in \Re^n} \|b - A\,x\|_2$$

may be very sensitive to small changes in A or b, and that some form of regularization is needed to produce a useful solution. One technique is *truncated* LS, where one computes the minimum norm solution, called the truncated singular value decomposition (*TSVD*) solution, to the LS problem

$$\min_{x \in \Re^n} \|b - A_k\,x\|_2.$$

An *SVD*-based sensitivity analysis of the *TSVD* solution is given in [8].

The purpose of this paper is to analyze (with respect to perturbations of A) the sensitivity of fundamental subspaces obtained from a *general* two-sided orthogonal decomposition, hereto referred as the *UMV* decomposition. Then we regularize the solution to the ill-conditioned LS problem by truncating the *UMV* decomposition and present perturbation theory for the minimum norm solution of the resulting LS problem.

It will be very convenient to use the following notation to represent two complete orthogonal decompositions where the middle matrix is either upper or lower triangular. The decompositions are denoted

$$A = U_R R V_R^T = [U_{Rk}\ U_{R0}\ U_{R\perp}]\, R\, [V_{Rk}\ V_{R0}]^T \tag{2}$$

and

$$A = U_L L V_L^T = [U_{Lk}\ U_{L0}\ U_{L\perp}]\, L\, [V_{Lk}\ V_{L0}]^T, \tag{3}$$

where R and L have the following block structure:

$$R = \begin{array}{c} \begin{matrix} k & n-k \end{matrix} \\ \begin{bmatrix} R_k & F \\ 0 & G \\ 0 & 0 \end{bmatrix} \begin{matrix} k \\ n-k \\ m-k, \end{matrix} \end{array} \qquad L = \begin{array}{c} \begin{matrix} k & n-k \end{matrix} \\ \begin{bmatrix} L_k & 0 \\ H & E \\ 0 & 0 \end{bmatrix} \begin{matrix} k \\ n-k \\ m-k. \end{matrix} \end{array} \tag{4}$$

Both R_k and G are upper triangular matrices, while L_k and E are lower triangular. Moreover, R_k and L_k are nonsingular. We shall let

$$A_{Rk} \equiv U_{Rk} R_k V_{Rk}^T \qquad \text{and} \qquad A_{Lk} \equiv U_{Lk} L_k V_{Lk}^T$$

denote rank-k matrix approximations to A. Stewart [15, 17] devised practical algorithms for computing such decompositions, called the rank-revealing *URV* and *ULV* algorithms. The rank-revealing nature of the algorithms is achieved by condition estimation, plane rotations, and deflation procedures. These algorithms are efficient whenever k is not too much smaller than $\min(m, n)$. Fierro and Hansen [8] show how more efficient algorithms can be used if the matrix A is low rank, i.e., $k \ll \min(m, n)$. These low rank revealing algorithms rely on principal singular vector estimation (via

the Power method or the Lanczos method, for example), orthogonal transformations (via plane rotations or Householder transformations), and deflation procedures.

In this paper we shall refer to the complete orthogonal decompositions in (2) and (3) simply as *URV* and *ULV* decompositions, respectively, and we will not be concerned with the details of the algorithm used to achieve the decompositions. The rank-revealing form of (2) and (3) provide the same rank and nearly all of the same subspace information as the *SVD*, but with lower computational complexity. Stewart [16] derived sharp bounds for estimating the singular values of A using a two-sided decomposition, and Fierro and Bunch [5] derived sharp a posteriori bounds for assessing the quality of their subspaces (cf. §2.1). These results prove that the *URV* and *ULV* decompositions not only reveal the numerical rank of the matrix but also provide good approximations to the singular subspaces as part of the factorization. In addition, they are attractive because they can be updated in $\mathcal{O}(n^2)$ flops, compared with $\mathcal{O}(n^3)$ flops for the *SVD*. This is particularly important in recursive problems that arise in signal processing applications. The efficiency, stability, parallel, and updating/downdating properties of the algorithms have stimulated many to investigate the feasibility of the decompositions for various applications. To date, the applicability has been examined in updating and downdating [2, 13, 17], subspace tracking [15], accuracy in approximating least squares solutions [7], an efficient and parallelizable total least squares algorithm [18], large-scale sparse factorizations [8, 14], direction-of-arrival estimation problems [1, 11], and information retrieval [3].

This paper is motivated by the promising possibilities of the *URV* and *ULV* decompositions in situations where the *SVD* is typically applied. For $M = R$ or L, we are primarily interested in a sensitivity analysis for $\mathcal{R}(V_{Mk})$ and $\mathcal{R}(V_{M0})$, which approximates the numerical rowspace and nullspace of A, respectively, and $\mathcal{R}(U_{Mk})$, which approximates the numerical range of A. These subspaces are relevant to applications that involve least squares, total least squares, matrix approximation, subspace tracking, etc. and thus it is important to identify the parameters that influence the sensitivity of the *URV* or *ULV*-based subspaces to perturbations.

The paper is organized as follows. In §2 we briefly review approximation properties of the previously mentioned *URV*- or *ULV*-based subspaces to illuminate the role of the block elements of the triangular matrices in the quality of the subspaces. Then we review perturbation bounds for the *SVD*-based subspaces by Wedin [19]. In §3 we identify the influencing parameters which provide insight into the *sensitivity* of the subspaces associated with a general two-sided orthogonal decomposition to perturbations. By a general two-sided orthogonal decomposition we mean the middle matrix, denoted M, has the 3×2 block partition

$$M = \begin{array}{c} \begin{matrix} k & n-k \end{matrix} \\ \begin{bmatrix} S_k & F \\ H & C \\ 0 & 0 \end{bmatrix} \begin{matrix} k \\ n-k \\ m-k. \end{matrix} \end{array}$$

The perturbation bounds illuminate the role of the block matrices of the middle matrix in the sensitivity of the subspaces. The bounds in §3 can be specialized to Wedin's well-known perturbation bounds in connection with the *SVD* [19]. In §4 we regularize the ill-conditioned LS problem by truncating the two-sided orthogonal decomposition and derive perturbation bounds for the minimum norm LS solution (which we call the *TUMV* solution). We include bounds in terms of subspace angles to give a clear perspective on the role of the subspaces in the sensitivity of the truncated

solution. The result can be specialized to Hansen's perturbation bounds for the *TSVD* solution [10]. In §5 we provide an example, and in §6 we summarize our primary results.

Throughout this paper $\|\cdot\|$ represents any orthogonally invariant norm, unless otherwise specified. $\sigma(D)$ denotes the set of singular values of matrix D, $\mathcal{R}(D)$ and $\mathcal{N}(D)$ denote the range (column space) and nullspace of the matrix D, respectively. $\sigma_{\min}(D)$ denotes the smallest singular value of D and $\sigma_{\max}(D)$ denotes the largest singular value of D. Finally, the superscript $\dagger$ denotes the pseudoinverse, while T denotes the transpose.

2. Review. The orthogonal projector onto a subspace $\mathcal{S}$ is denoted $P_{\mathcal{S}}$, and the projector onto its orthogonal complement is denoted $P_{\mathcal{S}}^{\perp} = I - P_{\mathcal{S}}$. For two equidimensional subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ we define

$$\sin\Theta(\mathcal{S}_1, \mathcal{S}_2) \equiv \|(I - P_{\mathcal{S}_1})P_{\mathcal{S}_2}\| = \|(I - P_{\mathcal{S}_2})P_{\mathcal{S}_1}\|.$$

In the 2-norm this represents the *distance* between the subspaces, cf. [9, p. 76]. If $\mathcal{S}_2$ is viewed as a *perturbation* of $\mathcal{S}_1$, then $\sin\Theta(\mathcal{S}_1, \mathcal{S}_2)$ characterizes the *sensitivity* of the subspace $\mathcal{S}_1$ to perturbations. In §2.1 we will consider the quality or closeness of the *URV*- and *ULV*-based subspaces to singular subspaces. In §2.2 we will review the well-known perturbation bounds for singular subspaces.

2.1. A posteriori error bounds. As mentioned earlier, the *URV* and *ULV* decompositions possess many nice properties that nearly demand it be considered as a possible substitute to the *SVD* in some applications. In many applications both rank and subspace information must be determined. Stewart [16] showed how one can infer the numerical rank of the matrix from such decompositions, and the bounds can easily be turned into a posteriori bounds. Fierro and Bunch [5] proved the following result which shows how one can determine the quality of the *URV*- and *ULV*-based subspaces as compared with the *SVD*-based subspaces. The bounds give insight to the role of the gap in the singular values of the diagonal blocks of the middle matrix as well as the norm of the off-diagonal blocks.

THEOREM 2.1 (Fierro and Bunch [5]). *Let A have the* SVD *in* (1) *and the* URV *and* ULV *decompositions as in* (2) *and* (3). *Assume $\|\cdot\| = \|\cdot\|_2$. If $\|E\| < \sigma_{\min}(L_k)$, then*

$$\sin\Theta(\mathcal{R}(A_{Lk}^T), \mathcal{R}(A_k^T)) \leq \frac{\|H\|\,\|E\|}{\sigma_{\min}^2(L_k) - \|E\|^2}$$

and

$$\frac{\|H\|}{\|L\| + \|E\|} \leq \sin\Theta(\mathcal{R}(A_{Lk}), \mathcal{R}(A_k)) \leq \frac{\|H\|}{\sigma_{\min}(L_k) - \|E\|}.$$

If $\|G\| < \sigma_{\min}(R_k)$, then

$$\frac{\|F\|}{\|R\| + \|G\|} \leq \sin\Theta(\mathcal{R}(A_{Rk}^T), \mathcal{R}(A_k^T)) \leq \frac{\|F\|}{\sigma_{\min}(R_k) - \|G\|}$$

and

$$\sin\Theta(\mathcal{R}(A_{Rk}), \mathcal{R}(A_k)) \leq \frac{\|F\|\,\|G\|}{\sigma_{\min}^2(R_k) - \|G\|^2}.$$

These bounds guarantee that as the off-diagonal blocks F or H decrease then the *URV*- or *ULV*-based subspaces correspondingly converge to their *SVD* counterparts. The bounds also reveal that there is a lower limit in the closeness of certain subspaces. In §3 these bounds will be extended to orthogonally invariant norms when we examine the sensitivity of *URV*- and *ULV*-based subspaces to perturbations in a more general setting.

2.2. *SVD* perturbation bounds. Now we shall review perturbation bounds in connection with the *SVD*. Let $\tilde{A} = A + \delta A$ represent a *perturbation* of A. Denote the *SVD* of $\tilde{A}$ by

$$\tilde{A} = \tilde{U}\tilde{\Sigma}\tilde{V}^T = [\tilde{U}_1\ \tilde{U}_2\ \tilde{U}_\perp]\tilde{\Sigma}[\tilde{V}_1\ \tilde{V}_2]^T \tag{5}$$

where $\tilde{\Sigma}$ is partitioned according to Σ. Defining $\tilde{A}_k \equiv \tilde{U}_1\tilde{\Sigma}_1\tilde{V}_1^T$, we wish to find good upper bounds for

$$\sin\Theta(\mathcal{R}(A_k^T), \mathcal{R}(\tilde{A}_k^T)) \quad \text{and} \quad \sin\Theta(\mathcal{R}(A_k), \mathcal{R}(\tilde{A}_k))$$

to obtain an idea of the sensitivity of these important subspaces associated with the *SVD*. These quantities are very important, for example, in characterizing the sensitivity of *TSVD* solutions; cf. §4.

We will need the *residuals* $A\tilde{X}_1 - \tilde{Y}_1\tilde{D}_1$ and $A^T\tilde{Y}_1 - \tilde{X}_1\tilde{D}_1^T$, where the columns of $\tilde{X}_1$ form an orthonormal basis for $\mathcal{R}(\tilde{A}_k^T)$, the columns of $\tilde{Y}_1$ form an orthonormal basis for $\mathcal{R}(\tilde{A}_k)$, and $\tilde{D}_1 \equiv \tilde{Y}_1^T\delta A\tilde{X}_1$. Note that

$$A\tilde{X}_1 - \tilde{Y}_1\tilde{D}_1 \;=\; -\delta A\tilde{X}_1 \quad \text{and} \quad A^T\tilde{Y}_1 - \tilde{X}_1\tilde{D}_1^T \;=\; -\delta A^T\tilde{Y}_1.$$

We now state the perturbation bounds in connection with the *SVD*.

THEOREM 2.2 (Wedin [19]). *Assume there exists a* $\delta > 0$ *and* $\alpha \geq 0$ *such that*

$$\sigma_{\min}(\tilde{\Sigma}_1) \geq \alpha + \delta \quad \textit{and} \quad \sigma_{\max}(\Sigma_2) \leq \alpha.$$

Take $\epsilon = \max\{\|\delta A\tilde{X}_1\|, \|\delta A^T\tilde{Y}_1\|\}$. *Then for every unitary invariant norm we have*

$$\sin\Theta(\mathcal{R}(A_k^T), \mathcal{R}(\tilde{A}_k^T)) \;\leq\; \frac{\epsilon}{\delta} \quad \textit{and} \quad \sin\Theta(\mathcal{R}(A_k), \mathcal{R}(\tilde{A}_k)) \;\leq\; \frac{\epsilon}{\delta}.$$

Using the definitions of ϵ and δ and perturbation theory of singular values, Theorem 2.2 can be quantified in terms of the size of the perturbation $\|\delta A\|$: If $\|\delta A\| < \sigma_k - \sigma_{k+1}$, then

$$\sin\Theta(\mathcal{R}(A_k^T), \mathcal{R}(\tilde{A}_k^T)) \leq \frac{\|\delta A\|}{\sigma_k - \sigma_{k+1} - \|\delta A\|} \text{ and} \tag{6}$$

$$\sin\Theta(\mathcal{R}(A_k), \mathcal{R}(\tilde{A}_k)) \leq \frac{\|\delta A\|}{\sigma_k - \sigma_{k+1} - \|\delta A\|}.$$

Therefore, the singular subspaces associated with the cluster of singular values $\{\sigma_i\}_{i=1}^k$ are relatively insensitive to small perturbations ($\|\delta A\| \ll \sigma_k - \sigma_{k+1}$). There is also a *complementary* version of Theorem 2.2; cf. [19] and §3.

3. Perturbation bounds for *UMV*-based subspaces. Let $\tilde{A} = A + \delta A$ have the corresponding complete orthogonal decomposition

$$\tilde{A} = \tilde{U}_R\tilde{R}\tilde{V}_R^T = [\tilde{U}_{Rk}\ \tilde{U}_{R0}\ \tilde{U}_{R\perp}]\,\tilde{R}\,[\tilde{V}_{Rk}\ \tilde{V}_{R0}]^T \tag{7}$$

where $\tilde{R}$ is partitioned according to (4). We want to estimate the sensitivity of the *URV*-based subspaces when we have estimates for $\|\delta A\|$, the gap between the least singular value of $\tilde{R}_k$ and the largest singular value of G, and the size of the off-diagonal blocks F and $\tilde{F}$. In a similar vein, if

$$\tilde{A} = \tilde{U}_L \tilde{L} \tilde{V}_L^T = [\tilde{U}_{Lk}\ \tilde{U}_{L0}\ \tilde{U}_{L\perp}]\, \tilde{L}\, [\tilde{V}_{Lk}\ \tilde{V}_{L0}]^T, \tag{8}$$

we are equally interested in the sensitivity of the *ULV*-based subspaces.

As we will see, the perturbation theory for the *URV*- and *ULV*-based subspaces can be derived with a single treatment if we consider the most general two-sided orthogonal decomposition, which we denote the *UMV* decomposition:

$$A = U_M M V_M^T = [U_{Mk}\ U_{M0}\ U_{M\perp}]\, M\, [V_{Mk}\ V_{M0}]^T. \tag{9}$$

Here, the "middle" matrix M has the following 3×2 block partition

$$M = \begin{array}{c} \begin{matrix} k & n-k \end{matrix} \\ \begin{bmatrix} S_k & F \\ H & C \\ 0 & 0 \end{bmatrix} \begin{matrix} k \\ n-k \\ m-k. \end{matrix} \end{array}$$

We do not place any restrictions on the block elements of M except that S_k is required to be nonsingular (w.l.o.g.). The *UMV* decomposition of the perturbed matrix $\tilde{A}$ is denoted

$$\tilde{A} = \tilde{U}_M \tilde{M} \tilde{V}_M^T = [\tilde{U}_{Mk}\ \tilde{U}_{M0}\ \tilde{U}_{M\perp}]\, \tilde{M}\, [\tilde{V}_{Mk}\ \tilde{V}_{M0}]^T \tag{10}$$

where $\tilde{M}$ is partitioned according to M. If we define the rank-k matrices

$$A_{Mk} \equiv U_{Mk} S_k V_{Mk}^T \quad \text{and} \quad \tilde{A}_{Mk} \equiv \tilde{U}_{Mk} \tilde{S}_k \tilde{V}_{Mk}^T, \tag{11}$$

it follows that

$$A = A_{Mk} + U_{M0} H V_{Mk}^T + U_{Mk} F V_{M0}^T + U_{M0} C V_{M0}^T$$

and

$$\tilde{A} = \tilde{A}_{Mk} + \tilde{U}_{M0} \tilde{H} \tilde{V}_{Mk}^T + \tilde{U}_{Mk} \tilde{F} \tilde{V}_{M0}^T + \tilde{U}_{M0} \tilde{C} \tilde{V}_{M0}^T.$$

We wish to derive good upper bounds for

$$\sin \Theta(\mathcal{R}(A_{Mk}^T), \mathcal{R}(\tilde{A}_{Mk}^T)) \quad \text{and} \quad \sin \Theta(\mathcal{R}(A_{Mk}), \mathcal{R}(\tilde{A}_{Mk})),$$

and it should be clear that good upper bounds for the *URV*- and *ULV*-based subspaces will immediately follow. As mentioned earlier, these quantities play an important role in understanding the sensitivity of truncated least squares solutions obtained by truncating the two-sided orthogonal decomposition; cf. §4.

In the spirit of Wedin [19] we will use the residuals $-\delta A \tilde{X}_{Mk}$ and $-\delta A^T \tilde{Y}_{Mk}$ in the upper bounds, where the columns of $\tilde{X}_{Mk}$ form an orthonormal basis for $\mathcal{R}(\tilde{A}_{Mk}^T)$ and the columns of $\tilde{Y}_{Mk}$ form an orthonormal basis for $\mathcal{R}(\tilde{A}_{Mk})$. If we define $\tilde{D}_{Mk} \equiv \tilde{Y}_{Mk}^T \delta A \tilde{X}_{Mk}$, then the *UMV* residuals $-\delta A \tilde{X}_{Mk}$ and $-\delta A^T \tilde{Y}_{Mk}$ are given by

$$\begin{aligned} -\delta A \tilde{X}_{Mk} &= (A \tilde{X}_{Mk} - \tilde{Y}_{Mk} \tilde{D}_{Mk}) - \tilde{U}_{M0} \tilde{H} \tilde{V}_{Mk}^T \tilde{X}_{Mk}, \\ -\delta A^T \tilde{Y}_{Mk} &= (A^T \tilde{Y}_{Mk} - \tilde{X}_{Mk} \tilde{D}_{Mk}^T) - \tilde{V}_{M0} \tilde{F}^T \tilde{U}_{Mk}^T \tilde{Y}_{Mk}. \end{aligned}$$

These residuals will play a crucial role in the following sensitivity analysis through parameter ϵ_M, defined by

$$\epsilon_M \equiv \max\left\{\|\delta A\tilde{X}_{Mk}\|, \|\delta A^T\tilde{Y}_{Mk}\|\right\}. \tag{12}$$

Note that ϵ_M is the maximum norm of the projection of δA into $\mathcal{R}(\tilde{X}_{Mk})$ or $\mathcal{R}(\tilde{Y}_{Mk})$, and

$$\epsilon_M \leq \|\delta A\|.$$

We are now ready to state the main result of this section.

THEOREM 3.1. *Let A and $\tilde{A}$ have the* UMV *decompositions as in* (9) *and* (10), *respectively. Assume there exists a $\delta_M > 0$ and $\alpha_M \geq 0$ such that*

$$\sigma_{\min}(\tilde{S}_k) \geq \alpha_M + \delta_M \qquad \text{and} \qquad \sigma_{\max}(C) \leq \alpha_M.$$

Take $\epsilon_M = \max\{\|\delta A\tilde{X}_{Mk}\|, \|\delta A^T\tilde{Y}_{Mk}\|\}$. Then for every unitary invariant norm we have

$$\sin\Theta(\mathcal{R}(A_{Mk}^T), \mathcal{R}(\tilde{A}_{Mk}^T)) \leq \frac{\|F\| + \|\tilde{F}\|}{\delta_M} + \frac{(\|H\| + \|\tilde{H}\|)}{\delta_M} \times \frac{\sigma_{\max}(C)}{\sigma_{\min}(\tilde{S}_k) + \sigma_{\max}(C)} + \frac{\epsilon_M}{\delta_M}$$

and

$$\sin\Theta(\mathcal{R}(A_{Mk}), \mathcal{R}(\tilde{A}_{Mk})) \leq \frac{\|H\| + \|\tilde{H}\|}{\delta_M} + \frac{(\|F\| + \|\tilde{F}\|)}{\delta_M} \times \frac{\sigma_{\max}(C)}{\sigma_{\min}(\tilde{S}_k) + \sigma_{\max}(C)} + \frac{\epsilon_M}{\delta_M}.$$

Proof. To bound $\sin\Theta(\mathcal{R}(A_{Mk}^T), \mathcal{R}(\tilde{A}_{Mk}^T))$, we proceed by first finding a tractable expression for the matrix $\tilde{V}_{Mk}^T V_{M0}$:

$$\begin{aligned}
\tilde{V}_{Mk}^T V_{M0} =& \tilde{S}_k^{-1}(\tilde{U}_{Mk}^T\tilde{A} - \tilde{F}\tilde{V}_{M0}^T)V_{M0} \\
=& \tilde{S}_k^{-1}(\tilde{U}_{Mk}^T(A + \delta A)V_{M0} - \tilde{F}\tilde{V}_{M0}^T V_{M0}) \\
=& \tilde{S}_k^{-1}(\tilde{U}_{Mk}^T U_{Mk}F + \tilde{U}_{Mk}^T U_{M0}C + \tilde{U}_{Mk}^T\delta A V_{M0} - \tilde{F}\tilde{V}_{M0}^T V_{M0}) \\
=& \tilde{S}_k^{-1}(\tilde{U}_{Mk}^T U_{Mk}F - \tilde{F}\tilde{V}_{M0}^T V_{M0}) + \tilde{S}_k^{-1}(\delta A^T\tilde{U}_{Mk})^T V_{M0} + \tilde{S}_k^{-1}\tilde{U}_{Mk}^T U_{M0}C \\
=& \tilde{S}_k^{-1}(\tilde{U}_{Mk}^T U_{Mk}F - \tilde{F}\tilde{V}_{M0}^T V_{M0}) + \tilde{S}_k^{-1}(\delta A^T\tilde{U}_{Mk})^T V_{M0} \\
&+ \tilde{S}_k^{-1}\tilde{U}_{Mk}^T[U_{M0}\ U_{M\perp}]\begin{bmatrix} C \\ 0 \end{bmatrix}.
\end{aligned}$$

It is easy to show $\|\delta A^T\tilde{U}_{Mk}\| = \|\delta A^T\tilde{Y}_{Mk}\|$. Further, we have

$$\|\tilde{U}_{Mk}^T[U_{M0}\ U_{M\perp}]\| = \|[U_{M0}\ U_{M\perp}]^T\tilde{U}_{Mk}\| = \sin\Theta(\mathcal{R}(A_{Mk}), \mathcal{R}(\tilde{A}_{Mk})).$$

Occasionally we shall use the following result in the proof (cf. [12]):

$$\|CD\| \leq \|C\|_2\,\|D\| \leq \|C\|\,\|D\|.$$

By taking norms in a straightforward way it follows that

$$\begin{aligned}
\sin\Theta(\mathcal{R}(A_{Mk}^T), \mathcal{R}(\tilde{A}_{Mk}^T)) \leq\ & \|\tilde{S}_k^{-1}\|_2\,(\|F\| + \|\tilde{F}\|) + \|\tilde{S}_k^{-1}\|_2\,\epsilon_M \\
&+ \|\tilde{S}_k^{-1}\|_2\,\|C\|_2\,\sin\Theta(\mathcal{R}(A_{Mk}), \mathcal{R}(\tilde{A}_{Mk})).
\end{aligned} \tag{13}$$

Now we need to bound $\sin\Theta(\mathcal{R}(A_{Mk}),\mathcal{R}(\tilde{A}_{Mk}))$, and we proceed by finding an expression for the matrix $P^{\perp}_{\mathcal{R}(A_{Mk})}\tilde{U}_{Mk}\tilde{U}^T_{Mk}$:

$$\begin{aligned}
P^{\perp}_{\mathcal{R}(A_{Mk})}\tilde{U}_{Mk}\tilde{U}^T_{Mk} &= P^{\perp}_{\mathcal{R}(A_{Mk})}\tilde{A}_{Mk}\tilde{A}^{\dagger}_{Mk}\\
&= P^{\perp}_{\mathcal{R}(A_{Mk})}(\tilde{A}-\tilde{U}_{M0}\tilde{H}\tilde{V}^T_{Mk}-\tilde{U}_{Mk}\tilde{F}\tilde{V}^T_{M0}-\tilde{U}_{M0}\tilde{C}\tilde{V}^T_{M0})\tilde{A}^{\dagger}_{Mk}\\
&= P^{\perp}_{\mathcal{R}(A_{Mk})}(A+\delta A-\tilde{U}_{M0}\tilde{H}\tilde{V}^T_{Mk})\tilde{A}^{\dagger}_{Mk}\\
&= P^{\perp}_{\mathcal{R}(A_{Mk})}(U_{M0}HV^T_{Mk}+U_{M0}CV^T_{M0}+\delta A-\tilde{U}_{M0}\tilde{H}\tilde{V}^T_{Mk})\tilde{A}^{\dagger}_{Mk}\\
&= (U_{M0}HV^T_{Mk}\tilde{V}_{Mk}+U_{M0}CV^T_{M0}\tilde{V}_{Mk}+P^{\perp}_{\mathcal{R}(A_{Mk})}\delta A\tilde{V}_{Mk}\\
&\qquad -P^{\perp}_{\mathcal{R}(A_{Mk})}\tilde{U}_{M0}\tilde{H})\tilde{S}_k^{-1}\tilde{U}_{Mk}\\
&= (U_{M0}HV^T_{Mk}\tilde{V}_{Mk}-P^{\perp}_{\mathcal{R}(A_{Mk})}\tilde{U}_{M0}\tilde{H})\tilde{S}_k^{-1}\tilde{U}_{Mk}\\
&\qquad +(P^{\perp}_{\mathcal{R}(A_{Mk})}\delta A\tilde{V}_{Mk}+U_{M0}CV^T_{M0}\tilde{V}_{Mk})\tilde{S}_k^{-1}\tilde{U}_{Mk}.
\end{aligned}$$

Therefore, since $\|\delta A\tilde{V}_{Mk}\| = \|\delta A\tilde{X}_{Mk}\|$, it follows that

$$\begin{aligned}
\sin\Theta(\mathcal{R}(A_{Mk}),\mathcal{R}(\tilde{A}_{Mk})) &\le (\|H\|+\|\tilde{H}\|)\,\|\tilde{S}_k^{-1}\|_2+\|\tilde{S}_k^{-1}\|_2\,\epsilon_M\\
&\quad +\|C\|_2\,\|\tilde{S}_k^{-1}\|_2\,\sin\Theta(\mathcal{R}(A^T_{Mk}),\mathcal{R}(\tilde{A}^T_{Mk})).
\end{aligned}\tag{14}$$

Substituting (14) into (13) we get

$$\begin{aligned}
\sin\Theta(\mathcal{R}(A^T_{Mk}),\mathcal{R}(\tilde{A}^T_{Mk})) &\le \|\tilde{S}_k^{-1}\|_2\,(\|F\|+\|\tilde{F}\|)+\|\tilde{S}_k^{-1}\|_2\,(1+\|C\|_2\,\|\tilde{S}_k^{-1}\|_2)\,\epsilon_M\\
&\quad +\|C\|_2\,\|\tilde{S}_k^{-1}\|_2^2\,(\|H\|+\|\tilde{H}\|)\\
&\quad +\|C\|_2^2\,\|\tilde{S}_k^{-1}\|_2^2\,\sin\Theta(\mathcal{R}(A^T_{Mk}),\mathcal{R}(\tilde{A}^T_{Mk})).
\end{aligned}\tag{15}$$

Solving for $\sin\Theta(\mathcal{R}(A^T_{Mk}),\mathcal{R}(\tilde{A}^T_{Mk}))$ and using the fact $\|\tilde{S}_k^{-1}\|_2/(\|\tilde{S}_k^{-1}\|_2+\|C\|_2) \le 1$, we get the final result

$$\begin{aligned}
\sin\Theta(\mathcal{R}(A^T_{Mk}),\mathcal{R}(\tilde{A}^T_{Mk})) &\le \frac{\|F\|+\|\tilde{F}\|}{\sigma_{\min}(\tilde{S}_k)-\sigma_{\max}(C)}+\frac{(\|H\|+\|\tilde{H}\|)\,\sigma_{\max}(C)}{\sigma^2_{\min}(\tilde{S}_k)-\sigma^2_{\max}(C)}\\
&\quad +\frac{\epsilon_M}{\sigma_{\min}(\tilde{S}_k)-\sigma_{\max}(C)}.
\end{aligned}\tag{16}$$

On the other hand, if we substitute (13) into (14) and solve for $\sin\Theta(\mathcal{R}(A_{Mk}),\mathcal{R}(\tilde{A}_{Mk}))$ instead we get

$$\begin{aligned}
\sin\Theta(\mathcal{R}(A_{Mk}),\mathcal{R}(\tilde{A}_{Mk})) &\le \frac{(\|H\|+\|\tilde{H}\|)\,\|\tilde{S}_k^{-1}\|_2}{1-\|C\|_2^2\,\|\tilde{S}_k^{-1}\|_2^2}+\frac{\|\tilde{S}_k^{-1}\|_2\,(1+\|C\|_2\,\|\tilde{S}_k^{-1}\|_2)\,\epsilon_M}{1-\|C\|_2^2\,\|\tilde{S}_k^{-1}\|_2^2}\\
&\quad +\frac{\|C\|_2\,\|\tilde{S}_k^{-1}\|_2^2\,(\|F\|+\|\tilde{F}\|)}{1-\|C\|_2^2\,\|\tilde{S}_k^{-1}\|_2^2}\\
&\le \frac{\|H\|+\|\tilde{H}\|}{\sigma_{\min}(\tilde{S}_k)-\sigma_{\max}(C)}+\frac{(\|F\|+\|\tilde{F}\|)\,\sigma_{\max}(C)}{\sigma^2_{\min}(\tilde{S}_k)-\sigma^2_{\max}(C)}\\
&\quad +\frac{\epsilon_M}{\sigma_{\min}(\tilde{S}_k)-\sigma_{\max}(C)}.
\end{aligned}\tag{17}$$

Again, we have used the mild overestimate $\|\tilde{S}_k^{-1}\|_2/(\|\tilde{S}_k^{-1}\|_2+\|C\|_2) \le 1$. The assumptions on α_M and δ_M imply

$$\frac{1}{\sigma_{\min}(\tilde{S}_k) - \sigma_{\max}(C)} \leq \frac{1}{\delta_M} \quad \text{and} \quad \frac{1}{\sigma^2_{\min}(\tilde{S}_k) - \sigma^2_{\max}(C)} \leq \frac{1}{\delta_M} \times \frac{1}{\sigma_{\min}(\tilde{S}_k) + \sigma_{\max}(C)}.$$

From this and (16), (17) the desired results immediately follow. □

Remark 1. If α_M is an upper bound for $\sigma_{\max}(C)$, then Theorem 3.1 has something meaningful to say for all perturbations δA such that $\alpha_M < \sigma_{\min}(\tilde{S}_k)$. An obvious choice for α_M is $\sigma_{\max}(C)$.

Remark 2. The bounds in Theorem 3.1 can be *tightened* by using $\sigma_{\min}(\tilde{S}_k) - \sigma_{\max}(C)$ instead of δ_M. Alternatively, the bounds can be overestimated by substituting $\sigma_{\min}(\tilde{S}_k) + \sigma_{\max}(C)$ with δ_M or $\sigma_{\min}(\tilde{S}_k) - \sigma_{\max}(C)$, and by bounding $\sigma_{\max}(C)/(\sigma_{\min}(\tilde{S}_k) + \sigma_{\max}(C))$ by 1.

Remark 3. Upper bounds for $1/(\sigma_{\min}(\tilde{S}_k) - \sigma_{\max}(C))$ are

$$\frac{1}{\sigma_{\min}(\tilde{S}_k) - \sigma_{\max}(C)} \leq \frac{1}{\sigma_{\min}(S_k) - \sigma_{\max}(C) - \|F\| - \|H\| - \|\tilde{F}\| - \|\tilde{H}\| - \|\delta A\|}$$

and

$$\frac{1}{\sigma_{\min}(\tilde{S}_k) - \sigma_{\max}(C)} \leq \frac{1}{\sigma_k - \sigma_{k+1} - \|F\| - \|H\| - \|\tilde{F}\| - \|\tilde{H}\| - \|\delta A\|}.$$

Remark 4. Theorem 3.1 specializes to Wedin's perturbation bounds for singular subspaces in Theorem 2.2: by setting $\|F\| = \|\tilde{F}\| = \|H\| = \|\tilde{H}\| = 0$, it immediately follows that

(a) from the definitions $\epsilon = \epsilon_M$ and $\delta = \delta_M$, and
(b) the corresponding subspaces coincide (e.g., $\mathcal{R}(A_k^T) = \mathcal{R}(A_{Mk}^T)$ and $\mathcal{R}(\tilde{A}_k) = \mathcal{R}(\tilde{A}_{Mk})$).

Remark 5. The bounds also permit a comparison between approximate numerical nullspaces of any two competing two-sided orthogonal decompositions. For example, if we consider the singular subspaces of A and either the *URV*- or *ULV*-based subspaces of A, then we can immediately deduce the upper bounds in Theorem 2.1. Or, we can compare singular subspaces of A with the corresponding *UMV*-based subspaces of $\tilde{A}$ to get

$$\begin{aligned}\sin\Theta(\mathcal{R}(A_k^T), \mathcal{R}(\tilde{A}_{Mk}^T)) &\leq \frac{\|\tilde{F}\| + \|\tilde{H}\| + \|\delta A\|}{\sigma_{\min}(\tilde{S}_k) - \sigma_{k+1}} \\ &\leq \frac{\|\tilde{F}\| + \|\tilde{H}\| + \|\delta A\|}{\sigma_k - \sigma_{k+1} - \|\tilde{F}\| - \|\tilde{H}\| - \|\delta A\|}\end{aligned}$$

and

$$\sin\Theta(\mathcal{R}(A_k),\mathcal{R}(\tilde{A}_{Mk})) \leq \frac{\|\tilde{F}\|+\|\tilde{H}\|+\|\delta A\|}{\sigma_{\min}(\tilde{S}_k)-\sigma_{k+1}} \leq \frac{\|\tilde{F}\|+\|\tilde{H}\|+\|\delta A\|}{\sigma_k-\sigma_{k+1}-\|\tilde{F}\|-\|\tilde{H}\|-\|\delta A\|}.$$

Using the definitions of ϵ_M and δ_M, and perturbation theory of singular values, Theorem 3.1 can be quantified in terms of an upper bound for $\max\{\|F\|, \|H\|, \|\tilde{F}\|, \|\tilde{H}\|\}$ and the size of the perturbation $\|\delta A\|$, as follows.

COROLLARY 3.2. *Using the notation of Theorem* 3.1, *suppose* τ *is a parameter that satisfies*

$$\max\left\{\|F\|, \|H\|, \|\tilde{F}\|, \|\tilde{H}\|\right\} \leq \tau.$$

If $4\tau + \|\delta A\| < \sigma_k - \sigma_{k+1}$, *then*

$$\sin\Theta(\mathcal{R}(A_{Mk}^T),\mathcal{R}(\tilde{A}_{Mk}^T)) \leq \frac{4\tau+\|\delta A\|}{\sigma_k-\sigma_{k+1}-4\tau-\|\delta A\|} \tag{18}$$

and

$$\sin\Theta(\mathcal{R}(A_{Mk}),\mathcal{R}(\tilde{A}_{Mk})) \leq \frac{4\tau+\|\delta A\|}{\sigma_k-\sigma_{k+1}-4\tau-\|\delta A\|}. \tag{19}$$

Therefore, the *UMV*-based subspaces $\mathcal{R}(A_{Mk}^T)$ and $\mathcal{R}(A_{Mk})$ are relatively insensitive to δA provided $\|\delta A\| \ll \sigma_k - \sigma_{k+1} - 4\tau$. In comparison with the singular subspace bounds in (6), Corollary 3 implies the subspaces associated with the *UMV* decomposition are only *slightly more sensitive* than the singular subspaces, provided the norm of the off-diagonal blocks of M and $\tilde{M}$ are "sufficiently" small with respect to $\|\delta A\|$, that is, $4\tau \ll \|\delta A\|$. This includes the *URV* and *ULV* decompositions, in which case 4τ can be replaced by 2τ.

Now we specialize Theorem 3.1 to the widely used *URV* and *ULV* decompositions for completeness and clarity, with particular focus on the ability of the *URV*- and *ULV*-based subspaces of $\tilde{A}$ in approximating the singular subspaces of A. Let the columns of $\tilde{X}_{Rk}$ form an orthonormal basis for $\mathcal{R}(\tilde{A}_{Rk}^T)$ and the columns of $\tilde{Y}_{Rk}$ form an orthonormal basis for $\mathcal{R}(\tilde{A}_{Rk})$, with an obvious corresponding meaning for $\tilde{X}_{Lk}$ and $\tilde{Y}_{Lk}$. Define ϵ_R and ϵ_L by

$$\epsilon_R \equiv \max\left\{\|\delta A\tilde{X}_{Rk}\|, \|\delta A^T\tilde{Y}_{Rk}\|\right\} \quad \text{and} \quad \epsilon_L \equiv \max\left\{\|\delta A\tilde{X}_{Lk}\|, \|\delta A^T\tilde{Y}_{Lk}\|\right\}. \tag{20}$$

COROLLARY 3.3. *Let* A *have the* SVD *as in* (1) *and let* $\tilde{A} = A + \delta A$ *have the* URV *and* ULV *decompositions as in* (7) *and* (8), *respectively. Define* ϵ_R *and* ϵ_L *as in* (20).

Let τ *be any number such that* $\|\tilde{F}\| \leq \tau$. *If* $\tau + \|\delta A\| < \sigma_k - \sigma_{k+1}$, *then*

$$\sin\Theta(\mathcal{R}(A_k^T),\mathcal{R}(\tilde{A}_{Rk}^T)) \leq \frac{\|\tilde{F}\|+\epsilon_R}{\sigma_k-\sigma_{k+1}-\|\tilde{F}\|-\|\delta A\|} \leq \frac{\tau+\|\delta A\|}{\sigma_k-\sigma_{k+1}-\tau-\|\delta A\|} \tag{21}$$

and

$$\sin\Theta(\mathcal{R}(A_k),\mathcal{R}(\tilde{A}_{Rk})) \le \frac{\|\tilde{F}\|\,\sigma_{k+1}}{\sigma_{\min}^2(\tilde{R}_k)-\sigma_{k+1}^2} + \frac{\epsilon_R}{\sigma_{\min}(\tilde{R}_k)-\sigma_{k+1}}$$

$$\le \frac{\tau+\|\delta A\|}{\sigma_k-\sigma_{k+1}-\tau-\|\delta A\|}. \tag{22}$$

Let τ be any number such that $\|\tilde{H}\| \le \tau$. If $\tau + \|\delta A\| < \sigma_k - \sigma_{k+1}$ then

$$\sin\Theta(\mathcal{R}(A_k^T),\mathcal{R}(\tilde{A}_{Lk}^T)) \le \frac{\|\tilde{H}\|\,\sigma_{k+1}}{\sigma_{\min}^2(\tilde{L}_k)-\sigma_{k+1}^2} + \frac{\epsilon_L}{\sigma_{\min}(\tilde{L}_k)-\sigma_{k+1}}$$

$$\le \frac{\tau+\|\delta A\|}{\sigma_k-\sigma_{k+1}-\tau-\|\delta A\|} \tag{23}$$

and

$$\sin\Theta(\mathcal{R}(A_k),\mathcal{R}(\tilde{A}_{Lk})) \le \frac{\|\tilde{H}\|+\epsilon_L}{\sigma_{\min}(\tilde{L}_k)-\sigma_{k+1}}$$

$$\le \frac{\tau+\|\delta A\|}{\sigma_k-\sigma_{k+1}-\tau-\|\delta A\|}. \tag{24}$$

By comparing (6) and (21)–(24), the important conclusion from Corollary 3 is that the *URV*- and *ULV*-based subspaces of $\tilde{A}$ are only *slightly more sensitive* to the perturbation δA than the singular subspaces of $\tilde{A}$, provided τ is sufficiently small with respect to $\|\delta A\|$; that is, $\tau \ll \|\delta A\|$.

The residuals $-\delta A\tilde{X}_{Mk}$ and $-\delta A^T\tilde{Y}_{Mk}$ in Theorem 3.1 are defined in terms of basis vectors for $\mathcal{R}(\tilde{A}_{Mk}^T)$ and $\mathcal{R}(\tilde{A}_{Mk})$. A similar result can be derived for residuals based on complementary basis vectors. Consider the *complementary* residuals $-\delta A\tilde{X}_{M0}$ and $-\delta A\tilde{Y}_{M0}$, where the columns of $\tilde{X}_{M0} \in \Re^{n\times(n-k)}$ form an orthonormal basis for $\mathcal{N}(\tilde{A}_{Mk})$ and the columns of $\tilde{Y}_{M0} \in \Re^{m\times(m-k)}$ form an orthonormal basis for $\mathcal{R}(\tilde{A}_{Mk})^{\perp}$. It is necessary to extend the columns of $\tilde{Y}_{M0}$ to more than just basis vectors for $\mathcal{R}(\tilde{U}_{M0})$ because the subspace sensitivity measure requires orthogonal complements. If we define $\tilde{D}_{M0} \equiv \tilde{Y}_{M0}^T\tilde{A}\tilde{X}_{M0}$ then

$$-\delta A\tilde{X}_{M0} = (A\tilde{X}_{M0} - \tilde{Y}_{M0}\tilde{D}_{M0}) - \tilde{U}_{Mk}\tilde{F}\tilde{V}_{M0}^T\tilde{X}_{M0},$$
$$-\delta A^T\tilde{Y}_{M0} = (A^T\tilde{Y}_{M0} - \tilde{X}_{M0}\tilde{D}_{M0}^T) - \tilde{V}_{Mk}\tilde{H}^T\tilde{U}_{M0}^T\tilde{Y}_{M0}.$$

Now we are ready to state the subspace bounds in terms of the complementary residuals, which extend Wedin's $\sin\Theta$ theorem with complementary residuals [19].

THEOREM 3.4. *Let A and $\tilde{A}$ have the* UMV *decompositions as in* (9) *and* (10), *respectively. Assume there exists a $\delta_M > 0$ and $\alpha_M \ge 0$ such that*

$$\sigma_{\min}(S_k) \ge \alpha_M + \delta_M \quad \textit{and} \quad \sigma_{\max}(\tilde{C}) \le \alpha_M.$$

Take $\epsilon_M = \max\left\{\|\delta A\tilde{X}_{M0}\|, \|\delta A^T\tilde{Y}_{M0}\|\right\}$. Then for every unitary invariant norm we have

$$\sin\Theta(\mathcal{R}(A_{Mk}^T),\mathcal{R}(\tilde{A}_{Mk}^T)) \le \frac{\|F\|+\|\tilde{F}\|}{\delta_M} + \frac{(\|H\|+\|\tilde{H}\|)}{\delta_M} \times \frac{\sigma_{\max}(\tilde{C})}{\sigma_{\min}(S_k)+\sigma_{\max}(\tilde{C})} + \frac{\epsilon_M}{\delta_M}$$

and

$$\sin\Theta(\mathcal{R}(A_{Mk}),\mathcal{R}(\tilde{A}_{Mk})) \leq \frac{\|H\|+\|\tilde{H}\|}{\delta_M} + \frac{(\|F\|+\|\tilde{F}\|)}{\delta_M} \times \frac{\sigma_{\max}(\tilde{C})}{\sigma_{\min}(S_k)+\sigma_{\max}(\tilde{C})} + \frac{\epsilon_M}{\delta_M}.$$

Proof. The proof is analogous to that of Theorem 3.1. □

4. *UMV* decompositions and regularization. In this section we regularize the solution to the ill-conditioned LS problem

$$\min_{x\in\Re^n} \|b - A\,x\|_2 \tag{25}$$

by truncating the two-sided orthogonal decomposition of A. Then we derive perturbation theory for the minimum norm solution of the resulting rank-deficient LS problem.

When A is very ill conditioned, the ordinary LS solution $x_{ols} = A^{\dagger}b = V\Sigma^{\dagger}U^T b$ may be very sensitive to errors in A and b in the sense that small errors in A or b may result in disproportionately large changes in the solution. The *TSVD* technique (cf. [9]) is commonly used to produce a less sensitive solution by computing by the minimum norm solution to the related LS problem

$$\min_{x\in\Re^n} \|b - A_k\,x\|_2. \tag{26}$$

The minimum norm LS solution to (26), called the *TSVD* solution, is given by

$$x_k = A_k^{\dagger}b = V_1\Sigma_1^{-1}U_1^T b. \tag{27}$$

Defining $\tilde{A} = A + \delta A$ and $\tilde{b} = b + \delta b$ as usual, we let

$$\tilde{x}_k = \tilde{A}_k^{\dagger}\tilde{b} = \tilde{V}_1\tilde{\Sigma}_1^{-1}\tilde{U}_1^T\tilde{b}$$

denote the *TSVD* solution to the perturbed truncated LS problem.

As an alternative to the *TSVD* technique, we consider truncating the two-sided orthogonal decomposition to obtain the following LS problem:

$$\min_{x\in\Re^n} \|b - A_{Mk}\,x\|_2. \tag{28}$$

The motivation here is that it is often less computationally demanding to compute A_{Mk} than A_k. The minimum norm LS solution to (28), called the *TUMV* solution, is given by

$$x_{Mk} = A_{Mk}^{\dagger}b = V_{Mk}S_k^{-1}U_{Mk}^T b, \tag{29}$$

and the *TUMV* solution to the corresponding perturbed LS problem is denoted

$$\tilde{x}_{Mk} = \tilde{A}_{Mk}^{\dagger}\tilde{b} = \tilde{V}_{Mk}\tilde{S}_k^{-1}\tilde{U}_{Mk}^T\tilde{b}. \tag{30}$$

The following theorem presents perturbation theory for the *TUMV* solution, i.e., upper bounds for the relative error $\|x_{Mk}-\tilde{x}_{Mk}\|/\|x_{Mk}\|$. The bounds give insight into the *sensitivity* of x_{Mk} to the perturbations δA and δb, as well as the decomposition itself. The derivation proceeds with two key ideas in mind:

• first, x_{Mk} and $\tilde{x}_{Mk}$ belong to $\mathcal{R}(V_{Mk}) = \mathcal{R}(A_{Mk}^T)$ and $\mathcal{R}(\tilde{V}_{Mk}) = \mathcal{R}(\tilde{A}_{Mk}^T)$, respectively;

• second, x_{Mk} and $\tilde{x}_{Mk}$ are the (unique) minimum norm solutions to the (compatible) overdetermined system of linear equations $A_{Mk}x = \mathcal{P}_{\mathcal{R}(A_{Mk})}b$ and $\tilde{A}_{Mk}x = \mathcal{P}_{\mathcal{R}(\tilde{A}_{Mk})}\tilde{b}$, respectively.

Therefore, it is natural to derive perturbation bounds which involve the familiar scalars

$$\sin\Theta(\mathcal{R}(A_{Mk}^T), \mathcal{R}(\tilde{A}_{Mk}^T)) \quad \text{and} \quad \sin\Theta(\mathcal{R}(A_{Mk}), \mathcal{R}(\tilde{A}_{Mk})).$$

For simplicity, let $\|\cdot\| = \|\cdot\|_2$.

THEOREM 4.1. *Let $\tilde{A} = A + \delta A$ and $\tilde{b} = b + \delta b$, and let x_{Mk} and $\tilde{x}_{Mk}$ be defined as in* (29) *and* (30), *respectively. Define the LS residual $r_{Mk} \equiv b - Ax_{Mk}$, $\rho_{Mk} \equiv \|r_{Mk} + U_{M0}HS_k^{-1}U_{Mk}^T b\|/\|b\|$, and the condition number $\Psi_{Mk} = \|A\|\,\|\tilde{A}_{Mk}^\dagger\|$. Assume there exists a $\delta_M > 0$ and $\alpha_M \geq 0$ such that*

$$\sigma_{\min}(\tilde{S}_k) \geq \alpha_M + \delta_M \quad \text{and} \quad \sigma_{\max}(C) \leq \alpha_M.$$

Then

$$\begin{aligned}\frac{\|x_{Mk} - \tilde{x}_{Mk}\|}{\|x_{Mk}\|} &\leq \sin\Theta(\mathcal{R}(A_{Mk}^T), \mathcal{R}(\tilde{A}_{Mk}^T)) \\ &\quad + \Psi_{Mk}\left(\frac{\|r_{Mk}\|/\|b\|}{\sqrt{1-\rho_{Mk}^2}} \times \sin\Theta(\mathcal{R}(A_{Mk}), \mathcal{R}(\tilde{A}_{Mk}))\right) \\ &\quad + \Psi_{Mk}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|/\|b\|}{\sqrt{1-\rho_{Mk}^2}} + \frac{\|F\| + \|\tilde{F}\|}{\|A\|}\right),\end{aligned}$$

where $\sin\Theta(\mathcal{R}(A_{Mk}^T), \mathcal{R}(\tilde{A}_{Mk}^T))$ and $\sin\Theta(\mathcal{R}(A_{Mk}), \mathcal{R}(\tilde{A}_{Mk}))$ are bounded according to Theorem 3.1. *Further, the numerator $\|\delta A\|$ can be replaced by ϵ_M in* (12) *to improve the bounds.*

Proof. We begin with

$$\begin{aligned}x_{Mk} - \tilde{x}_{Mk} &= A_{Mk}^\dagger b - \tilde{A}_{Mk}^\dagger \tilde{b} \\ &= (A_{Mk}^\dagger - \tilde{A}_{Mk}^\dagger)b - (A_{Mk}^\dagger A_{Mk} - \tilde{A}_{Mk}^\dagger \tilde{A}_{Mk})x_{Mk} - \tilde{A}_{Mk}^\dagger \delta b \\ &\quad + (A_{Mk}^\dagger A_{Mk} - \tilde{A}_{Mk}^\dagger \tilde{A}_{Mk})x_{Mk}. \qquad (31)\end{aligned}$$

After some algebraic manipulation it follows that

$$\begin{aligned}A_{Mk}^\dagger A_{Mk} - \tilde{A}_{Mk}^\dagger \tilde{A}_{Mk} = (A_{Mk}^\dagger - \tilde{A}_{Mk}^\dagger)A - V_{Mk}S_k^{-1}FV_{M0}^T + \tilde{V}_{Mk}\tilde{S}^{-1}\tilde{F}\tilde{V}_{M0}^T \\ -\tilde{A}_{Mk}^\dagger \delta A.\end{aligned}$$

Substituting this into (31), we get

$$\begin{aligned}x_{Mk} - \tilde{x}_{Mk} = (A_{Mk}^\dagger - \tilde{A}_{Mk}^\dagger)r_{Mk} - (V_{Mk}^T S_k^{-1} F V_{M0}^T - \tilde{V}_{Mk}\tilde{S}^{-1}\tilde{F}\tilde{V}_{M0}^T)x_{Mk} \\ - \tilde{A}_{Mk}^\dagger \delta A\, x_{Mk} - \tilde{A}_{Mk}^\dagger \delta b + (A_{Mk}^\dagger A_{Mk} - \tilde{A}_{Mk}^\dagger \tilde{A}_{Mk})x_{Mk},\end{aligned}$$

which leads to

$$\begin{aligned}\|x_{Mk} - \tilde{x}_{Mk}\| \leq \|(A_{Mk}^\dagger - \tilde{A}_{Mk}^\dagger)r_{Mk}\| + \|(V_{Mk}^T S_k^{-1} F V_{M0}^T - \tilde{V}_{Mk}\tilde{S}^{-1}\tilde{F}\tilde{V}_{M0}^T)x_{Mk}\| \\ + \|\tilde{A}_{Mk}^\dagger \delta A\, x_{Mk}\| + \|\tilde{A}_{Mk}^\dagger \delta b\| + \|x_{Mk}\| \sin\Theta(\mathcal{R}(A_{Mk}^T), \mathcal{R}(\tilde{A}_{Mk}^T)). \qquad (32)\end{aligned}$$

Now we shall bound the individual terms on the right. To bound the first term, note that $\tilde{A}^{\dagger}_{Mk} = \tilde{A}^{\dagger}_{Mk}\tilde{U}_{Mk}\tilde{U}^T_{Mk}$ and $r_{Mk} = P^{\perp}_{\mathcal{R}(U_{Mk})}b - U_{M0}HV^T_{Mk} = P^{\perp}_{\mathcal{R}(U_{Mk})}r_{Mk}$. Consequently,

$$\begin{aligned}
\|(A^{\dagger}_{Mk} - \tilde{A}^{\dagger}_{Mk})r_{Mk}\| &= \|\tilde{A}^{\dagger}_{Mk}r_{Mk}\| \\
&= \|\tilde{A}^{\dagger}_{Mk}P_{\mathcal{R}(\tilde{U}_{Mk})}P^{\perp}_{\mathcal{R}(U_{Mk})}r_{Mk}\| \\
&\leq \|\tilde{A}^{\dagger}_{Mk}\|\,\|r_{Mk}\|\,\sin\Theta(\mathcal{R}(A_{Mk}),\mathcal{R}(\tilde{A}_{Mk})) \qquad (33)\\
&\leq \Psi_{Mk}\frac{\|r_{Mk}\|}{\|U_{Mk}U^T_{Mk}b\|}\,\|x_{Mk}\|\,\sin\Theta(\mathcal{R}(A_{Mk}),\mathcal{R}(\tilde{A}_{Mk})). \qquad (34)
\end{aligned}$$

Now, it can be shown $r_{Mk} \perp \mathcal{R}(A_{Mk})$ and $b = r_{Mk} + U_{M0}HS_k^{-1}U^T_{MK}b + U_{Mk}U^T_{Mk}b$. Therefore, $\|U_{Mk}U^T_{Mk}b\|^2 = \|b\|^2\left(1 - \|r_{Mk} + U_{M0}HS_k^{-1}U^T_{MK}b\|^2/\|b\|^2\right)$. Hence,

$$\frac{1}{\|U_{Mk}U^T_{Mk}b\|} \leq \frac{1}{\|b\|} \times \frac{1}{\sqrt{1-\rho^2_{Mk}}},$$

where $\rho_{Mk} \equiv \|r_{Mk} + U_{M0}HS_k^{-1}U^T_{MK}b\|/\|b\|$, and

$$\|(A^{\dagger}_{Mk} - \tilde{A}^{\dagger}_{Mk})r_{Mk}\| \leq \Psi_{Mk}\frac{\|r_{Mk}\|/\|b\|}{\sqrt{1-\rho^2_{Mk}}}\,\|x_{Mk}\|\,\sin\Theta(\mathcal{R}(A_{Mk}),\mathcal{R}(\tilde{A}_{Mk})). \qquad (35)$$

We bound the second term according to

$$\|(V^T_{Mk}S_k^{-1}FV^T_{M0} - \tilde{V}_{Mk}\tilde{S}_k^{-1}\tilde{F}\tilde{V}^T_{M0})x_{Mk}\| \leq \Psi_{Mk}\left(\frac{\|F\| + \|\tilde{F}\|}{\|A\|}\right)\|x_{Mk}\|. \qquad (36)$$

To bound the third and fourth terms we have

$$\|\tilde{A}^{\dagger}_{Mk}\delta A\, x_{Mk}\| \leq \Psi_{Mk}\frac{\|\delta A\|}{\|A\|}\,\|x_{Mk}\| \qquad (37)$$

and

$$\|\tilde{A}^{\dagger}_{Mk}\,\delta b\| \leq \Psi_{Mk}\frac{\|\delta b\|/\|b\|}{\sqrt{1-\rho^2_{Mk}}}\,\|x_{Mk}\|. \qquad (38)$$

Taking into account (32)–(38) we see that the desired result immediately follows. Finally, the result can be strengthened as stated, since in (37) we see that

$$\begin{aligned}
\|\tilde{A}^{\dagger}_{Mk}\delta A\, x_{Mk}\| &= \|\tilde{V}_{Mk}\tilde{S}_k^{-1}\tilde{U}^T_{Mk}\delta A\, x_{Mk}\| \\
&\leq \|\tilde{S}_k^{-1}\|\,\|\delta A^T\tilde{U}_{Mk}\|\,\|x_{Mk}\| \\
&\leq \Psi_{Mk}\frac{\epsilon_M}{\|A\|}\,\|x_{Mk}\|. \qquad \square
\end{aligned}$$

Remark 6. The perturbation bound in Theorem 4.1 characterizes the role of the subspaces in the sensitivity of the *TUMV* solution x_{Mk}. Further, the upper bounds in Theorem 3.1 highlight the importance of the block matrices of the middle matrix in the sensitivity of x_{Mk}, providing additional insight.

Remark 7. By setting $\|F\| = \|\tilde{F}\| = \|H\| = \|\tilde{H}\| = 0$ we have also derived perturbation bounds for the *TSVD* solution: defining $\Psi_k \equiv \|A\|\,\|\tilde{A}_k^{\dagger}\|$, $r_k \equiv b - Ax_k$, and $\rho_k \equiv \|r_k\|/\|b\|$, we get

$$\begin{aligned}
\frac{\|x_k - \tilde{x}_k\|}{\|x_k\|} &\le \sin\Theta(\mathcal{R}(A_k^T), \mathcal{R}(\tilde{A}_k^T)) \\
&\quad + \Psi_k \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|/\|b\|}{\sqrt{1-\rho_k^2}} + \frac{\|r_k\|/\|b\|}{\sqrt{1-\rho_k^2}} \times \sin\Theta(\mathcal{R}(A_k), \mathcal{R}(\tilde{A}_k)) \right) \\
&\le \frac{\epsilon}{\delta} + \Psi_k \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|/\|b\|}{\sqrt{1-\rho_k^2}} + \frac{\|r_k\|/\|b\|}{\sqrt{1-\rho_k^2}} \times \frac{\epsilon}{\delta} \right) \\
&\le \frac{\|\delta A\|}{\sigma_k - \sigma_{k+1} - \|\delta A\|} \\
&\quad + \Psi_k \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|/\|b\|}{\sqrt{1-\rho_k^2}} + \frac{\|r_k\|/\|b\|}{\sqrt{1-\rho_k^2}} \times \frac{\|\delta A\|}{\sigma_k - \sigma_{k+1} - \|\delta A\|} \right) \\
&\le \frac{\eta_k}{1-\eta_k-\omega_k} + \frac{\kappa_k}{1-\eta_k} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|/\|b\|}{\sqrt{1-\rho_k^2}} + \frac{\|r_k\|/\|b\|}{\sqrt{1-\rho_k^2}} \times \frac{\eta_k}{1-\eta_k-\omega_k} \right), \qquad (39) \\
&\le \frac{2\eta_k}{1-\eta_k-\omega_k} + \frac{\kappa_k}{1-\eta_k} \left(\frac{\epsilon_b}{\sqrt{1-\rho_k^2}} + \frac{\rho_k}{\sqrt{1-\rho_k^2}} \times \frac{\eta_k}{1-\eta_k-\omega_k} \right), \qquad (40)
\end{aligned}$$

where we have defined

$$\epsilon_A \equiv \frac{\|\delta A\|}{\|A\|},\ \epsilon_b \equiv \frac{\|\delta b\|}{\|b\|},\ \kappa_k \equiv \|A\|\,\|A_k^{\dagger}\| = \frac{\sigma_1}{\sigma_k},\ \eta_k \equiv \kappa_k \epsilon_A,\ \text{and}\ \omega_k \equiv \frac{\sigma_{k+1}}{\sigma_k}.$$

Equation (39) is a well-known perturbation result[1] for the *TSVD* solution x_k cf. [10, Thm. 3.4]. Therefore, the perturbation bounds for the solution to the LS problem obtained by truncating the two-sided orthogonal decomposition specialize to the well-known *SVD*-based perturbation result.

Remark 8. The bound in Theorem 4.1 can be applied to obtain a bound for $\|x_k - \tilde{x}_{Mk}\|/\|x_k\|$.

For convenience we specialize the general perturbation result to the *URV* and *ULV* decompositions.

COROLLARY 4.2. *Using the notation in Theorem* 4.1, *the following bounds hold: If* $\sigma_{\max}(G) < \sigma_{\min}(\tilde{R}_k)$, *then*

$$\begin{aligned}
\frac{\|x_{Rk} - \tilde{x}_{Rk}\|}{\|x_{Rk}\|} &\le \frac{\|F\| + \|\tilde{F}\| + \|\delta A\|}{\sigma_{\min}(\tilde{R}_k) - \sigma_{\max}(G)} + \Psi_{Rk} \left(\frac{\|F\| + \|\tilde{F}\|}{\|A\|} + \frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|/\|b\|}{\sqrt{1-\rho_{Rk}^2}} \right) \\
&\quad + \Psi_{Rk} \frac{\|r_{Rk}\|/\|b\|}{\sqrt{1-\rho_{Rk}^2}} \left(\frac{(\|F\| + \|\tilde{F}\|)\,\sigma_{\max}(G)}{\sigma_{\min}^2(\tilde{R}_k) - \sigma_{\max}^2(G)} + \frac{\|\delta A\|}{\sigma_{\min}(\tilde{R}_k) - \sigma_{\max}(G)} \right).
\end{aligned}$$

If $\sigma_{\max}(E) < \sigma_{\min}(\tilde{L}_k)$, *then*

$$\frac{\|x_{Lk} - \tilde{x}_{Lk}\|}{\|x_{Lk}\|} \le \frac{(\|H\| + \|\tilde{H}\|)\,\sigma_{\max}(E)}{\sigma_{\min}^2(\tilde{L}_k) - \sigma_{\max}^2(E)} + \frac{\|\delta A\|}{\sigma_{\min}(\tilde{L}_k) - \sigma_{\max}(E)}$$

[1] The factor $\sqrt{1-\rho_k^2}$ in (39) represents a minor correction to [10, Thm. 3.4].

$$+\Psi_{Lk}\left(\frac{\|\delta A\|}{\|A\|}+\frac{\|\delta b\|/\|b\|}{\sqrt{1-\rho_{Lk}^2}}+\frac{\|r_{Lk}\|/\|b\|}{\sqrt{1-\rho_{Lk}^2}}\times\frac{\|H\|+\|\tilde{H}\|+\|\delta A\|}{\sigma_{\min}(\tilde{L}_k)-\sigma_{\max}(E)}\right).$$

As mentioned in §1, the *URV* and *ULV* decompositions can be used to solve LS problems (28) in a noisy environment in various domains such as signal processing, mainly because triangular systems are easy to solve and their structure is easier to preserve in updating and downdating problems. Often, x_k represents the exact or desired solution to the noise-free LS problem and the *TURV* and *TULV* solutions represent approximations to x_k derived from the noisy LS problem.

COROLLARY 4.3. *Using the notation in Theorem* 4.1, *define*

$$\epsilon_A \equiv \frac{\|\delta A\|}{\|A\|},\quad \epsilon_b \equiv \frac{\|\delta b\|}{\|b\|},\quad \kappa_k \equiv \|A\|\,\|A_k^{\dagger}\| = \frac{\sigma_1}{\sigma_k},\quad \eta_k \equiv \kappa_k\epsilon_A,\quad \omega_k \equiv \frac{\sigma_{k+1}}{\sigma_k},$$

$$\epsilon_{\tilde{F}} \equiv \frac{\|\tilde{F}\|}{\|\tilde{A}\|},\quad \eta_{\tilde{F}} \equiv \kappa_k\epsilon_{\tilde{F}},\quad \epsilon_{\tilde{H}} \equiv \frac{\|\tilde{H}\|}{\|\tilde{A}\|},\quad \eta_{\tilde{H}} \equiv \kappa_k\epsilon_{\tilde{H}},\quad \textit{and}\quad \rho_k \equiv \frac{\|r_k\|}{\|b\|}.$$

If $\epsilon_{\tilde{F}} + \|\delta A\| < \sigma_k - \sigma_{k+1}$, *then*

$$\frac{\|x_k - \tilde{x}_{Rk}\|}{\|x_k\|} \le \frac{2\eta_k + 2\eta_{\tilde{F}}}{1-\eta_k-\omega_k-\epsilon_{\tilde{F}}} + \frac{\kappa_k}{1-\eta_k-\epsilon_{\tilde{F}}}\left(\frac{\epsilon_b}{\sqrt{1-\rho_k^2}} + \frac{\rho_k}{\sqrt{1-\rho_k^2}}\left[\frac{\eta_k+\eta_{\tilde{F}}}{1-\eta_k-\omega_k-\epsilon_{\tilde{F}}}\right]\right).$$

If $\epsilon_{\tilde{H}} + \|\delta A\| < \sigma_k - \sigma_{k+1}$, *then*

$$\frac{\|x_k - \tilde{x}_{Lk}\|}{\|x_k\|} \le \frac{2\eta_k + 2\eta_{\tilde{H}}}{1-\eta_k-\omega_k-\epsilon_{\tilde{H}}} + \frac{\kappa_k}{1-\eta_k-\epsilon_{\tilde{H}}}\left(\frac{\epsilon_b}{\sqrt{1-\rho_k^2}} + \frac{\rho_k}{\sqrt{1-\rho_k^2}}\left[\frac{\eta_k+\eta_{\tilde{H}}}{1-\eta_k-\omega_k-\epsilon_{\tilde{H}}}\right]\right).$$

Equation (39) and Corollary 4.3 imply $\tilde{x}_{Rk}$ and $\tilde{x}_{Lk}$ are only *slightly more sensitive* than $\tilde{x}_k$ to the perturbation δA, provided $\epsilon_{\tilde{F}} \ll \|\delta A\|$ and $\epsilon_{\tilde{H}} \ll \|\delta A\|$, respectively.

5. Numerical example. We complete our discussion with a small numerical example. The singular values of the 20×6 matrix A are

$$\sigma(A) = \{10,\ 3,\ 1,\ 5\cdot 10^{-2},\ 10^{-4},\ 10^{-5}\},$$

and the numerical rank of A is $k = 4$ (arbitrarily chosen). We defined the right-hand side $b \in \Re^{20}$ by $b = A * [1,\ 1,\ 1,\ 1,\ 1,\ 1]^T$. The elements of $[\delta A\ \delta b]$ are from the normal distribution with mean zero and standard deviation σ, where

$$\sigma = 3\cdot 10^{-6},\ 8\cdot 10^{-6},\ 3\cdot 10^{-5},\ 8\cdot 10^{-5},\ 3\cdot 10^{-4},\ \text{or}\ 8\cdot 10^{-4}.$$

The *URV* and *ULV* decompositions for A and $\tilde{A}$ were computed by means of Stewart's algorithms, and the singular vector estimates needed in those algorithms were obtained with a single step of inverse iteration. As shown in [5], the quality of these estimates influences the size of the off-diagonal blocks (which in turn influence the subspace angles). A single step was enough in this simulation to produce sufficiently small off-diagonal blocks.

The values in Table 1 and Table 2 represent the average of 100 trials. Table 1 summarizes the mean values of the subspace angles (cf. Theorem 3.1):

$$\sin\Theta(\mathcal{R}(A_{Mk}^T), \mathcal{R}(\tilde{A}_{Mk}^T)) \quad \text{and} \quad \sin\Theta(\mathcal{R}(A_{Mk}), \mathcal{R}(\tilde{A}_{Mk}))$$

TABLE 1
Comparison of the subspace angles. All values represent the mean of 100 *trials.*

σ	$\sin\Theta(\mathcal{R}(A_k^T),\mathcal{R}(\tilde{A}_k^T))$	$\sin\Theta(\mathcal{R}(A_{Rk}^T),\mathcal{R}(\tilde{A}_{Rk}^T))$	$\sin\Theta(\mathcal{R}(A_{Lk}^T),\mathcal{R}(\tilde{A}_{Lk}^T))$
3.0e−06	8.0886e−05	8.0843e−05	8.0886e−05
8.0e−06	1.9946e−04	1.9955e−04	1.9946e−04
3.0e−05	7.9777e−04	7.9686e−04	7.9777e−04
8.0e−05	2.3027e−03	2.3055e−03	2.3027e−03
3.0e−04	7.9393e−03	7.9885e−03	7.9373e−03
8.0e−04	1.9140e−02	1.9967e−02	1.9119e−02
σ	$\sin\Theta(\mathcal{R}(A_k),\mathcal{R}(\tilde{A}_k))$	$\sin\Theta(\mathcal{R}(A_{Rk}),\mathcal{R}(\tilde{A}_{Rk}))$	$\sin\Theta(\mathcal{R}(A_{Lk}),\mathcal{R}(\tilde{A}_{Lk}))$
3.0e−06	2.3703e−04	2.3702e−04	2.3704e−04
8.0e−06	6.2721e−04	6.2721e−04	6.2716e−04
3.0e−05	2.3208e−03	2.3208e−03	2.3211e−03
8.0e−05	6.4126e−03	6.4126e−03	6.4128e−03
3.0e−04	2.2876e−02	2.2877e−02	2.2886e−02
8.0e−04	6.2716e−02	6.2742e−02	6.3042e−02

TABLE 2
Comparison of the relative errors for the truncated LS solutions. All values represent the mean of 100 *trials.*

σ	$\frac{\|x_4-\tilde{x}_4\|}{\|x_4\|}$	$\frac{\|x_{R4}-\tilde{x}_{R4}\|}{\|x_{R4}\|}$	$\frac{\|x_{L4}-\tilde{x}_{L4}\|}{\|x_{L4}\|}$
3.0e−06	8.0886e−05	8.0843e−05	8.0886e−05
8.0e−06	1.9946e−04	1.9955e−04	1.9946e−04
3.0e−05	7.9777e−04	7.9686e−04	7.9777e−04
8.0e−05	2.3027e−03	2.3055e−03	2.3027e−03
3.0e−04	7.9393e−03	7.9885e−03	7.9373e−03
8.0e−04	1.9140e−02	1.9967e−02	1.9119e−02

for the *SVD*, *URV*, and *ULV*. In this simulation the *URV*- and *ULV*-based subspaces are only slightly more sensitive to noise than the singular subspaces. Table 2 summarizes the mean values of $\|x_{Mk}-\tilde{x}_{Mk}\|/\|x_{Mk}\|$ (cf. Theorem 4.1) and the corresponding upper bound in terms of the subspace angles (cf. Theorem 3.1) for the *SVD*, *URV*, and *ULV*. In this simulation the *TURV* and *TULV* solutions are only slightly more sensitive to noise than the *TSVD* solution.

6. Conclusion. In this paper we derived perturbation bounds for the subspaces associated with a *general* two-sided (or complete) orthogonal decomposition (called the *UMV* decomposition) of a numerically rank-deficient matrix. The analysis implies the *UMV*-based subspaces are only *slightly more sensitive* to perturbations than singular subspaces, provided the off-diagonal blocks of the middle matrices M and $\tilde{M}$ are sufficiently small with respect to the size of the perturbation. Then we considered regularizing the solution to the ill-conditioned least squares problem by truncating the complete orthogonal decomposition and derived perturbation bounds for the resulting minimum norm least squares solution (called the *TUMV* solution). The analysis implies the *TUMV* solution is only *slightly more sensitive* to perturbations than the *TSVD* solution, provided the off-diagonal blocks of the middle matrices M and $\tilde{M}$ are sufficiently small with respect to the size of the perturbation. Finally, we showed how the new bounds can be specialized to well-known *SVD*-based perturbation results for singular subspaces and the *TSVD* solution.

Acknowledgments. The author wishes to thank James Bunch (UCSD), Per Christian Hansen (UNI•C), and the anonymous referees for helpful comments regarding the presentation of this paper.

REFERENCES

[1] G. ADAMS, M. F. GRIFFIN, AND G. W. STEWART, *Direction-of-arrival estimation using the rank-revealing URV decomposition*, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Washington, DC, 1991.

[2] J. BARLOW AND H. ZHA, *Stable algorithms for downdating two-sided orthogonal decompositions*, in SVD and Signal Processing III. Algorithms, Architectures and Applications, M. Moonen and B. De Moor, eds., Elsevier, New York, 1995.

[3] M. W. BERRY AND R. D. FIERRO, *Two-sided orthogonal decompositions for information retrieval applications*, Numer. Linear Algebra Appl., in press.

[4] C. H. BISCHOF, *Incremental condition estimation*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 312–322.

[5] R. D. FIERRO AND J. R. BUNCH, *Bounding the subspaces from rank revealing two-sided orthogonal decompositions*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 743–759.

[6] ——— *Perturbation theory for orthogonal projection methods with applications to least squares and total least squares*, Linear Algebra Appl., 234 (1996), pp. 71–96.

[7] R. D. FIERRO AND P. C. HANSEN, *Accuracy of TSVD solutions computed from rank revealing two-sided orthogonal decompositions*, Numer. Math., 70 (1995), pp. 453–471.

[8] ——— *Low rank revealing two-sided orthogonal decompositions*, Technical Report PAM 94-09, Department of Mathematics, California State University, San Marcos, CA, October 1994; Numer. Algorithms, submitted.

[9] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., The John Hopkins University Press, Baltimore, MD, 1989.

[10] P. C. HANSEN, *The truncated SVD as a method for regularization*, BIT, 27 (1987), pp. 534–553.

[11] K. J. R. LIU, D. P. O'LEARY, G. W. STEWART, AND Y.-J. J. WU, *URV ESPIRIT for tracking time-varying signals*, IEEE Trans. Signal Process., 42 (1994), pp. 3441–3449.

[12] L. MIRSKY, *Symmetric Gauge Functions and Unitarily Invariant Norms*, Quart. J. Math Oxford Ser. (2), 11 (1960), pp. 50–59.

[13] H. PARK AND L. ELDÉN, *Downdating the rank revealing URV decomposition*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 138–155.

[14] D. J. PIERCE, *A Sparse URL Rather than an URV Factorization*, Mathematics and Engineering Analysis Technical Report MEA-TR-203, Boeing Computer Services, Seattle, WA, 1992.

[15] G. W. STEWART, *An updating algorithm for subspace tracking*, IEEE Trans. Signal Process., 40 (1992), pp. 1535–1541.

[16] R. MATHIAS AND G. W. STEWART, *A block QR algorithm and the singular value decomposition*, Linear Algebra Appl., 182 (1992), pp. 91–100.

[17] G. W. STEWART, *Updating a rank revealing ULV decomposition*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 494–499.

[18] 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.

[19] P.-Å. WEDIN, *Perturbation bounds in connection with the singular value decomposition*, BIT, 12 (1973), pp. 99–111.

[20] ——— *Perturbation theory for pseudoinverses*, BIT, 13 (1973), pp. 217–232.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 2, pp. 401–425, April 1996

010

A JACOBI–DAVIDSON ITERATION METHOD FOR LINEAR EIGENVALUE PROBLEMS*

GERARD L. G. SLEIJPEN† AND HENK A. VAN DER VORST†

Abstract. In this paper we propose a new method for the iterative computation of a few of the extremal eigenvalues of a symmetric matrix and their associated eigenvectors. The method is based on an old and almost unknown method of Jacobi. Jacobi's approach, combined with Davidson's method, leads to a new method that has improved convergence properties and that may be used for general matrices. We also propose a variant of the new method that may be useful for the computation of nonextremal eigenvalues as well.

Key words. eigenvalues and eigenvectors, Davidson's method, Jacobi iterations, harmonic Ritz values

AMS subject classifications. 65F15, 65N25

1. Introduction.

Suppose we want to compute one or more eigenvalues and their corresponding eigenvectors of the $n \times n$ matrix A. Several iterative methods are available: Jacobi's diagonalization method [9], [22], the power method [9], the method of Lanczos [13], [22], Arnoldi's method [1], [25], and Davidson's method [4], [25], [3], [14], [17]. The latter method has been reported to be quite successful, most notably in connection with certain symmetric problems in computational chemistry [4], [5], [31]. The success of the method seems to depend quite heavily on the (strong) diagonal dominance of A.

The method of Davidson is commonly seen as an extension to Lanczos's method, but as Saad [25] points out, from the implementation point of view it is more related to Arnoldi's method. In spite of these relations, the success of the method is not well understood [25]. Some recent convergence results and improvements, as well as numerical experiments, are reported in [3], [14], [15], [17], [16], [18], [27].

Jacobi [12] proposed a method for eigenvalue approximation that essentially was a combination of (1) Jacobi rotations, (2) Gauss–Jacobi iterations, and (3) an almost forgotten method that we will refer to as Jacobi's orthogonal component correction (JOCC). Reinvestigation of Jacobi's ideas leads to another view on the method of Davidson, and this not only helps us explain the behavior of the method, it also leads to a new and robust method with superior convergence properties for nondiagonally dominant (unsymmetric) matrices as well. Special variants of this method are already known; see [18], [27] and our discussion in §4.1.

The outline of this paper is as follows. In §2 we briefly describe the methods of Davidson and Jacobi, and we show that the original Davidson's method may be viewed as an accelerated Gauss–Jacobi iteration method. Likewise, more recent approaches which include other preconditioners M can be interpreted as accelerated standard iteration methods associated with the splitting $A = M - N$.

In §3 we propose the new approach, which is essentially a combination of the JOCC approach and the method of Davidson for creating more general subspaces. The difference between this approach and Davidson's method may seem very subtle but it is fundamental. Whereas in Davidson's method accurate preconditioners M (accurate

* Received by the editors June 22, 1994; accepted for publication (in revised form) by A. Greenbaum May 22, 1995.

† Mathematical Institute, University of Utrecht, Budapestlaan 6, P.O. Box 80.010, Utrecht 3508 TA, The Netherlands (sleijpen@math.ruu.nl, vorst@math.ruu.nl).

in the sense that they approximate the inverse of the given operator very well) may lead to stagnation or very slow convergence, the new approach takes advantage of such preconditioners, even if they are exact inverses. It should be stressed that in this approach we do not precondition the given eigensystem (neither does Davidson), but we precondition an auxiliary system for the corrections for the eigen approximations. The behavior of the method is further discussed in §4. There we see that for a specific choice the speed of convergence of the approximated eigenvalue is quadratic (and for symmetric problems even cubic). In practice, this requires the exact solution of a correction equation, but as we will demonstrate by simple examples (§6), this may be relaxed. We suggest using approximate solutions for the correction equations. This idea may be further exploited for the construction of efficient inner–outer iteration schemes, or by using preconditioners similar to those suggested for the Davidson method.

In §5 we discuss the harmonic Ritz values, and we show how these can be used in combination with our new algorithm for the determination of "interior" eigenvalues. We conclude with some simple but illustrative numerical examples in §6. The new method has already found its way into more complicated applications in chemistry and plasma physics modeling.

2. The methods of Davidson and Jacobi. Jacobi and Davidson originally presented their methods for symmetric matrices, but as is well known and as we will do in our presentation, both methods can easily be formulated for nonsymmetric matrices.

2.1. Davidson's method. The main idea behind Davidson's method is the following one. Suppose we have some subspace K of dimension k, over which the projected matrix A has a Ritz value θ_k (e.g., θ_k is the largest Ritz value) and a corresponding Ritz vector u_k. Let us assume that an orthogonal basis for K is given by the vectors $v_1, v_2, \ldots, v_k$.

Quite naturally the problem of how to expand the subspace in order to find a successful update for u_k arises. To that end we compute the defect $r = Au_k - \theta_k u_k$. Then Davidson, in his original paper [4], suggests computing t from $(D_A - \theta_k I)t = r$, where D_A is the diagonal of the matrix A. The vector t is made orthogonal to the basis vectors $v_1, \ldots, v_k$, and the resulting vector is chosen as the new v_{k+1}, by which K is expanded.

It has been reported that this method can be quite successful in finding dominant eigenvalues of (strongly) diagonally dominant matrices. The matrix $(D_A - \theta_k I)^{-1}$ can be viewed as a preconditioner for the vector r. Davidson [6] suggests that his algorithm (more precisely, the Davidson–Liu variant of it) may be interpreted as a Newton–Raphson scheme, and this has been used as an argument to explain its fast convergence. It is tempting to also see the preconditioner as an approximation for $(A - \theta_k I)^{-1}$, and, indeed, this approach has been followed for the construction of more complicated preconditioners (see, e.g., [16], [3], [14], [17]). However, note that $(A - \theta_k I)^{-1}$ would map r onto u_k, and hence it would not lead to an expansion of our search space. Clearly, this is a wrong interpretation for the preconditioner.

2.2. The methods of Jacobi. In his paper of 1846 [12], Jacobi introduced a combination of two iterative methods for the computation of approximations of eigenvalues of a symmetric matrix.[1] He proposed the combination as an entity, but

[1] This came to our attention by reading A. den Boer's Master's thesis [7].

at present the methods are only used separately. The first method is well known and is referred to as the Jacobi method (e.g., §8.4 in [9]). It is based on Jacobi plane rotations, which are used to force the matrix A to diagonal dominance. We will refer to this method as Jacobi's diagonalisation method. The second method is much less well known and is related to the Davidson method. For ease of discussion we will call this second method the JOCC. It turns out that Davidson's method can be interpreted as an accelerated JOCC method, just as Arnoldi's method can be seen as an accelerated power method.

2.2.1. The JOCC method. Jacobi considered an eigenvalue problem as a system of linear equations for which his iterative linear solver [11], the Jacobi or Gauss–Jacobi iteration (e.g., §10.1 in [9]), might be applicable.

Suppose we have a diagonally dominant matrix A, of which $a_{1,1} = \alpha$ is the largest diagonal element. Then α is an approximation for the largest eigenvalue λ, and e_1 is an approximation for the corresponding eigenvector u. In modern matrix notation (which was unknown in Jacobi's time), his approach can be presented as follows.

Consider the eigenvalue problem

$$A \begin{bmatrix} 1 \\ z \end{bmatrix} = \begin{bmatrix} \alpha & c^T \\ b & F \end{bmatrix} \begin{bmatrix} 1 \\ z \end{bmatrix} = \lambda \begin{bmatrix} 1 \\ z \end{bmatrix}, \tag{1}$$

where F is a square matrix, α is a scalar, and b, c and z are vectors of appropriate size. We are interested in the eigenvalue λ that is close in some sense to α, and in the corresponding eigenvector $u = (1, z^T)^T$, with component z orthogonal to e_1. Problem (1) is equivalent with

$$\lambda = \alpha + c^T z, \tag{2}$$

$$(F - \lambda I)z = -b. \tag{3}$$

Jacobi proposed to solve (3) iteratively by his Jacobi iteration, with $z_1 = 0$, and an updated approximation for λ, using (2):[2]

$$\begin{cases} \theta_k &= \alpha + c^T z_k, \\ (D - \theta_k I) z_{k+1} &= (D - F) z_k - b, \end{cases} \tag{4}$$

where D is the diagonal of F (although θ_k is not a Ritz value, we have used it to characterize it as an eigenvalue approximation).

Jacobi solved, as is also customary now, the update y_k for z_k from the diagonal system rather than solving z_{k+1} directly. Therefore, a better representation of the JOCC method would be

$$\begin{cases} \mu_k = c^T y_k, \\ \theta_{k+1} = \theta_k + \mu_k, \\ z_{k+1} = z_k + y_k, \\ (D - \theta_{k+1} I) y_{k+1} = (D - F) y_k + \mu_k z_{k+1}, \end{cases} \tag{5}$$

with $z_1 = 0$, $\theta_1 = \alpha$, and $y_1 = -(D - \alpha I)^{-1} b$. However, in connection with Davidson's method, representation (4) simplifies the discussion.

[2] Actually, Jacobi updated the approximation of λ only at every even step. There is no explanation for why he did not update in the odd steps as well.

2.2.2. Short discussion on the JOCC method. Jacobi was well aware of the fact that the Jacobi iteration converges (fast) if the matrix is (strongly) diagonally dominant. Therefore, he proposed to perform a number of steps of the Jacobi diagonalization method in order to obtain a (strongly) diagonally dominant matrix before applying the JOCC method. Since he was interested in the eigenvalue closest to α, he did enough steps using the diagonalization method to obtain a diagonally dominant $F-\alpha I$ so that $F-\theta_k I$ was also diagonally dominant. This can be done provided that λ is a simple eigenvalue of A. The application of the diagonalization method can be viewed as a technique to improve the initial guess e_1, i.e., the given matrix is rotated so that e_1 is closer to the (rotated) eigenvector u. These rotations were done only at the start of the iteration process, and this process was carried out with fixed F and D.

However, note that in Jacobi's approach we are looking, at all stages, for the orthogonal complement to the *initial* approximation $u_1 = e_1$. We do not take into account that better approximations $u_k = (1, z_k^T)^T$ become available in the process, and that it may be more efficient to try to compute the orthogonal complement $u - (u^T u_k)u_k$. In the JOCC framework an improved approximation would have led to a similar situation as in (1), if we would have applied plane rotations on A, such that u_k would have been rotated to e_1 by these rotations. Therefore, what Jacobi did only in the first step of the process could have been done *at each step* of the iteration process. This is an exciting idea, since in Jacobi's approach the rotations were motivated by the desire to obtain stronger diagonal dominance, whereas our discussion suggests that one might take advantage of the information in the iteration process. Of course, this would have led to a different operator F in each step and this is an important observation for the formulation of our new algorithm.

2.3. Davidson's method as an accelerated JOCC method. We will apply Davidson's method for the same problem as before. In particular we will assume that A is as in (1) and that $u_1 = e_1$.

The eigenvector approximations produced in the kth step of JOCC, as well as in the Davidson method, are denoted by u_k. We assume that u_k is scaled such that its first coordinate is 1: $u_k = (1, z_k^T)^T$. Let θ_k be the associated approximation to the eigenvalue. It will be clear from the context to which process an approximation refers.

The residual is given by

$$r_k = (A - \theta_k I)u_k = \begin{bmatrix} \alpha - \theta_k + c^T z_k \\ (F - \theta_k I)z_k + b \end{bmatrix}. \tag{6}$$

Davidson proposes computing t_k from

$$(D_A - \theta_k I)t_k = -r_k, \tag{7}$$

where D_A is the diagonal of A. For the component $\widehat{y}_k := (0, y_k^T)^T$ of t_k orthogonal to u_1 it follows, with D the diagonal of F, that

$$(D - \theta_k I)y_k = -(F - \theta_k I)z_k - b = (D - F)z_k - (D - \theta_k I)z_k - b, \tag{8}$$

or equivalently,

$$(D - \theta_k I)(z_k + y_k) = (D - F)z_k - b. \tag{9}$$

Comparing (9) with (4), we see that $z_k + y_k$ is the z_{k+1} that we would have obtained with one step of JOCC starting at z_k. But after this point, Davidson's method is

an improvement over the JOCC method because instead of taking $\widehat{u}_{k+1} := (1, (z_k + y_k)^T)^T = u_k + \widehat{y}_k$ as the next approximating eigenvector (as in JOCC), Davidson suggests computing the Ritz vector of A with respect to the subspace computed so far (that is, over the subspace spanned by the old approximations $u_1, \ldots, u_k$ and the new $\widehat{u}_{k+1}$). Actually, Davidson selects the correction t_k, but

$$\operatorname{span}(u_1, \ldots, u_k, \widehat{u}_{k+1}) = \operatorname{span}(u_1, \ldots, u_k, t_k).$$

For the computation of the Ritz vector it is convenient to have an orthonormal basis, and that is precisely what is constructed in Davidson's method. This orthonormal basis $v_1, \ldots, v_{k+1}$ appears if one orthonormalizes $u_1, \ldots, u_k, t_k$ by Gram–Schmidt. The $(k+1)$th step in either JOCC or Davidson's method can be summarized as follows.

JOCC. Jacobi computes the component $\widehat{y}_k$ of t_k orthogonal to u_1 and takes $u_{k+1} = u_k + \widehat{y}_k$, $\theta_{k+1} = e_1^T A u_{k+1}$. Unlike Davidson's approach, Jacobi only computes components that are orthogonal to $u_1 = e_1$. However, in view of the orthogonalization step, the components in the u_1-direction (as well as in new directions) automatically vanish in Davidson's method.

Davidson's method. Davidson computes the component v_{k+1} of t_k orthogonal to $u_1, \ldots, u_k$ and takes for u_{k+1} and θ_{k+1} the Ritz vector, respectively, Ritz value, of A with respect to the space spanned by $v_1, \ldots, v_{k+1}$. Davidson exploits the complete subspace constructed so far, while Jacobi takes only a simple linear combination of the last vector z_k and the last correction y_k (or, taking the u_1-component into account, of u_1, the last vector u_k, and the last correction t_k). Although $\operatorname{span}(u_1, \ldots, u_k, \widehat{u}_{k+1}) = \operatorname{span}(u_1, \ldots, u_{k+1})$, the Davidson approximations do not span the same subspace as the Jacobi approximations since the eigenvalue approximations θ_k of Jacobi and of Davidson are different. Consequently, the methods have different corrections t_k and also different components orthogonal to the u_j.

Note that our formulation makes clear that Davidson's method also attempts to find the orthogonal update $(0, z^T)^T$ for the initial guess $u_1 = e_1$, and it does so by a clever orthogonalization procedure. However, just as in the JOCC approach, the process works with fixed operators (in particular D_A; other variants use different approximations for A) and not with operators associated with the orthogonal complement of the current approximation u_k (see also §2.2.2). This characterizes the difference between our algorithm (of §3) and Davidson's method.

3. The new Jacobi–Davidson iteration method. From now on we will allow the matrix A to be complex, and in order to express this we use the notation v^* for the complex conjugate of a vector (if complex), or the transpose (if real), and likewise for matrices.

As we have stressed in the previous section, the JOCC method and Davidson's method can be seen as methods that attempt to find the correction to some initially given eigenvector approximation. In fact, what we want is to find the orthogonal complement for our current approximation u_k with respect to the desired eigenvector u of A. Therefore, we are interested in seeing explicitly what happens in the subspace $u_k^{\perp}$.

The orthogonal projection of A onto that space is given by $B = (I - u_k u_k^*) A (I - u_k u_k^*)$ (we assume that u_k has been normalized). Note that for $u_k = e_1$, we have that F (cf. (1)) is the restriction of B with respect to $e_1^{\perp}$. It follows that

$$A = B + A u_k u_k^* + u_k u_k^* A - \theta_k u_k u_k^*. \tag{10}$$

When we are in search of an eigenvalue λ of A close to θ_k, then we want to have the correction $v \perp u_k$ to u_k such that

$$A(u_k + v) = \lambda(u_k + v),$$

or, after inserting (10) and realizing that $Bu_k = 0$,

$$(B - \lambda I)v = -r + (\lambda - \theta_k - u_k^* A v)u_k. \tag{11}$$

Since the left-hand side and r have no component in u_k, it follows that the factor for u_k must vanish, and hence v should satisfy

$$(B - \lambda I)v = -r. \tag{12}$$

We replace λ by the current approximation θ_k just as in JOCC and Davidson's method, but, unlike both methods, we propose to work with approximations for an operator (B) that varies from step to step. The resulting algorithm may be viewed as a combination of Jacobi's approach to look for the orthogonal complement of a given eigenvector approximation and the Davidson algorithm for expanding the subspace in which the eigenvector approximations are constructed. This explains our choice of the name Jacobi–Davidson for the new method. Note that we are free to use any method for the (approximate) solution of (12) and that it is not necessary to require diagonal dominance of B (or A).

Before we present our complete algorithm, we briefly review some different approaches.

1. If we approximate v simply by r, then we formally obtain the same results as with the Arnoldi method.
2. If we approximate v by $(D_A - \theta_k I)^{-1} r$, then we obtain the original Davidson method.
3. More recent suggestions made in [3], [14], [15], [17], [6] come down to better approximations for the inverse of $A - \theta_k I$, e.g., incomplete decompositions for this operator. However, as is well known, this is a risky approach (see [25], [3]) since the exact inverse of this operator leads to failure of the method.[3] Therefore the approximation should not be too accurate [25].
4. In our approach, we replace $B - \lambda I$ by $B - \theta_k I$, and one has to select suitable approximations $\tilde{t} \perp u_k$ for the solution of

   $$(B - \theta_k I)t = -r \quad \text{and} \quad t \perp u_k. \tag{13}$$

   This will lead to quadratical convergence if we take the exact solution t (see observation 3 in §4), which is in sharp contrast with what happens if one takes the exact solution of $(A - \theta_k I)t = -r$.
5. Another viewpoint on our modified Davidson algorithm is that it can be seen as an accelerated inexact shift and invert method (that is, the "invert" part may be inexact).

We do not know, for general systems, how to approximate a solution of (13) sufficiently well with modest computational effort. In most of our numerical experiments, we have constructed approximations by carrying out only a few steps of an iterative method (for instance, generalized minimum residual: GMRES [26]) in order

[3] Any progress in this case may be attributed to the effects of rounding errors.

ALGORITHM 1. *The Jacobi–Davidson method.*

1. **Start:** Choose an initial nontrivial vector v.
   - Compute $v_1 = v/\|v\|_2$, $w_1 = Av_1$, $h_{11} = v_1^* w_1$, set $V_1 = [v_1]$, $W_1 = [w_1]$, $H_1 = [h_{11}]$, $u = v_1$, $\theta = h_{11}$, compute $r = w_1 - \theta u$.
2. **Iterate:** Until convergence do:
3. **Inner Loop:** For $k = 1, \ldots, m-1$ do:
   - Solve (approximately) $t \perp u$,
   $$(I - uu^*)(A - \theta I)(I - uu^*)t = -r.$$
   - Orthogonalize t against V_k via modified Gram–Schmidt, and expand V_k with this vector to V_{k+1}.
   - Compute $w_{k+1} := Av_{k+1}$ and expand W_k with this vector to W_{k+1}.
   - Compute $V_{k+1}^* w_{k+1}$, the last column of $H_{k+1} := V_{k+1}^* A V_{k+1}$, and $v_{k+1}^* W_k$, the last row of H_{k+1} (only if $A \neq A^*$).
   - Compute the largest eigenpair (θ, s) of H_{k+1} (with $\|s\|_2 = 1$).
   - Compute the Ritz vector $u := V_{k+1}s$, compute $\widehat{u} := Au$ $(= W_{k+1}s)$, and the associated residual vector $r := \widehat{u} - \theta u$.
   - Test for convergence. Stop if satisfied.
4. **Restart:** Set $V_1 = [u]$, $W_1 = [\widehat{u}]$, $H_1 = [\theta]$, and goto 3.

to illustrate how powerful our approach is even when we solve the system only in very modest precision, but this is certainly not an optimal choice in many practical situations. However, our experiments illustrate that the better we approximate the solution $\tilde{v}$ of (13), the faster convergence we obtain (and no stagnation as in Davidson's method).

The algorithm for the improved Davidson method then becomes as in Algorithm 1 (in the style of [25]). We have skipped indices for variables that overwrite old values in an iteration step, e.g., u instead of u_k. We do not discuss implementation issues, but we note that the computational costs for Algorithm 1 are about the same as for Davidson's method (provided that the same amount of computational work is spent to approximate the solutions of the involved linear systems).

4. Further discussion. In this section we will discuss a convenient way of incorporating preconditioning in the Jacobi–Davidson method. We will also discuss relations with other methods, e.g., shift and invert techniques, and we will try to get some insight into the behavior of the new method in situations where Davidson's method or shift and invert methods work well. This will make the differences between these methods clearer.

In the Jacobi–Davidson method we must solve (13), or equivalently

$$(I - u_k u_k^*)(A - \theta_k I)(I - u_k u_k^*)t = -r \quad \text{and} \quad t \perp u_k \tag{14}$$

(see Algorithm 1, in which we have skipped the index k).

Equation (14) can be solved approximately by selecting some more easily invertible approximation for the operator $(I - u_k u_k^*)(A - \theta_k I)(I - u_k u_k^*)$, or by some (preconditioned) iterative method. If any approximation (preconditioner) is available, then this will most often be an approximation for $A - \theta_k I$.

However, the formulation in (14) is not very suitable for incorporating available approximations for $A - \theta_k I$. We will first discuss how to construct approximate solutions orthogonal to u_k straight from a given approximation for $A - \theta_k I$ (1-step approximation: §4.1). Then we will propose how to compute such an approximated solution efficiently by a preconditioned iterative scheme (iterative approximation: §4.2).

4.1. 1-step approximations. A more convenient formulation for (14) is obtained as follows. We are interested in determining $t \perp u_k$, and for this t we have that

$$(I - u_k u_k^*)t = t,$$

and then it follows from (14) that

$$(A - \theta_k I)t - \varepsilon u_k = -r \tag{15}$$

or

$$(A - \theta_k I)t = \varepsilon u_k - r.$$

When we have a suitable preconditioner $M \approx A - \theta_k I$ available, then we can compute an approximation $\tilde{t}$ for t:

$$\tilde{t} = \varepsilon M^{-1} u_k - M^{-1} r. \tag{16}$$

The value of ε is determined by the requirement that $\tilde{t}$ should be orthogonal to u_k:

$$\varepsilon = \frac{u_k^* M^{-1} r}{u_k^* M^{-1} u_k}. \tag{17}$$

Equation (16) leads to several interesting observations.

1. If we choose $\varepsilon = 0$, then we obtain the Davidson method (with preconditioner M). In this case $\tilde{t}$ will not be orthogonal to u_k.
2. If we choose ε as in (17), then we have an instance of the Jacobi–Davidson method. This approach has already been suggested in [18]. In that paper the method is obtained from a first-order correction approach for the eigensystem. Further experiments with this approach are reported in [27]. Note that this method requires two operations with the preconditioning matrix per iteration.
3. If $M = A - \theta_k I$, then (16) reduces to

$$t = \varepsilon(A - \theta_k I)^{-1} u_k - u_k.$$

Since t is made orthogonal to u_k afterwards, this choice is equivalent with $t = (A - \theta_k I)^{-1} u_k$. In this case the method is just mathematically equivalent to (accelerated) shift and invert iteration (with optimal shift). For symmetric A this is the (accelerated) inverse Rayleigh quotient method, which converges cubically [21]. In the unsymmetric case we have quadratical convergence [21], [19]. In view of the speed of convergence of shift and invert methods, it

may hardly be worthwhile to accelerate them in a "Davidson" manner: the overhead is significant and the gains may only be minor. Moreover, in finite precision arithmetic the vector $(A - \theta_k I)^{-1} u_k$ may make a very small angle with u_k, so that it will be impossible then to compute a significant orthogonal search direction.

4. If $M \neq A - \theta_k I$, then with $\tilde{t} = M^{-1} u_k$ we obtain an inexact shift and invert method with "Davidson" subspace acceleration. This method may be an attractive alternative for the previous one if the invert part cannot be carried out exactly. Also in this variant we have no orthogonality between $\tilde{t}$ and u_k. If M is a good approximation for $A - \theta_k I$ then $M^{-1} u_k$ may also make a very small angle with u_k, so that effective subspace expansion will be impossible (as in 3).

The methods suggested in the first and the third observation are well known, and the question arises whether we may gain anything by the less well known second alternative (or the fourth one).

To get some insight into this matter, we consider a situation for which Davidson's method converges rapidly, namely, when A is strongly diagonally dominant. We write

$$A = D_A + E,$$

where D_A denotes the diagonal of A, and we assume that $\|E\|_\infty$ is small compared with $\|D_A\|_\infty$ and that a_{11} is the largest diagonal element in absolute value (note that this also includes situations where only the largest diagonal element has relatively small off-diagonal elements in the same row and column).

We write $u_k = e_1 + f$ and assume that $\|f\|_\infty \ll \|e_1\| = 1$ (which is a natural assumption in this case).

Then for the coordinates $(r)_i$ of r it follows that

$$(r)_1 = (a_{11} - \theta_k) + (Eu_k)_1 + (a_{11} - \theta_k)(f)_1$$

and

$$(r)_i = (Eu_k)_i + (a_{ii} - \theta_k)(f)_i.$$

Since $\theta_k \approx a_{11}$, this means that the coordinates $(r)_i$ are not small relative to $(r)_1$. In the case that $f = 0$ we even have that $r = Eu_k$, and $(r)_1 = 0$ (since $r \perp u_k = e_1$).

With Davidson's method we obtain

$$\tilde{t} = (D_A - \theta_k I)^{-1} r = u_k + (D_A - \theta_k I)^{-1} E u_k,$$

and the part $(D_A - \theta_k I)^{-1} E u_k$ of this vector will not be very small compared to u_k (for $f = 0$ the component u_k even vanishes). This means that we may expect to recover this part in a large number of significant digits after orthogonalizing $\tilde{t}$ with respect to u_k, and this makes Davidson's method work well for diagonally dominant problems (since we expand the search space by a well-determined vector).

We have seen the effect of the component $(D_A - \theta_k I)^{-1} r$, and now we consider what happens with the component $(D_A - \theta_k I)^{-1} u_k$ in the Jacobi–Davidson method for this situation. To this end we compute $\tilde{t}$ as in observation 4 above:

$$\tilde{t} = (D_A - \theta_k I)^{-1} u_k.$$

For the coordinates of $\tilde{t}$ we have that

$$(\tilde{t})_i = \frac{(u)_i}{a_{ii} - \theta_k},$$

and we see that $\tilde{t}$ will make a very small angle with $u_k \approx e_1$ (since $a_{11} \approx \theta_k$). This implies that $\tilde{t} - (\tilde{t}^* u_k)u_k$ may have only a few significant digits, and then it may be a waste of effort to expand the subspace with this (insignificant) vector. However, in the Jacobi–Davidson method we compute the new vector as

$$\tilde{t} = \varepsilon(D_A - \theta_k I)^{-1}u_k - (D_A - \theta_k I)^{-1}r_k.$$

The factor ε is well determined, and note that for our strongly diagonally dominant model problem we have in good approximation that

$$\begin{aligned}\|\varepsilon(D_A - \theta_k I)^{-1}u_k\|_2 &\lesssim \frac{\|u_k\|_2\,\|(D_A - \theta_k I)^{-1}r\|_2}{\|u_k\|_2\,\|(D_A - \theta_k I)^{-1}u_k\|_2}\,\|(D_A - \theta_k I)^{-1}u_k\|_2 \\ &= \|(D_A - \theta_k I)^{-1}r\|_2.\end{aligned}$$

Furthermore, since $u_k \perp r$, we have that $\varepsilon(D_A - \theta_k I)^{-1}u_k$ and $(D_A - \theta_k I)^{-1}r$ are not in the same direction, and therefore there will be hardly any cancellation in the computation of $\tilde{t}$. This means that $\tilde{t}$ is well determined in finite precision and $\tilde{t} \perp u_k$.

Our discussion can be adapted for nondiagonally dominant matrices as well, when we restrict ourselves to situations where the approximations u_k and θ_k are sufficiently close to their limit values and where we have good preconditioners (e.g., inner iteration methods).

We will illustrate our observations by a simple example taken from [3]: Example 5.1. In that example the matrix A of dimension 1000 has diagonal elements $a_{j,j} = j$. The elements on the sub- and super-diagonal ($a_{j-1,j}$ and $a_{j,j+1}$) are all equal to 0.5, as well as the elements $a_{1,1000}$ and $a_{1000,1}$.

For this matrix we compute the largest eigenvalue ($\approx$ 1000.225641) with (a) the standard Lanczos method, (b) Davidson's method with diagonal preconditioning ($(D_A - \theta_k I)^{-1}$), and (c) the Jacobi–Davidson method with the same diagonal preconditioning, carried out as in (16).

With the same starting vector as in [3] we obtain, of course, the same results: a slowly converging Lanczos process, a much faster Davidson process, and Jacobi–Davidson is just as fast as Davidson in this case. The reason for this is that the starting vector $e_1 + e_{1000}$ for Davidson and $\approx e_1 + 2000e_{1000}$ for Lanczos are quite good for these processes, and the values for ε, which describe the difference between (b) and (c), are very small in this case. Shift and invert Lanczos with shift 1001.0 takes 5 steps for full convergence, whereas Jacobi–Davidson with exact inverse for $A - \theta_k I$ takes 3 steps.

In Table 1 we see the effects when we take a slightly different starting vector $u_1 = (0.01, 0.01, \ldots, 0.01, 1)^T$, that is, we have taken a starting vector which still has a large component in the dominating eigenvector. This is reflected by the fact that the Ritz value in the first step of all three methods is equal to $954.695\cdots$. In practical situations we will often not have such good starting vectors available. The Lanczos process converges slowly again, as might be expected for this uniformly distributed spectrum. In view of our discussion in §2.3 we may interpret the new starting vector as a good starting vector for a perturbed matrix A. This implies that the diagonal preconditioner may not be expected to be a very good preconditioner. This is reflected by the very poor convergence behavior of Davidson's method. The difference with the Jacobi–Davidson method is now quite notable (see the values of ε), and for this method we observe rather fast convergence again.

TABLE 1
Approximation errors $\lambda - \theta_k$.

Iteration	Lanczos	Davidson	Jacobi–Davidson	$\lvert\varepsilon\rvert$
0	0.45e+02	0.45e+02	0.45e+02	
1	0.56e+01	0.40e+02	0.25e+02	0.50e+02
2	0.16e+01	0.40e+02	0.74e+01	0.12e+03
3	0.71e+00	0.40e+02	0.15e+01	0.11e+02
4	0.43e+00	0.40e+02	0.14e+01	0.14e+01
5	0.32e+00	0.40e+02	0.55e–01	0.49e+00
6	0.26e+00	0.39e+02	0.13e–02	0.72e–01
7	0.24e+00	0.38e+02	0.29e–04	0.29e–02
8	0.22e+00	0.37e+02	0.33e–06	0.14e–03
9	0.21e+00	0.36e+02	0.25e–08	0.34e–05
10	0.20e+00	0.36e+02		
11	0.19e+00	0.35e+02		
12	0.19e+00	0.34e+02		
13	0.18e+00	0.33e+02		
14	0.17e+00	0.32e+02		
15	0.16e+00	0.31e+02		

Although this example may seem quite artificial, it displays quite nicely the behavior that we have seen in our experiments, as well as what we have tried to explain in our discussion.

In conclusion, Davidson's method works well in these situations where $\tilde{t}$ does not have a strong component in the direction of u_k. Shift and invert approaches work well if the component in the direction of u in u_k is strongly increased. However, in this case this component may dominate so strongly (when we have a good preconditioner) that it prevents us from reconstructing in finite precision arithmetic a relevant orthogonal expansion for the search space. In this respect, the Jacobi–Davidson method is a compromise between the Davidson method and the accelerated (inexact) shift and invert method, since the factor ε properly controls the influence of u_k and makes sure that we construct the orthogonal expansion of the subspace correctly. In this view Jacobi–Davidson offers the best of two worlds, and this will be illustrated by our numerical examples.

4.2. Iterative approximations. If a preconditioned iterative method is used to solve (14), then, in each step of the linear solver, a "preconditioning equation" has to be solved.

If M is some approximation of $A - \theta_k I$ then the projected matrix

$$M_d := (I - u_k u_k^*)\, M\, (I - u_k u_k^*)$$

can be taken as an approximation of $(I - u_k u_k^*)(A - \theta_k I)(I - u_k u_k^*)$ and, in each iterative step, we will have to solve an equation of the form $M_d z = y$, where y is some given vector orthogonal to u_k and $z \perp u_k$, has to be computed. Of course, z can be computed as (cf. (16)–(17))

$$(M_d^{-1} y =) \quad z = \alpha M^{-1} u_k - M^{-1} y \quad \text{with} \quad \alpha = \frac{u_k^* M^{-1} y}{u_k^* M^{-1} u_k}.$$

In this approach, we have to solve, except for the first iteration step, only one system involving M in each iteration step. The inner product $u_k^* M^{-1} u_k$, to be computed only once, can also be used in all steps of the iteration process for (14).

The use of a (preconditioned) Krylov subspace iteration method for (14) does not lead to the same result as when we apply this iterative method to the two equations in (16) separately. For instance, if p is a polynomial such that $p(A-\theta_k I) \approx (A-\theta_k I)^{-1}$ then, with $M^{-1} = p(A-\theta_k I)$, (16) can be used to find an approximate solution of (14) leading to

$$\tilde{t} = \varepsilon p(A-\theta_k I)u_k - p(A-\theta_k I)r = (I-u_k u_k^*)\, p\left(A-\theta_k I\right)\, (I-u_k u_k^*)r, \tag{18}$$

while using p directly for (14) would yield

$$\tilde{t} = p\left((I-u_k u_k^*)\,(A-\theta_k I)\,(I-u_k u_k^*)\right) r. \tag{19}$$

Clearly, these expressions are not identical. For Krylov subspace methods that automatically (and implicitly) determine such polynomials, the differences may be even more significant. Most importantly, such a method for (14) would be aiming for an approximation of the inverse of $(I-u_k u_k^*)(A-\theta_k I)\,(I-u_k u_k^*)$ on the space orthogonal to u_k, rather than for an approximation of $(A-\theta_k I)^{-1}$ as the method for (16) would do. If θ_k is an accurate approximation of the eigenvalue λ, $A-\theta_k I$ will be almost singular, while that will not be the case for the projected matrix $(I-u_k u_k^*)(A-\theta_k I)\,(I-u_k u_k^*)$ (as a map on $u_k^{\perp}$, if λ is simple). This means that the iterative solution of (14) may be easier than iteratively solving systems such as $(A-\theta_k I)z = y$.

By iteratively solving (14) we expect more stable results: by putting the intermediate approximations orthogonal to u_k (as, for instance, in (19)) we may hope to have less cancellation by rounding errors than when putting only the final approximation $\tilde{t}$ orthogonal to u_k (as, for instance, in (18)).

We cannot answer the question of how accurately (14) should be solved in order to have convergence for the Jacobi–Davidson method. Our experiences, as well as experiences reported in [2], seem to indicate that even a modest error reduction in the solution of (14) suffices and more work spent in this (inner) iteration for (14) often leads to a reduction in the number of Jacobi–Davidson iteration steps. For some numerical evidence, see §6.

5. Jacobi–Davidson with harmonic Ritz values. In the previous sections we have used the Galerkin approach for the approximation of eigenvectors and eigenvalues. In this approach, H_k is the orthogonal projection of the matrix A onto $\mathcal{V}_k = \{v_1, \ldots, v_k\}$, and its eigenvalues are called the Ritz values of A with respect to $\mathcal{V}_k$ [22]. The Ritz values converge monotonically to the extremal eigenvalues when A is symmetric. If A is nonsymmetric, the convergence is in general not monotonical, but the convergence behavior is still often quite regular with respect to the extremal eigenvalues. Interesting observations for the nonsymmetric case have been made in [24], [10].

For the "interior" (the non extremal) eigenvalues the situation is less clear. The convergence can be very irregular, even in the symmetric situation (due to rounding errors). This behavior makes it difficult to approximate interior eigenvalues or to design algorithms that select the correct Ritz values and handle rounding errors well (see, e.g., [24]).

In [14] the author suggested using a minimum residual approach for the computation of interior eigenvalues. We follow a slightly different approach which leads to identical results for symmetric matrices. In this approach, as we will show in the next section, we use orthogonal projections onto $A\mathcal{V}_k$. The obtained eigenvalue

approximations differ from the standard Ritz values for A. In a recent paper [20], these approximations were called *harmonic Ritz values*, and they were identified as inverses of Ritz approximations for the inverse of A. It was also shown in [20] that, for symmetric matrices, these harmonic Ritz values exhibit a monotonic convergence behavior with respect to the eigenvalues with smallest absolute value. This further supports the observation made in [14] that, for the approximation of "interior" eigenvalues (close to some $\mu \in \mathbb{C}$), more regular convergence behavior with the harmonic Ritz values (of $A - \mu I$) than with the Ritz values may be expected.

In [20] the harmonic Ritz values for symmetric matrices are discussed. The nonsymmetric case has been considered in [30], [8]. However, in all these papers the discussion is restricted to harmonic Ritz values of A with respect to Krylov subspaces. In [14] the harmonic Ritz values are considered for more general subspaces associated with symmetric matrices. The approach is based on a generalized Rayleigh–Ritz procedure, and it is pointed out in [14] that the harmonic Ritz values are to be preferred for the Davidson method when aiming for interior eigenvalues.

In connection with the Jacobi–Davidson method for unsymmetric matrices, we propose a slightly more general approach based on projections. To this end, as well as for the introduction of notations that we will also need later, we discuss to some extent the harmonic Ritz values in §5.1.

5.1. Harmonic Ritz values on general subspaces.

Ritz values. If $\mathcal{V}_k$ is a linear subspace of $\mathbb{C}^n$ then θ_k is a *Ritz value* of A with respect to $\mathcal{V}_k$ with *Ritz vector* u_k if

$$u_k \in \mathcal{V}_k,\ u_k \neq 0, \qquad Au_k - \theta u_k \perp \mathcal{V}_k. \tag{20}$$

How well the Ritz pair (θ_k, u_k) approximates an eigenpair (λ, w) of A depends on the angle between w and $\mathcal{V}_k$.

In practical computations one usually computes Ritz values with respect to a growing sequence of subspaces $\mathcal{V}_k$ (that is, $\mathcal{V}_k \subset \mathcal{V}_{k+1}$ and $\dim(\mathcal{V}_k) < \dim(\mathcal{V}_{k+1})$.

If A is normal, then any Ritz value is in the convex hull of the spectrum of A: any Ritz value is a mean (convex combination) of eigenvalues. For normal matrices, at least, this helps to explain the often regular convergence of extremal Ritz values with respect to extremal eigenvalues. For further discussions on the convergence behavior of Ritz values (for symmetric matrices), see [22], [29].

Harmonic Ritz values. A value $\widetilde{\theta}_k \in \mathbb{C}$ is a *harmonic Ritz value* of A with respect to some linear subspace $\mathcal{W}_k$ if $\widetilde{\theta}_k^{-1}$ is a Ritz value of A^{-1} with respect to $\mathcal{W}_k$ [20].

For normal matrices, $\widetilde{\theta}_k^{-1}$ is in the convex hull of the collection of λ^{-1}'s, where λ is an eigenvalue of A: any harmonic Ritz value is a harmonic mean of eigenvalues. This property explains their name and, at least for normal matrices, it explains why we may expect a more regular convergence behavior of harmonic Ritz values with respect to the eigenvalues that are closest to the origin.

Of course, we would like to avoid computing A^{-1} or solving linear systems involving A. The following theorem gives a clue about how that can be done.

THEOREM 5.1. *Let $\mathcal{V}_k$ be some k-dimensional subspace with basis $v_1, \ldots, v_k$. A value $\widetilde{\theta}_k \in \mathbb{C}$ is a harmonic Ritz value of A with respect to the subspace $\mathcal{W}_k := A\mathcal{V}_k$ if and only if*

$$A\widetilde{u}_k - \widetilde{\theta}_k \widetilde{u}_k \perp A\mathcal{V}_k \quad \text{for some} \quad \widetilde{u}_k \in \mathcal{V}_k,\ \widetilde{u}_k \neq 0. \tag{21}$$

If $w_1, \ldots, w_k$ *span* $A\mathcal{V}_k$ *then,*[4] *with*

$$V_k := [v_1|\cdots|v_k], \quad W_k := [w_1|\cdots|w_k], \quad \text{and} \quad \widetilde{H}_k := (W_k^* V_k)^{-1} W_k^* A V_k,$$

property (21) *is equivalent to*

$$\widetilde{H}_k s = \widetilde{\theta}_k s \quad \text{for some} \quad s \in \mathbb{C}^k,\ s \neq 0 \quad (\text{and} \quad \widetilde{u}_k = V_k s): \tag{22}$$

the eigenvalues of the $k \times k$ *matrix* $\widetilde{H}_k$ *are the harmonic Ritz values of* A.

Proof. By (20), $(\widetilde{\theta}_k^{-1}, A\widetilde{u}_k)$ is a Ritz pair of A^{-1} with respect to $A\mathcal{V}_k$ if and only if

$$(A^{-1} - \widetilde{\theta}_k^{-1} I) A\widetilde{u}_k \perp A\mathcal{V}_k$$

for a $\widetilde{u}_k \in \mathcal{V}_k$, $\widetilde{u}_k \neq 0$.

Since $(A^{-1} - \widetilde{\theta}_k^{-1} I) A\widetilde{u}_k = -\widetilde{\theta}_k^{-1}(A\widetilde{u}_k - \widetilde{\theta}_k \widetilde{u}_k)$ we have the first property of the theorem.

For the second part of the theorem, note that (21) is equivalent to

$$AV_k s - \widetilde{\theta}_k V_k s \perp \mathcal{W}_k \quad \text{for an} \quad s \neq 0, \tag{23}$$

which is equivalent to

$$W_k^* A V_k s - \widetilde{\theta}_k (W_k^* V_k) s = 0$$

or $\widetilde{H}_k s - \widetilde{\theta}_k s = 0$. □

We will call the vector $\widetilde{u}_k$ in (21) the *harmonic Ritz vector* associated with the *harmonic Ritz value* $\widetilde{\theta}_k$ and $(\widetilde{\theta}_k, \widetilde{u}_k)$ is a *harmonic Ritz pair.*

In the context of Krylov subspace methods (Arnoldi or Lanczos), $\mathcal{V}_k$ is the Krylov subspace $\mathcal{K}_k(A; v_1)$. The v_j are orthonormal and such that $v_1, \ldots, v_i$ span $\mathcal{K}_i(A; v_1)$ for $i = 1, 2, \ldots$. Then $AV_k = V_{k+1} H_{k+1,k}$, with $H_{k+1,k}$ a $(k+1) \times k$ upper Hessenberg matrix.

The elements of $H_{k+1,k}$ follow from the orthogonalization procedure for the Krylov subspace basis vectors. In this situation, with $H_{k,k}$ the upper $k \times k$ block of $H_{k+1,k}$, we see that

$$\begin{aligned}(W_k^* V_k)^{-1} W_k^* A V_k &= (H_{k+1,k}^* V_{k+1}^* V_k)^{-1} H_{k+1,1}^* V_{k+1}^* V_{k+1} H_{k+1,k} \\ &= {H_{k,k}^*}^{-1} H_{k+1,k}^* H_{k+1,k}.\end{aligned}$$

Since $H_{k+1,k} = H_{k,k} + \beta e_{k+1} e_k^*$, where β is equal to the element in position $(k+1, k)$ of $H_{k+1,k}$, the harmonic Ritz values can be computed from a matrix which is a rank-one update of $H_{k,k}$:

$$\widetilde{H}_k = {H_{k,k}^*}^{-1} H_{k+1,k}^* H_{k+1,k} = H_{k,k} + |\beta|^2 {H_{k,k}^*}^{-1} e_k e_k^*.$$

In [8], the author is interested in quasi-kernel polynomials (e.g., GMRES and quasi-minimal residual (QMR) polynomials). The zeros of these polynomials are harmonic Ritz values with respect to Krylov subspaces. This follows from Corollary 5.3 in [8], where these zeros are shown to be the eigenvalues of ${H_{k,k}^*}^{-1} H_{k+1,k}^* H_{k+1,k}$. However,

[4] If $A\mathcal{V}_k$ has dimension less than k, then this subspace contains an eigenvector of A; this situation is often referred to as a lucky breakdown. We do not consider this situation here.

in that paper these zeros are not interpreted as the Ritz values of A^{-1} with respect to some Krylov subspace.

In the context of Davidson's method we have more general subspaces $\mathcal{V}_k$ and $\mathcal{W}_k = A\mathcal{V}_k$. According to Theorem 5.1 we have to construct the matrix $\widetilde{H}_k$ (which will not be Hessenberg in general), and this can be accomplished by either constructing an orthonormal basis for $A\mathcal{V}_k$ (similar to Arnoldi's method) or by constructing bi-orthogonal bases for $\mathcal{V}_k$ and $A\mathcal{V}_k$ (similar to the bi-Lanczos method). We will consider this in more detail in §5.2.

5.2. The computation of the harmonic Ritz values.

5.2.1. Bi-orthogonal basis construction. In our algorithms, we expand the subspace $\mathcal{V}_k$ by one vector in each sweep of the iteration. We proceed as follows.

Suppose that $v_1, \ldots, v_k$ span $\mathcal{V}_k$ and that $w_1, \ldots, w_k$ span $A\mathcal{V}_k$, in such a way that, with $V_k := [v_1|\cdots|v_k]$ and $W_k := [w_1|\cdots|w_k]$,

$$AV_k = W_k$$

and

$$L_k := W_k^* V_k \quad \text{is lower triangular}$$

(in this case we say that W_k and V_k are bi-orthogonal).

According to Theorem 5.1 the harmonic Ritz values are the eigenvalues of

$$\widetilde{H}_k := L_k^{-1}\widehat{H}_k, \quad \text{where} \quad \widehat{H}_k := W_k^* W_k.$$

Hence, if $(\widetilde{\theta}_k, s)$ is an eigenpair of $\widetilde{H}_k$ then $(\widetilde{\theta}_k, V_k s)$ is a harmonic Ritz pair.

Let t be the vector by which we want to expand the subspace $\mathcal{V}_k$.

First, we bi-orthogonalize t with respect to V_k and W_k:

$$\widetilde{t} := t - V_k L_k^{-1} W_k^* t \quad \text{and} \quad v_{k+1} := \frac{\widetilde{t}}{\|\widetilde{t}\|_2}. \tag{24}$$

Then v_{k+1} is our next basis vector in the $\mathcal{V}$-space and we expand V_k by v_{k+1} to V_{k+1}: $V_{k+1} := [V_k|v_{k+1}]$.

Next, we compute $w_{k+1} := Av_{k+1}$, our next basis vector in the $A\mathcal{V}$-space, and we expand W_k to W_{k+1}.

Then the vector $w_{k+1}^* V_{k+1}$ is computed and L_k is expanded to L_{k+1} by this $k+1$ row vector. Finally, we compute $W_{k+1}^* w_{k+1}$ as the new column of $\widehat{H}_{k+1}$. By symmetry, we automatically have the new row of $\widehat{H}_{k+1}$.

Since $L_k^* = V_k^* A V_k$ and L_k^* is upper triangular, we see that L_k is diagonal if A is self-adjoint.

The formulation of the bi-orthogonalization step (24) does not allow for the use of modified Gram–Schmidt orthogonalization (due to the $k \times k$ matrix L_k). We can incorporate L_k into W_k by working with $\widetilde{W}_k := W_k L_k^{-*}$ instead of with W_k, and then modified Gram–Schmidt is possible:

$$v^{(1)} = t, \quad v^{(i+1)} = v^{(i)} - v_i \widetilde{w}_i^* v^{(i)} \quad (i = 1, \ldots, k-1), \quad \widetilde{t} = v^{(k)}.$$

However, in this approach we have to update $\widetilde{W}_k$, which would require k additional long vector updates.

5.2.2. Orthogonal basis construction. For stability reasons one might prefer to work with an orthogonal basis rather than with bi-orthogonal ones. In the context of harmonic Ritz values, an orthonormal basis for the image space $A\mathcal{V}_k$ is attractive:

$$
\begin{aligned}
w = At, \quad \widetilde{w} := w - W_k W_k^* w, \quad &\text{and} \quad w_{k+1} := \frac{\widetilde{w}}{\|\widetilde{w}\|_2}, \\
\widetilde{t} := t - V_k W_k^* w \quad &\text{and} \quad v_{k+1} := \frac{\widetilde{t}}{\|\widetilde{w}\|_2}.
\end{aligned} \tag{25}
$$

Then $W_k^* W_k = I$, $W_k = AV_k$, and (cf. Theorem 5.1)

$$
\widetilde{H}_k = (W_k^* V_k)^{-1} W_k^* A V_k = (W_k^* V_k)^{-1}. \tag{26}
$$

It is not necessary to invert $W_k^* V_k$ since the harmonic Ritz values are simply the inverses of the eigenvalues of $W_k^* V_k$. The construction of an orthogonal basis for $A\mathcal{V}_k$ can be done with modified Gram–Schmidt.

Finally, note that $\widetilde{H}_k^{-1} = W_k^* V_k = W_k^* A^{-1} W_k$, and we explicitly have the matrix of the projection of A^{-1} with respect to an orthonormal basis of $\mathcal{W}_k$. This again reveals how the harmonic Ritz values appear as inverses of Ritz values of A^{-1} with respect to $A\mathcal{V}_k$.

5.3. A restart strategy. In the Jacobi–Davidson algorithm, Algorithm 1, it is suggested, just as for the original Davidson method, to restart simply by taking the Ritz vector u_m computed so far as a new initial guess. However, the process may construct a new search space that has considerable overlap with the previous one; this phenomenon is well known for the restarted power method and the restarted Arnoldi (without deflation) and it may lead to a reduced speed of convergence or even to stagnation. One may try to prevent this by retaining part of the search space, i.e., by returning to step 3 of Algorithm 1 with a well chosen ℓ-dimensional subspace of the span of $v_1, \ldots, v_m$ for some $\ell > 1$. With our simple restart, we expect that the process will also construct vectors with strong components in directions of eigenvectors associated with eigenvalues close to the wanted eigenvalue. And this is just the kind of information that we have discarded at the point of restart.

This suggests a strategy of retaining ℓ Ritz vectors associated with the Ritz values closest to this eigenvalue as well (including the Ritz vector u_m that is the approximation for the desired eigenvector). In Algorithm 1, these would be the ℓ largest Ritz values. A similar restart strategy can be used for the harmonic Ritz values and, say, bi-orthogonalization: for the initial matrices V_ℓ and W_ℓ after restart we should take care that $W_\ell = AV_\ell$ and the matrices should be bi-orthogonal (i.e., $W_\ell^* V_\ell$ should be lower triangular).

5.4. The use of the harmonic Ritz values. According to our approaches for the computation of harmonic Ritz values in §5.2, there are two variants for an algorithm that exploits the convergence properties of the harmonic Ritz values toward the eigenvalues closest to the origin. Of course, these algorithms can also be used to compute eigenvalues that are close to some $\mu \in \mathbb{C}$. In that case one should work with $A - \mu I$ instead of A.

We start with the variant based on bi-orthogonalization.

5.4.1. Jacobi–Davidson with bi-orthogonal basis. When working with harmonic Ritz values, we have to be careful in applying Jacobi's expansion technique. If $(\widetilde{\theta}_k, \widetilde{u}_k)$ is a harmonic Ritz pair of A then $r = A\widetilde{u}_k - \widetilde{\theta}_k \widetilde{u}_k$ is orthogonal to $A\widetilde{u}_k$, whereas in our discussion about the new Jacobi–Davidson method with regular Ritz

values (cf. §3) the vector r was orthogonal to u_k. However, we can follow Jacobi's approach here as well by using a skew basis or a skew projection. The update for $\widetilde{u}_k$ should be in the space orthogonal to $\widehat{u} := A\widetilde{u}_k$. If λ is the eigenvalue of A closest to the harmonic Ritz value $\widetilde{\theta}_k$, then the optimal update is the solution v of

$$(B - \lambda I)v = -r \quad \text{where} \quad B := \left(I - \frac{\widetilde{u}_k \widehat{u}^*}{\widehat{u}^* \widetilde{u}_k}\right) A \left(I - \frac{\widetilde{u}_k \widehat{u}^*}{\widehat{u}^* \widetilde{u}_k}\right). \tag{27}$$

In our algorithm we will solve this equation (27) approximately. To be more precise, we solve approximately

$$\left(I - \frac{\widetilde{u}_k \widehat{u}^*}{\widehat{u}^* \widetilde{u}_k}\right) (A - \widetilde{\theta}_k I) \left(I - \frac{\widetilde{u}_k \widehat{u}^*}{\widehat{u}^* \widetilde{u}_k}\right) t = -r. \tag{28}$$

Note that $\widehat{u}$ can be computed without an additional matrix vector product with A since $\widehat{u} = A\widetilde{u}_k = AV_k s = W_k U_k s$ (if $W_k U_k = AV_k$, where U_k is a matrix of order k).

The above considerations lead to Algorithm 2.

5.4.2. Jacobi–Davidson with orthogonal basis. If we want to work with an orthonormal basis for $A\mathcal{V}_k$ then we may proceed as follows.

Let $v_1, v_2, \ldots, v_k$ be the Jacobi–Davidson vectors obtained after k steps. Then we orthonormalize the set $Av_1, Av_2, \ldots, Av_k$ (as in (25)). The eigenvalues of $\widetilde{H}_k$ (cf. (26)) are the harmonic Ritz values of A, and let $\widetilde{\theta}_k$ be the one of interest, with corresponding harmonic Ritz vector $\widetilde{u}_k = V_k s$ (s is the eigenvector of $\widetilde{H}_k$, normalized such that $\|A\widetilde{u}_k\|_2 = 1$).

Since $(\widetilde{\theta}_k^{-1}, A\widetilde{u}_k)$ is a Ritz pair of A^{-1} with respect to $A\mathcal{V}_k$, we have with $z := A\widetilde{u}_k$ that $w := A^{-1}z - \widetilde{\theta}_k^{-1} z$ is orthogonal to z, and although we do not have A^{-1} available, the vector w can be efficiently computed from

$$w = A^{-1}z - \widetilde{\theta}_k^{-1} z = A^{-1} AV_k s - \widetilde{\theta}_k^{-1} z = V_k s - \widetilde{\theta}_k^{-1} z. \tag{29}$$

The orthonormal basis set for $A\mathcal{V}_k$ should be expanded by a suitable vector At, which, according to our Jacobi–Davidson approach, is computed approximately from

$$(I - zz^*)(A^{-1} - \widetilde{\theta}_k^{-1} I)(I - zz^*)\, At = -w. \tag{30}$$

Also in this case we can avoid working with A^{-1}, since (30) reduces to

$$(I - zz^*)(A - \widetilde{\theta}_k I)(I - V_k s z^* A)\, t = \widetilde{\theta}_k w. \tag{31}$$

We may not expect any difference between the use of (28) and (31) when we use GMRES as the inner iteration, as we will see now.

Since $\|z\|_2 = 1$ and $w \perp z = A\widetilde{u}_k$, we see that $1 = z^* z = \widetilde{\theta}_k z^* \widetilde{u}_k$. Furthermore, we have that $\widetilde{\theta}_k w = \widetilde{\theta}_k \widetilde{u}_k - A\widetilde{u}_k = -r$.

It can be shown that the operator in the left-hand side of (28) and the one in the left-hand side of (31) differ only by a rank-one matrix of the form $r(2(A - \widetilde{\theta}_k I)^* \widetilde{u}_k)^*$. Therefore, the operators generate identical Krylov subspaces if, in both cases, r is the first Krylov subspace vector: Krylov subspace methods like GMRES(m) with initial approximation $x_1 = 0$ lead, in exact arithmetic, to identical approximate solutions when used to solve equations (28) and (31).

It is not yet clear which of the two approaches will be more efficient in practical situations. Much depends upon how the projected systems are approximately solved.

ALGORITHM 2. *The Jacobi–Davidson algorithm with harmonic Ritz values and bi-orthogonalization.*

1. **Start:** Choose an initial nontrivial vector v.
   - Compute $v_1 = v/\|v\|_2$, $w_1 = Av_1$, $l_{11} = w_1^* v_1$, $h_{11} = w_1^* w_1$, set $\ell = 1$, $V_1 = [v_1]$, $W_1 = [w_1]$, $L_1 = [l_{11}]$, $\widehat{H}_1 = [h_{11}]$, $\widetilde{u} = v_1$, $\widehat{u} = w_1$, $\widetilde{\theta} = h_{11}/l_{11}$, compute $r = \widehat{u} - \widetilde{\theta}\widetilde{u}$.
2. **Iterate:** Until convergence do:
3. **Inner loop:** For $k = \ell, \ldots, m-1$ do:
   - Solve approximately $t \perp \widehat{u}$,
   $$\left(I - \frac{\widetilde{u}\widehat{u}^*}{\widehat{u}^*\widetilde{u}}\right)(A - \widetilde{\theta}I)\left(I - \frac{\widetilde{u}\widehat{u}^*}{\widehat{u}^*\widetilde{u}}\right)t = -r.$$
   - Bi-orthogonalize t against V_k and W_k ($\widetilde{t} = t - V_k L_k^{-1} W_k^* t$, $v_{k+1} = \widetilde{t}/\|\widetilde{t}\|_2$) and expand V_k with this vector to V_{k+1}.
   - Compute $w_{k+1} := Av_{k+1}$ and expand W_k with this vector to W_{k+1}.
   - Compute $w_{k+1}^* V_{k+1}$, the last row vector of $L_{k+1} := W_{k+1}^* V_{k+1}$, compute $w_{k+1}^* W_{k+1}$, the last row vector of $\widehat{H}_{k+1} := W_{k+1}^* W_{k+1}$, its adjoint is the last column of $\widehat{H}_{k+1}$. $\widetilde{H}_{k+1} := L_{k+1}^{-1}\widehat{H}_{k+1}$.
   - Compute the smallest eigenpair $(\widetilde{\theta}, s)$ of $\widetilde{H}_{k+1}$.
   - Compute the harmonic Ritz vector $\widetilde{u} := V_{k+1}s/\|V_{k+1}s\|_2$, compute $\widehat{u} := A\widetilde{u}$ $(= W_{k+1}s/\|V_{k+1}s\|_2)$, and the associated residual vector $r := \widehat{u} - \widetilde{\theta}\widetilde{u}$.
   - Test for convergence. Stop if satisfied.
4. **Restart:** Choose an appropriate value for $\ell < m$.
   - Compute the smallest ℓ eigenvalues of $\widetilde{H}_m$ and the matrix Y with columns the associated eigenvectors. Orthogonalize Y with respect to L_m (cf. §5.2.1): $Y = ZR$ with R upper triangular and $Z^* L_m Z$ lower triangular.
   - Set $V_\ell := V_m Z$, $W_\ell := W_m Z$, $L_\ell := Z^* L_m Z$, $\widehat{H}_\ell := Z^* \widehat{H}_m Z$, and goto 3.

6. Numerical examples. The new algorithm has already been successfully applied in applications from chemistry (in which Davidson's method was the preferred one before) and magnetohydrodynamics (MHD) models. For reports on these experiences, see [2], [28].

The simple examples that we will present here should be seen as illustrations only of the new approach. The examples have been coded in MATLAB and have been executed on a SUN SPARC workstation in about 15 decimals working precision. Most of our examples are for symmetric matrices, since it was easier for us to check the behavior of the Ritz values in this case, but our codes did not take advantage of this fact for the generation of the matrices H_k.

Example 1. We start with a simple tridiagonal diagonally dominant matrix A,

with diagonal elements 2.4 and off-diagonal elements 1, of order $n = 100$. The starting vector is taken to be the vector with all 1's, scaled to unit length.

In Davidson's method, the suggested approach is to approximate the inverse of $A - \theta_k I$, and in this example we take just $(A - \theta_k I)^{-1}$, which is the best one can do if one wants to approximate well. Note that the standard choice $(D_A - \theta_k I)$ leads to (almost) Arnoldi's method, which converges only very slowly (for $i \ll n$) in this case.

In Figure 1 we have plotted the log of $|\lambda - \theta_k|$ as the dashed curve, and we see that this indeed almost leads to stagnation (some progress is made, since we have computed the inverse in floating point arithmetic). If, however, we use the Jacobi–Davidson method (as in §3), again with exact inversion of the projected operator $B - \theta_k I$, then we observe (the lower curve) very fast (cubical?) convergence, just as expected.

Of course, solving the involved linear systems exactly is usually too expensive, and it might be more realistic to investigate what happens if we take more crude approximations for the operators involved. For the Davidson method we take 5 steps of GMRES for the approximation of the solution of $(A - \theta_k I)v = r$, and for Jacobi-Davidson we also take 5 steps of GMRES for the approximate solution of the projected system (13). The results are given in Figure 2 (the dashed curve represents the results for the Davidson method).

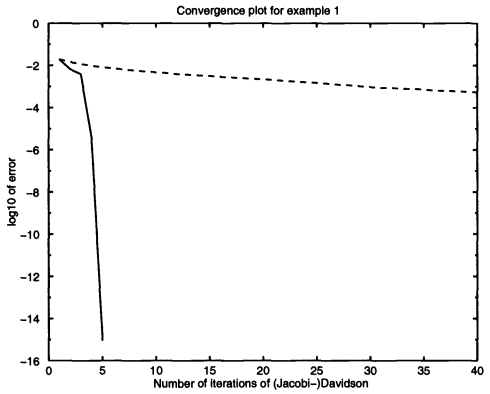


FIG. 1. *Convergence of Ritz values with exact inverses.*

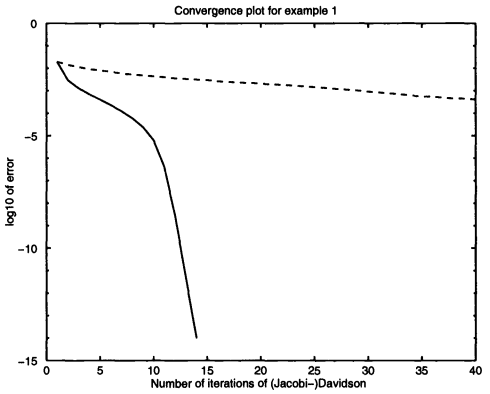


FIG. 2. *Convergence of Ritz values with approximate inverses.*

Again we see that for this moderately diagonally dominant matrix it is attractive to work with the Jacobi–Davidson method. Note that the 5 GMRES steps for the approximate inversion step are not sufficient to have quadratical convergence, but the linear convergence takes place with a very small convergence factor.

We also learn from this example that Krylov subspace methods may not be expected to make good preconditioners for Davidson's method: with a few steps one suffers from the fact that $A - \theta_k I$ has a very small eigenvalue, and if carried out to high precision (almost) stagnation is to be expected. The Jacobi–Davidson method does not have these problems, since the projected operator $B - \theta_k I$ does not have a small eigenvalue (unless the eigenvalue λ is close to some other eigenvalue of A).

Example 2. Our second example is still highly artificial, but here we try to mimic more or less what happens when a matrix is not diagonally dominant. The matrix A

is constructed as $B = Q_1 A Q_1$, with

$$A = \text{tridiag}(-1,\ 2,\ -1),$$

and Q_1 is a Householder transformation. The order of the matrices is 100. Note that the distribution of the eigenvalues at the upper and lower ends of the spectrum is not particularly favorable for Krylov subspace methods since they are not well separated in a relative sense.

For those who wish to repeat our experiments we add that the Householder vector h was chosen with elements $h_j = \sqrt{j + .45}$, $j = 1, 2, \ldots, 100$. The starting vector for the iteration algorithms was chosen as a vector with all elements equal to 1. Furthermore, we restarted the outer iterations after each 20 steps, which represents a usual strategy in practical situations.

The Davidson algorithm (with $D_A - \theta_k I$) needed 565 iteration steps to find the largest eigenvalue

$$\lambda = 3.9990325\ldots$$

to almost working precision.

In the Jacobi–Davidson algorithm we did the inner iterations, necessary for solving (13) approximately, with 5 steps of GMRES. This time we needed 65 outer iterations (i.e., 320 inner iteration steps). The inner iteration method (GMRES), as well as the number of steps (5 steps), has been chosen arbitrarily. In actual computations one may choose any appropriate means to approximate the solution of the projected system (13), such as, e.g., the incomplete LU (ILU).

Example 3. This example illustrates that our new algorithms (in §§3 and 5) may also be used for the computation of interior eigenvalues. In this example we compute an approximation for the eigenvalue of smallest absolute value. For this purpose the Jacobi–Davidson algorithm that uses Ritz values (§3) is modified; instead of computing the largest eigenpair (θ_k, u_k) of H_k we compute the one with smallest absolute value for the Ritz value.

Again we take a simple matrix: A is the 100×100 diagonal matrix with spectrum

$$\left\{ t^2 - 0.8 \mid t = \frac{j}{100},\ j = 1, \ldots, 100 \right\}.$$

All coordinates of the starting vector v_1 are equal to 1. We solve the projected equations approximately using 8 steps of GMRES. We do not use restarts for the (outer) iterations.

Note that our algorithms, like any Krylov subspace method, do not take advantage of the fact that A is diagonal as long as we do not use diagonal preconditioning. Indeed, with $D = Q^T A Q$ and $y = Qv$, we see that $\mathcal{K}_i(D; v) = Q^T \mathcal{K}_i(A; y)$, so that except for an orthogonal transformation the same subspaces are generated. In particular, the Rayleigh quotients, of which the Ritz values are the local extrema for symmetric matrices, are the same for both subspaces. This means that if the starting vectors are properly related as $y = Qv$, then the Krylov method with D and v leads to the same convergence history as the method with A and y.

We have also used the harmonic Ritz value variant of Jacobi–Davidson as in §5. Figures 3 and 5 show the convergence history of $\log_{10} \|r_k\|_2$, where the residual r_k, at step k, is either $Au_k - \theta_k u_k$ for the Ritz pair with Ritz value of smallest absolute value, or $A\widetilde{u}_k - \widetilde{\theta}_k \widetilde{u}_k$ for the harmonic Ritz pair with harmonic Ritz value of smallest

absolute value. Along the vertical axis we have $\log_{10} \|r_k\|_2$ and we have the iteration number k along the horizontal axis. Figures 4 and 6 show the convergence history of all (harmonic) Ritz values (indicated by $\cdot$) in the interval indicated along the vertical axis. Again, we have the number of iteration steps along the horizontal axis. We have marked the positions of the eigenvalues (with $+$) at the right of Figures 4 and 6.

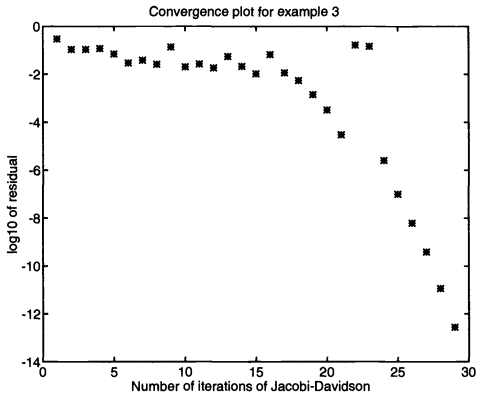


FIG. 3. *Convergence residuals using Ritz values.*

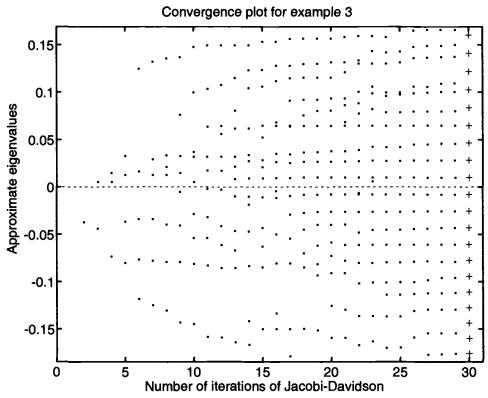


FIG. 4. *The Ritz values.*

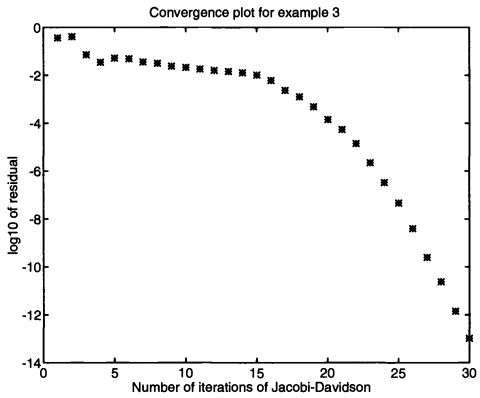


FIG. 5. *Convergence residuals using harmonic Ritz values.*

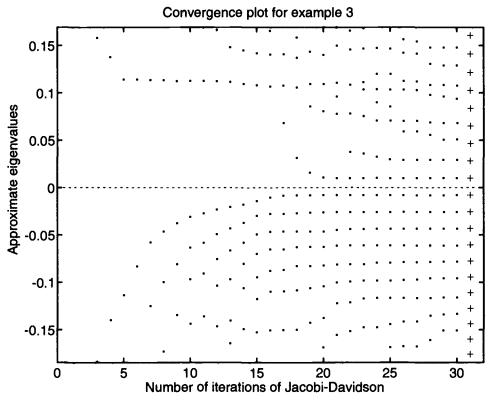


FIG. 6. *The harmonic Ritz values.*

For this example, the algorithm based on Ritz values (see Figures 3 and 4) converges about as fast as the algorithm based on harmonic Ritz values (see Figures 5 and 6), but the convergence history with harmonic Ritz values is much smoother. The difference in smoothness seems to also be typical for other examples. This fact can be exploited for the construction of restart strategies and stopping criteria. From experiments, we have learned that restarting for the "Ritz value" algorithm can be quite problematic; see also [14] for similar observations.

Example 4. In our previous examples the matrices A are symmetric. However, our algorithms are not restricted to the symmetric case and may also be used for the approximation of nonreal eigenvalues. In this example, we use complex arithmetic.

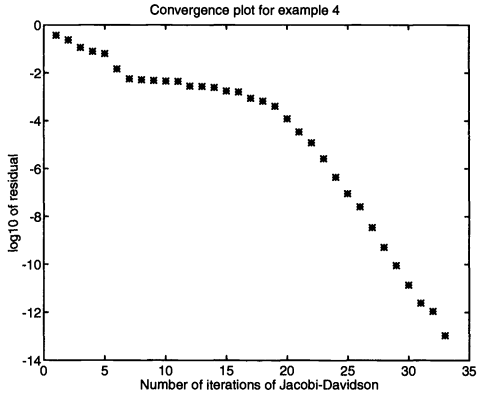


FIG. 7. *Convergence residuals using harmonic Ritz values.*

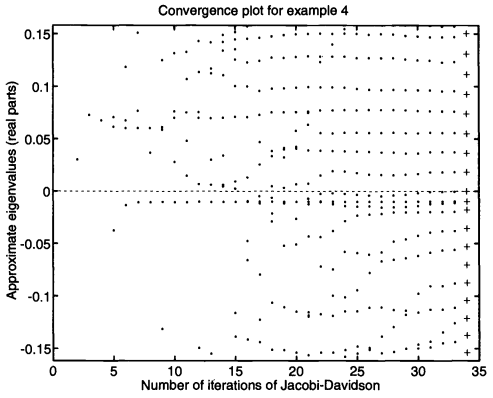


FIG. 8. *Real parts of harmonic Ritz values.*

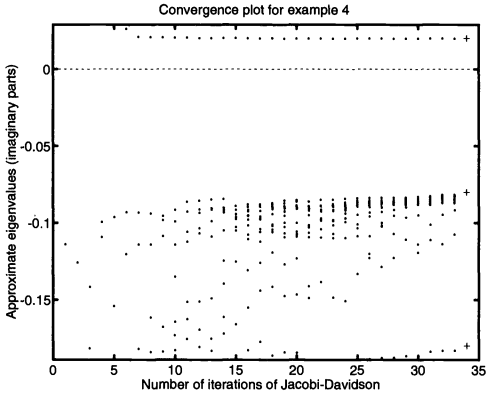


FIG. 9. *Imaginary parts of harmonic Ritz values.*

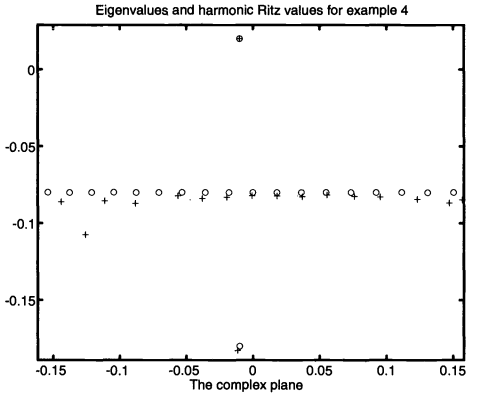


FIG. 10. *Harmonic Ritz values in* $\mathbb{C}$ *at step* 33.

For this example we have simply augmented the diagonal matrix of Example 3 by the two complex diagonal elements $0.8 + 0.1\,i$ and $0.8 - 0.1\,i$ to a matrix of order 102, and we have applied the harmonic Ritz value variant of our algorithm (§5) to the matrix $A - \mu I$ with shift $\mu = 0.81 + 0.08\,i$. Again, all coordinates of the starting vector are 1. Now we solve the projected equations approximately using 10 steps of GMRES.

Figures 7–9 show the convergence history of the residual vector (Figure 7), the real parts of the harmonic Ritz values (Figure 8), and the imaginary parts of the harmonic Ritz values (Figure 9). Figure 10 shows the harmonic Ritz values of order 33. In Figures 7–9, we have used the same symbols as in the previous example. In Figure 10, the harmonic Ritz values of step 33 of the Jacobi–Davidson iteration are represented by a $+$, while the eigenvalues of A are represented by $\circ$.

Clearly, the algorithm finds the eigenvalue $\lambda = 0.8 + 0.1\,i$ of A close to the shift μ, but also other ones, such as the conjugate of λ (which is quite far from the shift). This is typical for other experiments as well; usually a large number of (harmonic) Ritz values are converging in the Jacobi–Davidson method.

Example 5. In this example we experimentally compare the performances of the Davidson method, the Jacobi–Davidson method, and the accelerated shift and inexact invert (ASII) variant of observation 4 in §4.1, i.e., we expand our search space by (the orthogonal complement of) the approximate solution $\widetilde{t}$ of

$$\begin{aligned} (A-\theta_k I)t &= -r \qquad \text{(D)},\\ (I-u_k u_k^*)(A-\theta_k I)(I-u_k u_k^*)t &= -r \qquad \text{(JD)},\\ (A-\theta_k I)t &= u_k \qquad \text{(SI)}, \end{aligned}$$

respectively (cf. §4). We solve these equations approximately by m steps of full GMRES (with 0 as an initial guess). Since we are interested in the absolute smallest eigenvalue we take for θ_k the absolute smallest eigenvalue of $H_k = V_k^* A V_k$. The preconditioner M for GMRES is kept fixed throughout the iteration process. The systems (D) and (SI) are preconditioned by M^{-1}, while the projection $M_d := (I - u_k u_k^*)M(I - u_k u_k^*)$ is used as preconditioner for (JD). This means that for (JD) we have to solve equations of the form $M_d z = y$, where y is a given vector orthogonal to u_k. We follow the approach as indicated in §4.1 and we solve this equation by $z = \alpha M^{-1}u_k - M^{-1}y$ with $\alpha = u_k^* M^{-1} y / u_k^* M^{-1} u_k$.

We have applied the three methods—Davidson, Jacobi–Davidson, and SI—for a matrix from the Harwell–Boeing set of test matrices: A is the SHERMAN4 matrix shifted by 0.5 (we wish to compute the eigenvalue of the SHERMAN4 matrix that is closest to 0.5). A is of order $n = 1104$. All eigenvalues of A appear to be real and are in the interval $[0.030726, 66.497]$. The smallest eigenvalues are (in 5 decimal places): 0.030726, 0.084702, 0.27757, 0.39884, 0.43154, 0.58979, 0.63480, ..., so that with the given shift we are aiming at the fifth eigenvalue.

For M we have selected the ILU(2) decomposition of A. We have plotted the $\log_{10}$ of the norm of the residual versus the number of outer iteration steps (which is the dimension of the search space V_k): Figures 11, 12, and 13 show the results for, respectively, 5, 10, and 25 steps of GMRES.

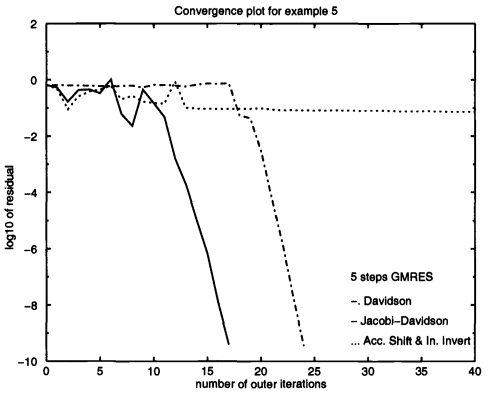


FIG. 11. *Using* 5 *steps of preconditioned* GMRES.

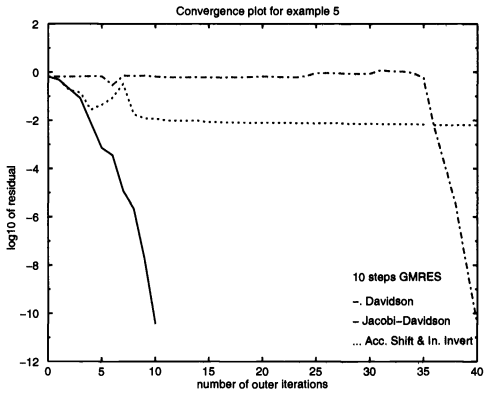


FIG. 12. *Using* 10 *steps of preconditioned* GMRES.

Larger values of m imply more accurate approximate solutions of the "expansion equations" (D), (JD), and (SI). In line with our discussions in §3 and our results in Example 1, we see that improving the approximation in Davidson's method slows the speed of convergence and it may even lead to stagnation (cf. the dash-dotted curves

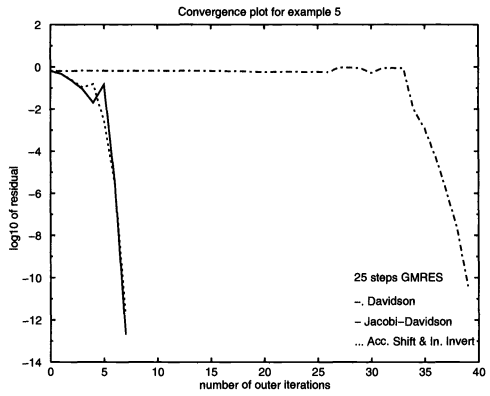


FIG. 13. *Using* 25 *steps of preconditioned* GMRES.

–.). As might be anticipated, for ASII we observe the opposite effect (cf. the dotted curves ...); the more precisely we solve (SI), the faster the method converges, while stagnation may occur if (SI) is not solved accurately enough. The speed of convergence of our Jacobi–Davidson method does not depend so much upon the accuracy of the approximate solutions of (JD) (cf. the solid curves —): the method converges faster than Davidson and ASII.

As argued in §4.1, ASII may be rather sensitive to rounding errors, especially if the expanding vector $\tilde{t}$ has a large component in the direction of u. For ASII, but also for Davidson, we had to apply modified Gram–Schmidt (mod-GS) twice to maintain sufficient orthogonality of V_k, while in Jacobi–Davidson this was not necessary. By doing mod-GS only once, the angle between the expansion vector $\tilde{t}$ and the already available search space may become too small to allow an accurate computation of the orthogonal component. In such a situation, it may help to apply mod-GS more often [23]. For the present example, twice was enough (but other examples, not reported here, required more mod-GS sweeps).

Acknowledgments. We thank Albert Booten for his careful reading of an early version of the manuscript. He was also helpful in deriving formula (31). We gratefully acknowledge helpful discussions with Diederik Fokkema on the subject of §4. He also provided the numerical data for Example 5. Axel Ruhe helped us to improve our presentation.

We are also grateful to the referees who made many valuable suggestions. One of the referees drew our attention to the references [16], [18], [27], which contain information that should have been more widely known.

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] J. G. L. BOOTEN, H. A. VAN DER VORST, P. M. MEIJER, AND H. J. J. TE RIELE, *A preconditioned Jacobi–Davidson method for solving large generalized eigenvalue problems*, Report NM-R9414, Department of Numerical Mathematics, CWI, Amsterdam, The Netherlands, 1994.

[3] M. CROUZEIX, B. PHILIPPE, AND M. SADKANE, *The Davidson method*, SIAM J. Sci. Comput., 15 (1994), pp. 62–76.

[4] E. R. DAVIDSON, *The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real symmetric matrices*, J. Comput. Phys., 17 (1975), pp. 87–94.
[5] ———, *Matrix eigenvector methods in quantum mechanics*, in Methods in Computational Molecular Physics, G. H. F. Diercksen and S. Wilson, eds., Reidel, Dordrecht, The Netherlands, 1983, pp. 95–113.
[6] ———, *Monster matrices: Their eigenvalues and eigenvectors*, Computers in Physics, 7 (1993), pp. 519–522.
[7] A. DEN BOER, *De Jacobi methode van 1845 tot 1990*, Master's thesis, Department of Mathematics, University of Utrecht, Utrecht, The Netherlands, 1991. (In Dutch.)
[8] R. W. FREUND, *Quasi-kernel polynomials and their use in non-Hermitian matrix iterations*, J. Comput. Appl. Math., 43 (1992), pp. 135–158.
[9] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., The John Hopkins University Press, Baltimore and London, 1989.
[10] Y. HUANG AND H. A. VAN DER VORST, *Some observations on the convergence behaviour of GMRES*, Tech. report 89-09, Delft University of Technology, Faculty of Tech. Math., Delft, The Netherlands, 1989.
[11] C. G. J. JACOBI, *Ueber eine neue Auflösungsart der bei der Methode der kleinsten Quadrate vorkommende linearen Gleichungen*, Astronom. Nachr., (1845), pp. 297–306.
[12] ———, *Ueber ein leichtes Verfahren, die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen*, J. Reine und Angew. Math., (1846), pp. 51–94.
[13] C. LANCZÔS, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. Nat. Bur. Standards, 45 (1950), pp. 255–282.
[14] R. B. MORGAN, *Computing interior eigenvalues of large matrices*, Linear Algebra Appl., 154/156 (1991), pp. 289–309.
[15] ———, *Generalizations of Davidson's method for computing eigenvalues of large nonsymmetric matrices*, J. Comput. Phys., 101 (1992), pp. 287–291.
[16] R. B. MORGAN AND D. S. SCOTT, *Generalizations of Davidson's method for computing eigenvalues of sparse symmetric matrices*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 817–825.
[17] ———, *Preconditioning the Lanczos algorithm for sparse symmetric eigenvalue problems*, SIAM J. Sci. Comput., 14 (1993), pp. 585–593.
[18] J. OLSEN, P. JØRGENSEN, AND J. SIMONS, *Passing the one-billion limit in full configuration-interaction (FCI) calculations*, Chemical Physics Letters, 169 (1990), pp. 463–472.
[19] A. M. OSTROWSKI, *On the convergence of the Rayleigh quotient iteration for the computation of characteristic roots and vectors. V*, Arch. Rational Mech. Anal., 3 (1959), pp. 472–481.
[20] C. C. PAIGE, B. N. PARLETT, AND H. A. VAN DER VORST, *Approximate solutions and eigenvalue bounds from Krylov subspaces*, Num. Lin. Alg. with Appl., 2 (1995), pp. 115–134.
[21] B. N. PARLETT, *The Rayleigh quotient iteration and some generalizations*, Math. Comp., 28 (1974), pp. 679–693.
[22] ———, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[23] A. RUHE, *Numerical aspects of Gram–Schmidt orthogonalization of vectors*, Linear Algebra Appl., 52/53 (1983), pp. 591–602.
[24] ———, *Rational Krylov algorithms for nonsymmetric eigenvalue problems. II. Matrix pairs*, Linear Algebra Appl., 197/198 (1994), pp. 283–295.
[25] Y. SAAD, *Numerical Methods for Large Eigenvalue Problems*, Manchester University Press, Manchester, UK, 1992.
[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] A. STATHOPOULOS, Y. SAAD, AND C. F. FISCHER, *Robust preconditioning of large sparse symmetric eigenvalue problems*, J. Comput. Appl. Math., 64 (1995), pp. 197–215.
[28] H. J. J. VAN DAM, J. H. VAN LENTHE, G. L. G. SLEIJPEN, AND H. A. VAN DER VORST, *An improvement of Davidson's iteration method; Applications to MRCI and MRCEPA calculations*, J. Comput. Chem., to appear.
[29] A. VAN DER SLUIS AND H. A. VAN DER VORST, *The convergence behavior of Ritz values in the presence of close eigenvalues*, Linear Algebra Appl., 88/89 (1987), pp. 651–694.
[30] M. B. VAN GIJZEN, *Iterative solution methods for linear equations in finite element computations*, Ph.D. thesis, Delft University of Technology, Delft, The Netherlands, 1994.
[31] J. H. VAN LENTHE AND P. PULAY, *A space-saving modification of Davidson's eigenvector algorithm*, J. Comput. Chem., 11 (1990), pp. 1164–1168.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 2, pp. 426–442, April 1996

011

GENERALIZED INVERSES OF DIFFERENTIAL-ALGEBRAIC OPERATORS*

PETER KUNKEL† AND VOLKER MEHRMANN‡

Abstract. In the theoretical treatment of linear differential-algebraic equations one must deal with inconsistent initial conditions, inconsistent inhomogeneities, and undetermined solution components. Often their occurrence is excluded by assumptions to allow a theory along the lines of differential equations. This paper aims at a theory that generalizes the well-known least squares solution of linear algebraic equations to linear differential-algebraic equations and that fixes a unique solution even when the initial conditions or the inhomogeneities are inconsistent or when undetermined solution components are present. For that a higher index differential-algebraic equation satisfying some mild assumptions is replaced by a so-called strangeness-free differential-algebraic equation with the same solution set. The new equation is transformed into an operator equation and finally generalized inverses are developed for the underlying differential-algebraic operator.

Key words. differential-algebraic equations, standard form, Moore–Penrose pseudoinverse, generalized inverse, least squares regularization

AMS subject classifications. 34A09, 47E05, 15A09, 58E25

1. Introduction.

We study the solution of linear differential-algebraic equations (DAEs)

$$E(t)\dot{x}(t) = A(t)x(t) + f(t) \tag{1}$$

with initial condition

$$x(a) = x_0, \tag{2}$$

where $t \in [a,b]$ and $E, A \in C([a,b], \mathbb{R}^{m,n})$, $f \in C([a,b], \mathbb{R}^m)$, $x_0 \in \mathbb{R}^n$. Here $C^r([t_0,t_1], \mathbb{R}^{m,n})$ denotes the set of r-times continuously differentiable functions from the interval $[t_0,t_1]$ to the vector space $\mathbb{R}^{m,n}$ of real $m \times n$ matrices.

Although problems of the form (1), (2) can easily be seen as generalizations of possibly under- or overdetermined systems of linear equations

$$Ax = b \tag{3}$$

with $A \in \mathbb{R}^{m,n}$, $b \in \mathbb{R}^m$, theoretical investigations of (1) mostly require the DAE to have a unique solution for consistent initial values x_0. This reduces the considerations not only to the case $m = n$ but also prohibits the occurrence of undetermined solution components. This, however, excludes DAEs that may be found, e.g., in the study of optimal control problems for descriptor systems; see [16].

In the theory of linear equations the problem of undetermined solution components or inconsistent right-hand sides is overcome by embedding (3) into the minimization problem

$$\frac{1}{2}\|x\|_2^2 = \min! \quad \text{s.t.} \quad \frac{1}{2}\|Ax - b\|_2^2 = \min!, \tag{4}$$

* Received by the editors May 16, 1994; accepted for publication (in revised form) by A. Bunse-Gerstner May 4, 1995. This work was supported by Deutsche Forschungsgemeinschaft, Research grant Me 790/5-1 Differentiell-algebraische Gleichungen.

† Fachbereich Mathematik, Carl von Ossietzky Universität, Postfach 2503, D-26111 Oldenburg, Germany.

‡ Fakultät für Mathematik, Technische Universität Chemnitz-Zwickau, D-09107 Chemnitz, Germany.

which has a unique solution in any case. This unique solution, also called *least squares solution*, can be written in the form

$$x = A^+ b \tag{5}$$

with the help of the Moore–Penrose pseudoinverse A^+ of A. A more detailed interpretation is that the matrix A induces a homomorphism $A: \mathbb{R}^n \to \mathbb{R}^m$ by $x \mapsto Ax$. For fixed A the mapping which maps b on the unique solution x of (4) is found to be linear. A matrix representation of this homomorphism with respect to canonical bases is then given by A^+. It is well known that A^+ satisfies the four Penrose axioms

$$\begin{array}{ll} \text{(a)} & AA^+A = A, \\ \text{(b)} & A^+AA^+ = A^+, \\ \text{(c)} & (AA^+)^T = AA^+, \\ \text{(d)} & (A^+A)^T = A^+A; \end{array} \tag{6}$$

see, e.g., [1, 6]. In turn, for given $A \in \mathbb{R}^{m,n}$, the four axioms fix a unique matrix $A^+ \in \mathbb{R}^{n,m}$, whose existence follows, for example, by the solvability of (4).

It is the aim of this paper to generalize this concept for linear equations to a large class of linear DAEs with variable coefficients. In more detail, we first replace the given problem by an equivalent (in the sense that there is a special one-to-one correspondence of the solution sets) so-called strangeness-free problem. For this we then develop an appropriate generalized inverse of some operator representation of the new problem in the spirit of the Moore–Penrose pseudoinverse. In particular, the explicit representation of this Moore–Penrose pseudoinverse mainly consists of the solution of a linear quadratic control problem.

In [11], Hanke treated similar questions in the context of integrable functions. In a Hilbert space setting he was able to show that, in general, the operators describing (1), (2) with $m = n$ are closable. He then gave a representation of the associated closed operator for which the Moore–Penrose pseudoinverse then exists but is in general not continuous. Finally he showed that it is indeed continuous for problems with (differentiation) index at most one and not continuous when the index exceeds one. In contrast to his approach, we replace a higher index problem by an equivalent strangeness-free problem, we work in spaces of continuous functions (i.e., we have no Hilbert space structure), we allow for undetermined solution components and nonsquare systems, and we give an explicit representation of the Moore–Penrose pseudoinverse, thus showing continuity of the pseudoinverse.

Note that besides the Moore–Penrose pseudoinverse one can find other kinds of generalized inverses when dealing with differential-algebraic equations or special cases of them. The so-called Drazin inverse, see, e.g., [7] or [6, Chap. 9], is used for equations with constant coefficients to give an explicit representation of the set of solutions and consistent initial values. This theory, however, is not extendable to the case of variable coefficients. In the theory of boundary value problems for linear ordinary differential equations, so-called generalized Green's functions are used; see, e.g., [22, Chap. III, §10]. These functions define operators that yield a specific solution for a given inhomogeneity even when the solution is not unique due to the choice of the boundary conditions. Note, however, that due to the unique solvability of initial value problems these operators are Fredholm operators; i.e., the given problem is essentially finite dimensional. In the present paper we only treat initial value problems for linear DAEs. But these allow for the presence of undetermined solution components such that the kernel of the associated operator may have infinite dimension. Thus the

operator is, in general, not Fredholm. The main focus of this paper, therefore, will be to handle this infinite dimensional kernel. The extension to boundary value problems seems to be possible but is beyond the scope of this paper.

The present paper is organized as follows. In §2, we give a standard form of DAEs required for the subsequent construction, thus specifying the class of DAEs we can treat in the theory to follow. The appropriate analytical context on the basis of dual systems is outlined in §3. We then treat two possible embeddings of (1), (2) into minimization problems in §4, both leading to generalizations of the Moore–Penrose pseudoinverse for matrices. Finally, we give some conclusions in §5.

2. Standard form of DAEs. In order to treat (1) as generalization of linear equations on the one hand as well as of differential equations on the other we must carefully select suitable definitions for solvability and related questions fitting to both extreme cases. Even finding an appropriate notion of solvability of (1) seems to be a hard problem. See, e.g., [2, 4, 5, 8, 10, 12] for different definitions of solvability in the context of DAEs. Many of them are orientated at properties of linear differential equations and ignore results known for the special case (3), one of which, for example, is that (3) is solvable (in the sense that there is a solution) if and only if $\operatorname{rank} A = \operatorname{rank}(A, b)$. In view of (1) the weakest possible meaning of a (strong or classical) solution without additional assumptions on the smoothness of the coefficients is given in the following definition.

DEFINITION 2.1. (a) *A function* $x \in C^1([a,b], \mathbb{R}^n)$ *is called a* solution *of* (1) *if and only if it satisfies* (1) *pointwise.*

(b) *The DAE* (1) *is called* solvable *and the inhomogeneity* f *is called* consistent *if and only if* (1) *has at least one solution.*

(c) *An initial condition* (2) *is called* consistent *if and only if* (1) *has a solution that satisfies* (2).

(d) *An initial value problem* (1), (2) *is called* (uniquely) solvable *if and only if there is a* (*unique*) *solution of* (1) *satisfying* (2).

Under certain circumstances it is possible and necessary to weaken the smoothness requirements for a solution. We shall come back to this point when it becomes important.

Unfortunately it seems to be impossible to deal with (1) in full generality. Without any further restrictions many undesired phenomena can occur. Compare the observations made in the following examples with the fact that linear differential equations, corresponding to E being pointwise nonsingular, are uniquely solvable for any continuous coefficients E, A, and f.

Example 2.2. Consider the singular differential equation

$$t\dot{x}(t) = f(t)$$

on $[-1, 1]$ with initial condition $x(-1) = 0$. For $\dot{x}$ to be continuous, we must require f to be continuous $f(0) = 0$ and f differentiable at $t = 0$. The unique solution is then given by

$$x(t) = \int_{-1}^{t} \frac{f(s)}{s}\, ds.$$

Example 2.3. Consider the so-called standard problem of index two

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix},$$

which has the unique solution

$$x_1(t) = -\dot{f}_2(t) - f_1(t),$$
$$x_2(t) = -f_2(t),$$

independent of the interval of interest. Obviously we must at least require f to be continuously differentiable on the entire interval to be able to write down the solution. Because of the special shape of E, which cancels the entry $\dot{x}_1$, we may be theoretically satisfied with this smoothness requirement, although the above definition would need f to be twice continuously differentiable.

Hence the set of possible inhomogeneities may be restricted even in the case of uniquely solvable problems by additional smoothness requirements or even by inner point conditions depending on the given matrix functions E and A. For a unified treatment we must therefore impose some restrictions on the functions E and A. It must, however, be clear that the remaining class of DAEs is reasonably large.

In [17, 19, 18, 21] it has been shown that under some constant rank and smoothness assumptions concerning the matrix functions E and A a given (higher index) DAE can be transformed in such a way that the set of solutions remains the same and the new equation is *strangeness-free.* The latter property can be defined in the following way.

DEFINITION 2.4. *The DAE* (1) *is called strangeness-free if there exist* $P \in C([a,b], \mathbb{R}^{m,m})$ *and* $Q \in C^1([a,b], \mathbb{R}^{n,n})$, *both pointwise orthogonal, such that we can transform* (1) *to the standard form*

$$\tilde{E}(t)\dot{\tilde{x}}(t) = \tilde{A}(t)\tilde{x}(t) + \tilde{f}(t), \tag{7}$$

where

$$\begin{aligned}
\tilde{E}(t) &= P(t)E(t)Q(t) = \begin{bmatrix} \Sigma_E(t) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \\
\tilde{A}(t) &= P(t)A(t)Q(t) - P(t)E(t)\dot{Q}(t) = \begin{bmatrix} A_{11}(t) & A_{12}(t) & A_{13}(t) \\ A_{21}(t) & \Sigma_A(t) & 0 \\ 0 & 0 & 0 \end{bmatrix}, \\
\tilde{x}(t) &= Q(t)^T x(t), \\
\tilde{f}(t) &= P(t)f(t)
\end{aligned} \tag{8}$$

with Σ_E *and* Σ_A *pointwise nonsingular and all blocksizes are allowed to be zero.*

Observe that this definition is more general than requiring the differentiation index (see, e.g., [5]) to be at most one. This is due to the occurrence of the third block row and column in $\tilde{E}$ and $\tilde{A}$ which yields an infinite dimensional solution space for the homogeneous problem $\tilde{E}(t)\dot{\tilde{x}}(t) = \tilde{A}(t)\tilde{x}(t)$. If these blocks are not present (implying $m = n$) the assumption of (1) being strangeness-free reduces to the assumption of (1) having differentiation index zero or one.

Note that the above transformation onto a strangeness-free DAE is numerically implementable; i.e., for a numerical treatment of (1) we may assume that the problem is strangeness-free; see [19]. At this point we should mention that the transformation procedure in [17]–[19] does not determine E, A, and f uniquely but up to multiplication by a pointwise orthogonal matrix function from the left. Thus we must take care that our approach does not depend on such transformations. We also remark that it is currently under investigation how far the necessary constant rank assumptions can

be relaxed if one still requires classical solutions or how weaker solvability concepts can be obtained by dropping further assumptions; see [3, 21].

In order to treat problems of the form (1), (2) that have no unique solution along the lines of the treatment of (3), a necessary condition is that in the uniquely solvable case the mapping which maps f on the unique solution x for fixed E and A is linear. In particular, we must have the trivial solution for $f \equiv 0$. Necessary for this is that the initial condition is homogeneous, i.e., that $x_0 = 0$. This, however, can be obtained without loss of generality by shifting $x(t)$ to $x(t) - x_0$ which changes the inhomogeneity from $f(t)$ to $f(t) + A(t)x_0$.

Summarizing this section, considering the current state of research, it seems reasonable to concentrate on those linear DAEs with homogeneous initial condition

$$E(t)\dot{x}(t) = A(t)x(t) + f(t), \quad x(a) = 0, \tag{9}$$

that are strangeness-free, i.e., that can be transformed into the standard form indicated by (7) and (8).

3. Dual systems. Following the lines of the construction of the Moore–Penrose pseudoinverse for matrices as sketched in the introduction, we must deal with homomorphisms between function spaces, preferably some linear spaces of continuous functions or appropriate subspaces. In view of (4) the norm of choice would be given by

$$\|x\| = \sqrt{(x,x)}, \quad (x,y) = \int_a^b x(t)^T y(t)\, dt. \tag{10}$$

Because spaces of continuous functions cannot be closed with respect to this norm, we are not in the pure setting of Banach spaces nor of Hilbert spaces. See [1, Chap. 8] for details on generalized inverses of operators on Hilbert spaces. In this section we therefore build up a scenario for defining a Moore–Penrose pseudoinverse which is general enough to be applicable in the setting of linear spaces of continuous functions.

Looking at (6) we find two essential ingredients in imposing the four Penrose axioms. These are the binary operation of matrix multiplication and the transposition of square matrices. In the language of mappings they must be interpreted as composition of homomorphisms (we shall still call it multiplication) and the adjoint of endomorphisms. While the first item is trivial in any setting, the notion of an adjoint is heavily based on the presence of a Hilbert space structure. The most general substitute we can find here is the concept of conjugates with respect to dual systems (pairs); cf. [14, Chap. IX].

DEFINITION 3.1. *Let (X, X^*) be a pair of (real) vector spaces equipped with a bilinear form $(\cdot,\cdot)\colon X \times X^* \to \mathbb{R}$.*

(a) *The pair (X, X^*) is called a* left dual system *if and only if $(x, x^*) = 0$ for all $x \in X$ implies $x^* = 0$.*

(b) *The pair (X, X^*) is called a* right dual system *if and only if $(x, x^*) = 0$ for all $x^* \in X^*$ implies $x = 0$.*

(c) *The pair (X, X^*) is called a* dual system *if and only if it is a left as well as a right dual system.*

It is common sense not to state the bilinear form explicitly. Requiring (X, X^*) to be some dual system therefore includes that there is a related fixed bilinear form with the above properties.

DEFINITION 3.2. *Let (X, X^*) be a left dual system and $A: X \to X$ be an endomorphism. An endomorphism $A^*: X^* \to X^*$ is called a* conjugate of A *if and only if*

$$(Ax, x^*) = (x, A^*x^*) \tag{11}$$

holds for all $x \in X$ and $x^* \in X^*$.

For a unique declaration of a Moore–Penrose pseudoinverse we of course need at least uniqueness of a conjugate. In addition we also need the inversion rule for the conjugate of a product.

LEMMA 3.3. *Let (X, X^*) be a left dual system and $A: X \to X$ be an endomorphism. There is at most one endomorphism $A^*: X^* \to X^*$ being conjugate to A. Let the endomorphisms $A^*, B^*: X^* \to X^*$ be conjugate to the endomorphisms $A, B: X \to X$. Then AB has a conjugate $(AB)^*$ which is given by*

$$(AB)^* = B^*A^*. \tag{12}$$

Proof. See, e.g., [14]. □

Observing that the third and fourth Penrose axioms in (6) require some endomorphisms to be self-conjugate, we must restrict to *self-dual systems*, i.e., to $X^* = X$. At this point we have everything prepared to define a Moore–Penrose pseudoinverse for an appropriate class of homomorphisms.

DEFINITION 3.4. *Let (X, X) and (Y, Y) be (left) dual systems and $D: X \to Y$ be a homomorphism. A homomorphism $D^+: Y \to X$ is called a* Moore–Penrose pseudoinverse *of D if and only if DD^+ and D^+D possess conjugates $(DD^+)^*$ and $(D^+D)^*$ and the relations*

$$\begin{array}{ll} \text{(a)} & DD^+D = D, \\ \text{(b)} & D^+DD^+ = D^+, \\ \text{(c)} & (DD^+)^* = DD^+, \\ \text{(d)} & (D^+D)^* = D^+D \end{array} \tag{13}$$

hold.

As for matrices, the four axioms (13) guarantee uniqueness of the Moore–Penrose pseudoinverse, whereas existence in general cannot be shown.

LEMMA 3.5. *Let (X, X) and (Y, Y) be (left) dual systems and $D: X \to Y$ be a homomorphism. Then D has at most one Moore–Penrose pseudoinverse $D^+: Y \to X$.*

Proof. Let $D^+, D^-: Y \to X$ be two Moore–Penrose pseudoinverses of D. Then we have

$$\begin{aligned} D^+ &= D^+DD^+ = D^+DD^-DD^+ \\ &= (D^+D)^*(D^-D)^*D^+ = (D^-DD^+D)^*D^+ \\ &= (D^-D)^*D^+ = D^-DD^+ = D^-(DD^+)^* \\ &= D^-(DD^-DD^+)^* = D^-(DD^+)^*(DD^-)^* \\ &= D^-DD^+DD^- = D^-DD^- = D^-. \end{aligned}$$ □

We finish this section with the remark that a Euclidean space X, i.e., a (real) vector space with an inner product, trivially forms a dual system (X, X) with itself.

4. Generalized inverses. According to (4) and (10) we consider the minimization problem

$$\frac{1}{2}\|x\|^2 = \min! \quad \text{s.t.} \quad \frac{1}{2}\|Dx - f\|^2 = \min! \tag{14}$$

with D defined by

$$Dx(t) = E(t)\dot{x}(t) - A(t)x(t) \tag{15}$$

from (9) or more explicitly

$$\frac{1}{2}\int_a^b \|x(t)\|_2^2\, dt = \min! \quad \text{s.t.} \quad \frac{1}{2}\int_a^b \|E(t)\dot{x}(t) - A(t)x(t) - f(t)\|_2^2\, dt = \min!. \tag{16}$$

In this form the specification of the problem is not complete. We still have to specify the appropriate spaces X and Y for D to act between. Requiring x to be continuously differentiable in general yields a continuous $f = Dx$. But even in the uniquely solvable case, f being continuous cannot guarantee the solution x to be continuously differentiable as the case $E \equiv 0$ shows.

We circumvent this problem by setting

$$\begin{aligned} X &= \{x \in C([a,b],\mathbb{R}^n) \mid \boldsymbol{E}^+\boldsymbol{E}x \in C^1([a,b],\mathbb{R}^n),\ \boldsymbol{E}^+\boldsymbol{E}x(a) = 0\}, \\ Y &= C([a,b],\mathbb{R}^m), \end{aligned} \tag{17}$$

and defining $D\colon X \to Y$ indirectly via the standard form (7) by

$$D = \boldsymbol{P}^T\tilde{D}\boldsymbol{Q}^T \tag{18}$$

where $\tilde{D}\colon \tilde{X} \to \tilde{Y}$ with

$$\tilde{D}\tilde{x}(t) = \tilde{E}(t)\dot{\tilde{x}}(t) - \tilde{A}(t)\tilde{x}(t) \tag{19}$$

and

$$\begin{aligned} \tilde{X} &= \{\tilde{x} \in C([a,b],\mathbb{R}^n) \mid \tilde{\boldsymbol{E}}^+\tilde{\boldsymbol{E}}\tilde{x} \in C^1([a,b],\mathbb{R}^n),\ \tilde{\boldsymbol{E}}^+\tilde{\boldsymbol{E}}\tilde{x}(a) = 0\}, \\ \tilde{Y} &= C([a,b],\mathbb{R}^m). \end{aligned} \tag{20}$$

To simplify notation, here and in the following we use bold letters to denote operators standing for pointwise application of the corresponding matrix function; e.g., $\boldsymbol{E}x(t) = E(t)x(t)$. Similarly, one has to interpret superscripts at such operators; e.g., $\boldsymbol{Q}^T x(t) = Q(t)^T x(t)$. In this way the matrix functions P and Q fix operators $\boldsymbol{P}\colon Y \to \tilde{Y}$ and $\boldsymbol{Q}\colon \tilde{X} \to X$. The latter property holds, because for $\tilde{x} \in \tilde{X}$ and $x = \boldsymbol{Q}\tilde{x}$ we get

$$\begin{aligned} \boldsymbol{E}^+\boldsymbol{E}x &= (\boldsymbol{P}^T\tilde{\boldsymbol{E}}\boldsymbol{Q}^T)^+(\boldsymbol{P}^T\tilde{\boldsymbol{E}}\boldsymbol{Q}^T)x \\ &= \boldsymbol{Q}\tilde{\boldsymbol{E}}^+\boldsymbol{P}\boldsymbol{P}^T\tilde{\boldsymbol{E}}\tilde{x} = \boldsymbol{Q}\tilde{\boldsymbol{E}}^+\tilde{\boldsymbol{E}}\tilde{x} \in C^1([a,b],\mathbb{R}^n), \end{aligned}$$

because $Q \in C^1([a,b],\mathbb{R}^{n,n})$ and P, Q are pointwise orthogonal, and hence $x \in X$.

Setting $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)$ according to the block structure of the standard form (8) and observing the special form of $\tilde{E}$, the condition $\tilde{x} \in \tilde{X}$ implies $\tilde{x}_1$ to be continuously differentiable. But only this part of $\tilde{x}$ actually appears on the right-hand side of (19). Thus (18) indeed defines an operator $D\colon X \to Y$ allowing the use of less smooth functions x compared with Definition 2.1. In addition, it is easy to see that D is a homomorphism. Compare this construction with the introduction of a so-called modified matrix pencil in [10] which also aimed at admitting less smooth solutions.

In accordance with the theory of differential equations, we call D a *differential-algebraic operator.*

Because

$$\|x\| = \|\boldsymbol{Q}^T x\| = \|\tilde{x}\|, \quad \|Dx - f\| = \|\boldsymbol{P}(\boldsymbol{P}^T\tilde{D}\boldsymbol{Q}^T x - f)\| = \|\tilde{D}\tilde{x} - \tilde{f}\|, \tag{21}$$

the minimization problem (14) transforms covariantly with the application of the operators $\boldsymbol{P}$ and $\boldsymbol{Q}$. Consequently, we can first solve the minimization problem for DAEs in standard form and then transform the solution back to get a solution of the original problem. Moreover, having found the Moore–Penrose pseudoinverse $\tilde{D}^+$ of $\tilde{D}$ the relation

$$D^+ = \boldsymbol{Q}\tilde{D}^+\boldsymbol{P} \tag{22}$$

immediately gives the Moore–Penrose pseudoinverse of D.

Inserting the explicit form of $\tilde{E}$ and $\tilde{A}$ into (16) for the transformed problem yields

$$\begin{aligned} &\frac{1}{2}\int_a^b (\tilde{x}_1(t)^T\tilde{x}_1(t) + \tilde{x}_2(t)^T\tilde{x}_2(t) + \tilde{x}_3(t)^T\tilde{x}_3(t))\,dt = \min! \\ &\qquad \text{s.t. } \frac{1}{2}\int_a^b (\tilde{w}_1(t)^T\tilde{w}_1(t) + \tilde{w}_2(t)^T\tilde{w}_2(t) + \tilde{w}_3(t)^T\tilde{w}_3(t))\,dt = \min! \end{aligned} \tag{23}$$

with

$$\begin{aligned} \tilde{w}_1(t) &= \Sigma_E(t)\dot{\tilde{x}}_1(t) - A_{11}(t)\tilde{x}_1(t) - A_{12}(t)\tilde{x}_2(t) - A_{13}(t)\tilde{x}_3(t) - \tilde{f}_1(t), \\ \tilde{w}_2(t) &= -A_{21}(t)\tilde{x}_1(t) - \Sigma_A(t)\tilde{x}_2(t) - \tilde{f}_2(t), \\ \tilde{w}_3(t) &= -\tilde{f}_3(t), \end{aligned} \tag{24}$$

where $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)$ and $\tilde{f} = (\tilde{f}_1, \tilde{f}_2, \tilde{f}_3)$ are partitioned according to the given block structure. For given $\tilde{f} \in \tilde{Y}$ minimization is to be taken over the whole of $\tilde{X}$ from (20) which can be written as

$$\tilde{X} = \{(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) \in C([a,b], \mathbb{R}^n) \mid \tilde{x}_1 \text{ continuously differentiable, } \tilde{x}_1(a) = 0\}. \tag{25}$$

The constraint is easily satisfied by choosing an arbitrary continuous function $\tilde{x}_3$, taking $\tilde{x}_1$ to be the solution of the linear initial value problem

$$\begin{aligned} \dot{\tilde{x}}_1(t) &= \Sigma_E(t)^{-1}[A_{11}(t) - A_{12}(t)\Sigma_A(t)^{-1}A_{21}(t)]\tilde{x}_1(t) \\ &\quad + \Sigma_E(t)^{-1}[A_{13}(t)\tilde{x}_3(t) + \tilde{f}_1(t) - A_{12}(t)\Sigma_A(t)^{-1}\tilde{f}_2(t)],\ \tilde{x}_1(a) = 0, \end{aligned} \tag{26}$$

and finally setting

$$\tilde{x}_2(t) = -\Sigma_A(t)^{-1}[A_{21}(t)\tilde{x}_1(t) + \tilde{f}_2(t)]. \tag{27}$$

Thus we remain with the problem of minimizing $\frac{1}{2}\|x\|^2$ under the constraints (26) and (27).

Observe that this is the place where pointwise invertibility of the matrices Σ_E, Σ_A is needed in order to satisfy the constraints. But it is clear that weak solvability or smooth completion of solutions may be applied to generalize our results.

Taking a closer look at system (23), one immediately recognizes a linear quadratic control problem where $\tilde{x}_3$ takes the role of the control. But compared with the standard problem of linear quadratic control, the constraints appear to be more general due to the occurrence of inhomogeneities. See, e.g., [20] and references therein for related results on the homogeneous linear quadratic control problem.

4.1. Linear quadratic control problems with inhomogeneities. Because the solution of linear quadratic control problems with inhomogeneities is an interesting topic by itself, we treat this problem with a new simplified and adapted notation which should not be mixed up with the one used so far.

THEOREM 4.1. *Let*

$$\begin{aligned} &A \in C([a,b],\mathbb{R}^{d,d}), \quad B \in C([a,b],\mathbb{R}^{d,k}), \quad C \in C([a,b],\mathbb{R}^{l,d}),\\ &f \in C([a,b],\mathbb{R}^{d}), \quad g \in C([a,b],\mathbb{R}^{l}). \end{aligned} \tag{28}$$

Then the linear quadratic control problem

$$\begin{aligned} &\frac{1}{2}\int_a^b (x(t)^T x(t) + y(t)^T y(t) + u(t)^T u(t))\, dt = \min!\\ &\qquad s.t.\ \dot{x}(t) = A(t)x(t) + B(t)u(t) + f(t),\ x(a) = 0,\\ &\qquad\quad y(t) = C(t)x(t) + g(t) \end{aligned} \tag{29}$$

possesses a unique solution $x \in C^1([a,b],\mathbb{R}^d)$, $y \in C([a,b],\mathbb{R}^l)$, $u \in C([a,b],\mathbb{R}^k)$. *This solution coincides with the corresponding part of the unique solution of the boundary value problem*

$$\begin{aligned} \dot{\lambda}(t) &= (I + C(t)^T C(t))x(t) - A(t)^T\lambda(t) + C(t)^T g(t),\ \lambda(b) = 0,\\ \dot{x}(t) &= A(t)x(t) + B(t)u(t) + f(t),\ x(a) = 0,\\ y(t) &= C(t)x(t) + g(t),\\ u(t) &= B(t)^T\lambda(t) \end{aligned} \tag{30}$$

which can be obtained by the successive solution of the initial value problems

(31)

$$\begin{aligned} \dot{P}(t) &= I + C(t)^T C(t) - P(t)A(t) - A(t)^T P(t) - P(t)B(t)B(t)^T P(t),\ P(b) = 0,\\ \dot{v}(t) &= C(t)^T g(t) - P(t)f(t) - A(t)^T v(t) - P(t)B(t)B(t)^T v(t),\ v(b) = 0,\\ \dot{x}(t) &= A(t)x(t) + B(t)B(t)^T(P(t)x(t) + v(t)) + f(t),\ x(a) = 0,\\ \lambda(t) &= P(t)x(t) + v(t),\\ y(t) &= C(t)x(t) + g(t),\\ u(t) &= B(t)^T\lambda(t). \end{aligned}$$

Proof. Eliminating y with the help of the algebraic constraint and using a Lagrangian multiplier λ (see, e.g., [13]), problem (29) is equivalent to (omitting arguments)

$$J[x,\dot{x},u,\lambda] = \int_a^b \left[\frac{1}{2}(x^T x + (Cx+g)^T(Cx+g) + u^T u) + \lambda^T(\dot{x} - Ax - Bu - f)\right] dt = \min!$$

with $x, \lambda \in C^1([a,b],\mathbb{R}^d)$, and $u \in C([a,b],\mathbb{R}^k)$. Variational calculus then yields

$$\begin{aligned} &J[x + \varepsilon\delta x, \dot{x} + \varepsilon\delta\dot{x}, u + \varepsilon\delta u, \lambda + \varepsilon\delta\lambda]\\ &\quad = \int_a^b \Big[\frac{1}{2}\big((x + \varepsilon\delta x)^T(x + \varepsilon\delta x) + (u + \varepsilon\delta u)^T(u + \varepsilon\delta u)\big)\\ &\qquad + (C(x + \varepsilon\delta x) + g)^T(C(x + \varepsilon\delta x) + g)\\ &\qquad + (\lambda + \varepsilon\delta\lambda)^T\big((\dot{x} + \varepsilon\delta\dot{x}) - A(x + \varepsilon\delta x) - B(u + \varepsilon\delta u) - f\big)\Big]\, dt \end{aligned}$$

$$
\begin{aligned}
&= J[x, \dot{x}, u, \lambda] \\
&\quad + \varepsilon\left[\lambda^T \delta x\big|_a^b + \int_a^b (x^T + (Cx+g)^T C - \lambda^T A - \dot{\lambda}^T)\delta x \, dt\right. \\
&\qquad \left. + \int_a^b (u^T - \lambda^T B)\delta u \, dt + \int_a^b \delta\lambda^T(\dot{x} - Ax - Bu - f) \, dt\right] \\
&\quad + \varepsilon^2\left[\frac{1}{2}\int_a^b (\delta x^T(I + C^T C)\delta x + \delta u^T \delta u) \, dt + \int_a^b \delta\lambda^T(\delta\dot{x} - A\delta x - B\delta u) \, dt\right]
\end{aligned}
$$

after sorting and integration by parts.

For (x, u, λ) to be a minimum, a necessary condition is that for all variations the coefficient of ε vanishes. This at once yields (30).

Now let $(x+\varepsilon\delta x, u+\varepsilon\delta u, \lambda+\varepsilon\delta\lambda)$ be a second minimum. Without loss of generality we have $\varepsilon > 0$. Then $(\delta x, \delta u, \delta\lambda)$ must solve the corresponding homogeneous problem. In particular, we must have

$$\delta\dot{x} = A\delta x + B\delta u.$$

Thus, in this case,

$$
\begin{aligned}
&J[x + \varepsilon\delta x, \dot{x} + \varepsilon\delta\dot{x}, \lambda + \varepsilon\delta\lambda, u + \varepsilon\delta u] \\
&\quad = J[x, \dot{x}, \lambda, u] + \varepsilon^2 \int_a^b \frac{1}{2}(\delta x^T(I + C^T C)\delta x + \delta u^T \delta u) \, dt.
\end{aligned}
$$

It follows that $\delta x \equiv 0$, $\delta u \equiv 0$, and consequently $\delta\lambda \equiv 0$. Hence, there is at most one solution of the linear quadratic control problem (29) and thus also of the boundary value problem (30).

To determine the unique solution of (30) we set

$$\lambda = Px + v, \quad \dot{\lambda} = P\dot{x} + \dot{P}x + \dot{v},$$

with some $P \in C^1([a,b], \mathbb{R}^{d,d})$, $v \in C^1([a,b], \mathbb{R}^d)$. Inserting into (30), we obtain

$$P\dot{x} + \dot{P}x + \dot{v} = (I + C^T C)x - A^T(Px + v) + C^T g$$

and

$$P\dot{x} = PAx + PBB^T(Px + v) + Pf.$$

Combining these equations, we obtain

$$
\begin{aligned}
&(PA + A^T P + PBB^T P - (I + C^T C) + \dot{P})x \\
&\qquad + (PBB^T v + Pf + A^T v - C^T g + \dot{v}) = 0.
\end{aligned}
$$

Now we choose P and v to be the solutions of the initial value problems

$$
\begin{aligned}
&\dot{P} = I + C^T C - PA - A^T P - PBB^T P, \ P(b) = 0, \\
&\dot{v} = C^T g - Pf - A^T v - PBB^T v, \ v(b) = 0.
\end{aligned}
$$

This choice is possible because the second equation is linear and the first equation is a Riccati differential equation of a kind for which one can show that a symmetric solution exists for any interval of the form $[a, b]$; see, e.g., [15, Chap. 10].

It remains to show that (31) indeed solves (30). This is trivial for the third and fourth equations. For the second equation we of course have $x(a) = 0$ but also

$$\begin{aligned}&\dot{x} - Ax - Bu - f\\ &\quad = Ax + BB^T Px + BB^T v + f - Ax - BB^T Px - BB^T v - f = 0.\end{aligned}$$

For the first equation we have $\lambda(b) = P(b)x(b) + v(b) = 0$ and also

$$\begin{aligned}&\dot{\lambda} - (I + C^T C)x + A^T\lambda - C^T g\\ &\quad = P\dot{x} + \dot{P}x + \dot{v} - (I + C^T C)x + A^T Px + A^T v - C^T g\\ &\quad = PAx + PBB^T Px + PBB^T v + Pf\\ &\qquad + (I + C^T C)x - PAx - A^T Px - PBB^T Px\\ &\qquad + C^T g - Pf - A^T v - PBB^T v - (I + C^T C)x + A^T Px + A^T v - C^T g = 0. \quad \square\end{aligned}$$

We remark here that the objective functional in a standard linear quadratic control problem often contains pointwise symmetric and positive definite matrix functions as additional parameters. Problem (29), however, represents no loss of generality because using the Cholesky decomposition of such matrix-valued functions, which is smooth, we can rescale the unknowns by linear transformations such that these matrix functions become pointwise identities.

4.2. The Moore–Penrose inverse of differential-algebraic operators. We now apply the results obtained for linear quadratic control problems with inhomogeneities to construct the Moore–Penrose inverse of a differential-algebraic operator.

COROLLARY 4.2. *Problem* (23) *with constraints* (26) *and* (27) *has a unique solution* $\tilde{x} \in \tilde{X}$.

Proof. The claim follows from Theorem 4.1 by the following substitutions (again without arguments)

$$\begin{aligned}&A = \Sigma_E^{-1}(A_{11} - A_{12}\Sigma_A^{-1}A_{21}),\ B = \Sigma_E^{-1}A_{13},\ C = -\Sigma_A^{-1}A_{21},\\ &f = \Sigma_E^{-1}(\tilde{f}_1 - A_{12}\Sigma_A^{-1}\tilde{f}_2),\ g = -\Sigma_A^{-1}\tilde{f}_2.\end{aligned}$$

The unique solution is then given in the form $\tilde{x} = (x, y, u)$. $\square$

We are now ready to define an appropriate operator $\tilde{D}^+ : \tilde{Y} \to \tilde{X}$ as follows. For $\tilde{f} = (\tilde{f}_1, \tilde{f}_2, \tilde{f}_3) \in \tilde{Y}$, the image $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) = \tilde{D}^+\tilde{f}$ shall be the unique solution of (23) with (26) and (27). Note that $\tilde{D}^+\tilde{f} \in \tilde{X}$ because $\tilde{x}_1$ as part x of (31) is continuously differentiable and $\tilde{x}_1(a) = 0$. Moreover, because the Riccati differential equation in (31) does not depend on the inhomogeneities, the operator $\tilde{D}^+$ is linear, hence a homomorphism.

THEOREM 4.3. *The operator* $\tilde{D}^+$, *defined as above, is the Moore–Penrose pseudoinverse of* $\tilde{D}$; *i.e., the endomorphisms* $\tilde{D}\tilde{D}^+$ *and* $\tilde{D}^+\tilde{D}$ *have conjugates such that* (13) *holds for* $\tilde{D}$ *and* $\tilde{D}^+$.

Proof. Let $\tilde{f} = (\tilde{f}_1, \tilde{f}_2, \tilde{f}_3) \in \tilde{Y}$ and $\tilde{D}\tilde{D}^+\tilde{f} = (\hat{f}_1, \hat{f}_2, \hat{f}_3)$. With (19) and the notation of Theorem 4.1 and Corollary 4.2 (for simplicity) we get

$$\begin{aligned}\hat{f}_1 &= \Sigma_E\dot{\tilde{x}}_1 - A_{11}\tilde{x}_1 - A_{12}\tilde{x}_2 - A_{13}\tilde{x}_3\\ &= \Sigma_E\dot{x} - A_{11}x - A_{12}y - A_{13}u\\ &= \Sigma_E(Ax + BB^T(Px + v) + f) - A_{11}x - A_{12}(Cx + g) - A_{13}B^T(Px + v)\\ &= A_{11}x - A_{12}\Sigma_A^{-1}A_{21}x + \Sigma_E BB^T(Px + v) + \tilde{f}_1 - A_{12}\Sigma_A^{-1}\tilde{f}_2\\ &\qquad - A_{11}x + A_{12}\Sigma_A^{-1}A_{21}x + A_{12}\Sigma_A^{-1}\tilde{f}_2 - \Sigma_E BB^T(Px + v) = \tilde{f}_1,\\ \hat{f}_2 &= -A_{21}\tilde{x}_1 - \Sigma_A\tilde{x}_2 = -A_{21}x - \Sigma_A y\\ &= \Sigma_A(Cx - y) = \Sigma_A(Cx - Cx - g) = \tilde{f}_2,\\ \hat{f}_3 &= 0,\end{aligned}$$

and $\tilde{D}\tilde{D}^+$ is obviously conjugate to itself. Since $\tilde{D}\tilde{D}^+$ projects onto the first two components and, because $\tilde{f}_3$ has no influence on the solution of (23), we also have $\tilde{D}^+\tilde{D}\tilde{D}^+ = \tilde{D}^+$. Because $\tilde{D}\tilde{x}$ has a vanishing third component for all $\tilde{x} \in \tilde{X}$, the projector $\tilde{D}\tilde{D}^+$ acts as an identity on $\tilde{D}$; i.e., $\tilde{D}\tilde{D}^+\tilde{D} = \tilde{D}$. The rest of the proof deals with the fourth Penrose axiom.

Let $\tilde{x} = (x, y, u) \in \tilde{X}$ and $\tilde{D}^+\tilde{D}\tilde{x} = (\hat{x}, \hat{y}, \hat{u})$. We must now apply $\tilde{D}^+$ to the inhomogeneity

$$\tilde{D}\tilde{x} = \begin{bmatrix} \Sigma_E\dot{x} - A_{11}x - A_{12}y - A_{13}u \\ -A_{21}x - \Sigma_A y \\ 0 \end{bmatrix}.$$

Therefore we must set

$$\begin{aligned} f &= \Sigma_E^{-1}(\Sigma_E\dot{x} - A_{11}x - A_{12}y - A_{13}u + A_{12}\Sigma_A^{-1}(A_{21}x + \Sigma_A y)) \\ &= \dot{x} - Ax - Bu, \\ g &= \Sigma_A^{-1}(A_{21}x + \Sigma_A y) \\ &= -Cx + y. \end{aligned}$$

Recalling that the solution P of the Riccati differential equation does not depend on the inhomogeneity, we must solve

$$\begin{aligned} &\dot{v} = C^T(-Cx + y) - A^T v - PBB^T v - P(\dot{x} - Ax - Bu),\ v(b) = 0, \\ &\dot{\hat{x}} = A\hat{x} + BB^T(P\hat{x} + v) + (\dot{x} - Ax - Bu),\ \hat{x}(a) = 0, \\ &\hat{y} = C\hat{x} - Cx + y, \\ &\hat{u} = B^T(P\hat{x} + v). \end{aligned}$$

Setting $v = w - Px$, $\dot{v} = \dot{w} - P\dot{x} - \dot{P}x$, we obtain

$$\begin{aligned} \dot{w} &= P\dot{x} + (I + C^TC)x - PAx - A^TPx - PBB^TPx \\ &\quad - C^TCx + C^Ty - A^Tw + A^TPx - PBB^Tw + PBB^TPx \\ &\quad - P\dot{x} + PAx + PBu \\ &= -(A^T + PBB^T)w + (x + C^Ty + PBu),\ w(b) = 0. \end{aligned} \tag{32}$$

Let $W(t, s)$ be the Wronskian matrix belonging to $A + BB^TP$ in the sense that

$$\dot{W}(t, s) = (A + BB^TP)W(t, s),\ W(s, s) = I.$$

Then $W(t, s)^{-T}$ is the Wronskian matrix belonging to $-(A + PBB^T)$. With the help of $W(t, s)$ we can represent the solution of the initial value problem (32) in the form

$$w = \int_b^t W(t, s)^{-T}(x + C^Ty + PBu)\, ds,$$

or

$$v = -Px + \int_b^t W(t, s)^{-T}(x + C^Ty + PBu)\, ds.$$

Here, and in the following, the arguments which must be inserted start with t, and a Wronskian matrix changes it from the first to the second argument.

Setting $\hat{x} = x + z$, we obtain

$$\begin{aligned} \dot{z} &= -\dot{x} + Ax + Az + BB^TPx + BB^TPz \\ &\quad + BB^Tw - BB^TPx + \dot{x} - Ax - Bu \\ &= (A + BB^TP)z + (BB^Tw - Bu),\ z(a) = 0, \end{aligned}$$

or

$$z = \int_a^t W(t,s)(BB^T w - Bu)\, ds.$$

Thus we get $(\hat{x}, \hat{y}, \hat{u})$ according to

$$\hat{x} = x + z,\ \hat{y} = y + Cz,\ \hat{u} = B^T(Pz + w).$$

In addition, now let $(\bar{x}, \bar{y}, \bar{u}) \in \tilde{X}$ be given and $\tilde{D}^+\tilde{D}(\bar{x}, \bar{y}, \bar{u}) = (\hat{\bar{x}}, \hat{\bar{y}}, \hat{\bar{u}})$. Then we have

$$\begin{aligned}
&\int_a^b (\bar{x}^T\hat{x} + \bar{y}^T\hat{y} + \bar{u}^T\hat{u})\, dt \\
&\quad = \int_a^b \Bigg[\bar{x}^T x + \bar{x}^T \int_a^t W(t,s)(BB^T \int_b^s W(s,r)^{-T}(x + C^T y + PBu)\, dr - Bu)\, ds \\
&\qquad + \bar{y}^T y + \bar{y}^T C \int_a^t W(t,s) \left(BB^T \int_b^s W(s,r)^{-T}(x + C^T y + PBu)\, dr - Bu\right) ds \\
&\qquad + \bar{u}^T B^T P \int_a^t W(t,s) \left(BB^T \int_b^s W(s,r)^{-T}(x + C^T y + PBu)\, dr - Bu\right) ds \\
&\qquad + \bar{u}^T B^T \int_b^t W(t,s)^{-T}(x + C^T y + PBu)\, ds\Bigg]\, dt \\
&\quad = \int_a^b (\bar{x}^T x + \bar{y}^T y)\, dt \\
&\qquad - \int_a^b \int_a^t (\bar{x}^T + \bar{y}^T C + \bar{u}^T B^T P) W(t,s) Bu\, ds\, dt \\
&\qquad + \int_a^b \int_b^t \bar{u}^T B^T W(t,s)^{-T}(x + C^T y + PBu)\, ds\, dt \\
&\qquad + \int_a^b \int_b^t \int_b^s (\bar{x}^T + \bar{y}^T C + \bar{u}^T B^T P) W(t,s) B \\
&\qquad\qquad \cdot B^T W(s,r)^{-T}(x + C^T y + PBu)\, dr\, ds\, dt.
\end{aligned}$$

By transposition and changing the order of the integrations, we finally find

$$\int_a^b (\bar{x}^T\hat{x} + \bar{y}^T\hat{y} + \bar{u}^T\hat{u})\, dt = \int_a^b (x^T\hat{\bar{x}} + y^T\hat{\bar{y}} + u^T\hat{\bar{u}})\, dt,$$

which is nothing else than that $\tilde{D}^+\tilde{D}$ is conjugate to itself. □

It follows immediately that (22) yields the Moore–Penrose pseudoinverse of D. That is, we have shown the existence and uniqueness of an operator D^+ satisfying (13) and thus fixed a unique classical least squares solution for a large class of DAEs (inculding higher index problems) with possibly inconsistent initial data or inhomogeneities or free solution components.

4.3. A (1,2,3)-inverse. Using D^+ for solving DAEs with undetermined solutions components, however, bears at least two disadvantages. First, the undetermined component $\tilde{x}_3$ need not satisfy the given initial value and, second, instead of an initial value problem we must solve a boundary value problem, which means that values of the coefficients in future times influence the solution at the present time.

A simple way out of this problem is to choose the undetermined part to be zero. In the following we shall investigate this approach in the context of generalized inverses.

To do this we consider the matrix functions given by

$$F(t) = (I - E(t)E(t)^+)A(t)(I - E(t)^+E(t)) \tag{33}$$

and

$$\Pi(t) = E(t)^+E(t) + F(t)^+F(t). \tag{34}$$

Transforming to standard form, we then find (omitting arguments)

$$\begin{aligned}\tilde{F} &= (I - \tilde{E}\tilde{E}^+)\tilde{A}(I - \tilde{E}^+\tilde{E}) \\ &= (I - PEQQ^TE^+P^T)(PAQ - PE\dot{Q})(I - Q^TE^+P^TPEQ) \\ &= P(I - EE^+)(A - E\dot{Q}Q^T)(I - E^+E)Q \\ &= P(I - EE^+)A(I - E^+E)Q = PFQ.\end{aligned}$$

Thus F transforms like E and therefore

$$\begin{aligned}\tilde{\Pi} &= \tilde{E}^+\tilde{E} + \tilde{F}^+\tilde{F} \\ &= Q^TE^+P^TPEQ + Q^TF^+P^TPFQ \\ &= Q^T(E^+E + F^+F)Q = Q^T\Pi Q.\end{aligned}$$

A simple calculation now yields

$$\tilde{\Pi} = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

This in particular shows that Π is pointwise an orthogonal projector. Note that $I - \tilde{\Pi}$ indeed projects onto the undetermined component $\tilde{x}_3$. Hence, we are led to the problem

$$\begin{aligned}&\frac{1}{2}\int_a^b \|(I - \Pi(t))x(t)\|_2^2\, dt = \min! \\ &\qquad \text{s.t. } \frac{1}{2}\int_a^b \|E(t)\dot{x}(t) - A(t)x(t) - f(t)\|_2^2\, dt = \min!\end{aligned} \tag{35}$$

replacing (16). The preceding results say that again the problem transforms covariantly with the application of $\boldsymbol{P}$ and $\boldsymbol{Q}$ so that we only need to solve (35) for DAEs in standard form. Because (35) here implies $\tilde{x}_3 = 0$ by construction, we remain with a reduced DAE that is uniquely solvable. We can therefore carry over all results obtained so far as long as they do not depend on the specific choice of $\tilde{x}_3$. Recognizing that this choice was utilized only for the fourth Penrose axiom, we find that (35) fixes a so-called (1,2,3)-inverse $\tilde{D}^-$ of $\tilde{D}$ satisfying the axioms (13 a, b, c). Keeping the spaces as before, we arrive at the following result.

THEOREM 4.4. *The operator $\tilde{D}^-$ defined by* (35) *is a $(1,2,3)$-inverse of $\tilde{D}$; i.e., the endomorphism $\tilde{D}\tilde{D}^-$ has a conjugate such that* (13 a, b, c) *hold for $\tilde{D}$ and $\tilde{D}^-$.*

Again defining the operator D^- by $D^- = \boldsymbol{Q}\tilde{D}^-\boldsymbol{P}$ then gives a (1,2,3)-inverse of the operator D. We finish this part with a number of remarks and an example for the application of the presented theory.

Remark 1. In the case $A_{13} \equiv 0$ (including A_{13} empty, i.e., no corresponding block in the standard form), we immediately have $D^- = D^+$. Observing that for $E \equiv 0$ the existence of a standard form (8) requires $\operatorname{rank} A(t)$ to be constant on $[a, b]$, we

find $D^+ = D^- = -\boldsymbol{A}^+$. In particular, this shows that both D^+ and D^- are indeed generalizations of the Moore–Penrose pseudoinverse of matrices.

Remark 2. The boundedness of the linear operators $D: X \to Y$ and $D^+, D^-: Y \to X$ where X and Y are seen as the given linear spaces equipped with the norms $\|x\|_X = \|x\|_{L_2} + \|d/dt(\boldsymbol{E}^+\boldsymbol{E}x)\|_{L_2}$ and $\|y\|_Y = \|y\|_{L_2}$ allows for their extension to the closure of X and Y with respect to these norms; see, e.g., [9, Lemma 4.3.16]. In particular, Y becomes the Hilbert space $L_2([a,b],\mathbb{R}^m)$. Other choices of the norms are possible as well.

Remark 3. For the numerical calculation of a solution of a given DAE represented by the operators D^+ or D^-, one has to discretize (16) or (35). Using fixed stepsize $h = (b-a)/N$, $N \in \mathbb{N}$, one would choose discrete spaces X_h and Y_h of finite sequences $\{x_\nu\}_{\nu=0}^N$ and $\{f_\nu\}_{\nu=0}^N$. Thus by discretization we come back to a finite dimensional problem of the form (3) where we know how to compute generalized inverses. But any numerical scheme will couple $x_{\nu+1}$ at least with x_ν due to replacing the derivative by some difference approximation. Because a (1,2,3)-inverse of a lower block triangular matrix is in general not lower block triangular, it is not clear whether there is a (1,2,3)-inverse such that x_ν does not depend on values of the coefficients at points in the future.

Example 4.5. Consider the initial value problem

$$\begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix}, \quad \begin{bmatrix} x_1(t_0) \\ x_2(t_0) \end{bmatrix} = \begin{bmatrix} x_{10} \\ x_{20} \end{bmatrix}.$$

The DAE of this problem has strangeness-index one (note that in contrast the differentiation index is not defined); see [17]. To obtain a strangeness-free DAE with the same solution space according to [19], we compute

$$M = \begin{bmatrix} E & 0 \\ \dot{E} - A & E \end{bmatrix}, \quad N = \begin{bmatrix} A & 0 \\ \dot{A} & 0 \end{bmatrix}, \quad g = \begin{bmatrix} f \\ \dot{f} \end{bmatrix}$$

and obtain (with shifted initial values)

$$M(t) = \left[\begin{array}{cc|cc} -t & t^2 & 0 & 0 \\ -1 & t & 0 & 0 \\ \hline 0 & 2t & -t & t^2 \\ 0 & t & -1 & t \end{array}\right], \quad N(t) = \left[\begin{array}{cc|cc} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right], \quad g(t) = \left[\begin{array}{c} f_1(t) - x_{10} \\ f_2(t) - x_{20} \\ \hline \dot{f}_1(t) \\ \dot{f}_2(t) \end{array}\right].$$

Because $\operatorname{rank} M(t) = 2$ for all $t \in \mathbb{R}$, the procedure in [19] reduces to the computation of an orthogonal projection onto the corange of $M(t)$ given, e.g., by

$$Z(t)^T = \frac{1}{\sqrt{1+t^2}} \left[\begin{array}{cc|cc} 1 & -t & 0 & 0 \\ 0 & 0 & 1 & -t \end{array}\right].$$

Now replacing E, A, and f by Z^TM, Z^TN, and Z^Tg yields the strangeness-free DAE

$$\begin{aligned} 0 &= \frac{1}{\sqrt{1+t^2}}(x_1(t) + tx_2(t) + f_1(t) - x_{01} - tf_2(t) + tx_{20}), \\ 0 &= \frac{1}{\sqrt{1+t^2}}(\dot{f}_1(t) - t\dot{f}_2(t)), \end{aligned}$$

together with homogeneous initial conditions. Denoting the coefficient functions again by E, A, and f, we have $E(t) = 0$ and

$$A(t) = \frac{1}{\sqrt{1+t^2}} \begin{bmatrix} -1 & t \\ 0 & 0 \end{bmatrix}, \quad f(t) = \frac{1}{\sqrt{1+t^2}} \begin{bmatrix} f_1(t) - x_{01} - tf_2(t) + tx_{20} \\ \dot{f}_1(t) - t\dot{f}_2(t) \end{bmatrix}.$$

According to Remark 1, the least squares solution of the latter DAE is given by $x = -\boldsymbol{A}^+ f$. Shifting back we obtain as the least squares solution of the given original problem

$$x(t) = -\frac{1}{\sqrt{1+t^2}} \begin{bmatrix} -1 & 0 \\ t & 0 \end{bmatrix} \frac{1}{\sqrt{1+t^2}} \begin{bmatrix} f_1(t) - x_{01} - tf_2(t) + tx_{20} \\ \dot{f}_1(t) - t\dot{f}_2(t) \end{bmatrix} + \begin{bmatrix} x_{01} \\ x_{02} \end{bmatrix}$$

or

$$x(t) = \frac{1}{1+t^2} \begin{bmatrix} f_1(t) - x_{01} - tf_2(t) + tx_{20} \\ -t(f_1(t) - x_{01} - tf_2(t) + tx_{20}) \end{bmatrix} + \begin{bmatrix} x_{01} \\ x_{02} \end{bmatrix}.$$

5. Conclusions. Considering linear DAEs as common generalization of linear ordinary differential equations and linear algebraic equations, our aim in this paper was to define a counterpart of a least squares solution in the case of inconsistent data and/or nonuniquely solvable problems. For this, we followed the approach taken for linear algebraic equations. In particular, we embedded DAEs of a certain type into a minimization problem which was then shown to be uniquely solvable. The corresponding solution operator turned out to satisfy axioms of Penrose type in a general setting of conjugates with respect to some dual systems. In this sense we defined least squares solutions of a large class of DAEs or, in other words, Moore–Penrose pseudoinverses of the corresponding differential-algebraic operators.

REFERENCES

[1] A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications*, John Wiley & Sons, New York, 1974.

[2] K. E. Brenan, S. L. Campbell, and L. R. Petzold, *Numerical Solution of Initial-Value Problems in Differential Algebraic Equations*, Elsevier, North Holland, New York, 1989.

[3] R. Byers, P. Kunkel, and V. Mehrmann, *Regularization of Linear Descriptor Systems with Variable Coefficients*, Tech. report, Fakultät für Mathematik, Technische Universität Chemnitz-Zwickau, D-09107 Chemnitz, Fed. Rep. Germany, 1994.

[4] S. L. Campbell, *Singular Systems of Differential Equations*, Pitman, San Francisco, 1980.

[5] ———, *The numerical solution of higher index linear time varying singular systems of differential equations*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 334–348.

[6] S. L. Campbell and C. D. Meyer, *Generalized Inverses of Linear Transformations*, Pitman, London, 1979.

[7] S. L. Campbell, C. D. Meyer, and N. J. Rose, *Applications of the Drazin inverse to linear systems of differential equations with singular constant coefficients*, SIAM J. Appl. Math., 31 (1976), pp. 411–425.

[8] S. L. Campbell and L. R. Petzold, *Canonical forms and solvable singular systems of differential equations*, SIAM J. Algebraic Disc. Meth., 4 (1983), pp. 517–521.

[9] R. Engelking, *General Topology*, Polish Scientific Publishers, Warszawa, Poland, 1977.

[10] E. Griepentrog and R. März, *Differential-Algebraic Equations and Their Numerical Treatment*, Teubner Verlag, Leipzig, 1986.

[11] M. Hanke, *Linear differential-algebraic equations in spaces of integrable functions*, J. Differential Equations, 79 (1989), pp. 14–30.

[12] B. Hansen, *Comparing Different Concepts to Treat Differential Algebraic Equations*, Tech. Report 220, Sektion Mathematik, Humboldt-Universität, Berlin, 1989.

[13] M. R. Hestenes, *Calculus of Variations and Optimal Control Theory*, John Wiley & Sons, New York, 1966.

[14] H. Heuser, *Funktionalanalysis*, 3rd ed., B. G. Teubner, Stuttgart, 1992.

[15] H. W. Knobloch and H. Kwakernaak, *Lineare Kontrolltheorie*, Springer-Verlag, Berlin, 1985.

[16] P. Kunkel and V. Mehrmann, *Numerical solution of differential algebraic Riccati equations*, Linear Algebra Appl., 137/138 (1990), pp. 39–66.

[17] P. KUNKEL AND V. MEHRMANN, *Canonical forms for linear differential-algebraic equations with variable coefficients*, J. Comput. Appl. Math., 56 (1994), pp. 225–251.

[18] ———, *A new look at pencils of matrix valued functions*, Linear Algebra Appl., 212/213 (1994), pp. 215–248.

[19] ———, *A new class of discretization methods for the solution of linear differential-algebraic equations*, SIAM J. Numer. Anal., 33 (1996), to appear.

[20] V. MEHRMANN, *The Autonomous Linear Quadratic Control Problem*, Springer-Verlag, Berlin, 1991.

[21] P. J. RABIER AND W. C. RHEINBOLDT, *Classical and Generalized Solutions of Time-Dependent Linear Differential Algebraic Equations*, Tech. report, Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, PA, 1993.

[22] W. T. REID, *Ordinary Differential Equations*, John Wiley & Sons, New York, 1971.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 2, pp. 443–459, April 1996

SOME RESULTS ON STRUCTURE PREDICTION IN SPARSE *QR* FACTORIZATION*

ESMOND G. NG† AND BARRY W. PEYTON†

Abstract. In QR factorization of an $m \times n$ matrix A ($m \geq n$), the orthogonal factor Q is often stored implicitly as an $m \times n$ lower trapezoidal matrix W, known as the Householder matrix. When the sparsity of A is to be exploited, the factorization is often preceded by a symbolic factorization step, which computes a data structure in which the nonzero entries of W and R are computed and stored. This is achieved by computing an upper bound on the nonzero structure of these factors, based solely on the nonzero structure of A. In this paper we use a well-known upper bound on the nonzero structure of W to obtain an upper bound on the nonzero structure of Q.

Let U be the matrix consisting of the first n columns of Q. One interesting feature of the new bound is that the bound on W's structure is identical to the lower trapezoidal part of the bound on U's structure. We show that if A is strong Hall and has no zero entry on its main diagonal, then the bounds on the nonzero structures of W and U are the smallest possible based solely on the nonzero structure of A. We then use this result to obtain corresponding smallest upper bounds in the case where A is weak Hall, is in block upper triangular form, and has no zero entry on its main diagonal. Finally, we show that one can always reorder a weak Hall matrix into block upper triangular form so that there is no increase in the fill incurred by the QR factorization.

Key words. structure prediction, QR factorization, Hall property, block upper triangular form, elimination tree, sparse matrix computations

AMS subject classifications. 15A23, 65F05, 65F50

1. Introduction.

Let $A = [a_{ij}]$ be an $m \times n$ matrix with $m \geq n$, and assume that A has full column rank. Consider the reduction of A to upper triangular form using orthogonal factorization:

$$A = Q \begin{bmatrix} R \\ O \end{bmatrix}, \tag{1}$$

where Q is $m \times m$ orthogonal and R is $n \times n$ upper triangular. Assume that both Q and R are needed and that Householder transformations [11] are used to compute the factorization. The orthogonal factor Q can then be stored implicitly in the Householder matrix W, which is an $m \times n$ lower trapezoidal matrix, each column of which contains a vector used to construct a Householder transformation [2].

If A is sparse and its zero entries are to be exploited, the factorization is often preceded by a *symbolic factorization* step, which computes a data structure in which first the nonzero entries of A are inserted and subsequently the nonzero entries of R and W are computed and stored. The primary purpose of the symbolic factorization step is to predict which factor entries will be zero and which will be nonzero, *based solely on the nonzero structure of A*. George and Ng [9] predict the nonzero structure of W and R by applying a *symbolic Householder* procedure to the nonzero structure of A.

Consider the 7×5 matrix A in Figure 1, with each nonzero entry represented by an "×" and each zero entry represented by a blank. The predicted fill due to

* Received by the editors May 22, 1992; accepted for publication (in revised form) by J. W. H. Liu April 28, 1995. This work was supported in part by the Applied Mathematical Sciences Research Program, Office of Energy Research, U. S. Department of Energy contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc., and in part by the Institute for Mathematics and Its Applications at the University of Minnesota with funds provided by the National Science Foundation.

† Mathematical Sciences Section, Oak Ridge National Laboratory, P. O. Box 2008, Oak Ridge, TN 37831-6367 (esmond@msr.epm.ornl.gov, peyton@msr.epm.ornl.gov).

$$A = \begin{bmatrix} \times & & & \times \\ & \times & \times & \times \\ & \times & \times & \times \\ & & & \times & \times \\ \times & & & & \times \\ & & & \times & \times \\ & & & & \times \end{bmatrix} \qquad F = \begin{bmatrix} \times & & & \times \\ & \times & \times & + & \times \\ & \times & \times & \times & + \\ & & & \times & \times \\ \times & & & & \times \\ & & & \times & \times \\ & & & & \times \end{bmatrix}$$

FIG. 1. *A is the original matrix; F is the filled matrix generated by the symbolic Householder procedure. (A filled entry is denoted by +.)*

the first Householder transformation is obtained by (1) determining the rows of A with a nonzero entry in column one and then (2) replacing the nonzero structure of each of these rows with the union of these sets. The first symbolic Householder transformation thus replaces rows one and five with the union of $\{1,5\}$ from row one and $\{1,5\}$ from row five, producing no change in the nonzero structure. The predicted fill due to the second Householder transformation is obtained by removing the first row and column from the matrix and then applying the same row-merge operation to the remaining 6×4 matrix. This replaces $\{2,3,5\}$ in row two and $\{2,3,4\}$ in row three with $\{2,3,4,5\}$, thus introducing a "+" into positions $(2,4)$ and $(3,5)$ of the matrix. (The row and column numbers refer to the original matrix, not the reduced matrix.) The reader should verify that recursive application of this process to the remaining 5×3 submatrix predicts no additional fill. The final result of the symbolic Householder procedure is the filled matrix F in Figure 1. The lower trapezoidal part of F predicts the nonzero structure of W; the upper triangular part of F predicts the nonzero structure of R. For further details consult George and Ng [9].

We will now make the notion of "predicted nonzero structure" more precise. For any $p \times q$ matrix M let

$$\text{Struct}(M) := \{(i,j) : m_{ij} \neq 0\}.$$

Let $\bar{\mathcal{W}}$ and $\bar{\mathcal{R}}$ be the sets of nonzero positions predicted by the symbolic Householder procedure (as described in [9]) for W and R, respectively. The usefulness of the symbolic Householder procedure comes from the fact that $\text{Struct}(W) \subseteq \bar{\mathcal{W}}$ and $\text{Struct}(R) \subseteq \bar{\mathcal{R}}$, *no matter what values occupy the nonzero positions of A*. For any factor we will refer to a predicted set of nonzero positions with this property as an *upper bound* on the nonzero structure of the factor.

In §2 we extend $\bar{\mathcal{W}}$ in a fairly straightforward manner to obtain an upper bound $\bar{\mathcal{Q}}$ on $\text{Struct}(Q)$. This is done by forming the appropriate "symbolic product" of the Householder transformations, with the nonzero structure of each transformation based directly on $\bar{\mathcal{W}}$. George, Liu, and Ng [8] introduced an implicit representation of $\bar{\mathcal{W}}$ in terms of $\text{Struct}(A)$ and the *elimination forest* of $\bar{\mathcal{R}}$. In §2 we show that a simple extension of this implicit representation of $\bar{\mathcal{W}}$ results in an implicit representation of $\bar{\mathcal{Q}}$. To incorporate columns $n+1, n+2, \ldots, m$ into $\bar{\mathcal{Q}}$, we extend the elimination forest of $\bar{\mathcal{R}}$ to include nodes $n+1, n+2, \ldots, m$.

Let U be the matrix consisting of the first n columns of Q, and let $\bar{\mathcal{U}}$ be the upper bound on $\text{Struct}(U)$ consisting of the first n "columns" of $\bar{\mathcal{Q}}$. The following is an immediate and interesting consequence of the main result in §2: *the lower trapezoidal pattern $\bar{\mathcal{W}}$ and the lower trapezoidal part of $\bar{\mathcal{U}}$ are identical.*

For each of the matrices W, U, and R, there exists a unique upper bound on its nonzero structure which is the smallest possible bound based solely on $\text{Struct}(A)$.

Consider the Householder matrix W. Let Λ be the set of $m \times n$ full-rank matrices M for which Struct(M) = Struct(A), and let W_M be the Householder matrix of $M \in \Lambda$. The *least upper bound* on Struct(W) is given by $\cup_{M \in \Lambda}$ Struct(W_M). It should be readily apparent that this is the unique smallest upper bound on Struct(W) that can be obtained based solely on Struct(A).[1] We can obtain unique least upper bounds on Struct(U) and Struct(R) in a similar fashion.

In §§3 and 4 we are concerned with conditions under which the symbolic Householder procedure can be used to obtain least upper bounds on Struct(W) and Struct(U). It is well known that A can be reordered into *block upper triangular form* by row and column permutations [5, 18], where the diagonal blocks are square, except for the last, which may be rectangular. We will call A a *strong Hall* matrix if its block upper triangular form is trivial (i.e., has one block). We will call A a *weak Hall* matrix if its block upper triangular form is nontrivial (i.e., has more than one block). It is also well known that the reordering into block upper triangular form can be chosen so that the resulting matrix has no zero entries on its main diagonal [4, 5].

Coleman, Edenbrandt, and Gilbert [1] showed that $\bar{\mathcal{R}}$ is the least upper bound on Struct(R) whenever A is in block upper triangular form. (Their result is based on a symbolic Givens procedure, which also produces $\bar{\mathcal{R}}$ whenever A is in block upper triangular form.) Using results from Hare et al. [13], we show in §3 that if A is strong Hall and has no zero entry on its main diagonal, then $\bar{\mathcal{W}}$ and $\bar{\mathcal{U}}$ are least upper bounds on Struct(W) and Struct(U), respectively. In §4 we deal with the case where A is weak Hall, is in block upper triangular form, and has no zero entries on its main diagonal. We show that for such a matrix the least upper bounds on Struct(W) and Struct(U) have a block diagonal form, each block of which can be obtained by applying the symbolic Householder procedure to the corresponding strong Hall submatrix on the block main diagonal of A. We contend that from the practitioner's point of view, restricting our attention to matrices having this special form is both adequate and quite natural. To further strengthen our case, we show in §4 that one can always reorder the rows and columns of A into block upper triangular form so that there is no increase in the fill incurred by the QR factorization.

2. The nonzeros of Q predicted by symbolic Householder. A well-known algorithm for computing the orthogonal factorization in (1) uses a sequence of Householder transformations [11] to reduce the $m \times n$ matrix A to upper triangular form:

$$H^n H^{n-1} \cdots H^1 A = \begin{bmatrix} R \\ O \end{bmatrix}.$$

Since each Householder transformation H^k is symmetric and orthogonal, the orthogonal factor can then be expressed as the product

$$Q = H^1 H^2 \cdots H^n. \tag{2}$$

Each Householder transformation H^k has the form $I - w_k w_k^T$ for some m-vector w_k of the form

$$\begin{bmatrix} 0 \\ \alpha_k \\ x_k \end{bmatrix},$$

[1] We could let Λ contain all full-rank $m \times n$ matrices M for which Struct$(M) \subseteq$ Struct(A), since removing pairs from Struct(A) may remove pairs from, but will never add pairs to, the least upper bound on Struct(W).

where α_k is a nonzero scalar, x_k is an $(m-k)$-vector, and the first $k-1$ entries of w_k are zero.[2] The vector w_k is often called a *Householder vector*. The orthogonal factor Q can therefore be represented *implicitly* by the $m \times n$ lower trapezoidal matrix

$$W = \begin{bmatrix} w_1 & w_2 & \cdots & w_n \end{bmatrix},$$

commonly called the *Householder matrix*.

In this section we compute an upper bound $\bar{\mathcal{Q}}$ on Struct(Q) by forming a *symbolic product* based on (2). To make the notion of a symbolic product more precise we need the following notation. For any $p \times q$ matrix M, its ith row and jth column are denoted by M_{i*} and M_{*j}, respectively. Let $\bar{\mathcal{M}}$ be an upper bound on Struct(M), so that Struct$(M) \subseteq \bar{\mathcal{M}}$. "Row i" of $\bar{\mathcal{M}}$ is the set

$$\bar{\mathcal{M}}_{i*} := \{j : (i,j) \in \bar{\mathcal{M}}\};$$

similarly, "column j" of $\bar{\mathcal{M}}$ is the set

$$\bar{\mathcal{M}}_{*j} := \{i : (i,j) \in \bar{\mathcal{M}}\}.$$

Suppose that M is a $p \times q$ matrix and N is a $q \times r$ matrix, and let $\bar{\mathcal{M}}$ and $\bar{\mathcal{N}}$ be upper bounds on Struct(M) and Struct(N), respectively. We define the *symbolic product* of $\bar{\mathcal{M}}$ and $\bar{\mathcal{N}}$ as

$$\bar{\mathcal{M}}\bar{\mathcal{N}} := \{(i,j) : (i,k) \in \bar{\mathcal{M}} \text{ and } (k,j) \in \bar{\mathcal{N}} \text{ for some } k, 1 \le k \le q\}.$$

Observe that $\bar{\mathcal{M}}\bar{\mathcal{N}}$ is an upper bound on Struct(MN).

In this section we form a symbolic product based on (2), namely,

$$\bar{\mathcal{Q}} := \bar{\mathcal{H}}^1 \bar{\mathcal{H}}^2 \cdots \bar{\mathcal{H}}^n, \tag{3}$$

where $\bar{\mathcal{H}}^k$, $1 \le k \le n$, is the upper bound on Struct(H^k) obtained directly from column $\bar{\mathcal{W}}_{*k}$, that is,

$$\bar{\mathcal{H}}^k := \text{Struct}(I) \cup (\bar{\mathcal{W}}_{*k} \times \bar{\mathcal{W}}_{*k}). \tag{4}$$

It should be readily apparent that $\bar{\mathcal{Q}}$ is an upper bound on Struct(Q).

2.1. Background. An $m \times n$ matrix A, with $m \ge n$, has full *structural rank* n if there exists a full-rank $m \times n$ matrix M for which Struct(M) = Struct(A). The matrix A is a *Hall* matrix if every subset of k columns has nonzero entries in at least k rows. It is well known that A has structural rank n if and only if it is a Hall matrix [4, 6]; we therefore restrict our attention to Hall matrices in this paper.

For any $m \times n$ Hall matrix A, there exists an $m \times m$ permutation matrix P_m for which every diagonal entry of $P_m A$ is nonzero [4, 12, 16]. Since $P_m A = P_m QR$, it follows that reordering the rows of A merely reorders the rows of the orthogonal factor. George, Liu, and Ng [8, 9] apply the symbolic Householder procedure to Struct(A) only *after* the rows of A have been permuted to obtain a zero-free diagonal. One reason for this restriction is that a zero on the diagonal of A may create needless fill in $\bar{\mathcal{W}}$ or $\bar{\mathcal{R}}$ (or both). This problem can be avoided by incorporating "row pivoting" into the symbolic Householder procedure, which, however, makes the procedure more complicated. Further details on the desirability of a zero-free diagonal can be found

[2] It is convenient to include H^n in all cases, even though $w_n^T = \begin{bmatrix} 0 & \alpha_n \end{bmatrix}$ whenever $m = n$.

in George, Liu, and Ng [8, 9]. Henceforth we shall assume that A is an $m \times n$ Hall matrix ($m \geq n$) with a zero-free diagonal.

In §2.3 we introduce an implicit representation of $\bar{\mathcal{Q}}$ that is similar to the implicit representation of $\bar{\mathcal{W}}$ introduced in George, Liu, and Ng [8]. For $1 \leq k \leq n$, if $\bar{\mathcal{R}}_{k*} - \{k\} \neq \emptyset$, then we let

$$\rho(k) := \min\{j : j \in \bar{\mathcal{R}}_{k*} - \{k\}\};$$

otherwise, $\rho(k)$ will be undefined. It is easy to see that ρ is the *parent* function of a forest. We shall call this forest the *elimination forest of* $\bar{\mathcal{R}}$. (Liu [15] surveys elimination forests in detail.) The elimination forest of $\bar{\mathcal{R}}$ may consist of one or more trees. For each tree there is exactly one node r for which $\rho(r)$ is undefined, and it is called the *root* of the tree. Note that the nodes of the forest are numbered by a topological ordering, so that $k < \rho(k)$ for each nonroot k in the forest.

If A_{i*} is a nonzero row, then we let $f(i)$ be the column index of the first nonzero entry in that row, that is, we let

$$f(i) := \min\{j : a_{ij} \neq 0\};$$

otherwise, we let $f(i) = m + 1$.[3] Note that $f(i) \leq i$ whenever $1 \leq i \leq n$ because A has a zero-free diagonal. Furthermore, whenever $n + 1 \leq i \leq m$ we have $f(i) \leq n$ if and only if A_{i*} is a nonzero row. It is straightforward to show that if A_{i*} is a zero row (i.e., if $f(i) = m+1$), then W_{i*} is a zero row and $\bar{\mathcal{W}}_{i*} = \emptyset$. Theorem 2.1 identifies each nonempty $\bar{\mathcal{W}}_{i*}$ with a path in the elimination forest of $\bar{\mathcal{R}}$.

THEOREM 2.1 (see George, Liu, and Ng [8]). *Let A be an $m \times n$ Hall matrix ($m \geq n$) with a zero-free diagonal, and let ρ be the parent function of the elimination forest of $\bar{\mathcal{R}}$. Assume that A_{i*} is a nonzero row, and let $f(i)$ be the column index of the first nonzero entry in A_{i*}. Whenever $1 \leq i \leq n$, we have*

$$\bar{\mathcal{W}}_{i*} = \{\, f(i), \rho(f(i)), \rho(\rho(f(i))), \ldots, i \,\};$$

whenever $n + 1 \leq i \leq m$, we have

$$\bar{\mathcal{W}}_{i*} = \{\, f(i), \rho(f(i)), \rho(\rho(f(i))), \ldots, r \,\},$$

where r is a root in the elimination forest of $\bar{\mathcal{R}}$. □

2.2. Extending the elimination forest. The nodes in the elimination forest of $\bar{\mathcal{R}}$ are labeled from 1 to n. To include columns $n + 1, n + 2, \ldots, m$ in our implicit representation of $\bar{\mathcal{Q}}$, we need to add nodes $n+1, n+2, \ldots, m$ to the elimination forest. We first prove a lemma needed to ensure that the modifications result in a well-defined forest and that the modified forest indeed extends, rather than merely replaces, the original forest.

LEMMA 2.2. *Let A be an $m \times n$ Hall matrix ($m \geq n$) with a zero-free diagonal, and let r and s be two distinct roots in the elimination forest of $\bar{\mathcal{R}}$. It then follows that column $\bar{\mathcal{W}}_{*r}$ is disjoint from column $\bar{\mathcal{W}}_{*s}$. Moreover, for every root r in the elimination forest of $\bar{\mathcal{R}}$, we have $\bar{\mathcal{W}}_{*r} - \{r\} \subseteq \{n + 1, n + 2, \ldots, m\}$.*

Proof. Without loss of generality, assume that $r < s$. By way of contradiction, assume that $\bar{\mathcal{W}}_{*r} \cap \bar{\mathcal{W}}_{*s} \neq \emptyset$, and choose $i \in \bar{\mathcal{W}}_{*r} \cap \bar{\mathcal{W}}_{*s}$. We then have $r, s \in \bar{\mathcal{W}}_{i*}$; hence by Theorem 2.1 row $\bar{\mathcal{W}}_{i*}$ has the form

$$\{\, f(i), \rho(f(i)), \ldots, r, \rho(r), \ldots, s, \ldots, t \,\},$$

[3] The reason for this particular choice of a "null" value for $f(i)$ will become apparent in §2.3.

with $f(i) \leq r < \rho(r) \leq s \leq t$. Contrary to hypothesis, the node r clearly is not a root in the elimination forest of $\bar{\mathcal{R}}$. This contradiction suffices to prove the first part of the result.

To prove the second part, again let r be the root of a tree in the elimination forest of $\bar{\mathcal{R}}$. Suppose that column $\bar{\mathcal{W}}_{*r}$ is given by

$$\bar{\mathcal{W}}_{*r} = \{i_1, i_2, \ldots, i_p\}, \tag{5}$$

where $i_1 < i_2 < \cdots < i_p$ and $p \geq 1$. Clearly $i_1 = r$ and $i_p \leq m$. Furthermore, whenever $p \geq 2$, we have $i_2 > n$, for were it the case that $r = i_1 < i_2 \leq n$, the action of the rth row-merge step on the zero-free diagonal would place $i_2 > r$ in row $\bar{\mathcal{R}}_{r*}$, which is impossible since r is the root of a tree in the elimination forest of $\bar{\mathcal{R}}$. Consequently, every row index other than r in column $\bar{\mathcal{W}}_{*r}$ is taken from $\{n+1, n+2, \ldots, m\}$. □

We extend the elimination tree rooted at r by adding a chain from the old root $r = i_1 \in \bar{\mathcal{W}}_{*r}$ to the new root $i_p \in \bar{\mathcal{W}}_{*r}$ (see (5)). That is, the parent of the new root i_p will be undefined, and for $1 \leq s \leq p-1$ we define the parent of i_s to be i_{s+1}. Each tree in the elimination forest is modified in this manner. Lemma 2.2 ensures that these chains are disjoint, and the modified forest thus remains a forest. Note that node k $(n+1 \leq k \leq m)$ is not yet included in the modified forest if and only if A_{k*} is a zero row; whenever A_{k*} is zero the parent of k remains undefined, and this completes the *extended elimination forest of* $\bar{\mathcal{R}}$.

For convenience, the path in the (original or extended) elimination forest from a vertex k to the root of the tree to which it belongs will be called the *root path* from k.

2.3. Forming the symbolic product. In this section we will introduce an implicit representation of the symbolic product $\bar{\mathcal{Q}} := \bar{\mathcal{H}}^1 \bar{\mathcal{H}}^2 \cdots \bar{\mathcal{H}}^n$ in terms of the "first nonzero" indices $f(i)$, $1 \leq i \leq m$, and the extended elimination forest of $\bar{\mathcal{R}}$. Specifically, we will prove the following theorem.

THEOREM 2.3. *Let A be an $m \times n$ Hall matrix $(m \geq n)$ with a zero-free diagonal. If A_{i*} is a zero row, then $\bar{\mathcal{Q}}_{i*} = \{i\}$. If on the other hand A_{i*} is a nonzero row, then $\bar{\mathcal{Q}}_{i*}$ is the root path from $f(i)$ in the extended elimination forest of $\bar{\mathcal{R}}$.*

Let M be an $m \times p$ matrix, and let $\bar{\mathcal{M}}$ be an upper bound on Struct(M). Based on (4), the upper bound $\bar{\mathcal{B}} = \bar{\mathcal{H}}^k \bar{\mathcal{M}}$ on Struct$(H^k M)$ can be obtained by performing a row-merge step on $\bar{\mathcal{M}}$ that forms row $\bar{\mathcal{B}}_{i*}$ as follows: if $i \notin \bar{\mathcal{W}}_{*k}$ then $\bar{\mathcal{H}}^k_{i*} = \{i\}$, whence $\bar{\mathcal{B}}_{i*} = \bar{\mathcal{M}}_{i*}$; if on the other hand $i \in \bar{\mathcal{W}}_{*k}$ then $\bar{\mathcal{H}}^k_{i*} = \bar{\mathcal{W}}_{*k}$, whence

$$\bar{\mathcal{B}}_{i*} = \bigcup_{r \in \bar{\mathcal{W}}_{*k}} \bar{\mathcal{M}}_{r*} .$$

To prove Theorem 2.3 we will examine the sequence of symbolic products

$$\bar{\mathcal{Q}}^k = \bar{\mathcal{H}}^k \bar{\mathcal{H}}^{k+1} \cdots \bar{\mathcal{H}}^n, \qquad k = n, n-1, \ldots, 1.$$

This approach requires a sequence of "first nonzero" index sets, defined as follows. Given k, $1 \leq k \leq n$, choose i for which $k \leq i \leq m$. Whenever

$$\bar{\mathcal{W}}_{i*} \cap \{k, k+1, \ldots, n\} \neq \emptyset,$$

we write

$$f_k(i) := \min\{j : j \in \bar{\mathcal{W}}_{i*} \cap \{k, k+1, \ldots, n\}\};$$

otherwise, we write $f_k(i) := m+1$. (When $f_k(i) \leq n$, it identifies the "first Householder vector" $\bar{\mathcal{W}}_{*j}$ used in forming $\bar{\mathcal{Q}}^k$ that has a nonzero entry in the ith position.) The following theorem expresses each symbolic product $\bar{\mathcal{Q}}^k$, $1 \leq k \leq n$, in terms of the first nonzero indices $f_k(i)$ and the extended elimination forest of $\bar{\mathcal{R}}$.

THEOREM 2.4. *Let A be an $m \times n$ Hall matrix ($m \geq n$) with a zero-free diagonal. Choose indices k and i for which $1 \leq k \leq n$ and $k \leq i \leq m$. If $f_k(i) = m+1$, then $\bar{\mathcal{Q}}^k_{i*} = \{i\}$. If on the other hand $f_k(i) \leq n$, then $\bar{\mathcal{Q}}^k_{i*}$ is the root path from $f_k(i)$ in the extended elimination forest of $\bar{\mathcal{R}}$.*

Proof. We prove the result by induction on k, where $k = n, n-1, \ldots, 1$. For the base step $k = n$ we have

$$\bar{\mathcal{Q}}^n = \bar{\mathcal{H}}^n = \mathrm{Struct}(I) \cup (\bar{\mathcal{W}}_{*n} \times \bar{\mathcal{W}}_{*n}).$$

Choose a row $\bar{\mathcal{Q}}^n_{i*}$ where $n \leq i \leq m$. There are two cases to consider: $i \notin \bar{\mathcal{W}}_{*n}$ and $i \in \bar{\mathcal{W}}_{*n}$. If $i \notin \bar{\mathcal{W}}_{*n}$, clearly $f_n(i) = m+1$ and $\bar{\mathcal{Q}}^n_{i*} = \{i\}$, giving us the result in this case. If $i \in \bar{\mathcal{W}}_{*n}$, then $f_n(i) = n$ and $\bar{\mathcal{Q}}^n_{i*} = \bar{\mathcal{W}}_{*n}$. Since n is necessarily a root in the elimination forest of $\bar{\mathcal{R}}$, it follows from the definition of the extended forest that $\bar{\mathcal{Q}}^n_{i*}$ is a root path from $f_n(i)$ in the extended elimination forest of $\bar{\mathcal{R}}$. This completes the argument for the base case.

Before proceeding with the induction step, note that for each pair of indices i and k such that $1 \leq i < k \leq n$, we have $\bar{\mathcal{H}}^k_{i*} = \bar{\mathcal{H}}^k_{*i} = \{i\}$. We leave it for the reader to verify that as a consequence of the preceding observation we have $\bar{\mathcal{Q}}^k_{i*} = \bar{\mathcal{Q}}^k_{*i} = \{i\}$ for $1 \leq i < k \leq n$. We thus restrict our attention to $\bar{\mathcal{Q}}^k \cap \{k, \ldots, m\} \times \{k, \ldots, m\}$.

Assume that the result holds for $\bar{\mathcal{Q}}^{k+1}$ and consider the symbolic product $\bar{\mathcal{Q}}^k = \bar{\mathcal{H}}_k \bar{\mathcal{Q}}^{k+1}$. Choose a row $\bar{\mathcal{Q}}^k_{i*}$ where $k \leq i \leq m$. There are two cases to consider: $i \notin \bar{\mathcal{W}}_{*k}$ and $i \in \bar{\mathcal{W}}_{*k}$. Assume that $i \notin \bar{\mathcal{W}}_{*k}$. Since $i \notin \bar{\mathcal{W}}_{*k}$, applying the appropriate row-merge step to $\bar{\mathcal{Q}}^{k+1}$ to obtain $\bar{\mathcal{Q}}^k$ gives us $\bar{\mathcal{Q}}^k_{i*} = \bar{\mathcal{Q}}^{k+1}_{i*}$. There are two subcases to consider: $f_{k+1}(i) = m+1$ and $f_{k+1}(i) \leq n$. Let $f_{k+1}(i) = m+1$. Since $i \notin \bar{\mathcal{W}}_{*k}$, it follows that $f_k(i) = m+1$ too. From the induction hypothesis, we have $\bar{\mathcal{Q}}^k_{i*} = \bar{\mathcal{Q}}^{k+1}_{i*} = \{i\}$, which completes the proof for this subcase. Now let $f_{k+1}(i) \leq n$. Since $i \notin \bar{\mathcal{W}}_{*k}$, it follows that $f_k(i) = f_{k+1}(i)$. Since $\bar{\mathcal{Q}}^k_{i*} = \bar{\mathcal{Q}}^{k+1}_{i*}$, it follows from the induction hypothesis that the result holds for this subcase.

Now we consider the more interesting case where $i \in \bar{\mathcal{W}}_{*k}$. There are three subcases to consider:

1. $i > k$ and $f_{k+1}(i) = m+1$,
2. $i > k$ and $f_{k+1}(i) \leq n$, and finally
3. $i = k$.

Suppose that $i > k$ and $f_{k+1}(i) = m+1$. Note that $f_k(i) = k$; moreover, from the induction hypothesis, we have $\bar{\mathcal{Q}}^{k+1}_{i*} = \{i\}$. The fact that $f_{k+1}(i) = m+1$ and $f_k(i) = k$ implies that k is maximum among the members of $\bar{\mathcal{W}}_{i*}$. It follows from Theorem 2.1 that $k = i$ or k is a root in the elimination forest of $\bar{\mathcal{R}}$, and since $i > k$, the latter must hold true. Choose $\ell \in \bar{\mathcal{W}}_{*k} - \{k\}$. From Lemma 2.2 it follows that $n+1 \leq \ell \leq m$. Now, the forest path in Theorem 2.1 which characterizes $\bar{\mathcal{W}}_{\ell *}$ must terminate at a root, and since $k \in \bar{\mathcal{W}}_{\ell *}$, that root must be k. It follows that $f_{k+1}(\ell) = m+1$, and consequently by the induction hypothesis, $\bar{\mathcal{Q}}^{k+1}_{\ell *} = \{\ell\}$. Recall moreover that $\bar{\mathcal{Q}}^{k+1}_{k*} = \{k\}$. Applying the appropriate row-merge step to $\bar{\mathcal{Q}}^{k+1}$ to obtain $\bar{\mathcal{Q}}^k$ thus gives us $\bar{\mathcal{Q}}^k_{i*} = \bar{\mathcal{W}}_{*k}$. By the definition of the extended forest, $\bar{\mathcal{Q}}^k_{i*}$ is a root path from $f_k(i)$. This completes the proof for this subcase.

Now suppose that $i > k$ and $f_{k+1}(i) \leq n$. It follows from the induction hypothesis that $\bar{\mathcal{Q}}^{k+1}_{i*}$ is the root path from $f_{k+1}(i)$. Note that $k = f_k(i) < f_{k+1}(i) \leq n$.

Consequently, from Theorem 2.1 and the definition of $f_{k+1}(i)$ it follows that $f_{k+1}(i) = \rho(k)$. The key observation is that $f_{k+1}(i)$ is independent of i; we thus have $\bar{\mathcal{Q}}_{\ell *}^{k+1} = \bar{\mathcal{Q}}_{i*}^{k+1}$ for every $\ell \in \bar{\mathcal{W}}_{*k} - \{k\}$. Recall moreover that $\bar{\mathcal{Q}}_{k*}^{k+1} = \{k\}$. Applying the appropriate row-merge step to $\bar{\mathcal{Q}}^{k+1}$ to obtain $\bar{\mathcal{Q}}^{k}$ establishes that $\bar{\mathcal{Q}}_{i*}^{k}$ is the root path from $f_k(i) = k$ in the extended forest, which proves the result for this case.

Finally, assume that $i = k$. We know that $k \in \bar{\mathcal{W}}_{*k}$, $\bar{\mathcal{Q}}_{k*}^{k+1} = \{k\}$, and $f_k(k) = k$. From the argument in the preceding paragraph, applying the appropriate row-merge step to $\bar{\mathcal{Q}}^{k+1}$ to obtain $\bar{\mathcal{Q}}^{k}$ establishes that $\bar{\mathcal{Q}}_{k*}^{k}$ is the root path from $f_k(k) = k$ in the extended forest, which completes the proof. □

Since $\bar{\mathcal{Q}}^1 = \bar{\mathcal{Q}}$, Theorem 2.3 follows from Theorem 2.4: the case of a zero row A_{i*} in Theorem 2.3 is covered by the case $f_1(i) = m + 1$ in Theorem 2.4; the case of a nonzero row A_{i*} in Theorem 2.3 is covered by the case $f_1(i) \leq n$ in Theorem 2.4. Furthermore, whenever $f_1(i) \leq n$ we have $f_1(i) = f(i)$, as required.

Recall that the matrix U consists of the first n columns of Q. The first part of the next theorem describes the upper bound $\bar{\mathcal{U}}$ on Struct(U) given by Theorem 2.3; the second part states a relationship between $\bar{\mathcal{W}}$ and $\bar{\mathcal{U}}$ which follows directly from Theorems 2.1 and 2.3.

COROLLARY 2.5. *Let A be an $m \times n$ Hall matrix ($m \geq n$) with a zero-free diagonal, and let ρ be the parent function of the elimination forest of $\bar{\mathcal{R}}$.*

(a) *If A_{i*} is a zero row, then $\bar{\mathcal{U}}_{i*} = \emptyset$. If A_{i*} is a nonzero row, then $\bar{\mathcal{U}}_{i*}$ is the root path of $f(i)$ in the elimination forest of $\bar{\mathcal{R}}$.*

(b) *The lower trapezoidal pattern $\bar{\mathcal{W}}$ and the lower trapezoidal part of $\bar{\mathcal{U}}$ are identical, that is,*

$$\bar{\mathcal{W}} = \bar{\mathcal{U}} - \{(i,j) : 1 \leq i \leq n \text{ and } i+1 \leq j \leq n\}. \qquad \square$$

3. Least upper bounds: The strong Hall case. The bounds $\bar{\mathcal{W}}$, $\bar{\mathcal{U}}$, and $\bar{\mathcal{R}}$ may not be least upper bounds on Struct(W), Struct(U), and Struct(R), respectively. Coleman, Edenbrandt, and Gilbert [1] showed that $\bar{\mathcal{R}}$ is the least upper bound on Struct(R) whenever A is either (1) strong Hall or (2) weak Hall, but in block upper triangular form. (Strong Hall and weak Hall matrices are defined below.) In this section we prove analogous results for $\bar{\mathcal{W}}$ and $\bar{\mathcal{U}}$ in the case where A is strong Hall. We will look at the weak Hall case in §4.

The following defines a strong Hall matrix: an $m \times n$ Hall matrix A with $m \geq n$ is *strong Hall* if every $m \times k$ submatrix, $1 \leq k < n$, has at least $k + 1$ nonzero rows. It is well known that strong Hall matrices are precisely those that have trivial block upper triangular form. (A Hall matrix that does not satisfy the strong Hall property is called *weak Hall*.) In §3.1 we give a result due to Hare et al. [13], which describes the least upper bound on Struct(U) for *any Hall matrix* A. In §3.2 we use their result to show that if A is strong Hall and has a zero-free diagonal, then $\bar{\mathcal{W}}$ and $\bar{\mathcal{U}}$ are least upper bounds on Struct(W) and Struct(U), respectively.

3.1. A previous sparsity analysis for Hall matrices. Let $\mathcal{U}_{\text{lub}}$ denote the least upper bound on Struct(U). To determine $\mathcal{U}_{\text{lub}}$, Hare et al. [13] found it most useful to consider QR factorization obtained via the Gram–Schmidt procedure. As a natural consequence they examined $\mathcal{U}_{\text{lub}}$ one column at a time, starting with the first and ending with the last.

A key concept used in their analysis is the notion of a *Hall set*. Let A be an $m \times n$ Hall matrix ($m \geq n$). A *Hall set* of size p is a set of p columns from A such that the $m \times p$ matrix formed by these columns has *exactly* p nonzero rows [13]. It is easy to show that the union of two distinct Hall sets is also a Hall set. It follows that every set

of columns from A has a unique Hall set (possibly empty) of maximum cardinality. Let C_k^H be the maximum cardinality Hall set associated with the submatrix consisting of the first k columns of A, where $1 \le k \le n$; let R_k^H be the set of nonzero rows associated with C_k^H. The two sets C_k^H and R_k^H will be referred to as a set of *Hall columns* and a set of *Hall rows*, respectively.[4] Henceforth we will use $A[k]$ to denote the submatrix containing the first k columns of A.

Hare et al. associate with each submatrix $A[k]$ a bipartite graph $B_k = (X_k, Y_k, E_k)$ that describes the nonzero structure of $A[k]$ *with the Hall rows* R_{k-1}^H *and Hall columns* C_{k-1}^H *removed.* More specifically, the graph B_k is defined as follows:

$$
\begin{aligned}
Y_k &:= \{j : 1 \le j \le k \text{ and } j \notin C_{k-1}^H\},\\
X_k &:= \{i : 1 \le i \le m, i \notin R_{k-1}^H, \text{ and } \exists\, j \in Y_k \text{ such that } a_{ij} \neq 0\},\\
E_k &:= \big\{\{i,j\} : i \in X_k, j \in Y_k, \text{ and } a_{ij} \neq 0\big\}.
\end{aligned}
$$

Now, consider the set of row indices $F_k := \{1, 2, \ldots, m\} - X_k - R_{k-1}^H$. It follows that $i \in F_k$ if and only if row i of $A[k]$ is zero. The sets F_k, R_{k-1}^H, and X_k clearly partition the set of row indices $\{1, 2, \ldots, m\}$ into three sets. We further partition the vertex set X_k into two sets, as follows. Let P_k be the subset of X_k that lies in the connected component of B_k to which column vertex k belongs, and let $D_k = X_k - P_k$. That is, D_k contains the row vertices in X_k that are *disconnected* from the last column vertex k, while P_k contains the row vertices in X_k that are connected by a *path* to the column vertex k.

The following result describes how this four-way partition characterizes column k of $\mathcal{U}_{\text{lub}}$. It is worth noting that possession of a zero-free diagonal plays no role in the result.

THEOREM 3.1 (see Hare et al. [13]). *Let A be an $m \times n$ Hall matrix with $m \ge n$. For $1 \le k \le n$, let R_{k-1}^H be the set of Hall rows associated with the first $k-1$ columns of A, and let F_k, D_k, and P_k be defined as above. We then have $(i,k) \in \mathcal{U}_{\text{lub}}$ if and only if $i \in P_k$; equivalently, $(i,k) \notin \mathcal{U}_{\text{lub}}$ if and only if $i \in F_k \cup D_k \cup R_{k-1}^H$.* □

Theorem 3.1 places each row index i for which $(i,k) \notin \mathcal{U}_{\text{lub}}$ into one of three categories. First, observe that if $i \in F_k$, then $a_{ij} = 0$ for $1 \le j \le k$. The Gram–Schmidt procedure ensures that column U_{*k} is a linear combination of the columns of $A[k]$; hence $u_{ik} = 0$ for $i \in F_k$, as the theorem asserts.

Second, consider $i \in R_{k-1}^H$. The Gram–Schmidt procedure also ensures that column U_{*k} must be orthogonal to every column of $A[k-1]$; in particular, it is orthogonal to the columns in C_{k-1}^H. Now, the Hall columns in C_{k-1}^H span a subspace of dimension $|C_{k-1}^H| = |R_{k-1}^H|$, and for every vector x in this space we have $x_i = 0$ for $i \notin R_{k-1}^H$. In consequence, any vector y such that $y_i \neq 0$ for some $i \in R_{k-1}^H$ cannot be orthogonal to every vector in this space. Thus, $u_{ik} = 0$ for $i \in R_{k-1}^H$, as the theorem states. This is perhaps the key insight in Hare et al. [13].

Finally, for $i \in D_k$ we will not argue that $u_{ik} = 0$, as we did for the previous two cases. The argument is longer and more technical, and we thus refer the reader to Hare et al. [13] for these details. However, we do prove that $u_{ik} = 0$ for $i \in D_k$ under the restrictions imposed on A in the next subsection.

To complete the proof of Theorem 3.1, Hare et al. showed that for each $i \in P_k$, there exists an assignment of values to the nonzero positions of A such that $u_{ik} \neq 0$, and consequently $(i,k) \in \mathcal{U}_{\text{lub}}$ if and only if $i \in P_k$. Pothen [17] later proved that

[4] It is sometimes convenient to treat R_k^H and C_k^H as sets of row and column indices, respectively.

there exists an assignment of values to the nonzero positions of A such that $u_{ij} \neq 0$ for every $(i,j) \in \mathcal{U}_{\text{lub}}$.

Hare et al. [13] also describe $\mathcal{R}_{\text{lub}}$, the least upper bound on Struct(R).

THEOREM 3.2 (see Hare et al. [13]). *For any $m \times n$ Hall matrix A with $m \geq n$, we have $\mathcal{R}_{\text{lub}} = \mathcal{U}_{\text{lub}}^T \mathcal{A}$, where $\mathcal{A} = \text{Struct}(A)$ and $\mathcal{U}_{\text{lub}}^T = \{(j,i) : (i,j) \in \mathcal{U}_{\text{lub}}\}$.* □

3.2. Least upper bounds from symbolic Householder. Let A be an $m \times n$ strong Hall matrix ($m \geq n$) with a zero-free diagonal. To prove that $\bar{\mathcal{U}}$ is the least upper bound on Struct(U), it suffices to show that $\bar{\mathcal{U}} \subseteq \mathcal{U}_{\text{lub}}$. Toward that end, choose $(i,k) \notin \mathcal{U}_{\text{lub}}$, so that by Theorem 3.1, $i \in F_k \cup D_k \cup R_{k-1}^H$. Since A is a strong Hall matrix, it follows that $R_{k-1}^H = C_{k-1}^H = \emptyset$. Consequently, the following two results suffice to show that $(i,k) \notin \bar{\mathcal{U}}$.

LEMMA 3.3. *Let A be an $m \times n$ strong Hall matrix ($m \geq n$) with a zero-free diagonal. Suppose $(i,k) \notin \mathcal{U}_{\text{lub}}$ and let F_k be as defined in* §3.1. *If $i \in F_k$, then $(i,k) \notin \bar{\mathcal{U}}$.*

Proof. Suppose that $i \in F_k$. It then follows that $a_{ij} = 0$ for $1 \leq j \leq k$. Consequently, $k < f(i)$, which by Corollary 2.5 ensures that $(i,k) \notin \bar{\mathcal{U}}$. □

LEMMA 3.4. *Let A be an $m \times n$ strong Hall matrix ($m \geq n$) with a zero-free diagonal. Suppose $(i,k) \notin \mathcal{U}_{\text{lub}}$ and let D_k be as defined in* §3.1. *If $i \in D_k$, then $(i,k) \notin \bar{\mathcal{U}}$.*

Proof. Suppose that $i \in D_k$. It follows that $i \in X_k$, and thus we have $f(i) \leq k$; moreover, since i is disconnected in B_k from $k \in Y_k$, it follows that $f(i) \neq k$, whence $f(i) < k$. To show that $(i,k) \notin \bar{\mathcal{U}}$, it is sufficient, according to Corollary 2.5, to show that k is not an ancestor of $f(i)$ in the elimination forest of $\bar{\mathcal{R}}$.

Consider the symmetric positive definite matrix $M = A^T A$ and its Cholesky factor L. Let $\bar{\mathcal{L}}$ denote the upper bound on Struct(L) computed by the symbolic Cholesky procedure (see George and Liu [7]). Coleman, Edenbrandt, and Gilbert [1] showed that if A is strong Hall, then $\bar{\mathcal{R}} = \bar{\mathcal{L}}^T$, where $\bar{\mathcal{L}}^T = \{(j,i) : (i,j) \in \bar{\mathcal{L}}\}$. It follows that the elimination forest of $\bar{\mathcal{R}}$ and the elimination forest of $\bar{\mathcal{L}}$ are identical. Let $G(M) = (Y_n, E')$ be the adjacency graph of M, i.e., the graph for which there is an edge joining $s,t \in Y_n$ if and only if $m_{st} \neq 0$.[5] Liu has shown [14, Lem. 2.3] that for $s < t$, vertex t is an ancestor of s in the elimination forest of $\bar{\mathcal{L}}$ if and only if they are connected by a path in the subgraph of $G(M)$ induced by $Y_t = \{1, 2, \ldots, t\}$.

Now, membership of i in D_k implies that there exists no path in B_k from $k \in Y_k$ to $i \in X_k$. Since $\{f(i), i\} \in E_k$, there is also no path in B_k from $k \in Y_k$ to $f(i) \in Y_k$. Thus, to prove the result it suffices to show that the absence of a path in B_k from k to $f(i)$ implies the absence of a path from k to $f(i)$ in the subgraph of $G(M)$ induced by Y_k.

Toward that end, suppose that there is a path

$$(f(i), j_1, j_2, \ldots, j_\tau, k)$$

in $G(M)$ such that $j_t < k$ for $1 \leq t \leq \tau$. (Recall that $f(i) < k$, as well.) It is trivial to verify that $G(M)$ is the graph on Y_n with edge set E' consisting of precisely the edges necessary to make each vertex set $\{j : a_{rj} \neq 0\}$, $1 \leq r \leq m$, a clique in the graph (i.e., pairwise adjacent in the graph). Consequently, if $\{j, j'\} \in E'$ with $j < j' \leq k$, then there exists some $i \in X_k$ for which $a_{ij} \neq 0$ and $a_{ij'} \neq 0$, and therefore (j, i, j') is

[5] Here, since A is strong Hall, and thus $C_n^H = \emptyset$, we can use the same vertex set Y_n in both graphs $G(M)$ and B_n.

a path in B_k. It follows that there exists a path

$$(f(i), i_1, j_1, i_2, j_2, \ldots, i_\tau, j_\tau, i_{\tau+1}, k)$$

in B_k such that $j_t \in Y_k$ and $j_t < k$ for $1 \leq t \leq \tau$ and $i_t \in X_k$ for $1 \leq t \leq \tau + 1$. This concludes the proof. □

With these two results and the discussion preceding them we have proven the following result.

THEOREM 3.5. *If A is an $m \times n$ strong Hall matrix ($m \geq n$) with a zero-free diagonal, then $\bar{\mathcal{U}}$ is the least upper bound on* Struct(U). □

Having shown that $\bar{\mathcal{U}} = \mathcal{U}_{\text{lub}}$ in the strong Hall case, we can now compare the characterization of this set given in Theorem 3.1 with that given in part (a) of Corollary 2.5. The following result shows how the classification of zero entries in Theorem 3.1 can be expressed in terms of the first nonzero indices $f(i)$ and the elimination forest of $\bar{\mathcal{R}}$.

COROLLARY 3.6. *Suppose that A is an $m \times n$ strong Hall matrix ($m \geq n$) with a zero-free diagonal, and let F_k and D_k be as defined in* §3.1. *Moreover, let $f(i)$ be the column index of the first nonzero in row i of A. The following statements then hold true:*

1. *$i \in F_k$ if and only if $k < f(i)$.*
2. *$i \in D_k$ if and only if $f(i) < k$ and k is not an ancestor of $f(i)$ in the elimination forest of $\bar{\mathcal{R}}$.*

Proof. The result follows immediately from Theorem 3.5 and the proofs of Lemmas 3.3 and 3.4. □

Finally, we show that $\bar{\mathcal{W}}$ is the least upper bound on Struct(W) whenever A is a strong Hall matrix with a zero-free diagonal.

COROLLARY 3.7. *If A is an $m \times n$ strong Hall matrix ($m \geq n$) with a zero-free diagonal, then $\bar{\mathcal{W}}$ is the least upper bound on* Struct(W).

Proof. Let $\mathcal{W}_{\text{lub}}$ be the least upper bound on Struct(W). As in (3), we can obtain an upper bound $\mathcal{Q}^*$ on Struct(Q) by forming a symbolic product of upper bounds on the nonzero structures of Householder transformations—upper bounds obtained directly from $\mathcal{W}_{\text{lub}}$ rather than $\bar{\mathcal{W}}$. Let $\mathcal{U}^*$ comprise the first n columns of $\mathcal{Q}^*$. For the same reasons that the lower trapezoidal pattern $\bar{\mathcal{W}}$ matches the lower trapezoidal part of $\bar{\mathcal{U}}$ (part (b) of Corollary 2.5), $\mathcal{W}_{\text{lub}}$ matches the lower trapezoidal part of $\mathcal{U}^*$.

Since $\mathcal{W}_{\text{lub}} \subseteq \bar{\mathcal{W}}$ it suffices to show that $\bar{\mathcal{W}} \subseteq \mathcal{W}_{\text{lub}}$. By way of contradiction, assume that $(i,j) \in \bar{\mathcal{W}} - \mathcal{W}_{\text{lub}}$. Since $\bar{\mathcal{W}}$ matches the lower trapezoidal part of $\bar{\mathcal{U}}$ and $\mathcal{W}_{\text{lub}}$ matches the lower trapezoidal part of $\mathcal{U}^*$, we have $(i,j) \in \bar{\mathcal{U}} - \mathcal{U}^*$. By Theorem 3.5, $\bar{\mathcal{U}}$ is the least upper bound on Struct(U), whence $\bar{\mathcal{U}} \subseteq \mathcal{U}^*$, contrary to $\bar{\mathcal{U}} - \mathcal{U}^* \neq \emptyset$. From this contradiction we thus have $\bar{\mathcal{W}} \subseteq \mathcal{W}_{\text{lub}}$, which gives us the result. □

4. Least upper bounds: The weak Hall case. Let A be an $m \times n$ Hall matrix with $m \geq n$ and assume that A is a weak Hall matrix, so that it has a nontrivial block upper triangular form (i.e., more than one block) [18]. Weak Hall matrices are precisely the matrices for which nonempty Hall sets play a key role in the sparsity analysis of their QR factorizations. The impact of Hall sets on Struct(Q) (and hence Struct(U)) can be described in very simple terms whenever A is in block upper triangular form.

In §4.1 we show that for any weak Hall matrix that has a zero-free diagonal and is in block upper triangular form, the least upper bounds on Struct(W) and Struct(U) have a block diagonal form, each block of which can be obtained by applying

the symbolic Householder procedure to the corresponding strong Hall diagonal block (i.e., submatrix) in A. It is natural then to consider how reordering into block upper triangular form influences the fill incurred by the QR factorization. In §4.2 we show that one can always reorder the rows and columns of A into block upper triangular form so that there is no increase in the fill incurred by the QR factorization.

4.1. Least upper bounds from symbolic Householder. Assume that the $m \times n$ weak Hall matrix A $(m \geq n)$ is in block upper triangular form

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1p} \\ O & A_{22} & \cdots & A_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & A_{pp} \end{bmatrix}, \tag{6}$$

where $p \geq 2$, and for $1 \leq k \leq p-1$ the submatrix A_{kk} is an $n_k \times n_k$ matrix that has the strong Hall property. The submatrix A_{pp} is an $(n_p + m - n) \times n_p$ matrix that also has the strong Hall property. (Whenever A is square, A_{pp} is square; whenever A is strictly rectangular, A_{pp} is strictly rectangular.) Again we assume that A has a zero-free diagonal.

For each k, $1 \leq k \leq p$, let the QR factorization of the strong Hall submatrix A_{kk} be given by

$$A_{kk} = Q_{kk} R_{kk}.$$

(Note that R_{pp} may be upper trapezoidal.) Let $U_{kk} = Q_{kk}$ for each k, where $1 \leq k \leq p-1$, and let U_{pp} comprise the first n_p columns of Q_{pp}. Consider A (in (6)) and the block diagonal matrix

$$D = \begin{bmatrix} A_{11} & O & \cdots & O \\ O & A_{22} & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & A_{pp} \end{bmatrix}.$$

The QR factorization of A is given by

$$A = QR = \begin{bmatrix} Q_{11} & O & \cdots & O \\ O & Q_{22} & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & Q_{pp} \end{bmatrix} \begin{bmatrix} R_{11} & Q_{11}^T A_{12} & \cdots & Q_{11}^T A_{1p} \\ O & R_{22} & \cdots & Q_{22}^T A_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & R_{pp} \end{bmatrix},$$

and the QR factorization of D is given by

$$D = QR_D = \begin{bmatrix} Q_{11} & O & \cdots & O \\ O & Q_{22} & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & Q_{pp} \end{bmatrix} \begin{bmatrix} R_{11} & O & \cdots & O \\ O & R_{22} & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & R_{pp} \end{bmatrix}.$$

The key observation is that the block diagonal matrix Q is the orthogonal factor of both A and D.

Let W_{kk} be the Householder matrix associated with A_{kk}. We leave it for the reader to verify that the matrix

$$W = \begin{bmatrix} W_{11} & O & \cdots & O \\ O & W_{22} & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & W_{pp} \end{bmatrix}$$

is the Householder matrix associated with both A and D.

Now we will use the preceding observations to describe least upper bounds on Struct(U) and Struct(W), which are furthermore obtained via the symbolic Householder procedure. Consider the upper bound $\bar{\mathcal{W}}_{kk}$ on Struct(W_{kk}) obtained by applying the symbolic Householder procedure to Struct(A_{kk}); let $\bar{\mathcal{Q}}_{kk}$ ($\bar{\mathcal{U}}_{kk}$) be the upper bound on Struct(Q_{kk}) (Struct(U_{kk})) obtained by forming the appropriate symbolic product based on $\bar{\mathcal{W}}_{kk}$. Taking advantage of the block diagonal form of Q, we can obtain an upper bound on Struct(Q) by using $\bar{\mathcal{Q}}_{kk}$ as an upper bound on Struct(Q_{kk}) for each k, $1 \le k \le p$. Since each submatrix A_{kk} is strong Hall, by Theorem 3.5 we can obtain the least upper bound on Struct(U) by using $\bar{\mathcal{U}}_{kk}$ as an upper bound on Struct(U_{kk}) for each k, $1 \le k \le p$, and by Corollary 3.7 we obtain the least upper bound on Struct(W) by using $\bar{\mathcal{W}}_{kk}$ as an upper bound on Struct(W_{kk}) for each k, $1 \le k \le p$.

It is worth noting that the block diagonal form of Q and U can also be argued from Theorem 3.1, the second paragraph after Theorem 3.1, and the following lemma. (We leave the details to the reader.)

LEMMA 4.1. *Let A be in block upper triangular form as in* (6), *and let column A_{*j} be a column in the kth block column of A, where $1 \le k \le p$. If $k = 1$, then $C^H_{j-1} = R^H_{j-1} = \emptyset$. If $2 \le k \le p$, then $C^H_{j-1} = R^H_{j-1} = \{1, 2, \ldots, \ell - 1\}$, where $A_{*\ell}$ is the first column in the kth block column of A.*

Proof. Let $k = 1$. Since A_{11} is strong Hall, any set of columns $S \subseteq A[j-1]$ has nonzero entries in at least $|S| + 1$ rows in A_{11}, whence $C^H_{j-1} = R^H_{j-1} = \emptyset$.

Now consider k where $2 \le k \le p$, and let $A_{*\ell}$ be the first column in the kth block column in A. Due to the block upper triangular form of A, the set $S = \{1, 2, \ldots, \ell - 1\}$ is clearly a Hall set contained in $A[j-1]$. Since any Hall set contained in $A[j-1]$ is also contained in its largest Hall set C^H_{j-1}, it follows that $S \subseteq C^H_{j-1}$. Let T be a nonempty subset of $\{\ell, \ell + 1, \ldots, j - 1\}$. The columns in S have nonzero entries in rows $1, 2, \ldots, |S|$ of A. Because A_{kk} is strong Hall, the columns in T have nonzero entries in at least $|T| + 1$ rows in A_{kk}. Consequently, the columns in $|S \cup T|$ have nonzero entries in at least $|S \cup T| + 1$ rows in A. It follows that $C^H_{j-1} = S$. Due to the block upper triangular form of A, it is clear that $R^H_{j-1} = C^H_{j-1} = S$. This completes the proof. □

4.2. Obtaining a block upper triangular form that limits fill. In this subsection we look at how reordering a weak Hall matrix into block upper triangular form influences the number of nonzero entries in the triangular and orthogonal factors. Assume that A is an $m \times n$ weak Hall matrix ($m \ge n$) with a zero-free diagonal. Let $\hat{A}$ be the same matrix after its rows and columns have been reordered so that it has a zero-free diagonal and is in block upper triangular form. The matrix $\hat{A}$ will have the

following form:

$$\hat{A} = \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} & \cdots & \hat{A}_{1p} \\ O & \hat{A}_{22} & \cdots & \hat{A}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & \hat{A}_{pp} \end{bmatrix}, \tag{7}$$

where $p \geq 2$, and $\hat{A}_{kk}$ is square whenever $1 \leq k \leq p-1$. Let $\hat{Q}$ and $\hat{R}$, respectively, be the orthogonal and triangular factors of $\hat{A}$, and let $\hat{U}$ comprise the first n columns of $\hat{Q}$. In this subsection we show that if the column order in $\hat{A}$ is "consistent" with the column order in A, then

1. $|\hat{\mathcal{U}}_{\text{lub}}| \leq |\mathcal{U}_{\text{lub}}|$, where $\hat{\mathcal{U}}_{\text{lub}}$ is the least upper bound on Struct$(\hat{U})$ and $\mathcal{U}_{\text{lub}}$ is the least upper bound on Struct(U), and
2. $|\hat{\mathcal{R}}_{\text{lub}}| \leq |\mathcal{R}_{\text{lub}}|$, where $\hat{\mathcal{R}}_{\text{lub}}$ is the least upper bound on Struct$(\hat{R})$ and $\mathcal{R}_{\text{lub}}$ is the least upper bound on Struct(R).

We now define what we mean by the term "consistent." Let $\alpha : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ be the permutation that maps the position of each column in A to its new position in $\hat{A}$, and likewise let $\beta : \{1, 2, \ldots, m\} \to \{1, 2, \ldots, m\}$ be the permutation that maps the position of each row in A to its new position in $\hat{A}$. The column order in $\hat{A}$ is said to be *consistent* with the column order in A if the individual columns within a block column of $\hat{A}$ occur in the same order in which they are found in A. For example, for the first block column in $\hat{A}$ we must have

$$\alpha^{-1}(1) < \alpha^{-1}(2) < \alpha^{-1}(3) < \cdots < \alpha^{-1}(n_1),$$

where n_1 is the number of columns in the first block column of $\hat{A}$.

Our first goal is to prove that $|\hat{\mathcal{U}}_{\text{lub}}| \leq |\mathcal{U}_{\text{lub}}|$. We will use the following relationship between Hall sets and block upper triangular form in our proof.

LEMMA 4.2. *Let $\hat{A}$ be in block upper triangular form as shown in* (7)*, and let S be a set of Hall columns in $\hat{A}$. If S contains a column in the kth block column of $\hat{A}$, then S contains every column in the kth block column of $\hat{A}$.*

Proof. For a set of Hall columns S in $\hat{A}$, let S_k, $1 \leq k \leq p$, contain the columns of S taken from the kth block column of $\hat{A}$. Note that $S_1, S_2, \ldots, S_p$ form a partition of S. We argue that S_k is either empty or contains every column in the kth block column of $\hat{A}$. Suppose, to the contrary, that S_k contains some, but not all, columns in the kth block column of $\hat{A}$. To prove the result it suffices to show that S has nonzero entries in at least $|S| + 1$ rows of $\hat{A}$.

Consider the columns in S_ℓ where $1 \leq \ell \leq p$ and $\ell \neq k$. Since $\hat{A}_{\ell\ell}$ is (strong) Hall, the columns in S_ℓ will have nonzero entries in at least $|S_\ell|$ rows of $\hat{A}_{\ell\ell}$. Since S_k contains some, but not all, columns in the kth block column, the columns in S_k will have nonzero entries in at least $|S_k| + 1$ rows of $\hat{A}_{kk}$. The columns in S consequently have nonzero entries in at least $|S| + 1$ rows of $\hat{A}$, and this concludes the proof. □

THEOREM 4.3. *Let A be a weak Hall matrix with a zero-free diagonal, and let $\hat{A}$ be the same matrix after it has been reordered so that it has a zero-free diagonal and is in block upper triangular form. Moreover, assume that the column order in $\hat{A}$ is consistent with the column order in A. It follows then that $|\hat{\mathcal{U}}_{\text{lub}}| \leq |\mathcal{U}_{\text{lub}}|$.*

Proof. Assume that $(r, s) \in \hat{\mathcal{U}}_{\text{lub}}$, and let $i = \beta^{-1}(r)$ and $j = \alpha^{-1}(s)$. To prove the result, it suffices to show that $(i, j) \in \mathcal{U}_{\text{lub}}$.

Recall the bipartite graph $B_j = (X_j, Y_j, E_j)$ associated with $A[j]$ in §3.1, and recall the row index set P_j which specifies the pairs (ℓ, j) included in column j of $\mathcal{U}_{\text{lub}}$

(see Theorem 3.1). Let $\hat{B}_s = (\hat{X}_s, \hat{Y}_s, \hat{E}_s)$ be the bipartite graph associated with $\hat{A}[s]$, and let $\hat{P}_s$ be the row index set which specifies the pairs (t, s) included in column s of $\hat{\mathcal{U}}_{\text{lub}}$. Since $(r, s) \in \hat{\mathcal{U}}_{\text{lub}}$, it follows from Theorem 3.1 that $r \in \hat{P}_s$; hence there exists a path from $s \in \hat{Y}_s$ to $r \in \hat{X}_s$ in $\hat{B}_s$. To show that $(i, j) \in \mathcal{U}_{\text{lub}}$ it suffices to show that $\hat{B}_s$ is isomorphic to a subgraph of B_j under the row permutation β and the column permutation α, for if this were the case there then would exist a path from $j \in Y_j$ to $i \in X_j$ in B_j, which would imply that $i \in P_j$. It would then follow from Theorem 3.1 that $(i, j) \in \mathcal{U}_{\text{lub}}$, as desired.

Suppose column $\hat{A}_{*s}$ lies in the kth block column of $\hat{A}$, where $1 \leq k \leq p$. Let $\hat{A}_{*t}$ be the first column in the kth block column of $\hat{A}$. If follows from Lemma 4.1 that the Hall columns and Hall rows of $\hat{A}[s-1]$ are $\hat{C}^H_{s-1} = \hat{R}^H_{s-1} = \{1, 2, \ldots, t-1\}$, and consequently we have $\hat{Y}_s = \{t, t+1, \ldots, s\}$. To show that $\hat{B}_s$ is isomorphic to a subgraph of B_j under the reordering, it then suffices to show that $\alpha^{-1}(q) \in Y_j$ for every $q \in \hat{Y}_s$ and $\beta^{-1}(q) \in X_j$ for every $q \in \hat{X}_s$. Equivalently, it suffices to show that $\alpha^{-1}(q) \notin C^H_{j-1}$ for every $q \in \hat{Y}_s$ and $\beta^{-1}(q) \notin R^H_{j-1} \cup F_j$ for every $q \in \hat{X}_s$. (Recall that $Y_j = \{1, 2, \ldots, j\} - C^H_{j-1}$ and $X_j = \{1, 2, \ldots, m\} - R^H_{j-1} - F_j$.)

Clearly, a set S of columns taken from A is a set of Hall columns if and only if $\alpha(S)$ is a set of Hall columns in $\hat{A}$. Now let S be a set of Hall columns taken from $A[j-1]$. Since the column A_{*j} is excluded from S, clearly the column $\hat{A}_{*s}$ is excluded from the set of Hall columns $\alpha(S)$. Hence, by Lemma 4.2, every column $\hat{A}_{*q}$, $t \leq q < s$, is excluded from $\alpha(S)$. In consequence, $\alpha^{-1}(q) \notin S$ for every $q \in \hat{Y}_s$. In particular, $\alpha^{-1}(q)$ does not belong to the set of Hall columns C^H_{j-1} for every $q \in \hat{Y}_s$. Hence $\alpha^{-1}(q) \in Y_j$ for every $q \in \hat{Y}_s$, as required.

Using a similar argument, $\beta^{-1}(q) \notin R^H_{j-1}$ for every $q \in \hat{X}_s$. It remains to show that $\beta^{-1}(q) \notin F_j$ for every $q \in \hat{X}_s$. By way of contradiction, assume that there exists a $q \in \hat{X}_s$ such that $\beta^{-1}(q) \in F_j$. The consistent ordering implies that row $\hat{A}_{q*}$ has zero entries from column 1 through column s, which means that $q \in \hat{F}_s$, and hence $q \notin \hat{X}_s$, contrary to the assumption. This concludes the proof. □

COROLLARY 4.4. *Let A be a weak Hall matrix with a zero-free diagonal, and let $\hat{A}$ be the same matrix after it has been reordered so that it has a zero-free diagonal and is in block upper triangular form. Moreover, assume that the column order in $\hat{A}$ is consistent with the column order in A. It follows then that $|\hat{\mathcal{R}}_{\text{lub}}| \leq |\mathcal{R}_{\text{lub}}|$.*

Proof. Choose i and j for which $1 \leq i, j \leq n$, and let $r = \alpha(i)$ and $s = \alpha(j)$. Moreover, choose i and j so that $(r, s) \in \hat{\mathcal{R}}_{\text{lub}}$. It follows from Theorem 3.2 that $\hat{\mathcal{R}}_{\text{lub}} = \hat{\mathcal{U}}^T_{\text{lub}} \hat{\mathcal{A}}$, where $\hat{\mathcal{A}} = \text{Struct}(\hat{A})$ and $\hat{\mathcal{U}}^T_{\text{lub}} = \{(q, t) : (t, q) \in \hat{\mathcal{U}}_{\text{lub}}\}$. Thus, there exists k, $1 \leq k \leq m$, so that for $t = \beta(k)$ we have $(r, t) \in \hat{\mathcal{U}}^T_{\text{lub}}$ and $(t, s) \in \hat{\mathcal{A}}$. Clearly then $(k, j) \in \mathcal{A}$, where $\mathcal{A} = \text{Struct}(A)$. By the proof of Theorem 4.3, $(r, t) \in \hat{\mathcal{U}}^T_{\text{lub}}$ implies that $(i, k) \in \mathcal{U}^T_{\text{lub}}$. Since by Theorem 3.2 we have $\mathcal{R}_{\text{lub}} = \mathcal{U}^T_{\text{lub}} \mathcal{A}$, it follows that $(i, j) \in \mathcal{R}_{\text{lub}}$, which proves the result. □

5. Concluding remarks. In this paper we used a recent sparsity analysis of the QR factorization [13] to better understand the structural upper bounds generated by the symbolic Householder procedure—especially the upper bound $\bar{\mathcal{W}}$ on Struct(W), where W is the $m \times n$ Householder matrix associated with an $m \times n$ matrix A [8, 9]. To bridge the gap between the recent analysis and the symbolic Householder procedure, we used a symbolic product based on $\bar{\mathcal{W}}$ to compute a new upper bound $\bar{\mathcal{Q}}$ on Struct(Q), where Q is the $m \times m$ orthogonal factor. Moreover, by extending the representation of $\bar{\mathcal{W}}$ introduced in George, Liu, and Ng [8], we obtained an implicit representation of $\bar{\mathcal{Q}}$ in terms of an extended elimination forest of $\bar{\mathcal{R}}$ and a set of "first

nonzero" parameters $f(i)$ associated with Struct(A).

We then let U be the matrix comprising the first n columns of Q, and we let $\bar{\mathcal{U}}$ be the upper bound on Struct(U) obtained directly from $\bar{\mathcal{Q}}$. We showed that the lower trapezoidal pattern $\bar{\mathcal{W}}$ and the lower trapezoidal part of $\bar{\mathcal{U}}$ coincide. As a result, we can better quantify the added costs incurred by computing and storing the orthogonal factor Q (or U), rather than computing and storing the Householder matrix W. Gilbert, Ng, and Peyton [10] have analyzed these added costs for certain standard finite element matrices.

Coleman, Edenbrandt, and Gilbert [1] showed that the upper bound $\bar{\mathcal{R}}$ generated by the symbolic Householder procedure is the least upper bound on Struct(R) whenever A is strong Hall. As might be expected, we were able to show that $\bar{\mathcal{U}}$ is the least upper bound on Struct(U) and $\bar{\mathcal{W}}$ is the least upper bound on Struct(W) whenever A is strong Hall and has a zero-free main diagonal.

We showed how to use the symbolic Householder procedure to obtain least upper bounds on the block diagonal patterns Struct(W) and Struct(U) that arise whenever A is weak Hall, has a zero-free main diagonal, and is in block upper triangular form. Finally, we showed how to reorder the rows and columns of A into block upper triangular form without increasing the fill incurred by the QR factorization.

We contend that, from the practitioner's point of view, focusing our attention on weak Hall matrices in block upper triangular form and with a zero-free main diagonal is adequate and quite natural. Efficient algorithms for finding a zero-free diagonal [3] and for reordering a matrix into block upper triangular form [5, 18] have long been used by the sparse matrix research community. It is interesting to consider whether these tools might provide the basis for a more sophisticated variant of the symbolic Householder procedure which retains the efficiency of the original and computes least upper bounds on Struct(W), Struct(U), and Struct(R) for any $m \times n$ Hall matrix A with $m \geq n$. If such an algorithm is possible, we believe that it would further clarify the connections between the analysis in Hare et al. [13] and the efficient techniques and data structures used in practice.

Acknowledgment. This paper has been improved a great deal by suggestions and constructive criticism provided by the referees.

REFERENCES

[1] T. COLEMAN, A. EDENBRANDT, AND J. GILBERT, *Predicting fill for sparse orthogonal factorization*, J. Assoc. Comput. Mach., 33 (1986), pp. 517–532.

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

[3] I. DUFF, *Algorithm* 575. *Permutations for a zero-free diagonal*, ACM Trans. Math. Software, 7 (1981), pp. 387–390.

[4] ———, *On algorithms for obtaining a maximum transversal*, ACM Trans. Math. Software, 7 (1981), pp. 315–330.

[5] I. DUFF AND J. REID, *Algorithm* 529. *Permutations to block triangular form*, ACM Trans. Math. Software, 4 (1978), pp. 189–192.

[6] A. GEORGE AND F. GUSTAVSON, *A new proof on permuting to block triangular form*, Tech. report RC 8238, IBM T. J. Watson Research Center, Yorktown Heights, NY, 1980.

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

[8] A. GEORGE, J. W. H. LIU, AND E. G.-Y. NG, *A data structure for sparse QR and LU factorizations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 100–121.

[9] A. GEORGE AND E. G.-Y. NG, *Symbolic factorization for sparse Gaussian elimination with partial pivoting*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 877–898.

[10] J. GILBERT, E. NG, AND B. PEYTON, *Separators and structure prediction in sparse orthogonal factorization*, Tech. report CSL-93-15, Xerox Palo Alto Research Center, Palo Alto, CA, 1993, submitted to Linear Algebra Appl.

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

[12] P. HALL, *On representatives of subsets*, J. Lond. Math. Soc., 10 (1935), pp. 26–30.

[13] D. HARE, C. JOHNSON, D. OLESKY, AND P. VAN DEN DRIESSCHE, *Sparsity analysis of the QR factorization*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 655–669.

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

[15] ———, *The role of elimination trees in sparse factorization*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 134–172.

[16] L. LOVASZ AND M. PLUMMER, *Matching Theory*, North–Holland, Amsterdam, 1986.

[17] A. POTHEN, *Predicting the structure of sparse orthogonal factors*, Linear Algebra Appl., 194 (1993), pp. 183–204.

[18] A. POTHEN AND C.-J. FAN, *Computing the block triangular form of a sparse matrix*, ACM Trans. Math. Software, 16 (1990), pp. 303–324.

IN MEMORIAM
ROBERT C. THOMPSON
1931–1995

The matrix theory community was shocked and deeply saddened by the untimely death of Bob Thompson on December 10, 1995. He was awaiting a heart transplant that had recently become necessary.

After growing up near Vancouver, British Columbia and receiving his bachelor's and master's degrees from the University of British Columbia (UBC), Bob received his Ph.D. from Caltech in 1960. He was Olga Taussky Todd's first official student. After returning to the faculty at UBC for three years, Bob moved to the University of California at Santa Barbara (UCSB), where he spent the remainder of his career. At Santa Barbara, he began a long-term professional relationship with Marvin Marcus that included collaborative research, the founding of the journal *Linear and Multilinear Algebra* (now one of the three main journals of matrix theory), and the founding of the Institute for the Interdisciplinary Application of Algebra and Combinatorics. With the arrival of other prominent colleagues, including Ky Fan, Eugene Johnsen, Henryk Minc, and, later, Morris Newman, Santa Barbara became for several years the world's mecca for research in matrix theory. During this period, Santa Barbara did much to focus attention upon the subject of matrix theory and to promote the high-level research that has been the foundation of the subject's vigorous, world-wide renaissance. Important meetings and other special activities were hosted, and UCSB was a place for sabbaticals and other visits; assistance and inspiration were given to young researchers (such as this author), and many of the Ph.D. students trained at UCSB (Bob himself had 11) have become important contributors to the field. Several newer, strong centers of matrix research in other countries, such as Israel, Hong Kong, Portugal, and Spain, can trace their intellectual roots to Santa Barbara.

Bob published more than 120 papers and a number of other items (including four undergraduate textbooks) during his career. He was serving as an editor of this journal at the time of his death. His interests were very broad and, like many researchers, his work went through stages and changes in taste, so much so that it is impossible to briefly categorize in any accurate way. Bob read a great deal of matrix theory and actually listened carefully to virtually all talks at the meetings he attended, so he knew the subject very broadly. He was often able to make helpful suggestions, even about topics on which he had no interest in working. His early work was especially algebraic, often dealing with his thesis area (a favorite of Taussky Todd's), which was multiplicative matrix commutators (and their products) over arbitrary fields. This very detailed work answered nearly all major questions in the subject and showed a hallmark of Bob's work: a willingness and ability to make unusually elaborate algebraic calculations in order to answer a question. It was not that he didn't appreciate external or efficient, implicit tools if they were available. Quite the contrary—Bob was a major proponent of employing other parts of mathematics useful in matrix theory. But he almost always discovered or convinced himself of important ideas through very complicated calculations.

A unifying theme of the broad middle part of Bob's publishing career was the drive to discover and understand the exact relationship among particular fundamental matrix parameters. If necessary conditions were obvious or known, a proof of sufficiency often involved very intricate constructions. For example, Bob's work on invariant factors, including the Sá–Thompson inequalities (separate papers), became very well known and attracted attention to his work in the systems and control community. This period included a major influence from and collaboration with Morris Newman, often involving number theoretic issues in integral matrices. Other examples included

the relationship between diagonal entries and singular values, the diagonal entries of normal matrices, and a major effort—motivated by Lidskii's claims—to prove A. Horn's conjectures about the eigenvalues of a sum of two Hermitian matrices. It was most intriguing to Bob when an unusual condition turned up, such as the possibility of a subtracted smallest term in what otherwise appeared to be a majorization relationship. He wrote multiple numbered series of papers in this period, and there is still a wealth of not-well-enough-known information to be found in his nine-paper series on "principal submatrices." Readers can get to know Bob by reading his amusing and thought-provoking *American Mathematical Monthly* piece (*Amer. Math. Monthly*, Vol. 90, pp. 661–668) "Author vs. Referee...." It contains professional, as well as mathematical, insights and is a good example of some of Bob's interests, described above.

Most recently, Bob returned to one of his favorite areas: generalizations of the field of values/numerical range. Inspired in part by the many questions raised by a 1950 paper of Kippenhahn, he was working very hard on the quaternionic field of values. Bob rarely spoke about the same piece of work twice, but his fascination with the quaternionic field was evidenced by the fact that he spoke about this subject frequently in the last several major talks that he gave. Among the many services Bob did to research was to help dispel the misinformed view that linear algebra is simple and uninteresting. He often worked on difficult problems and, as much as anyone, showed that core matrix theory is laden with deeply challenging and intellectually compelling problems that are fundamentally connected to many parts of mathematics, perhaps more so than other subfields of mathematics. The body of Bob's work was honored with his 1988 Johns Hopkins Summer Lecture Series and his recent (unfortunately posthumous) ILAS Hans Schneider Prize in Linear Algebra.

Bob will surely be missed as an innovative researcher and expert resource, but his grace and style in the community will be missed just as much. He was always encouraging to others and never jealous; he simply worked hard to solve difficult problems—not just to publish—and he was always happy to acknowledge the role of others. His talks were fresh and informative and, though quiet, Bob always maintained a good sense of humor in matters both casual and professional.

Bob embodied a tradition of cooperation, respect, and the desire to advance knowledge in all aspects of matrix theory. In Bob's memory, let's hope that tradition will continue to prevail.

Charles R. Johnson
College of William and Mary

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 3, pp. 465–469, July 1996

ANY NONINCREASING CONVERGENCE CURVE IS POSSIBLE FOR GMRES*

ANNE GREENBAUM†, VLASTIMIL PTÁK‡, AND ZDENĚK STRAKOŠ‡

Abstract. Given a nonincreasing positive sequence $f(0) \geq f(1) \geq \cdots \geq f(n-1) > 0$, it is shown that there exists an n by n matrix A and a vector r^0 with $\|r^0\| = f(0)$ such that $f(k) = \|r^k\|$, $k = 1, \ldots, n-1$, where r^k is the residual at step k of the GMRES algorithm applied to the linear system $Ax = b$, with initial residual $r^0 = b - Ax^0$. Moreover, the matrix A can be chosen to have any desired eigenvalues.

Key words. GMRES, Krylov subspace, Krylov residual space

AMS subject classifications. 65F10, 65F15

1. Introduction. The GMRES algorithm [2] is a popular iterative technique for solving large sparse nonsymmetric (non-Hermitian) linear systems. Let A be an n by n nonsingular matrix and b an n-dimensional vector (both may be complex). To solve a linear system $Ax = b$, given an initial guess x^0 for the solution, the algorithm constructs successive approximations x^k, $k = 1, 2, \ldots$, from the affine spaces

$$x^0 + \text{span}\{r^0, Ar^0, \ldots, A^{k-1}r^0\}, \tag{1}$$

where $r^0 \equiv b - Ax^0$ is the initial residual. The approximations are chosen to minimize the Euclidean norm of the residual vector $r^k \equiv b - Ax^k$, i.e.,

$$\|r^k\| = \min_{u \in AK_k(A, r^0)} \|r^0 - u\|, \tag{2}$$

where $K_k(A, r^0) = \text{span}\{r^0, Ar^0, \ldots, A^{k-1}r^0\}$ is the kth Krylov subspace generated by A and r^0. We call $AK_k(A, r^0)$ the kth Krylov residual subspace.

In a previous paper [1] it was shown that any convergence curve that can be generated by the GMRES algorithm can be generated by the algorithm applied to a matrix having any desired eigenvalues. This is in marked contrast to the situation for normal matrices, where the eigenvalues of the matrix, together with the initial residual, completely determine the GMRES convergence curve. This dramatically illustrates the fact that when highly nonnormal matrices are allowed, eigenvalue information alone cannot guarantee fast convergence of GMRES.

The residual norms of successive GMRES approximations are nonincreasing since the residuals are being minimized over a set of expanding subspaces. The question arises, however, as to whether every nonincreasing sequence of residual norms is possible for the GMRES algorithm applied to some linear system. The question from [1] is extended in the following way: Given a nonincreasing positive sequence $f(0) \geq f(1) \geq \cdots \geq f(n-1) > 0$ and a set of nonzero complex numbers $\{\lambda_1, \ldots, \lambda_n\}$,

* Received by the editors September 30, 1994; accepted for publication (in revised form) by R. Freund August 23, 1995.

† Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012 (greenbaum@nyu.edu). The work of this author was supported by NSF contract INT 9218024. Part of this work was performed while this author was visiting the Institute of Computer Science, Czech Academy of Sciences.

‡ Institute of Computer Science, Czech Academy of Sciences, Pod vodarenskov věží 2, 18207 Praha 8, Czech Republic (ptak@uivt.cas.cz and strakos@uivt.cas.cz). This work was supported by AS CR grant A230401.

is there an n by n matrix A having eigenvalues $\lambda_1, \ldots, \lambda_n$ and an initial residual r^0 with $\|r^0\| = f(0)$, such that the GMRES algorithm applied to the linear system $Ax = b$, with initial residual r^0, generates approximations x^k such that $\|r^k\| = f(k)$, $k = 1, \ldots, n-1$? In this paper we answer this question affirmatively and show how to construct such a matrix and initial residual. The presented construction is very simple; it is not derived from the considerations described in [1]. Moreover, for a given convergence behavior, we characterize all the matrices and initial residuals for which GMRES generates the prescribed sequence of residual norms.

Note that the assumption $f(n-1) > 0$ means that the related GMRES procedure does not converge to the exact solution until the step n and the dimensions of both $K_n(A, r^0)$ and $AK_n(A, r^0)$ are equal to n. Using that assumption will simplify the notation; the modification of the results to the general case is straightforward.

Throughout the paper we assume exact arithmetic.

2. Constructing a problem with a given convergence curve and any prescribed nonzero eigenvalues. In this section, we construct a matrix A and a right-hand side b, solving the question formulated in the introduction without using the results from [1].

We start with a simple analysis of some properties of the desired solution. Since the residual vectors generated by the GMRES algorithm applied to a linear system $Ax = b$, with initial guess x^0, are completely determined by the matrix A and the initial residual r^0, we can assume without loss of generality that the initial guess x^0 is zero and the right-hand side vector b is the initial residual. We will refer to this procedure as GMRES (A, b). Suppose that A and b represent the unknown matrix and right-hand side. Let $\mathcal{W} = \{w^1, \ldots, w^n\}$ be an orthonormal basis for the Krylov residual space $AK_n(A, b)$ such that $\text{span}\{w^1, \ldots, w^j\} = AK_j(A, b)$, $j = 1, 2, \ldots, n$, and let W be the matrix with the orthonormal columns $(w^1, \ldots, w^n)$. From the minimization property (2) it is clear that b can be expanded as

$$b = \sum_{j=1}^{n} \langle b, w^j \rangle w^j, \tag{3}$$

where $|\langle b, w^j \rangle| = \sqrt{\|r^{j-1}\|^2 - \|r^j\|^2}$, $r^0 = b$, $\|r^n\| = 0$. Given a nonincreasing positive sequence $f(0) \geq f(1) \geq \cdots \geq f(n-1) > 0$, define $f(n) \equiv 0$ and the differences $g(k)$ by

$$g(k) = \sqrt{(f(k-1))^2 - (f(k))^2}, \quad k = 1, \ldots, n. \tag{4}$$

The conditions $\|b\| = f(0), \|r^j\| = f(j)$, $j = 1, 2, \ldots, n-1$, will then be satisfied if the coordinates of b in the basis $\mathcal{W}$ are determined by the prescribed sequence of residual norms,

$$W^* b = (g(1), \ldots, g(n))^T. \tag{5}$$

Let $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, $\lambda_j \neq 0$, $j = 1, 2, \ldots, n$, be a set of nonzero points in the complex plane. Consider the monic polynomial

$$z^n - \sum_{j=0}^{n-1} \alpha_j z^j = (z - \lambda_1)(z - \lambda_2) \ldots (z - \lambda_n). \tag{6}$$

Clearly, $\alpha_0 \neq 0$.

Construction of the matrix A and the right-hand side b is straightforward. The idea is the following. Matrix A can be considered as a linear operator on the n-dimensional Hilbert space C^n. We denote this operator by $\mathcal{A}$; its matrix representation in the standard basis $\mathcal{E} = \{e_1, \ldots, e_n\}$ gives the desired matrix A:

$$\mathcal{A}^{\mathcal{E}} = A.$$

$\mathcal{A}$ is uniquely determined by its values on any set of basis vectors.

Let $\mathcal{V} = \{v^1, \ldots, v^n\}$ be *any* orthonormal basis in C^n, and let V be the matrix with the orthonormal columns $(v^1, \ldots, v^n)$. Let b satisfy

$$V^* b = (g(1), \ldots, g(n))^T \tag{7}$$

(note that given any b with $\|b\| = f(0)$, V can be chosen or, alternatively, given V, b can be chosen). Since $g(n)$ is nonzero, the set of vectors $\mathcal{B} = \{b, v^1, \ldots, v^{n-1}\}$ is linearly independent and also forms a basis for C^n. Let B be the matrix with columns $(b, v^1, \ldots, v^{n-1})$. Then the operator $\mathcal{A}$ is simply determined by the equations

$$\begin{aligned}
\mathcal{A}b &\stackrel{def}{=} v^1, \\
\mathcal{A}v^1 &\stackrel{def}{=} v^2, \\
&\vdots \\
\mathcal{A}v^{n-2} &\stackrel{def}{=} v^{n-1}, \\
\mathcal{A}v^{n-1} &\stackrel{def}{=} \alpha_0 b + \alpha_1 v^1 + \cdots + \alpha_{n-1} v^{n-1}.
\end{aligned} \tag{8}$$

Its matrix representation in the basis $\mathcal{B}$ is

$$\mathcal{A}^{\mathcal{B}} = \begin{pmatrix} 0 & \ldots & 0 & \alpha_0 \\ 1 & & 0 & \alpha_1 \\ & \ddots & \vdots & \vdots \\ & & 1 & \alpha_{n-1} \end{pmatrix}, \tag{9}$$

which is the companion matrix corresponding to the set of eigenvalues Λ. Finally, the matrix A is given by

$$A = \mathcal{A}^{\mathcal{E}} = B \mathcal{A}^{\mathcal{B}} B^{-1}. \tag{10}$$

Summarizing, we have proved the following theorem.

THEOREM 2.1. *Given a nonincreasing positive sequence* $f(0) \geq f(1) \geq \cdots \geq f(n-1) > 0$ *and a set of nonzero complex numbers* $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, *there exists a matrix* A *with eigenvalues* $\lambda_1, \lambda_2, \ldots, \lambda_n$ *and a right-hand side* b *with* $\|b\| = f(0)$ *such that the residual vectors* r^k *at each step of GMRES* (A, b) *satisfy* $\|r^k\| = f(k)$, $k = 1, 2, \ldots, n-1$.

It is obvious that the whole subject can be formulated in terms of linear operators and operator equations on a finite-dimensional Hilbert space.

For any chosen orthonormal basis $\mathcal{V}$, the matrix A and the right-hand side b can be constructed via (6), (9), (10) and (4), (7).

3. Characterization of all the matrices and right-hand sides for which GMRES generates the prescribed sequence of residual norms. In [1] it was shown that many different matrices can generate the same Krylov residual spaces. We start with a slightly generalized formulation of the theorem from [1].

THEOREM 3.1. *Let $E_1 \subset E_2 \subset \cdots \subset E_n$ be a sequence of subspaces of C^n, where E_j is of dimension j, $j = 1, 2, \ldots, n$, and let b be any n-dimensional vector. By $\mathcal{W} = \{w^1, \ldots, w^n\}$ we denote an orthonormal basis of E_n such that* span $\{w^1, \ldots, w^j\} = E_j$, *$j = 1, 2, \ldots, n$ and by W we denote the matrix with orthonormal columns $(w^1, \ldots, w^n)$. Let $\mathcal{A}$ be any nonsingular linear operator on E_n represented by its matrix A in the standard basis $\mathcal{E}$, $A = \mathcal{A}^{\mathcal{E}}$. Then $AK_j(A, b) = E_j$, $j = 1, 2, \ldots, n$, if and only if $\langle b, w^n \rangle \neq 0$ and the operator $\mathcal{A}$ has in the basis $\mathcal{W}$ matrix*

$$\mathcal{A}^{\mathcal{W}} = R\hat{H},$$

where R is any nonsingular upper triangular matrix and

$$\hat{H} = \begin{pmatrix} 0 & \ldots & 0 & 1/\langle b, w^n \rangle \\ 1 & & 0 & -\langle b, w^1 \rangle / \langle b, w^n \rangle \\ & \ddots & \vdots & \vdots \\ 0 & \ldots & 1 & -\langle b, w^{n-1} \rangle / \langle b, w^n \rangle \end{pmatrix}. \tag{11}$$

Proof. See Theorem 2.2 of [1]. □

As a consequence we obtain the following theorem.

THEOREM 3.2. *Given a nonincreasing positive sequence $f(0) \geq f(1) \geq \cdots \geq f(n-1) > 0$, the residual vectors r^k at each step of GMRES (A, b) satisfy $\|r^k\| = f(k)$, $k = 1, 2, \ldots, n-1$, if and only if A is of the form $A = W R \hat{H} W^*$ and b satisfies $W^* b = (g(1), \ldots, g(n))^T$, where W is a unitary matrix, R is a nonsingular upper triangular matrix, $\hat{H}$ is defined in* (11), *and $g(1), \ldots, g(n)$ are defined in* (4).

Proof. It is easy to see that for any nonsingular matrix C and orthonormal matrix Q, GMRES (QCQ^*, b) generates the same sequence of residual norms as GMRES $(C, Q^* b)$. Combining this observation with Theorem 3.1 finishes the proof. □

Thus, *all* matrices A and right-hand side vectors b for which GMRES (A, b) generates the required residual norms must be such that A is of the form $W R \hat{H} W^*$, where $\hat{H}$ is given by (11) and b satisfies (5) for some orthonormal matrix W. Conversely, for all matrix-vector pairs A, b of this form, GMRES (A, b) does indeed generate residual vectors with the required norms.

If we take, using the notation from (4), (6),

$$R = \begin{pmatrix} 1 & 0 & \ldots & 0 & \alpha_1 + \alpha_0 g(1) \\ 0 & 1 & & 0 & \alpha_2 + \alpha_0 g(2) \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & & & 1 & \alpha_{n-1} + \alpha_0 g(n-1) \\ 0 & 0 & \ldots & 0 & \alpha_0 g(n) \end{pmatrix}, \tag{12}$$

then $\hat{H} R$ is a companion matrix corresponding to the eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$. Since the matrix $\hat{H} R$ is similar to $R\hat{H}$, it follows that, with this choice of R, the matrix $A = W R \hat{H} W^*$ has eigenvalues $\lambda_1, \ldots, \lambda_n$, and so such a matrix can be constructed with any desired eigenvalues.

Note that for the simplest choice $W = I$, $b = (g(1), g(2), \ldots, g(n))^T$, the matrices $\hat{H}$ (11), resp. R (12), are identical to the matrices B^{-1}, resp. $B\mathcal{A}^{\mathcal{B}}$, from the previous

section,

$$B^{-1} \equiv \hat{H} = \begin{pmatrix} 0 & 0 & \dots & 0 & 1/f(n-1) \\ 1 & 0 & \dots & 0 & -g(1)/f(n-1) \\ & \ddots & & \vdots & \vdots \\ & & 1 & 0 & -g(n-2)/f(n-1) \\ & & & 1 & -g(n-1)/f(n-1) \end{pmatrix}, \tag{13}$$

and A is given by $R\hat{H}$. Emphasizing the fact that *any* nonincreasing convergence curve can be considered, these simple formulas form a useful tool for constructing numerical examples.

4. Conclusions and open questions. The results of this paper and [1] clearly demonstrate that eigenvalues are *not* the relevant quantities in determining the behavior of GMRES for nonnormal matrices. Any nonincreasing convergence curve can be obtained with GMRES applied to a matrix having any desired eigenvalues. Different quantities on which to base a convergence analysis have been suggested by others (for example, [4], [5]). It remains an open problem to determine the most appropriate set of system parameters for describing the behavior of GMRES. Another open problem is to determine what convergence curves are possible for the *envelope* of GMRES [3]. That is, if one does not consider a particular initial residual but instead considers the worst possible initial residual for each step k, $\max_{\|r^0\|=1} \|r^k\|$, $k = 1, \dots, n-1$, where the vectors r^k are generated by GMRES(A,r^0), then the sequence of norms must again be nonincreasing, but not every nonincreasing sequence is possible. It remains an open problem to characterize the possible sequences.

REFERENCES

[1] A. Greenbaum and Z. Strakoš, *Matrices that generate the same Krylov residual spaces*, in Recent Advances in Iterative Methods, G. Golub, A. Greenbaum, and M. Luskin, eds., Springer-Verlag, New York, 1994, pp. 95–118.

[2] 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.

[3] A. Greenbaum and L. N. Trefethen, *GMRES/CR and Arnoldi/Lanczos as matrix approximation problems*, SIAM J. Sci. Comput., 15 (1994), pp. 359–368.

[4] A. Greenbaum, *Norms of Functions of Matrices*, Courant Institute Tech. report 645, Courant Institute of Mathematical Sciences, New York, 1993.

[5] L. N. Trefethen, *Pseudospectra of matrices*, in Numerical Analysis 1991, D. F. Griffiths and G. A. Watson, eds., Longman Scientific and Technical, Essex, UK, 1992, pp. 234–266.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 3, pp. 470–488, July 1996

002

A BLOCK-GTH ALGORITHM FOR FINDING THE STATIONARY VECTOR OF A MARKOV CHAIN*

DIANNE P. O'LEARY† AND YUAN-JYE JASON WU‡

Abstract. Grassman, Taksar, and Heyman have proposed an algorithm for computing the stationary vector of a Markov chain. Analysis by O'Cinneide confirmed the results of numerical experiments, proving that the GTH algorithm computes an approximation to the stationary vector with low relative error in each component. In this work, we develop a block form of the GTH algorithm, more efficient on high-performance architectures, and show that it too produces a vector with low relative error. We demonstrate the efficiency of the algorithm on vector processors and on workstations with hierarchical memory.

Key words. Markov chain, GTH algorithm, stationary vector, relative error bounds

AMS subject classifications. 65F15, 60J10, 65G05

1. Introduction.

We consider the problem of computing the steady state distribution of a finite, discrete time, irreducible Markov chain. Equivalently, we seek the left eigenvector π corresponding to the eigenvalue 1 of a stochastic matrix P:

$$\pi P = \pi, \quad \pi e = 1, \quad Pe = e, \quad 0 \le p_{ij}, \quad i,j = 1,2,\ldots,n, \tag{1}$$

where e is the column vector of ones.

Grassman, Taksar, and Heyman [3] used probability theory to develop an algorithm (the *GTH algorithm*) for computing π by successively reducing the state space. The algorithm works with the *generator* matrix $G = P - I$ having zero row sums. It proved to be surprisingly accurate in numerical experiments and was later recognized as a variant of Gaussian elimination. The key difference is that the main diagonal element of the triangular factor is computed as the negative sum of the computed off-diagonal elements, and thus the row sum property is preserved. O'Cinneide [4] later analyzed the GTH algorithm, showing that the computed vector π has low relative error in each component.

No single algorithm runs at peak efficiency on each of the wide variety of computer architectures in current use. For some architectures, a simple count of arithmetic operations provides an accurate prediction of performance. For machines with vector pipelines and multilevel memories, however, the number of loads and stores of data can be a more critical factor. For parallel architectures, the data layout and communication patterns are crucial.

A common approach to algorithm design is to consider a parameterized family of algorithms that can be tuned to different architectures. *Block-matrix algorithms* provide one such parameterization, and their use is widespread in portable libraries such as LAPACK. There is a considerable body of literature on the error analysis of such block algorithms. Backward error bounds are established, for example, in [2]. The O'Cinneide bounds for GTH are much stronger than these results, since

* Received by the editors January 13, 1994; accepted for publication (in revised form) by L. Kaufman August 25, 1995. This work was supported by NSF grant 91-15568. Access to the Cray and Convex machines was provided by the National Institute for Standards and Technology.

† Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 (oleary@cs.umd.edu).

‡ Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439-4801 (jwu@mcs.anl.gov).

properties of the matrix G allowed him to obtain forward error bounds independent of the condition number of the matrix.

The purpose of our work is to define a *block-GTH algorithm* (§2), analyze its error properties (§3) to obtain results analogous to those of O'Cinneide, and determine the performance of the algorithm on various architectures (§4).

2. The block-GTH algorithm. Consider an irreducible *generator* G of dimension $n \times n$; i.e., G is a matrix with nonnegative off-diagonal elements and row sums equal to zero. We seek the row vector π satisfying

$$\pi G = 0, \quad \pi e = 1.$$

The GTH algorithm reduces G to lower triangular form. It is an iterative process, working with a matrix G_k of dimension $(n-k) \times (n-k)$ at the kth stage—a generator from which k states have been eliminated. Let $G_0 = G$, and partition as

$$G_k = \begin{bmatrix} A_k & B_k \\ C_k & D_k \end{bmatrix}, \tag{2}$$

where A_k is the $(1,1)$ element of G_k and B_k is the remaining part of the first row. Then if $p_k G_k = 0$, it is also true that

$$0 = p_k G_k \begin{bmatrix} 1 & -A_k^{-1} B_k \\ 0 & I \end{bmatrix} = p_k \begin{bmatrix} A_k & 0 \\ C_k & D_k - C_k A_k^{-1} B_k \end{bmatrix}.$$

Define

$$G_{k+1} = D_k - C_k A_k^{-1} B_k. \tag{3}$$

Note from (2) that $[C_k \ D_k]e = 0$ and $[A_k \ B_k]e = 0$, and so

$$[0, G_{k+1}]e = [C_k \ D_k]e - C_k A_k^{-1} [A_k \ B_k]e = 0.$$

Furthermore, the sign pattern is preserved, and G_{k+1} is a generator [4].

If we have a nonzero row vector p_{k+1} satisfying $p_{k+1} G_{k+1} = 0$, then the nonzero row vector defined by

$$p_k = \begin{bmatrix} -p_{k+1} C_k A_k^{-1} & p_{k+1} \end{bmatrix} \tag{4}$$

satisfies $p_k G_k = 0$. Thus, we have reduced the original problem to that of solving $p_{k+1} G_{k+1} = 0$, a problem with one fewer state.

The main difference between the GTH algorithm and standard Gaussian elimination is in the computation of A_k. In Gaussian elimination, this element is accumulated as a result of the updates (3). In the GTH algorithm, A_k is computed as the negative sum of the off-diagonal elements in the first row of G_k. A minor difference between the algorithms is that the GTH algorithm is usually formulated so that the last state (rather than the first one) is the first to be eliminated, but in this work we will eliminate the first state first, as in Gaussian elimination.

These relations form the basis for the GTH algorithm, which we now state more formally.

ALGORITHM GTH

FACTORIZATION PHASE

1. Let $G_0 = G$.

2. For $k = 0, 1, \ldots, n-2$

2.1. Partition G_k as in (2), where A_k is calculated as $A_k = -B_k e$.

2.2. Define G_{k+1} by (3).

End for.

BACKSUBSTITUTION PHASE

3. Let $p_{n-1} = 1$.

4. For $k = n-2, n-3, \ldots, 0$

4.1 Define p_k by (4).

End for.

5. Renormalize $\pi = p_0/(p_0 e)$.

The LU factors of G can be defined using quantities computed in the factorization phase of the algorithm:

$$G = \begin{array}{|c|c|c|c|} \hline A_0 & 0 & & \\ \hline & A_1 & 0 & \\ \hline C_0 & C_1 & \ddots & \\ \hline & & & A_{n-1} \\ \hline \end{array} \qquad \begin{array}{|c|c|c|c|} \hline 1 & A_0^{-1}B_0 & & \\ \hline & 1 & A_1^{-1}B_1 & \\ \hline 0 & 0 & \ddots & \\ \hline & & & 1 \\ \hline \end{array}$$

and we will make use of this fact later.

The GTH algorithm is easy to implement and numerically stable, but its efficiency on certain computer architectures can be disappointing. Notice, for example, that the (n, n) element of G is accessed and updated $n-1$ distinct times. It is well known that block-oriented algorithms can reduce the memory traffic for elimination algorithms, so we now direct our attention to developing a block-GTH algorithm.

The basis of the block-GTH algorithm is a *block* partitioning of the matrix G_k: we partition as in (2), but now A_k is an $l \times l$ matrix, rather than a single element. Similarly, B_k has l rows. The block size l can be tuned to achieve improved efficiency on various architectures, as discussed in §4. The generator G_{k+l} and its eigenvector p_{k+l} are expressed in terms of G_k and p_k by formulas similar to (3) and (4):

$$G_{k+l} = D_k - C_k A_k^{-1} B_k \,, \tag{5}$$

$$p_k = \begin{bmatrix} -p_{k+l} C_k A_k^{-1} & p_{k+l} \end{bmatrix} . \tag{6}$$

Rather than division by a scalar, (5) and (6) now require solution of linear systems involving the blocks A_k. This can easily be done using an LU factorization of these blocks.

The other main implementation issue is the correction of the main diagonal elements of A_k. To avoid memory traffic, we wish to do this with minimal access to the elements of B_k. Notice that the matrix

$$H_k = \begin{bmatrix} A_k & B_k e \\ 0 & 0 \end{bmatrix} \tag{7}$$

is also a generator, and the diagonal corrections that would be generated in step 2 of the GTH algorithm applied to this matrix are the same as those that GTH would generate for the original problem at the corresponding steps. For instance, after elimination in the first row of H_k, the elements $h_i \equiv (B_k e)_i$ are updated as

$$\bar{h}_i = h_i - \frac{\bar{a}_{i1} h_1}{\bar{a}_{11}} \tag{8}$$

$$= \sum_j b_{ij} - \frac{\bar{a}_{i1} \sum_j b_{1j}}{\bar{a}_{11}}$$

$$= \sum_j \left(b_{ij} - \frac{\bar{a}_{i1} b_{1j}}{\bar{a}_{11}} \right), \tag{9}$$

where $\bar{a}_{i1}$ is the updated value of a_{i1}, j ranges over the column indices in B_k, and $i = 2, \ldots, l$. Equation (9) shows that the update to the row sum vector $B_k e$ in (8) is mathematically equivalent to taking the row sum after correcting the matrix B in (9).

This is the basis for the block-GTH algorithm. For convenience in notation, we assume that l evenly divides n, although varying block sizes can be easily handled.

ALGORITHM BLOCK-GTH-I

FACTORIZATION PHASE

1. Let $G_0 = G$. Given an integer l between 1 and n, let $\widehat{n} = n/l$.

2. For $k = 0, l, \ldots, (\widehat{n} - 1)l$

2.1. Partition G_k as in (2), where A_k is an $l \times l$ matrix.

2.2. Apply the factorization phase of algorithm GTH to the matrix H_k defined by (7).

2.3. If $k \neq (\widehat{n} - 1)l$, define G_{k+l} by (5):

$$G_{k+l} = D_k - C_k(A_k^{-1} B_k),$$

where the factors from 2.2 are used to compute the expression in parentheses.

End for.

BACKSUBSTITUTION PHASE

3. Let $p_{(\widehat{n}-1)l}$ be computed from the backsubstitution phase of the GTH algorithm applied to $G_{(\widehat{n}-1)l}$.

4. For $k = (\widehat{n} - 2)l, (\widehat{n} - 3)l, \ldots, 0$

4.1. Define p_k by (6), again using the factors of A_k.

End for.

5. Renormalize $\pi = p_0/(p_0 e)$.

As an alternative to block-GTH-I, which relies on a block lower triangular factor with A_k on the main diagonal, we can compute a standard LU factorization of G:

$$G = \begin{array}{|c|c|c|} \hline L_0 & \multicolumn{2}{c|}{0} \\ \hline & L_l & 0 \\ \cline{2-3} C_0U_0^{-1} & C_lU_l^{-1} & \ddots \\ \cline{3-3} & & L_{(\widehat{n}-1)l} \\ \hline \end{array} \quad \begin{array}{|c|c|c|} \hline U_0 & \multicolumn{2}{c|}{L_0^{-1}B_0} \\ \hline & U_l & L_l^{-1}B_l \\ \cline{2-3} 0 & 0 & \ddots \\ \cline{3-3} & & U_{(\widehat{n}-1)l} \\ \hline \end{array} .$$

This algorithm takes the following form.

ALGORITHM BLOCK-GTH-II

FACTORIZATION PHASE

1. Let $G_0 = G$. Given an integer l between 1 and n, let $\widehat{n} = n/l$.

2. For $k = 0, l, \ldots, (\widehat{n}-1)l$

2.1. Partition G_k as in (2), where A_k is a $l \times l$ matrix.

2.2. Apply the factorization phase of algorithm GTH to the matrix H_k defined by (7), applying the same updates to C_k (i.e., computing $C_kU_k^{-1}$).

2.3. If $k \neq (\widehat{n}-1)l$, define G_{k+l} by (5):

$$G_{k+l} = D_k - (C_kU_k^{-1})(L_k^{-1}B_k).$$

End for.

BACKSUBSTITUTION PHASE

3–5. Use the backsubstitution phase of algorithm GTH, organizing the computations by single rows or by blocks of l rows.

3. Error analysis. As we mentioned before, the left eigenvector computed by the GTH algorithm has a small entry-wise relative error bound. Our next task is a rounding error analysis for the block-GTH algorithm in order to demonstrate that it preserves this error property.

Let us introduce some notation first. We use the special symbols $\langle\gamma\rangle$ from Appendix 3 of [5]. Let $\mathbf{u}$ be the unit roundoff in floating-point arithmetic. Then we write

$$\langle\gamma\rangle = \frac{(1+a_1)(1+a_2)\cdots(1+a_\alpha)}{(1+b_1)(1+b_2)\cdots(1+b_\beta)}$$

whenever $|a_i| \leq \mathbf{u}$, $|b_i| \leq \mathbf{u}$, and $\alpha + \beta = \gamma$. The $\langle\gamma\rangle$ symbols satisfy the relations

$$\langle\gamma\rangle\ \langle\alpha\rangle = \langle\gamma+\alpha\rangle$$

and

$$\frac{\langle\gamma\rangle}{\langle\alpha\rangle} = \langle\gamma+\alpha\rangle$$

and make floating-point expressions simple and clear. Let us denote the floating-point operators with a "hat." The error analysis of floating-point operations is based on the following rules:

1. $\langle\alpha\rangle\ a \mathbin{\widehat{\pm}} \langle\beta\rangle\ b = \langle\alpha+1\rangle\ a \pm \langle\beta+1\rangle\ b$,
2. $\langle\alpha\rangle\ a \mathbin{\widehat{*}} \langle\beta\rangle\ b = \langle\alpha+\beta+1\rangle\ a * b$,

3. $\langle\alpha\rangle\ a\ \widehat{/}\ \langle\beta\rangle\ b = \langle\alpha+\beta+1\rangle\ a/b$.

Note that these rules also hold if we interchange the operators on the left-hand side with those on the right-hand side. A fundamental property upon which we heavily rely is that if no cancellation occurs in forming a sum or difference (i.e., if the two operands have the same sign), then

$$\langle\alpha\rangle\ a\ \widehat{\pm}\ \langle\beta\rangle\ b = \langle\max(\alpha,\beta)+1\rangle\ (a \pm b).$$

The following theorem gives the error bounds for the GTH algorithm.

THEOREM 1 (see O'Cinneide [4]). *For any stochastic matrix P of order n with stationary vector π, the accuracy of the left eigenvector $\tilde{\pi}$ computed by the GTH algorithm using floating-point arithmetic is characterized by*

$$\tilde{\pi}_i = \langle 2\phi(n)+n\rangle\ \pi_i,\ \ i = 1,\ldots,n,$$

where $\phi(n) = (2n^3 + 6n^2 - 8n)/3$. Furthermore, if $(2\phi(n)+n)\mathbf{u} \leq .1$, then

$$\frac{|\tilde{\pi}_i - \pi_i|}{\pi_i} \leq 1.06(2\phi(n)+n)\mathbf{u}\ .$$

The formula for $\phi(n)$ is derived by induction [4] and makes use of a theorem of Tweedie [6], which says that if two irreducible generators G and $\tilde{G}$ of order m have the property that $\tilde{g}_{ij} = \langle\alpha\rangle\ g_{ij},\ i \neq j$, then their eigenvectors satisfy

$$\tilde{p}_i = \langle 2m\alpha\rangle\ p_i\ ,\ \ i = 1,\ldots,m\ .$$

The proof strategy is shown in Figure 1.

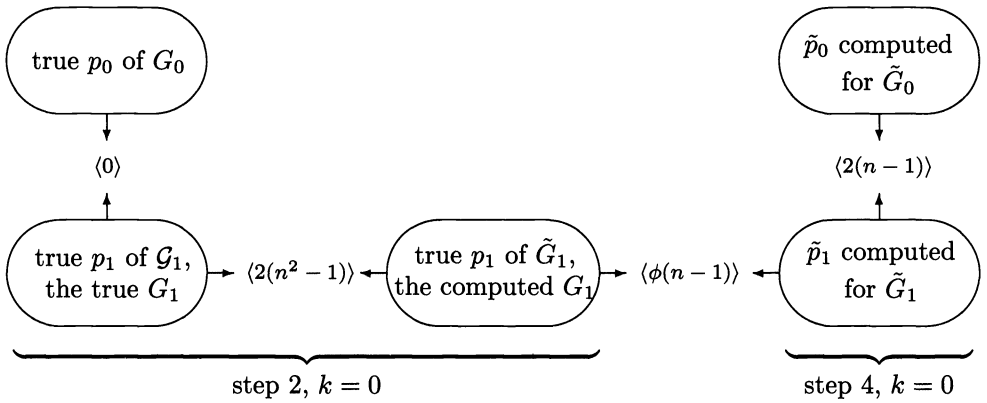


FIG. 1.

The bound for step 2 comes from verifying that given a generator G_0 in algorithm GTH, the relative error for the off-diagonal entries of the computed generator G_1 is $\langle n+1\rangle$. Since the generator G_1 is of order $n-1$, then by Tweedie's result the true eigenvector of computed G_1 has component-wise error bounded by

$$\langle 2(n+1)(n-1)\rangle\ = \langle 2(n^2-1)\rangle\ .$$

The bound for step 4 results from direct calculation. Combining these error bounds gives the recursion

$$\phi(n) = \phi(n-1) + 2n^2 + 2n - 4, \tag{10}$$

with the initial condition $\phi(1) = 0$.

This proof strategy yields a valid although far too pessimistic bound for the block-GTH algorithm. Suppose that we eliminate l states instead of 1. By an error analysis similar to [4], the error introduced into the eigenvector p_l in step 2 for the block algorithm is bounded by

$$\langle l^2(3^l - 1)\hat{n}^2 + \mathcal{O}(l^2 3^l \hat{n})\rangle \, .$$

We denote the error bound by $\psi(\hat{n} - 1)$. With the initial condition $\psi(1) = \phi(l)$, we have

$$\begin{aligned}\psi(\hat{n}) &= \frac{l^2}{3}(3^l - 1)\hat{n}^3 + \mathcal{O}((l3^l + l^2)\hat{n}^2) \\ &= \frac{3^l - 1}{3l} n^3 + \mathcal{O}(\frac{3^l}{l} n^2 + n^2) \, ,\end{aligned}$$

which is not tight for large block size l because of the exponential term. Therefore, a more delicate analysis is necessary.

We obtain a polynomial error bound for the computed eigenvector in GTH by repeatedly applying Tweedie's theorem to the generators resulting from eliminating one state only. This suggests reconsidering the error bound for one iteration in step 2 of block-GTH by accumulating error bounds when one state is eliminated instead of calculating the error bound for the eigenvector after eliminating l states. Our next task is to define the generators that are implicit in the intermediate steps of the block-GTH algorithm and determine the error bound for their off-diagonal entries. The proof strategy for block-GTH is shown in Figure 2.

Suppose that we have a given generator G_0 of order n. We need to determine an error bound $\psi_l(\hat{n})$ with polynomial growth in each iteration in step 2, where $\hat{n} = \lceil n/l \rceil$. Let $\tilde{G}_0 = G_0$. The generator $\tilde{G}_l$ is defined by the block-GTH algorithm, so we need to define the following generators.

$\mathcal{G}_k$, the generator of size $n-k$ that has the same eigenvector as $\tilde{G}_{k-1}$, for $k = 1, \ldots, l$.

$\tilde{G}_k$, the computed generator of size $n - k$, $k = 1, \ldots, l - 1$.

Since the block-GTH algorithm is closely related to GTH, it is useful to define $\mathcal{G}_k$ to be the generator resulting from eliminating the first state from $\tilde{G}_{k-1}$ by GTH using exact arithmetic. Note that the definition of $\mathcal{G}_1$ is the same for GTH and block-GTH (since G_0 is the same for both), but generators $\mathcal{G}_2, \ldots, \mathcal{G}_l$ differ for the two algorithms because they are defined in terms of the computed quantities $\tilde{G}_1, \ldots, \tilde{G}_{l-1}$. Our goal, then, is to study $\tilde{G}_k$ for the block-GTH algorithm and show that its eigenvector is close to the eigenvector of $\mathcal{G}_k$.

Throughout the following paragraphs, index k will vary between 1 and $l - 1$. A scalar with superscript k will denote a result after eliminating k states from G_0. An operator with a "hat" uses floating-point arithmetic. Since the error bound strongly depends on the specific computational formulas, we analyze the error by strictly following the order of operations in the block-GTH algorithm. We will derive an error bound for block-GTH-II. The bound for block-GTH-I is derived in the appendix.

Suppose that we partition the generator $\tilde{G}_k$ (including G_0) as

$$\tilde{G}_k = \begin{bmatrix} A_k & B_k \\ C_k & D_k \end{bmatrix} \, ,$$

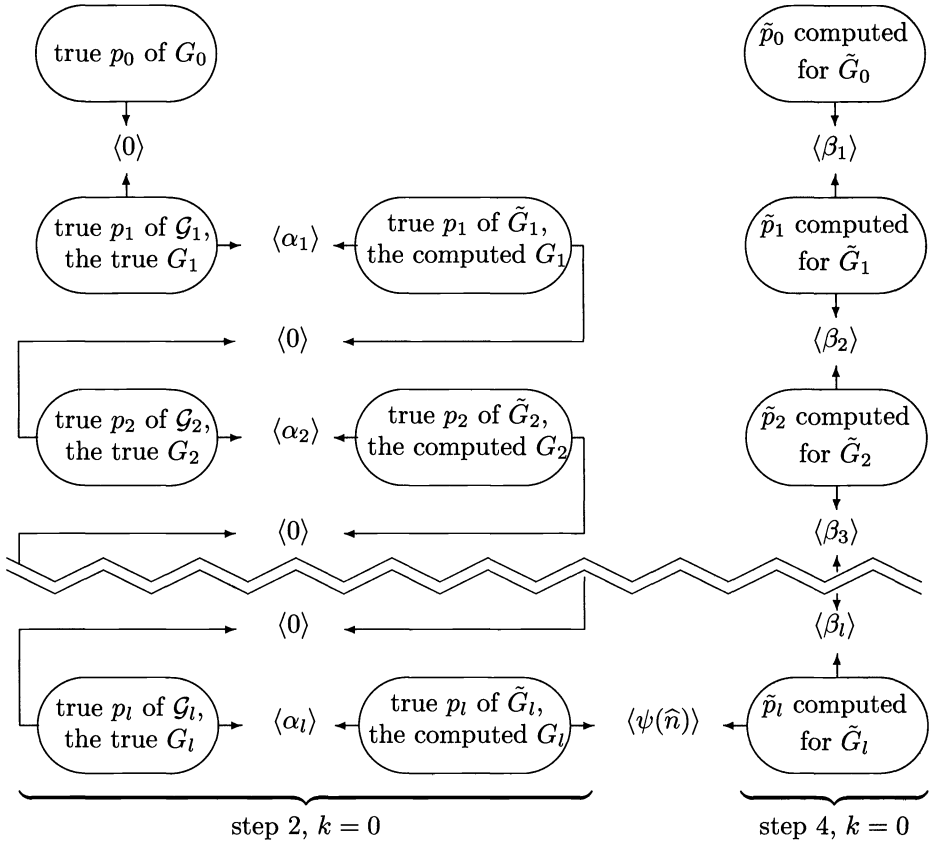


FIG. 2.

where A_k is of order $l - k$. Note that we define the order of A_k in a different way from the partition (2) for block-GTH in §2. It is convenient to index the elements of A_k as

$$\begin{bmatrix} a^k_{k+1,k+1} & \cdots & a^k_{k+1,l} \\ \vdots & \ddots & \vdots \\ a^k_{l,k+1} & \cdots & a^k_{ll} \end{bmatrix}. \tag{11}$$

Let $h^0 = B_0 e$. In step 2.2, we apply GTH to the matrix

$$H_0 = \begin{bmatrix} A_0 & h^0 \\ 0 & 0 \end{bmatrix}$$

using the following computations: for $k < i, j \leq l$,

$$s_k = -a^{k-1}_{kk} = a^{k-1}_{k,k+1} \mathbin{\hat{+}} \cdots \mathbin{\hat{+}} a^{k-1}_{kl} \mathbin{\hat{+}} h^{k-1}_k , \tag{12}$$

$$a^k_{ij} = a^{k-1}_{ij} \mathbin{\hat{+}} a^{k-1}_{ik} \mathbin{\hat{*}} (a^{k-1}_{kj} \mathbin{\hat{/}} s_k) , \tag{13}$$

$$h^k_i = h^{k-1}_i \mathbin{\hat{+}} a^{k-1}_{ik} \mathbin{\hat{*}} (h^{k-1}_k \mathbin{\hat{/}} s_k) , \tag{14}$$

and $s_l = -a^{l-1}_{ll} = h^{l-1}_l$.

The next task is to determine error bounds for the off-diagonal entries of A_k relative to corresponding entries of $\mathcal{G}_k$. Let $\langle \gamma_k \rangle$ be the error bound for s_k. From (12) we see that this bound arises from the error bound for h_k^{k-1} plus $l-k$ additions, and from (13) we conclude that the error for the off-diagonal entries of A_k is bounded by $\langle \gamma_k + 3 \rangle$ (since the entries in A_{k-1} have no error relative to $\mathcal{G}_k$). We determine γ_k by studying the h_i^k. Initially, h^0 has a component-wise error bound $\langle n-l-1 \rangle$ coming from $n-l-1$ additions. From (12), we have an error bound $\langle n-2 \rangle$ for s_1 relative to the sum of the off-diagonal entries of the first row of G_0. From (14), we have

$$\begin{aligned} h_i^1 &= h_i^0 \mathbin{\widehat{+}} a_{i1}^0 \mathbin{\widehat{*}} (h_1^0 \mathbin{\widehat{/}} s_1) \\ &= \widehat{\sum_j} b_{ij}^0 \mathbin{\widehat{+}} a_{i1}^0 \mathbin{\widehat{*}} \left[\left(\widehat{\sum_j} b_{1j}^0 \right) \mathbin{\widehat{/}} s_1 \right], \end{aligned}$$

where the summation is taken over the $n-l$ column indices of B_0. By using the rules of floating-point operations, we have

$$\begin{aligned} h_i^1 &= \langle n-l-1 \rangle \sum_j b_{ij}^0 \mathbin{\widehat{+}} \langle (n-l-1)+2 \rangle \sum_j a_{i1}^0 * (b_{1j}^0 / s_1) \\ &= \langle n-l-1 \rangle \sum_j b_{ij}^0 \mathbin{\widehat{+}} \langle (n-l-1)+4 \rangle \sum_j a_{i1}^0 \mathbin{\widehat{*}} (b_{1j}^0 \mathbin{\widehat{/}} s_1) \\ &= \langle (n-l-1)+5 \rangle \sum_j [b_{ij}^0 + a_{i1}^0 \mathbin{\widehat{*}} (b_{1j}^0 \mathbin{\widehat{/}} s_1)] \\ &= \langle (n-l-1)+6 \rangle \sum_j [b_{ij}^0 \mathbin{\widehat{+}} a_{i1}^0 \mathbin{\widehat{*}} (b_{1j}^0 \mathbin{\widehat{/}} s_1)] . \end{aligned} \tag{15}$$

Let us define the entries of B_1 by

$$b_{ij}^1 = b_{ij}^0 \mathbin{\widehat{+}} a_{i1}^0 \mathbin{\widehat{*}} (b_{1j}^0 \mathbin{\widehat{/}} s_1) . \tag{16}$$

Then we obtain

$$h_i^1 = \langle (n-l-1)+6 \rangle \sum_j b_{ij}^1 .$$

For $i=2$, the above equation and (12) imply that s_2, the (1,1) entry of $\tilde{G}_1$, has error bound $\langle (n-3)+6 \rangle$ relative to the (1,1) entry of $\mathcal{G}_1$.

For the next update, we have

$$\begin{aligned} h_i^2 &= h_i^1 \mathbin{\widehat{+}} a_{i2}^1 \mathbin{\widehat{*}} (h_2^1 \mathbin{\widehat{/}} s_2) \\ &= \langle (n-l-1)+6 \rangle \sum_j b_{ij}^1 \mathbin{\widehat{+}} \langle (n-l-1)+8 \rangle \sum_j a_{i2}^1 * (b_{2j}^1 / s_2) , \end{aligned}$$

which is similar to the first line of (15), so we can define B_2 and directly derive

$$h_i^2 = \langle (n-l-1)+12 \rangle \sum_j b_{ij}^2 .$$

Then s_3, the (1,1) entry of $\tilde{G}_2$, has error bound $\langle (n-4)+2*6 \rangle$ relative to the (1,1) entry of $\mathcal{G}_2$. Continuing this process, we define matrix B_k as

$$\begin{bmatrix} b_{k+1,1}^k & \cdots & b_{k+1,n-l}^k \\ \vdots & \ddots & \vdots \\ b_{l1}^k & \cdots & b_{n-l,n-l}^k \end{bmatrix} , \tag{17}$$

and we have an error bound for s_k (including s_l) as $\langle (n-k-1)+6(k-1)\rangle \equiv \gamma_k$. Therefore, we have shown that there is an error bound $\langle \gamma_k + 3\rangle$ for each off-diagonal entry of A_k and for each entry of B_k, relative to the corresponding element in $\mathcal{G}_k$. The matrix C_0 has been updated in a similar way during step 2.2, and by defining

$$C_k = \begin{bmatrix} c^k_{1,k+1} & \cdots & c^k_{1l} \\ \vdots & \ddots & \vdots \\ c^k_{n-l,k+1} & \cdots & c^k_{n-l,l} \end{bmatrix} \tag{18}$$

to be the matrix after k updates, it is easy to see that the entry-wise relative error between C_k and the corresponding entries of $\mathcal{G}_k$ is also bounded by $\langle \gamma_k + 3\rangle$.

Next, we consider the matrix D_k. After finishing step 2.2, we have an LU factorization of the matrix A_0 as

$$A_0 = L\,U = \begin{bmatrix} -s_1 & & & \\ a^0_{21} & -s_2 & & 0 \\ \vdots & \vdots & \ddots & \\ a^0_{l1} & a^1_{l2} & \cdots & -s_l \end{bmatrix} \begin{bmatrix} 1 & (-a^0_{12} \mathbin{\widehat{/}} s_1) & \cdots & (-a^0_{1l} \mathbin{\widehat{/}} s_1) \\ & 1 & \cdots & (-a^1_{2l} \mathbin{\widehat{/}} s_2) \\ 0 & & \ddots & \vdots \\ & & & 1 \end{bmatrix}.$$

The computations for $\tilde{G}_l$ can be expressed as

$$\tilde{G}_l = D_0 - (C_0 U^{-1})(L^{-1}B_0)\,.$$

Now, let us focus on the computation of the entry (i,j) of $\tilde{G}_l$, where $1 \le i,j \le n-l$. We need to compute $(L^{-1}B_0)$ first. Let b be the jth column of the matrix B_0, and let x be the jth column of the matrix $(L^{-1}B_0)$. Then the solution to the triangular linear system $Lx = b$ is computed as

$$\begin{aligned} x_1 &= b^0_{1j} \mathbin{\widehat{/}} (-s_1)\,, \\ x_m &= (b^0_{mj} \mathbin{\widehat{-}} a^0_{m1} \mathbin{\widehat{*}} x_1 \mathbin{\widehat{-}} \cdots \mathbin{\widehat{-}} a^{m-2}_{m,m-1} \mathbin{\widehat{*}} x_{m-1}) \mathbin{\widehat{/}} (-s_m) \\ &= b^{m-1}_{mj} \mathbin{\widehat{/}} (-s_m), \quad m = 2,\ldots,l\,. \end{aligned}$$

To obtain $\tilde{G}_l$ in block-GTH-II, we have

$$\begin{aligned} (i,j) \text{ entry of } \tilde{G}_l &= d^0_{ij} \mathbin{\widehat{-}} [\text{row } i \text{ of } (C_0U^{-1})] \mathbin{\widehat{*}} [\text{column } j \text{ of } (L^{-1}B_0)] \\ &= d^0_{ij} \mathbin{\widehat{+}} [c^0_{i1} \mathbin{\widehat{*}} (b^0_{1j} \mathbin{\widehat{/}} s_1) \mathbin{\widehat{+}} \cdots \mathbin{\widehat{+}} c^{l-1}_{il} \mathbin{\widehat{*}} (b^{l-1}_{lj} \mathbin{\widehat{/}} s_l)] \\ &= \langle l\rangle\, [d^0_{ij} + c^0_{i1} \mathbin{\widehat{*}} (b^0_{1j} \mathbin{\widehat{/}} s_1) + \cdots + c^{l-1}_{il} \mathbin{\widehat{*}} (b^{l-1}_{lj} \mathbin{\widehat{/}} s_l)]\,. \end{aligned} \tag{19}$$

If we define

$$d^k_{ij} = d^{k-1}_{ij} + c^{k-1}_{ik} \mathbin{\widehat{*}} (b^{k-1}_{kj} \mathbin{\widehat{/}} s_k)\,, \tag{20}$$

then (19) becomes

$$\begin{aligned} (i,j) \text{ entry of } \tilde{G}_l &= \langle l\rangle\, [d^{l-1}_{ij} + c^{l-1}_{il} \mathbin{\widehat{*}} (b^{l-1}_{lj} \mathbin{\widehat{/}} s_l)] \\ &= \langle l+2\rangle\, [d^{l-1}_{ij} + c^{l-1}_{il} * (b^{l-1}_{lj}/s_l)] \\ &= \langle \gamma_l + l + 2\rangle \ (i,j) \text{ entry of } \mathcal{G}_l\,, \end{aligned} \tag{21}$$

where the γ_l comes from the error bound for s_l.

Note that (20) also gives us the definition of D_k as

$$\begin{bmatrix} d_{11}^k & \cdots & d_{1,n-l}^k \\ \vdots & \ddots & \vdots \\ d_{n-l,1}^k & \cdots & d_{n-l,n-l}^k \end{bmatrix}, \tag{22}$$

with entry-wise error bound $\langle \gamma_k + 2 \rangle$. From (11), (17), (18), and (22), we have defined $\tilde{G}_k$ and shown that its off-diagonal entry-wise error can be bounded by $\langle \gamma_k + 3 \rangle$. By applying Tweedie's theorem, for $k = 1, \ldots, l-1$, the component-wise error bound for the eigenvector of $\tilde{G}_k$ relative to the eigenvector of $\mathcal{G}_k$ is $\langle \alpha_k \rangle = \langle 2(n-k)(\gamma_k + 3) \rangle$.

From (21), we have the off-diagonal entry-wise error bound $\langle \gamma_l + l + 2 \rangle$ for $\tilde{G}_l$ computed by block-GTH-II. Since $\tilde{G}_l$ is of order l, by applying Tweedie's theorem, we have $\langle \alpha_l \rangle = \langle 2(n-l)(\gamma_l + l + 2) \rangle$ as a component-wise error bound between the eigenvectors of $\tilde{G}_l$ and $\mathcal{G}_l$. Therefore, the error bound accumulated in one iteration of the factorization phase of block-GTH-II is bounded by

$$\begin{aligned} \sum_{k=1}^{l} \alpha_k &= \left[\sum_{k=1}^{l-1} 2(n-k)(\gamma_k + 3) \right] + 2(n-l)(\gamma_l + l + 2) \\ &= \frac{1}{3}[6ln^2 + (12l^2 - 6l - 6)n - (10l^3 + 9l^2 - 13l)]. \end{aligned} \tag{23}$$

As for the backsubstitution phase, we can also express the computations in the form

$$q = p_l(C_0 U^{-1}) ,$$

where p_l is the computed eigenvector of $\tilde{G}_l$. Note that $p_0 = [qL^{-1} \;\; p_l]$ is the eigenvector of G_0.

The type II process uses the same backsubstitution process as the GTH algorithm. Thus the component-wise error bound $\langle \beta_k \rangle$ between the computed vectors $\tilde{p}_k$ and $\tilde{p}_{k-1}$ in block-GTH-II is $\langle \gamma_k + (n-k+1) \rangle$, where the bound $\langle n-k+1 \rangle$ comes from one multiplication, $n-k-1$ additions, and one division. The error bound for one iteration in step 4 of block-GTH-II is

$$\begin{aligned} \sum_{k=1}^{l} \beta_k &= \sum_{k=1}^{l} [\gamma_k + n - k + 1] \\ &= 2ln + 2(l^2 - 2l) . \end{aligned} \tag{24}$$

Combining (23) and (24), we have established a polynomial error bound for block-GTH-II: we have

$$\begin{aligned} \widehat{\psi}_l(n) &= \psi_l(\widehat{n}) \\ &= \frac{1}{3}\left[2n^3 + \left(9l - \frac{3}{l} \right) n^2 - (3l^2 + 3l + 2)n - (6l^3 - 9l^2 + 3l) \right], \end{aligned} \tag{25}$$

with $\psi_l(1) = \phi(l)$.

Therefore, we have the following analogue to Theorem 1.

THEOREM 2. *For any stochastic matrix P of order n with stationary vector π, the accuracy of the left eigenvector $\widehat{\pi}$ computed by the block-GTH algorithms* I *and* II *with block size l $(1 \le l \le n)$ using floating-point arithmetic is characterized by*

$$\widehat{\pi}_i = \langle 2\widehat{\psi}_l(n) + n \rangle \, \pi_i, \quad i = 1, \ldots, n,$$

where $\widehat{\psi}_l(n)$ is defined by (25) *or* (A.31). *Furthermore, if $(2\widehat{\psi}_l(n) + n)\mathbf{u} \le .1$, then*

$$\frac{|\widehat{\pi}_i - \pi_i|}{\pi_i} \le 1.06(2\widehat{\psi}_l(n) + n)\mathbf{u} \, .$$

Note that for $1 \le l \le n$, $\widehat{\psi}_l(n) \ge \phi(n)$, with equality holding at $l = 1$ and n. However, we can sharpen $\widehat{\psi}_l(n)$ in several places. For example, if l does not evenly divide n, then the initial value $\psi_l(1)$ is less than $\phi(l)$. For $l \le n/2$, the error bound $\langle \gamma_k \rangle$ for s_k can be reduced to $\langle n - l \rangle$, which is independent of k, by summing (12) from left to right.

4. Performance of the block-GTH algorithm. In this section, we consider the implementation of the block-GTH algorithm and discuss some numerical results comparing the performance of GTH and block-GTH. Experiments were performed with single precision IEEE arithmetic on a DECstation 3100 (DEC) and a SUN SPARCstation 2 (SUN), each with a 64k byte cache memory; a Convex C3820 (Convex); and a Cray Y-MP4D/2/16 (CRAY).

We implemented the algorithm using standard software as much as possible, primarily the basic linear algebra subroutines from the BLAS collection [1]. On the SUN and DEC machines, we used Fortran versions of the BLAS; on the Convex and Cray, we used the manufacturer-supplied versions. The standard Fortran compilers (f77, f77, fc, and cf77) were used with default levels of optimization. We summarize the machine configurations and computing environments in Table 1.

TABLE 1

Machine	Operating system	Processor	Compiler	Word length
DEC	ULTRIX V4.1	1	f77 V3.2	32-bit
SUN	SUN4c_OS413A	1	f77 SC2.0.1	32-bit
Convex	ConvexOS 10.1	1	fc version 7.0	64-bit
CRAY	UNICOS 7.0.4.3	1	cf77 Release 6.0	64-bit

The principal implementation task is SGTHLU, which computes the LU factorization of the matrix A_k for type I block-GTH ($[A_k^T \; C_k^T]^T$ for type II). Note that this subroutine performs the standard GTH algorithm when the block size l equals the order n of the original generator.

The major time-consuming modules of the algorithm are shown in Table 2.

TABLE 2

Step	Routine	Source	Function
2.2	SGTHLU	Uses Level-1 BLAS	Apply GTH
2.3	STRSM	Level-3 BLAS	Update B_k
2.3	SGEMM	Level-3 BLAS	Update D_k

Since the block-GTH algorithm is a variant of Gaussian elimination, the complexity is of order n^3 and the factorization phase dominates the computational time. In

our implementation, the cost of factorization is

$$\frac{1}{6}(4n^3 + 3n^2 - 7n) ,$$

independent of the block size l.

For our numerical experiments, we defined the generator of order n to be a circulant matrix

$$G = \begin{bmatrix} \alpha & \beta & \cdots & \gamma & \gamma \\ \gamma & \alpha & \ddots & & \gamma, \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \gamma & & \ddots & \alpha & \beta \\ \beta & \gamma & \cdots & \gamma & \alpha \end{bmatrix} ,$$

with

$$\begin{aligned} \alpha &= -0.01, \\ \beta &= 0.0002, \\ \gamma &= 0.0098/(n-2) . \end{aligned}$$

It can be shown that this generator has a simple eigenvalue 0 with left eigenvector

$$\pi = (1/n)\, e .$$

We tested only the type II block-GTH algorithm.

First we examined the accuracy of the block-GTH algorithm. We set $n = 400$ and varied the block size as $l = 1, 2, \ldots, 49, 50$ and then $l = 60, 80, \ldots, 400$. Table 3 shows the resulting rounding errors. As predicted by the theory, the errors do not have strong dependence on block size: the errors produced by the block-GTH algorithm varied between .87 and 1.5 times the errors produced by the GTH algorithm.

TABLE 3
Rounding errors resulting from use of the block-GTH algorithm with different block sizes for a generator of order 400.

	Block size	Average difference	Largest difference	Largest rela. difference
	20	3.5700e−09	1.8000e−08	7.2000e−06
	40	3.8925e−09	1.8000e−08	7.2000e−06
D	60	4.2275e−09	1.7000e−08	6.8000e−06
E	80	3.9475e−09	2.2000e−08	8.8000e−06
C	100	5.1225e−09	2.0000e−08	8.0000e−06
&	120	3.4000e−09	1.9000e−08	7.6000e−06
S	140	3.3600e−09	1.8000e−08	7.2000e−06
U	160	3.0025e−09	1.6000e−08	6.4000e−06
N	180	3.2050e−09	1.9000e−08	7.6000e−06
	200	3.1250e−09	1.9000e−08	7.6000e−06
	400	3.4175e−09	1.7000e−08	6.8000e−06

Figure 3 shows the total CPU times for our implementation of the factorization phase of the block-GTH algorithm and the time taken by its three dominant subroutines (SGTHLU, STRSM, and SGEMM) as the block size changes. Although the

number of operations is independent of block size, the total execution time on the SUN and DEC machines had a significant drop around block size 40, due to efficient utilization of the cache memory. This timing phenomenon is primarily due to the behavior of SGEMM. This routine computes the matrix $\tilde{G}_l$ column-wise by computing linear combinations of columns of the matrix C_k. Since Fortran stores matrices column by column, we can expect that SGEMM will perform best if for all k, the matrix C_k and some portion of D_k and B_k fit in the cache memory.

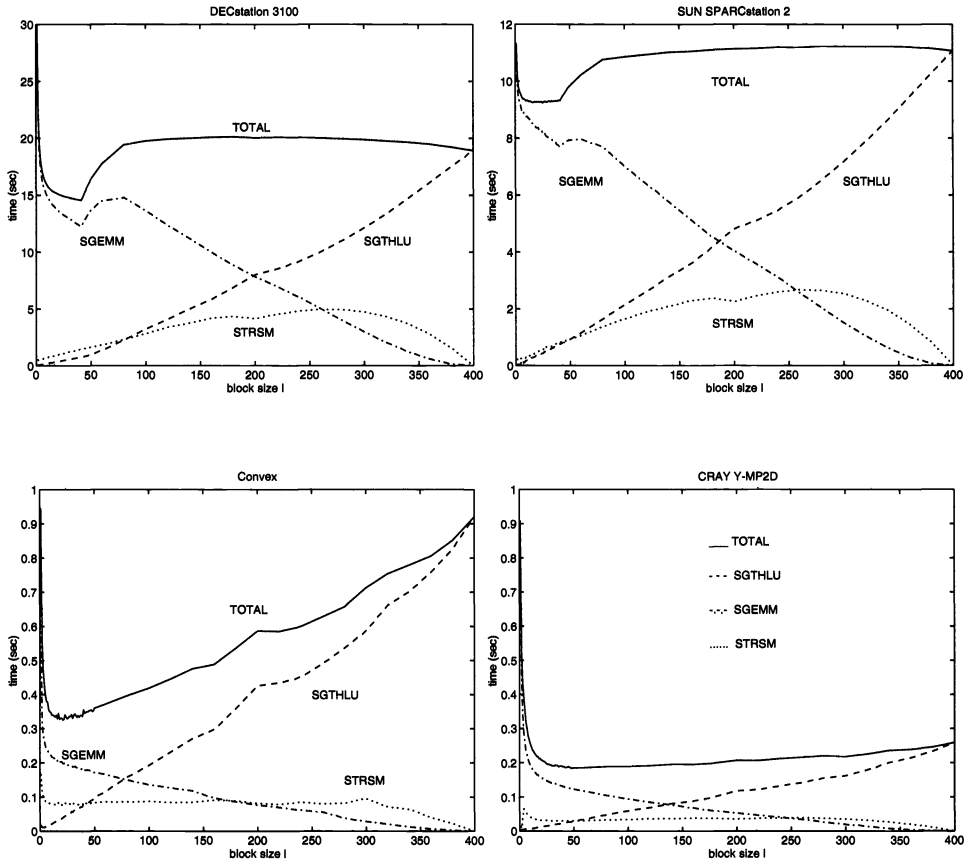


FIG. 3. *Block-GTH time as a function of block size for a generator of order* 400.

The biggest of the C matrices is C_0, of size $(n-l)l$. Each column of B_k has size l, and the D_0 matrix has columns of length $n-l$. Therefore, we predict that the optimal block size should occur at the largest integer l satisfying

$$(n-l)\,l + l + (n-l) \leq \frac{\text{cache memory size}}{\text{word size}}\,.$$

For the DEC and SUN, the cache capacity is 64k/4=16k words. The actual optimal block size also depends on other features of the machine architecture such as page size, cache line size, etc.

Next, we ran numerical experiments on generators of different orders. We varied the block size in increments of 1 until well past the predicted optimum, and then in

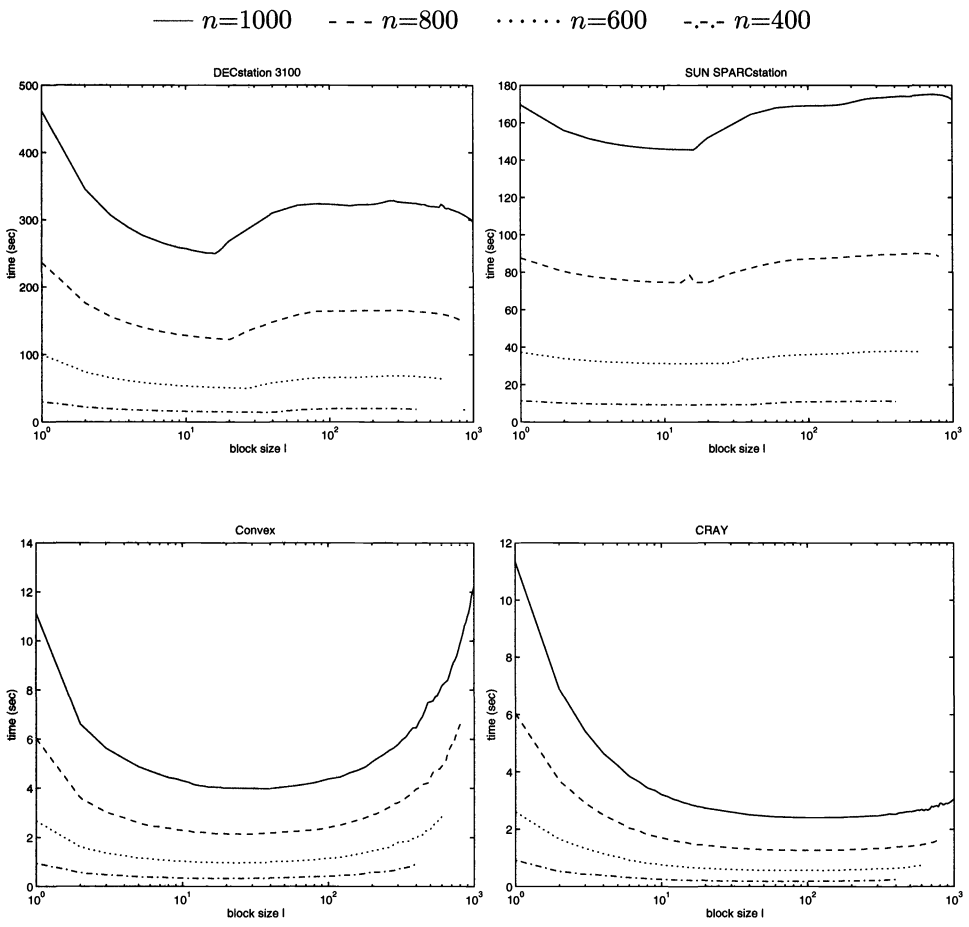


FIG. 4. *Block-GTH time as a function of block size for generators of order* 400, 600, 800, *and* 1000.

increments of 20. Figure 4 shows the total time as a function of block size. Table 4 gives the timings, predicted and actual optimal block size, and the speedup, defined by

$$\text{speedup} = \frac{\text{the time for the GTH algorithm}}{\text{the best time for the block-GTH algorithm}}.$$

On the SUN, the timing gain for the block-GTH algorithm over the standard GTH algorithm is 18–20%, while it is 19–30% on the DEC machine. The predictions of optimal block sizes were quite accurate for the DEC, but were overestimates for the SUN. Using the predicted optimal block size on the SUN gave timing gains of 16–19%, not much less than the actual optimal.

On the Convex, a block size of 21, independent of the order of the matrix, performs quite well, while on the Cray, the performance varies only slightly for a large range of block sizes, with the optimal size about 12% of n.

Further timing gains could be achieved by using level-2 or level-3 BLAS in the implementation of SGTHLU.

TABLE 4
Timings, optimal block size, and speedup for generators of order $n = 400, 600, 800,$ *and* 1000.

	Problem size n	GTH time (sec)	Block Size Predicted optimal	Block Size Actual optimal	Block-GTH Time (sec)	Block-GTH Mflop rate	Speedup
DEC	400	18.91	44	41	14.55	2.9	1.30
	600	64.18	28	27	50.11	2.9	1.28
	800	151.72	20	20	122.49	2.8	1.24
	1000	296.83	16	16	249.94	2.7	1.19
SUN	400	11.07	44	25	9.24	4.6	1.20
	600	37.35	28	19	31.23	4.6	1.20
	800	88.56	20	17	74.45	4.6	1.19
	1000	172.25	16	16	145.57	4.6	1.18
Convex	400	0.92		21	0.33	130	2.82
	600	2.85		24	0.96	150	2.95
	800	6.61		21	2.13	160	3.10
	1000	12.24		40	3.98	168	3.07
CRAY	400	0.26		48	0.18	238	1.41
	600	0.75		60	0.56	258	1.32
	800	1.64		100	1.27	269	1.29
	1000	3.06		120	2.41	277	1.27

5. Conclusions. It is necessary to use block algorithms in order to attain good utilization of vector processors and cache memory. In this work we have shown that the GTH algorithm has a block implementation that can achieve a considerable increase in efficiency without sacrificing accuracy. Future work will deal with the parallel implementation of the algorithm.

Appendix A. Derivation of the error bound for the block-GTH-I algorithm. The block-GTH-I differs from block-GTH-II only at step 2.3 for computing $\tilde{G}_l$ and in the backsubstitution phase, so the definitions for A_k and B_k in (11) and (17) remain the same for $k = 1, \ldots, l-1$. Thus we have an error bound $\langle \gamma_k + 3 \rangle$ for each off-diagonal entry of A_k and for each entry of B_k.

The computations of $\tilde{G}_l$ for block-GTH-I can be expressed as

$$\tilde{G}_l = D_0 - C_0(U^{-1}(L^{-1}B_0)) .$$

We save the original matrix C_0, so there is one more linear system to solve for computing $U^{-1}x$. Let y be the jth column of the matrix $U^{-1}(L^{-1}B_0)$. Applying backsubstitution to the triangular linear system $U\,y = x$, we have

$$\begin{aligned} y_l &= x_l \ , \\ y_m &= x_m \ \hat{+}\ (a_{ml}^{m-1} \ \hat{/}\ s_m) \ \hat{*}\ y_l \ \hat{+}\ \cdots \ \hat{+}\ (a_{m,m+1}^{m-1} \ \hat{/}\ s_m) \ \hat{*}\ y_{m+1}, \\ & \qquad m = l-1, \ldots, 1 \ . \end{aligned}$$

Note that all x_m are negative, so all y_m are negative and there is no cancellation. To obtain $\tilde{G}_l$, we have

$$\begin{aligned} (i,j) \text{ entry of } \tilde{G}_l &= d_{ij}^0 \ \hat{-}\ [\text{row } i \text{ of } C_0] \ \hat{*}\ [\text{column } j \text{ of } U^{-1}(L^{-1}B_0)] \\ &= d_{ij}^0 \ \hat{-}\ [c_{i1}^0 \ \hat{*}\ y_1 \ \hat{+}\ c_{i2}^0 \ \hat{*}\ y_2 \ \hat{+}\ \cdots \ \hat{+}\ c_{il}^0 \ \hat{*}\ y_l] \\ &= \langle l+1 \rangle \ [d_{ij}^0 - c_{i1}^0 * y_1 - c_{i2}^0 * y_2 - \cdots - c_{il}^0 * y_l] \ . \end{aligned} \tag{A.26}$$

By expanding y_1, we have

(i,j) entry of $\tilde{G}_l$

$$
\begin{aligned}
&= \langle l+1\rangle \{d_{ij}^0 - c_{i1}^0 * [(-b_{1j}^0 \mathbin{\widehat{/}} s_1) \mathbin{\widehat{+}} (a_{1l}^0 \mathbin{\widehat{/}} s_1) \mathbin{\widehat{*}} y_l \mathbin{\widehat{+}} \cdots \mathbin{\widehat{+}} (a_{12}^0 \mathbin{\widehat{/}} s_1) \mathbin{\widehat{*}} y_2] \\
&\qquad -c_{il}^0 * y_l - \cdots - c_{i2}^0 * y_2\} \\
&= \langle l+1\rangle \{d_{ij}^0 + \langle l\rangle [c_{i1}^0 * (b_{1j}^0 \mathbin{\widehat{/}} s_1) - c_{i1}^0 * (a_{1l}^0 \mathbin{\widehat{/}} s_1) * y_l - \cdots \\
&\qquad -c_{i1}^0 * (a_{12}^0 \mathbin{\widehat{/}} s_1) * y_2] - c_{il}^0 * y_l - \cdots - c_{i2}^0 * y_2\} \\
&= \langle (l+1)+l\rangle \{d_{ij}^0 + c_{i1}^0 * (a_{1l}^0 \mathbin{\widehat{/}} s_1) - [c_{il}^0 + c_{i1}^0 * (b_{1j}^0 \mathbin{\widehat{/}} s_1)] * y_l - \cdots \\
&\qquad -[c_{i2}^0 + c_{i1}^0 * (a_{12}^0 \mathbin{\widehat{/}} s_1)] * y_2\} .
\end{aligned}
$$

The updated matrix C_1 is not explicitly present, but we can define

$$
c_{im}^1 = c_{im}^0 + c_{i1}^0 * (a_{1m}^0 \mathbin{\widehat{/}} s_1), \quad m = 2, \ldots, l . \tag{A.27}
$$

From (20) and (A.27), the update becomes

$$
\text{entry of } \tilde{G}_l = \langle (l+1)+l\rangle \, [d_{ij}^1 + c_{il}^1 * y_l + \cdots + c_{i2}^1 * y_2] ,
$$

which is similar to (A.26). Therefore, we can expand $y_m, m = 2, \ldots, l-1$, one by one and repeat the same process to obtain

$$
\begin{aligned}
&(i,j) \text{ entry of } \tilde{G}_l \\
&= \langle (l+1)+l+\cdots+2\rangle \, [d_{ij}^{l-1} + c_{il}^{l-1} * y_l] \\
&= \left\langle \sum_{m=2}^{l+1} m \right\rangle \, [d_{ij}^{l-1} + c_{il}^{l-1} * (b_{lj}^{l-1} \mathbin{\widehat{/}} s_l)] \\
&= \left\langle 1 + \sum_{m=2}^{l+1} m \right\rangle \, [d_{ij}^{l-1} + c_{il}^{l-1} * (b_{lj}^{l-1}/s_l)] \\
&= \left\langle \gamma_l + \sum_{m=1}^{l+1} m \right\rangle \, (i,j) \text{ entry of } \mathcal{G}_l ,
\end{aligned} \tag{A.28}
$$

where the γ_l comes from s_l.

Note that from (A.27), the matrix C_k defined in (18) has an entry-wise error bound $\langle \gamma_k + 1\rangle$. Again, from (11), (17), (18), and (22), we have defined $\tilde{G}_k$ and shown that its off-diagonal entry-wise error can be bounded by $\langle \gamma_l + 3\rangle$. From (A.28), the entry-wise error bound for $\tilde{G}_l$ computed by block-GTH-I relative to $\mathcal{G}_l$ is $\langle \gamma_l + \sum_{m=1}^{l+1} m\rangle$. By applying Tweedie's theorem, the error accumulated for one iteration in step 2 of block-GTH-I is bounded by

$$
\begin{aligned}
\sum_{k=1}^{l} \alpha_k &= \left\lceil \sum_{k=1}^{l-1} 2(n-k)(\gamma_k + 3) \right\rceil + 2(n-l) \left(\gamma_l + \sum_{m=1}^{l+1} m \right) \\
&= \frac{1}{3}[6ln^2 + (15l^2 - 3l - 12)n - (13l^3 + 12l^2 - 19l)] .
\end{aligned} \tag{A.29}
$$

For the backsubstitution phase, let $p_l = [p_{l1} \cdots p_{l,n-l}]$ be the computed eigenvector of $\tilde{G}_l$. Then we have the vector

$$
\left[\widehat{\sum_i} p_{li} \mathbin{\widehat{*}} c_{i1}^0 \;\; \cdots \;\; \widehat{\sum_i} p_{li} \mathbin{\widehat{*}} c_{il}^0 \right]
$$

resulting from computing $(p_l C)$ first. Then the solution to the triangular linear system $qU = p_l C$ is

$$q_1 = \widehat{\sum_i} p_{li} \,\hat{*}\, c_{i1}^0 ,$$

$$q_k = \widehat{\sum_i} p_{li} \,\hat{*}\, c_{ik}^0 \,\hat{+}\, (a_{1k}^0 \,\hat{/}\, s_1) \,\hat{*}\, q_1 \,\hat{+}\, \cdots \,\hat{+}\, (a_{k-1,k}^{k-2} \,\hat{/}\, s_{k-1}) \,\hat{*}\, q_{k-1} ,$$

$$k = 2, \ldots, l .$$

From the first equation, we have $q_1 = \delta_1 \sum_i p_{li} * c_{i1}^0$, with $\delta_1 = n - l$. To find an error bound for q_k in the second equation, suppose that we have $q_m = \langle \delta_m \rangle \sum_i p_{li} * c_{im}^{m-1}$. We know that δ_m is an increasing function of m, so

$$\begin{aligned} q_k = \langle n - l \rangle \sum_i p_{li} * c_{ik}^0 \,\hat{+}\, \langle \delta_1 + 1 \rangle \sum_i [(a_{1k}^0 \,\hat{/}\, s_1) * p_{li} * c_{i1}^0] \\ \hat{+}\, \langle \delta_2 + 1 \rangle \sum_i [p_{li} * (a_{2k}^1 \,\hat{/}\, s_2) * c_{i2}^1] \,\hat{+}\, \cdots \\ \hat{+}\, \langle \delta_{k-1} + 1 \rangle \sum_i [p_{li} * (a_{k-1,k}^{k-2} \,\hat{/}\, s_{k-1}) * c_{i,k-1}^{k-2}] . \end{aligned}$$

Note that the floating-point additions are carried out from left to right. Using (A.27), we have

$$\begin{aligned} q_k &= \langle \delta_1 + 2 \rangle \left[\sum_i p_{li} * c_{ik}^0 + \sum_i p_{li} * (a_{1k}^0 \,\hat{/}\, s_1) * c_{i1}^0 \right] \\ &\quad \hat{+}\, \langle \delta_2 + 1 \rangle \sum_i [p_{li} * (a_{2k}^1 \,\hat{/}\, s_2) * c_{i2}^1] \,\hat{+}\, \cdots \\ &\quad \hat{+}\, \langle \delta_{k-1} + 1 \rangle \sum_i [p_{li} * (a_{k-1,k}^{k-2} \,\hat{/}\, s_{k-1}) * c_{i,k-1}^{k-2}] \\ &= \langle \delta_1 + 2 \rangle \sum_i p_{li} * c_{ik}^1 \,\hat{+}\, \langle \delta_2 + 1 \rangle \sum_i [p_{li} * (a_{2k}^1 \,\hat{/}\, s_2) * c_{i2}^1] \,\hat{+}\, \cdots \\ &\quad \hat{+}\, \langle \delta_{k-1} + 1 \rangle \sum_i [p_{li} * (a_{k-1,k}^{k-2} \,\hat{/}\, s_{k-1}) * c_{i,k-1}^{k-2}] . \end{aligned}$$

Since all δ_m are integers and $\delta_m + 1 \leq \delta_{m+1}$, we can repeat the same process and obtain

$$q_k = \langle \delta_{k-1} + 2 \rangle \sum_i p_{li} * c_{ik}^{k-1} .$$

Therefore, $\delta_k = \delta_{k-1} + 2$, and the solution to this recursion is $\delta_k = n - l + 2(k-1)$.

By an analysis similar to that for block-GTH-II, we have $\langle \gamma_k + \delta_k + (l - k + 1) \rangle$ as the component-wise error bound $\langle \beta_k \rangle$ for the computed p_k for block-GTH-I. The error bound for one iteration in step 4 of block-GTH-I is

$$\begin{aligned} \sum_{k=1}^{l} \beta_k &= \sum_{k=1}^{l} [\gamma_k + n + k - 1] \\ &= 2ln + 3l^2 - 5l . \end{aligned} \tag{A.30}$$

Finally, combining (A.29) and (A.30), we obtain a polynomially-growing error bound $\widehat{\psi}_l$ for block-GTH-I:

$$\begin{aligned}\widehat{\psi}_l(n) &= \psi_l(\widehat{n}) \\ &= \frac{1}{3}\left[2n^3 + \left(\frac{21l}{2} - \frac{6}{l} + \frac{3}{2}\right)n^2 - \left(\frac{9l^2}{2} + \frac{3l}{2} + 2\right)n - (6l^3 - 6l^2)\right]. \end{aligned} \tag{A.31}$$

This error bound agrees with (25), the error bound for block-GTH-II, in its highest order term.

Acknowledgments. We are grateful for the help of Ron Boisvert, James da Silva, Olafur Gudmundsson, Bill Pugh, and the referees.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. D. Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, *LAPACK Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.

[2] J. W. Demmel and N. J. Higham, *Stability of Block Algorithms with Fast Level 3 BLAS*, Tech. report, LAPACK Working Note 22, CS-90-110, University of Tennessee, Knoxville, TN, July 1990.

[3] W. K. Grassmann, M. I. Taksar, and D. P. Heyman, *Regenerative analysis and steady state distributions*, Oper. Res., 33 (1985), pp. 1107–1116.

[4] C. A. O'Cinneide, *Entrywise perturbation theory and error analysis for Markov chains*, Numer. Math., 65 (1993), pp. 109–120.

[5] G. W. Stewart, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[6] R. L. Tweedie, *Perturbations of countable Markov chains and processes*, Ann. Inst. Statist. Math., 32 (1980), pp. 283–290.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 3, pp. 489–508, July 1996

003

ON THE SPECTRAL RADIUS OF (0,1)-MATRICES WITH 1'S IN PRESCRIBED POSITIONS*

RICHARD A. BRUALDI† AND SUK-GEUN HWANG‡

Abstract. Let n and d be positive integers with $1 \leq d \leq n(n-1)/2$. We investigate the maximum and minimum spectral radii of a $(0,1)$-matrix of order n that has 1's on and below its main diagonal and d additional 1's. If $d \leq 4$ we determine all matrices of this type that have the maximum spectral radius. For general d we prove an asymptotic result that severely limits the structure of matrices with maximum spectral radius. For $d \leq n$, we determine the minimum spectral radius.

Key words. spectral radius, (0,1)-matrices

AMS subject classifications. 15A18, 15A48

1. Introduction.

Let n^2 not necessarily distinct nonnegative real numbers be given. A fundamental theorem of Schwarz [Sc64] asserts that the largest spectral radius of a matrix of order n whose entries are the n^2 given numbers occurs among those matrices in which the entries in each row and in each column are nonincreasing. Indeed let A be any matrix of order n whose entries are the given numbers. Then there exists a permutation matrix Q such that independently rearranging the entries of each row of QAQ^T to have nonincreasing order and then doing the same with the entries in each column results in a matrix A^* in which the entries in each row and in each column are nonincreasing, such that the spectral radius of A^* is at least as large as that of A. (Schwarz's argument was given for positive real numbers, but the conclusions apply to nonnegative numbers by a continuity argument.) Analogous conclusions hold for the smallest spectral radius; in particular, the smallest spectral radius occurs among those matrices in which the entries in each row are nonincreasing and the entries in each column are nondecreasing. Motivated by these results, Brualdi and Hoffman [BH85] considered the problem of determining the largest spectral radius for a $(0,1)$-matrix of order n with a prescribed number d of 1's (and thus $n^2 - d$ 0's). They proved that if $d = k^2$ or $d = k^2 + 1$, then the largest spectral radius equals k. Let $d = k^2 + t$, where $t \leq 2k$. Confirming (asymptotically) a conjecture of Brualdi and Hoffman, Friedland [Fr85] determined the largest spectral radius for all n if $t = 2k$ and for all sufficiently large n otherwise. Brualdi and Solheid [BS87] considered the corresponding minimum spectral radius problem and determined the minimum spectral radius for $(0,1)$-matrices of order n with at least $n^2 - \lfloor n/2 \rfloor \lceil n/2 \rceil$ 1's (an alternative proof of this result is contained in [Ch90]) and in all other cases bounded the minimum spectral radius between two consecutive integers.

In this paper we begin an investigation of a generalization of the above problems where 1's are prescribed in certain positions and d additional 1's can be put in any of the remaining positions. Let $\sigma(X)$ denote the number of 1's contained in a (0,1)-

* Received by the editors May 10, 1995; accepted for publication (in revised form) by R. Horn October 2, 1995.

† Department of Mathematics, University of Wisconsin–Madison, Madison, WI 53706 (brualdi@math.wisc.edu). The research of this author was partially supported by NSF grant DMS-9123318 and DMS-9424346.

‡ Department of Mathematics Education, Kyungpook University, Taegu 702-701, Republic of Korea. The research of this author was supported by a grant from TGRC-Kosef under the international cooperation program. This paper was written while this author was visiting the University of Wisconsin–Madison.

matrix X, and let $\rho(X)$ denote the spectral radius of X. A precise statement of our general problem is the following.

PROBLEM 1.1. *Let B be a $(0,1)$-matrix of order n having s 1's, and let d be a positive integer with $d \leq n^2 - s$. Let $\mathcal{A}(B,d)$ denote the set of all $(0,1)$-matrices A of order n such that $B \leq A$ (entrywise) and $\sigma(A - B) = d$. Determine the largest spectral radius $\overline{\rho}(\mathcal{A}(B,d))$(respectively, smallest spectral radius $\tilde{\rho}(\mathcal{A}(B,d))$) of a matrix in $\mathcal{A}(B,d)$ and all the matrices A for which $\rho(A) = \overline{\rho}(\mathcal{A}(B,d))$(respectively, $\rho(A) = \tilde{\rho}(\mathcal{A}(B,d))$).*

The specific matrix B that we consider in this paper is the $(0,1)$-matrix Δ_n of order n with 1's everywhere on and below its main diagonal. Thus, for instance,

$$\Delta_5 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

For brevity, we will write $\mathcal{A}(n,d)$ in place of $\mathcal{A}(\Delta_n, d)$.

In our investigations we shall make important use of the Perron–Frobenius theory of nonnegative matrices (matrices each of whose entries is nonnegative) [BP79, Ga59], and for this we need the notions of reducibility and irreducibility of a matrix. Let $A = [a_{ij}]$ be a matrix of order n. Then A is *reducible* provided there exists a permutation matrix P such that

$$PAP^T = \begin{bmatrix} A_1 & O \\ A_{21} & A_2 \end{bmatrix},$$

where A_1 and A_2 are square, nonvacuous matrices. Let $D(A)$ be the digraph of A, that is, the digraph with vertices $\{1, 2, \ldots, n\}$ in which there is an arc from i to j if and only if $a_{ij} \neq 0$. Then, as is well known, A is irreducible if and only if the digraph $D(A)$ is strongly connected [BR91]. In general, there exists a permutation matrix Q such that

$$QAQ^T = \begin{bmatrix} A_1 & O & \cdots & O \\ A_{21} & A_2 & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ A_{r1} & A_{r2} & \cdots & A_r \end{bmatrix}, \tag{1}$$

where $r \geq 1$ and $A_1, A_2, \ldots, A_r$ are the *irreducible components* of A. The irreducible components of A are uniquely determined up to simultaneous permutations of their rows and columns.

We use I_n to denote the identity matrix of order n and J_n to denote the all 1's matrix of order n. In general, J denotes an all-1's matrix whose size is determined from context. If $B_1, B_2, \ldots, B_r$ are arbitrary square matrices, we define the block triangular matrix $B_1 \# B_2 \# \cdots \# B_r$ by

$$B_1 \# B_2 \# \cdots \# B_r = \begin{bmatrix} B_1 & O & \cdots & O \\ J & B_2 & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ J & J & \cdots & B_r \end{bmatrix}.$$

LEMMA 1.2. *Let A be a $(0,1)$-matrix of order n such that $\Delta_n \leq A$.*

(a) *Then* $A = A_1 \# \cdots \# A_r$, *where* $r \geq 1$, *and* $A_1, \ldots, A_r$ *are the irreducible components of* A.

(b) *If* A *is irreducible, then* A *has an irreducible, principal submatrix* B *of order* $n-1$ *such that* $\Delta_{n-1} \leq B$.

Proof. (a) In the digraph associated with the matrix A we have the path $n \to n-1 \to \cdots \to 2 \to 1$. This implies that the irreducible components of A can be obtained by partitioning A (as opposed to having first to permute simultaneously the rows and columns of A as in (1)), and (a) follows.

(b) Since $\Delta_n \leq A$, we have $\Delta_{n-1} \leq A(j \mid j)$ for $j = 1, 2, \ldots, n$. Since A is irreducible and $\Delta_n \leq A$, it easily follows that for some $k \geq 1$ there exist integers $i_1, \ldots, i_k$ such that $1 = i_1 < i_2 < \cdots < i_k \leq n-1$ such that row i_1 has a 1 in column j_1 where $j_1 > i_1$, $i_2 \leq j_1$ and row i_2 has a 1 in column j_2 where $j_2 > i_2, \ldots, i_k \leq j_{k-1}$ and row i_k has a 1 in column $j_k = n$. If $j_{k-1} = n-1$, then $A(n \mid n)$ is irreducible. If $j_{k-1} \neq n-1$, then $i_k < n-1$ and $A(n-1 \mid n-1)$ is irreducible. □

We now summarize certain parts of the Perron–Frobenius theory as applied to a nonnegative matrix A of order n.

(PFI) $\rho(A)$ is an eigenvalue of A and there exists a nonnegative eigenvector x such that $Ax = \rho(A)x$. If A is irreducible, x is a positive vector.

(PFII) Let the row sums of A be $r_1, \ldots, r_n$. Then

$$\min\{r_1, \ldots, r_n\} \leq \rho(A) \leq \max\{r_1, \ldots, r_n\}.$$

If A is irreducible and not all the row sums are equal, both of these inequalities are strict.

(PFIII) Let z be a positive vector. Then

$$Az \leq rz \text{ implies } \rho(A) \leq r$$

and

$$Az \geq rz \text{ implies } \rho(A) \geq r.$$

For each of these assertions, if A is irreducible, then $\rho(A) = r$ if and only if $Az = rz$.

(PFIV) Let C also be a nonnegative matrix and assume that $A \leq C$. Then $\rho(A) \leq \rho(C)$ with strict inequality if either A or C is irreducible and $A \neq C$.

(PFV) If A' is a proper principal submatrix of A, then $\rho(A') \leq \rho(A)$ with strict inequality if A is irreducible.

The following lemma is a special case of a method of Schwarz [Sc64] (see also Lemma 2.1 of [BS86]) used to derive the results mentioned earlier. Let E_{ij} denote a square (0,1)-matrix with a 1 in position (i, j) and 0's elsewhere; the order of E_{ij} is taken from the context in which it is used. Also e_i denotes a column vector with a 1 in position i and 0's elsewhere.

LEMMA 1.3. *Let* $A = [a_{ij}]$ *be a* $(0,1)$*-matrix of order* n *with a positive eigenvector* $x = (x_1, \ldots, x_n)^T$ *corresponding to* $\rho(A)$. *Assume that* $a_{pq} = 0$ *and* $a_{pr} = 1$ *for some* p, q, r. *Let* $C = A + E_{pq} - E_{pr}$. *Then*

(a) $\rho(C) \geq \rho(A)$ (*respectively,* $\rho(C) \leq \rho(A)$) *if* $x_q \geq x_r$ (*respectively,* $x_q \leq x_r$);

(b) *if* C *is irreducible, then* $\rho(C) > \rho(A)$ (*respectively,* $\rho(C) < \rho(A)$) *if* $x_q > x_r$ (*respectively,* $x_q < x_r$).

Proof. We have

$$Cx = Ax + E_{pq}x - E_{pr}x = \rho(A)x + (x_q - x_r)e_p.$$

Thus we have $Cx \geq \rho(A)x$ (respectively, $Cx \leq \rho(A)x$) if $x_q \geq x_r$ (respectively, $x_q \leq x_r$), and assertion (a) follows from (PFIII). If $x_q \neq x_r$, then $Cx \neq \rho(A)x$ and thus if C is irreducible, assertion (b) also follows from (PFIII). □

We conclude this introduction by defining the following operation that will be useful in what follows. We denote the symmetric permutation matrix of order n with 1's on the back diagonal, that is, in those positions (i,j) with $i+j=n+1$, by L_n. If X is a matrix of order n, then we let

$$\widehat{X} = L_n X^T L_n,$$

the matrix obtained by *flipping* X over the back diagonal. Since $\widehat{X}$ is permutation similar to X^T, $\widehat{X}$ has the same determinant, trace, and spectrum as X. If X is a (0,1)-matrix, we have $\sigma(\widehat{X}) = \sigma(X)$, and $\widehat{X} \in \mathcal{A}(n,d)$ if and only if $X \in \mathcal{A}(n,d)$.

2. Maximizing the spectral radius over $\mathcal{A}(n,d)$. Throughout this section n and d denote positive integers with $d \leq n(n-1)/2$. Recall that $\mathcal{A}(n,d)$ denotes the set of all $(0,1)$-matrices such that $\Delta_n \leq A$ and $\sigma(A-\Delta_n) = d$. Let

$$\rho_{n,d} = \max\{\rho(A) : A \in \mathcal{A}(n,d)\},$$

and let

$$\mathcal{G}(n,d) = \{A : A \in \mathcal{A}(n,d), \rho(A) = \rho_{n,d}\}.$$

The determination of $\rho_{n,d}$ and characterization of the matrices in $\mathcal{G}(n,d)$ for all n and d is an apparently difficult problem. In this section we consider the cases $d = 1,2,3,$ and 4 (all n) and arbitrary d with n sufficiently large.

We first observe the following. For each $n \geq 2$, let $B_n = \Delta_n + E_{1n}$ and $\beta_n = \rho(B_n)$. Then because

$$\det \left[\begin{array}{c|c} \begin{matrix} 1 \\ \vdots \\ 1 \end{matrix} & \Delta_{n-2} - xI_{n-2}, \\ \hline 1 & \begin{matrix} 1 & \cdots & 1 \end{matrix} \end{array}\right] = x^{n-2},$$

we have that $f_n(x) = (-1)^n \det(B_n - xI_n)$ satisfies

$$f_n(x) = (x-1)^n - x^{n-2},$$

and β_n is the largest real root of $f_n(x) = 0$.

LEMMA 2.1. *We have $\beta_n < \beta_{n+1}$ and $\lim_{n\to\infty} \beta_n = \infty$.*

Proof. Let λ be any positive solution of $f_n(x) = 0$. Then

$$\begin{aligned} f_{n+1}(\lambda) &= (\lambda-1)^{n+1} - \lambda^{n-1} \\ &= (\lambda-1)^{n+1} - (\lambda-1)\lambda^{n-2} + (\lambda-1)\lambda^{n-2} - \lambda^{n-1} \\ &= (\lambda-1)((\lambda-1)^n - \lambda^{n-2}) + \lambda^{n-2}(\lambda - 1 - \lambda) \\ &= -\lambda^{n-2} < 0. \end{aligned}$$

In particular we have $f_{n+1}(\beta_n) < 0$. Thus $f_{n+1}(x)$ has a root that is larger than β_n and therefore $\beta_{n+1} > \beta_n$.

The fact that $\lim_{n\to\infty} \beta_n = \infty$ follows from the observation that

$$(\beta_n - 1)^2 \left(\frac{\beta_n - 1}{\beta_n}\right)^{n-2} = 1. \qquad \square$$

COROLLARY 2.2. *We have* $\lim_{n\to\infty} \rho_{n,d} = \infty$.

LEMMA 2.3. *Let* n' *and* d' *also be positive integers with* $d < d' \leq n'(n'-1)/2$ *and* $n < n'$. *Then* $\rho_{n,d} < \rho_{n',d'}$.

Proof. It suffices to prove the lemma for the case $n' = n+1$ and $d' = d+1$. Let $A \in \mathcal{G}(n,d)$ and

$$A^* = \left[\begin{array}{ccc|c} & A & & O \\ \hline 1 & \cdots & 1 & 1 \end{array}\right], \quad B = \left[\begin{array}{ccc|c} & A & & e_1 \\ \hline 1 & \cdots & 1 & 1 \end{array}\right].$$

Then $\rho(A^*) = \rho(A)$, $A^* \leq B$, and $B \in \mathcal{A}(n+1, d+1)$. The matrix B is clearly irreducible and it follows from (PFIV) that $\rho(A) < \rho(B)$. Therefore $\rho_{n,d} < \rho_{n+1,d+1}$. □

By (a) of Lemma 1.2, it follows that a matrix in $\mathcal{A}(n,d)$ is irreducible if and only if it does not have a k by $n-k$ zero submatrix in its upper right corner for any integer k with $1 \leq k \leq n-1$. This implies that $A \in \mathcal{A}(n,d)$ is irreducible if and only if its flip $\widehat{A}$ is irreducible.

THEOREM 2.4. *Every matrix in* $\mathcal{G}(n,d)$ *is irreducible.*

Proof. Suppose to the contrary that there is an integer n such that for some d, $\mathcal{G}(n,d)$ contains a reducible matrix A, and let n be the smallest such integer. By the above remark there is an integer k with $1 \leq k \leq n-1$ such that

$$A = \begin{bmatrix} U & O \\ J & V \end{bmatrix},$$

where U is a square matrix of order k. Since $\rho(A) = \rho_{n,d}$, either $\rho(U) = \rho_{n,d}$ or $\rho(V) = \rho_{n,d}$. Lemma 2.3 implies that either $U = \Delta_k$ or $V = \Delta_{n-k}$. After flipping, if necessary, we may assume that $\rho(U) = \rho_{n,d}$ and $V = \Delta_{n-k}$ and hence that $\rho_{k,d} = \rho_{n,d}$. The minimality of n implies that U is irreducible.

Let $x = (y^T, z^T)^T$ be a nonnegative eigenvector of A corresponding to $\rho_{n,d}$, where $y = (x_1, \ldots, x_k)^T$ and $z = (x_{k+1}, \ldots, x_n)^T$. Then

$$\begin{aligned} Uy &= \rho_{n,d} y, \\ Jy + \Delta_{n-k} z &= \rho_{n,d} z. \end{aligned}$$

If $y = 0$, then $z \neq 0$ and $\Delta_{n-k} z = \rho_{n,d} z$ implies that $\rho_{n,d}$ is an eigenvalue of $\Delta_{n,k}$. This is a contradiction since 1 is the only eigenvalue of Δ_{n-k} and clearly $\rho_{n,d} \geq 2$. Hence $y \neq 0$ and y is a nonnegative eigenvector of U corresponding to its spectral radius. Since U is irreducible, y is a positive vector. The above equations now imply that z, and hence x, are positive vectors. Since $Ax = \rho_{n,d} x$, we have

$$\rho_{n,d}(x_n - x_j) \geq x_{k+1} + \cdots + x_n > 0 \qquad (j = 1, 2, \ldots, k)$$

so that

$$x_j < x_n \qquad (j = 1, 2, \ldots, k).$$

Since U is irreducible, there exists an integer t with $2 \leq t \leq k$ such that $a_{1t} = 1$. The matrix $B = A - E_{1t} + E_{1n}$ is an irreducible matrix in $\mathcal{A}(n,d)$ and by Lemma 1.3, $\rho(B) > \rho(A)$, which is a contradiction. □

The following corollary improves Lemma 2.3.

COROLLARY 2.5. *If* $(n,d) \leq (n',d')$ *and* $(n,d) \neq (n',d')$, *then* $\rho_{n,d} < \rho_{n',d'}$.

Proof. It suffices to prove that $\rho_{n,d} < \rho_{n+1,d}$ and that $\rho_{n,d} < \rho_{n,d+1}$ if $d < n(n-1)/2$. Let $A \in \mathcal{G}(n,d)$, which by Theorem 2.4 is irreducible. The matrix

$$B = \left[\begin{array}{ccc|c} & A & & O \\ \hline 1 & \cdots & 1 & 1 \end{array}\right]$$

is a reducible matrix in $\mathcal{A}(n+1,d)$ and hence by Theorem 2.4, $\rho_{n,d} = \rho(A) = \rho(B) < \rho_{n+1,d}$.

Now assume that $d < n(n-1)/2$. Then A has a zero entry, say $a_{pq} = 0$. The matrix $A' = A + E_{pq}$ is a matrix in $\mathcal{A}(n,d+1)$ and by (PFIV), we have $\rho_{n,d} = \rho(A) < \rho(A') \leq \rho_{n,d+1}$. □

We now show that $\mathcal{G}(n,1)$ contains a unique matrix and determine $\rho_{n,1}$ asymptotically.

THEOREM 2.6. *The following hold.*

(a) $\mathcal{G}(n,1) = \{\Delta_n + E_{1n}\}$ *for all* $n \geq 2$.

(b) $\rho_{n,1} \asymp \frac{n}{2\log n}$.

Proof. Since the only irreducible matrix in $\mathcal{A}(n,1)$ is $\Delta_n + E_{1n}$, assertion (a) is a consequence of Theorem 2.4. Moreover, it follows from our earlier calculations that $\rho_{n,1}$ satisfies $(x-1)^n = x^{n-2}$. Putting $w = 1/x$ and differentiating twice we obtain

$$(1-w)^{n-2} = \frac{2}{n(n-1)}.$$

This yields

$$\log(1-w) = \frac{1}{n-2}\log\frac{2}{n(n-1)},$$

from which it follows that

$$1 - w = \exp\left(\frac{1}{n-2}\log\frac{2}{n(n-1)}\right) = \exp\left(\frac{-\log n(n-1)/2}{n-2}\right).$$

Hence

$$1 - w \asymp \exp\left(\frac{-\log n^2}{n}\right) = \exp\frac{-2\log n}{n} \asymp 1 - \frac{2\log n}{n}$$

and (b) follows. □

For each matrix X, we let $r_i(X)$ denote row i of X and $c_j(X)$ denote column j. The following lemma is useful in investigating the structure of matrices in $\mathcal{G}(n,d)$.

LEMMA 2.7. *Let* $A = [a_{ij}]$ *be a matrix in* $\mathcal{G}(n,d)$.

(a) *For integers* p *and* q *with* $2 \leq p,q \leq n$, *if* $r_p(A) \leq r_q(A)$ *but* $r_p(A) \neq r_q(A)$, *then* $a_{ip} \leq a_{iq}$ *for* $i = 1,\ldots,\min\{p,q\} - 1$.

(b) *For integers* p *and* q *with* $1 \leq p,q \leq n-1$, *if* $c_p(A) \leq c_q(A)$ *but* $c_p(A) \neq c_q(A)$, *then* $a_{pj} \leq a_{qj}$ *for* $j = \max\{p,q\} + 1,\ldots,n$.

Proof. Let $x = (x_1,\ldots,x_n)^T$ be a nonnegative eigenvector of A corresponding to its spectral radius $\rho(A)$. By Theorem 2.4, A is irreducible and hence x is a positive vector. Since $r_p(A) \leq r_q(A)$, we have

$$\rho(A)x_q - \rho(A)x_p = (r_q(A) - r_p(A))x > 0$$

and hence $x_q > x_p$. Assertion (a) now follows from Lemma 1.3. Assertion (b) follows by applying assertion (a) to $\widehat{A}$. □

THEOREM 2.8. *If $d \leq n-2$ and A is a matrix in $\mathcal{G}(n,d)$, then $a_{1n}=1$.*

Proof. Suppose to the contrary that $a_{1n}=0$. By Theorem 2.4, A is irreducible and hence $a_{1t}=1$ for some t with $2 \leq t \leq n-1$. Since then $a_{1t} > a_{1n}$ and since $r_n(A)$ is the all-1's vector, (a) of Lemma 2.7 (with $p=t$ and $q=n$) implies that $r_t(A)=r_n(A)$. In particular, $a_{tn}=1$. Now applying (b) of Lemma 2.7 (with $p=t$ and $q=1$), we conclude that $c_t(A)=c_1(A)$ and hence $c_t(A)$ is the all-1's vector. Therefore

$$d=\sigma(A-\Delta_n) \geq \sum_{j=t+1}^{n} a_{tj} + \sum_{i=1}^{t-1} a_{it} = n-1,$$

which is a contradiction. □

In maximizing (or minimizing) the spectral radius over some class of nonnegative matrices, the order of the components of a positive eigenvector corresponding to the spectral radius of a matrix under consideration generally plays a crucial role. In many cases, such an order can be assumed in any desired form by simultaneous permutations of the rows and the columns. The class $\mathcal{A}(n,d)$ is not invariant under simultaneous row and column permutations, but nonetheless we are able to compare some of the components of a positive eigenvector of a matrix in $\mathcal{G}(n,d)$.

LEMMA 2.9. *Let numbers $r, b_0, b_1, \ldots, b_p$ be given, and let numbers $a_0, a_1, \ldots, a_p$ be defined by*

$$a_0=b_0,\ a_i=ra_{i-1}+b_i \qquad (i=1,\ldots,p).$$

Then

$$\sum_{k=0}^{p} a_k = \sum_{k=0}^{p} \left(\sum_{i=0}^{p-k} r^i \right) b_k. \tag{2}$$

Proof. Equation (2) follows from the observation that

$$a_k = r^k b_0 + r^{k-1} b_1 + \cdots + r b_{k-1} + b_k \qquad (k=0,1,\ldots,p). \qquad \square$$

LEMMA 2.10. *Let A be a matrix in $\mathcal{G}(n,d)$, and let $x=(x_1,\ldots,x_n)^T$ be a positive eigenvector of A corresponding to $\rho=\rho(A)$. Let $y=(y_1,\ldots,y_n)^T=(A-\Delta_n)x$, and let $r=(\rho-1)/\rho$. Then for each pair of integers l,m with $1 \leq l < m \leq n$, we have*

$$\rho(x_m - x_l) \geq \frac{1-r^{m-l}}{1-r} x_m - y_l. \tag{3}$$

Proof. Since $\rho x = Ax = \Delta_n x + y$, we have

$$\rho x_i = (x_1 + \cdots + x_i) + y_i \qquad (i=1,2,\ldots,n),$$

which implies that

$$\rho(x_i - x_{i-1}) = x_i + y_i - y_{i-1} \qquad (i=2,3,\ldots,n).$$

Hence

$$x_{i-1} = r x_i + \frac{1}{\rho}(-y_i + y_{i-1}) \qquad (i=2,3,\ldots,n). \tag{4}$$

Let $p = m - l - 1$. Then

$$\rho(x_m - x_l) = \sum_{j=l+1}^{m} x_j + y_m - y_l = \sum_{i=0}^{p} x_{m-i} + y_m - y_l. \tag{5}$$

We now apply Lemma 2.9 with $a_0 = b_0 = x_m$ and with $a_i = x_{m-i}$ and $b_i = (-y_{m-i+1} + y_{m-i})/\rho$ for $i = 1, 2, \ldots, p$. Since $a_i = ra_{i-1} + b_i\,(i = 1, 2, \ldots, p)$, (2) gives

$$\begin{aligned}\sum_{i=0}^{p} x_{m-i} &= \left(\sum_{i=0}^{p} r^i\right) x_m + \frac{1}{\rho}\sum_{k=1}^{p}\left(\sum_{i=0}^{p-k} r^i\right)(-y_{m-k+1} + y_{m-k}) \\ &= \left(\sum_{i=0}^{p} r^i\right) x_m + \frac{1}{\rho}\left[\left(\sum_{i=0}^{p-1} r^i\right)(-y_m) + \sum_{k=1}^{p-1} r^{p-k} y_{m-k} + y_{m-p}\right] \\ &\geq \frac{1 - r^{p+1}}{1 - r} x_m - \frac{1}{\rho}\frac{1 - r^p}{1 - r} y_m. \qquad (6)\end{aligned}$$

Therefore from (5) and (6) we get

$$\rho(x_m - x_l) \geq \frac{1 - r^{p+1}}{1 - r} x_m + \left(1 - \frac{1}{\rho}\frac{1 - r^p}{1 - r}\right) y_m - y_l.$$

The inequality (3) now follows since $p + 1 = m - l$ and

$$1 - \frac{1}{\rho}\frac{1 - r^p}{1 - r} = 1 - (1 - r^p) = r^p > 0. \qquad \square$$

We now introduce some new notation. If M is a square matrix of order n, then $M(x)$ denotes the matrix $M - xI_n$. In addition, if $i_1, \ldots, i_t$ and $j_1, \ldots, j_t$ are each sets of t distinct elements of $\{1, 2, \ldots, n\}$, then $M(i_1, \ldots, i_t \mid j_1, \ldots, j_t)$ denotes the square matrix of order $n - t$ obtained from M by deleting rows $i_1, \ldots, i_t$ and columns $j_1, \ldots, j_t$.

LEMMA 2.11. *Let $A = [a_{ij}]$ be a matrix in $\mathcal{A}(n, d)$ such that $a_{13} + a_{14} + \cdots + a_{1n} = d$ and $a_{1,n-1} = a_{1n} = 1$. Let $B = A - E_{1,n-1} + E_{2n} = [b_{ij}]$. Then A and B have the same spectrum.*

Proof. Expanding the determinant of $B(x)$ along its first row and noticing that $b_{12} = 0$ we get

$$\begin{aligned}\det B(x) &= (1 - x)\det B(x)(1 \mid 1) + \sum_{j=3}^{n}(-1)^{1+j} b_{1j} \det B(x)(1 \mid j) \\ &= (1 - x)\det B(x)(1 \mid 1) \\ &\quad + \sum_{j=3}^{n}(-1)^{1+j} a_{1j} \det B(x)(1 \mid j) - (-1)^n \det B(x)(1 \mid n - 1).\end{aligned}$$

Let $\tilde{C}_n = \Delta_n + (1 - x)\sum_{i=1}^{n-1} E_{i,i+1}$. We then have

(i) $\det B(x)(1 \mid 1) = \det A(x)(1 \mid 1) + (-1)^n \det \tilde{C}_{n-2}$,

(ii) $\det B(x)(1 \mid j) = \det A(x)(1 \mid j)$ $(j = 3, \ldots, n)$, and

(iii) $\det B(x)(1 \mid n - 1) = \det A(x)(1 \mid n - 1) = (1 - x)\det \tilde{C}_{n-2}$.

The validity of (ii) follows from the fact that for each $j = 3, \ldots, n$, the first two columns of $B(x)(1,2 \mid j,n)$ are identical and hence $\det B(x)(1,2 \mid j,n) = 0$. Using (i)–(iii) in the above equation for $\det B(x)$ and the observation that $a_{12} = 0$, we obtain

$$\det B(x) = (1-x)\det A(x)(1 \mid 1) + \sum_{j=3}^{n}(-1)^{1+j}a_{1j}\det A(x)(1 \mid j) = \det A(x). \qquad \square$$

The following lemma can be proved in a similar way; hence we omit the proof.

LEMMA 2.12. *Let*

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

Then the two matrices

$$\Delta_n + \begin{bmatrix} O & J_2 \\ O & O \end{bmatrix} \quad \text{and} \quad \Delta_n + \begin{bmatrix} O & K \\ O & O \end{bmatrix}$$

have the same spectrum.

Now we are ready to determine the structure of matrices in $\mathcal{G}(n,d)$ for $d = 2, 3,$ and 4.

THEOREM 2.13. *The following hold.*

(a) $\mathcal{G}(n,2) = \{\Delta_n + E_{1n} + E_{1,n-1}, \Delta_n + E_{1n} + E_{2n}\}$ *for all* $n \geq 4$.

(b) $\mathcal{G}(n,3) = \{\Delta_n + E_{1,n-1} + E_{1n} + E_{2n}\}$ *for all* $n \geq 5$.

(c) $\mathcal{G}(n,4) = \{\Delta_n + E_{1,n-1} + E_{1n} + E_{2n} + E_{pq} : (p,q) = (1,n-2), (2,n-1), (3,n)\}$ *for all* $n \geq 7$.

Proof. Before proceeding with the individual cases (a)–(c), we first consider some common conclusions.

Let $A = [a_{ij}]$ be a matrix in $\mathcal{G}(n,d)$ where $2 \leq d \leq 4$, and let $Z = A - \Delta_n = [z_{ij}]$. The restrictions on n in (a)–(c) allow us to conclude from Theorem 2.8 that $z_{1n} = 1$. Suppose that there exist integers s and t with $s > 1$ and $t < n$ such that $z_{st} = 1$. If $z_{sn} = 0$, then Lemma 2.7 implies that $r_t(A) = (1, \ldots, 1)$ and hence $z_{tn} = 1$. We conclude that either $z_{sn} = 1$ or $z_{tn} = 1$, and hence $z_{2n} + z_{3n} + \cdots + z_{n-1,n} \geq 1$. Similarly, we get $z_{1s} = 1$ or $z_{1t} = 1$ and that $z_{12} + z_{13} + \cdots + z_{1,n-1} \geq 1$. But then Z has at least four 1's and hence $d \geq 4$. Therefore if $d \leq 3$, we have $Z(1 \mid n) = O$.

Case $d = 2, n \geq 4$. We have $z_{1n} = 1$, and by flipping if necessary we may assume that $z_{1t} = 1$ for some t with $2 \leq t \leq n-1$. Since $r_2(A) \leq r_3(A) \leq \cdots \leq r_n(A)$ and no two of these rows are identical, Lemma 2.7 implies that $t > n-2$ and hence $t = n-1$. Thus $Z = E_{1,n-1} + E_{1n}$ and (a) now follows.

Case $d = 3, n \geq 5$. We have $z_{1n} = 1$ and partition Z as

$$Z = \begin{bmatrix} u & 1 \\ W & v \end{bmatrix}, \tag{7}$$

where $u = (z_{11}, z_{12}, \ldots, z_{1,n-1})$ and $v = (z_{2n}, z_{3n}, \ldots, z_{nn})^T$. Suppose that $v = 0$. Then arguing as in the previous case, Lemma 2.7 implies that $z_{1,n-2} = z_{1,n-1} = z_{1n} = 1$. Let $B = A - E_{1,n-1} + E_{2n} = [b_{ij}]$. By Lemma 2.11, B is in $\mathcal{G}(n,3)$. Since $n \geq 5$ and $r_{n-2}(B) \leq r_{n-1}(B)$ but $r_{n-2}(B) \neq r_{n-1}(B)$ and since $b_{1,n-2} = 1 > 0 = b_{1,n-1}$, we contradict Lemma 2.7. Thus $v \neq 0$ and in a similar way we conclude that $u \neq 0$, and hence u and v each contain exactly one 1.

We now suppose that $z_{1,n-1} = 0$ and obtain a contradiction. Then $z_{1t} = 1$ for some t with $2 \leq t \leq n-2$. If $z_{tn} = 0$, then the fact that $a_{1t} = z_{1t} = 1 > 0 = z_{1,n-1} =$

$a_{1,n-1}$ leads to a contradiction of Lemma 2.7. Hence $z_{tn} = 1$. If $n = 5$, then up to flipping A is one of the two matrices

$$A_1 = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

But a calculation[1] shows that $\rho(A_1) = 3.5115\ldots$ and $\rho(A_2) = 3.5346\ldots$ while $\rho(A) \geq \rho(\Delta_5 + E_{14} + E_{15} + E_{25}) = 3.5605\ldots$, giving a contradiction. Now assume that $n \geq 6$. By flipping A if necessary we may assume that $t \leq \lfloor (n+1)/2 \rfloor$. Let $x = (x_1, x_2, \ldots, x_n)^T$ be a positive eigenvector of A corresponding to $\rho = \rho(A)$. Then $\rho(x_n - x_{n-1}) = x_n$ so that $x_{n-1} = rx_n$ where $r = (\rho - 1)/\rho$. Applying Lemma 2.10 we get

$$\begin{aligned} \rho(x_{n-1} - x_t) &\geq \frac{1 - r^{n-1-t}}{1 - r} x_{n-1} - x_n \\ &= (1 + r + \cdots + r^{n-2-t}) x_{n-1} - x_n \\ &\geq (r(1+r) - 1) x_n \end{aligned}$$

since $n \geq 6$ implies that $n - 2 - t \geq n - 2 - \lfloor (n+1)/2 \rfloor \geq 1$. Now $\rho \geq \rho_{6,3} > \rho_{6,1} = \rho(\Delta_6 + E_{16}) = 3.1479\ldots > 3$, and thus

$$r(1+r) - 1 > \left(\frac{2}{3}\right)^2 + \frac{2}{3} - 1 > 0.$$

Hence $x_{n-1} > x_t$. Let $C = A - E_{1t} + E_{1,n-1}$. Then C is a matrix in $\mathcal{A}(n,3)$ and $\rho(C) > \rho(A)$ by Lemma 1.3, contradicting the fact that A is in $\mathcal{G}(n,3)$. This contradiction implies that $z_{1,n-1} = 1$, and in a similar way we get $z_{2,n} = 1$. Therefore $Z = E_{1,n-1} + E_{1n} + E_{2n}$ and (b) follows.

Case $d = 4, n \geq 7$. As in the previous case, we can show that in the partition (7) of Z we have $u \neq 0$ and $v \neq 0$. We consider two subcases.

Subcase $W \neq O$. Since u and v are not zero and since $d = 4$, we see that u, v, and W each contain exactly one 1. In particular, there exists a unique pair of integers (s,t) with $s > 1$ and $t < n$ such that $z_{st} = 1$. By Lemma 2.7, either $z_{sn} = 1$ or $r_t(A) = (1, \ldots, 1)$. We first show that $t = n - 1$. Suppose to the contrary that $t \leq n-2$. If $z_{sn} = 0$, then $r_t(A) = (1, \ldots, 1)$ contradicts the fact that W contains only one 1. If $z_{sn} = 1$, then $r_t(A) \leq r_{t+1}(A), r_t(A) \neq r_{t+1}(A)$, and $a_{st} = 1 > 0 = a_{s,t+1}$ together contradict Lemma 2.7. Thus $t = n - 1$ and in a similar way we conclude that $s = 2$. Since either $z_{2n} = 1$ or $r_{n-1}(A) = (1, \ldots, 1)$ and either $z_{1,n-1} = 1$ or $c_2(A) = (1, \ldots, 1)$, we now conclude that up to flipping, Z equals one of the matrices

$$Z_0 = \begin{bmatrix} O & J_2 \\ O & O \end{bmatrix}, \quad Z_1 = \left[\begin{array}{cc|c|cc} 0 & 1 & & 0 & 1 \\ 0 & 0 & O & 1 & 0 \\ \hline \multicolumn{2}{c|}{O} & O & \multicolumn{2}{c}{O} \\ \hline \multicolumn{2}{c|}{O} & O & 0 & 1 \\ & & & 0 & 0 \end{array}\right], \quad Z_2 = \left[\begin{array}{cc|c|cc} 0 & 1 & & 0 & 1 \\ 0 & 0 & O & 1 & 1 \\ \hline \multicolumn{2}{c|}{O} & O & \multicolumn{2}{c}{O} \end{array}\right].$$

[1] We relied on MATLAB for all our calculations.

If $Z = Z_1$, then we contradict Lemma 2.7 since $a_{12} > a_{1,n-1}$ and since $r_2(A) \leq r_{n-1}(A)$ holds without equality. Thus $Z \neq Z_1$. We next show that $Z \neq Z_2$. This is true for $n = 7$ and 8 since

$$\rho(\Delta_7 + Z_2) = 4.3504\ldots < 4.3846\ldots = \rho(\Delta_7 + Z_0)$$

and

$$\rho(\Delta_8 + Z_2) = 4.6133\ldots < 4.6762\ldots = \rho(\Delta_8 + Z_0).$$

Assume that $n \geq 9$, and suppose to the contrary that $Z = Z_2$. Let

$$Z_3 = \left[\begin{array}{cc|c|cc} 0 & 1 & & 1 & 1 \\ 0 & 0 & O & 0 & 1 \\ \hline \multicolumn{2}{c|}{O} & O & \multicolumn{2}{c}{O} \end{array}\right],$$

and let $A_3 = \Delta_3 + Z_3$. Then A_3 is in $\mathcal{A}(n,4)$. Since $A_3 = A + E_{1,n-1} - E_{2,n-1}$ and $c_1(A) = c_2(A)$, A_3 can be obtained from A by permuting rows 1 and 2 and columns 1 and 2. Hence $\rho(A_3) = \rho(A)$ and A_3 is in $\mathcal{G}(n,4)$. Let $x = (x_1, \ldots, x_n)$ be a positive eigenvector of A_3 corresponding to $\rho = \rho(A_3) = \rho(A)$. Then we have $x_{n-1} = rx_n$ and $x_{n-2} = rx_{n-1}$, and hence $x_{n-2} = r^2 x_n$, where $r = (\rho - 1)/\rho$. We also have from Lemma 2.10 that

$$\begin{aligned} \rho(x_{n-2} - x_2) &\geq \frac{1 - r^{n-4}}{1 - r} x_{n-2} - x_n \\ &= (r^2(1 + r + \cdots + r^{n-5}) - 1)x_n \\ &\geq (r^2 + r^3 + r^4 - 1)x_n, \end{aligned}$$

from which we get $x_{n-2} > x_2$ because

$$r^2 + r^3 + r^4 > \left(\frac{3}{4}\right)^2 + \left(\frac{3}{4}\right)^3 + \left(\frac{3}{4}\right)^4 > 1,$$

since $\rho > 4$. But then, as in (c), $A_3 - E_{12} + E_{1,n-2}$ is a matrix in $\mathcal{A}(n,4)$ with $\rho(A_3 - E_{12} + E_{1,n-2}) > \rho(A)$, which is a contradiction. Thus $Z \neq Z_2$, and we conclude that $Z = Z_0$ in this subcase; that is, $A = \Delta_n + E_{1,n-1} + E_{1,n} + E_{2,n-1} + E_{2,n}$.

Subcase $W = O$. Since $u \neq 0$ and $v \neq 0$, we have $\sigma(u) = 2$ and $\sigma(v) = 1$ or $\sigma(u) = 1$ and $\sigma(v) = 2$. After flipping if necessary we assume that $\sigma(u) = 2$ and $\sigma(v) = 1$. We first observe that Lemma 2.7 implies that if for some integers p and j with $2 \leq p < j \leq n - 1$ we have $z_{1p} = 1$ and $z_{1j} = 0$, then $z_{pn} = 1$. Let the two 1's in u be in positions k and t where $2 \leq k < t \leq n - 1$. Then $t = n - 1$ for otherwise $z_{kn} = z_{tn} = 1$, contradicting $\sigma(v) = 1$.

We now show that $k = n - 2$. Suppose, to the contrary, that $k \leq n - 3$; hence by the above, $z_{kn} = 1$. Consider first the case $4 \leq k \leq n - 3$. Then by flipping A we obtain $\widehat{A} = [\hat{a}_{ij}]$, where

$$c_n(\widehat{A}) = (1, 1, 0, z_{1,n-3}, \ldots, z_{1k} = 1, \ldots, z_{14}, 0, 0, 1)^T \quad \text{and} \quad \hat{a}_{1k} = 1, \hat{a}_{1,n-1} = 0.$$

Let $\widehat{x} = (\hat{x}_1, \ldots, \hat{x}_n)^T$ be a positive eigenvector of $\widehat{A}$ corresponding to $\rho = \rho(\widehat{A}) = \rho(A)$. As before let $r = (\rho - 1)/\rho$ and let $\widehat{y} = (\widehat{A} - \Delta_n)\widehat{x} = (\hat{y}_1, \ldots, \hat{y}_n)^T$. Then

$\hat{y}_k = \hat{x}_n$ because $r_k(\widehat{A} - \Delta_n) = (0, \ldots, 0, 1)$. Since also $\hat{x}_{n-1} = r\hat{x}_n$, we get from Lemma 2.10 that

$$\begin{aligned}\rho(\hat{x}_{n-1} - \hat{x}_k) &\geq \left(r\frac{1 - r^{n-1-k}}{1-r} - 1\right)\hat{x}_n \\ &\geq (r + r^2 - 1)\hat{x}_n > 0,\end{aligned}$$

$k \leq n-3$, and $\rho > 3$. But then by Lemma 1.3, the matrix $\widehat{A} - E_{1k} + E_{1,n-1}$ is a matrix in $\mathcal{A}(n,4)$ whose spectral radius is greater than ρ, contradicting $A \in \mathcal{G}(n,4)$. We now obtain a contradiction in the remaining case $2 \leq k \leq 3$. Let W_2 and W_3 be the square matrices of order n defined by

$$W_2 = \left[\begin{array}{cc|c|cc} 0 & 1 & & 1 & 1 \\ 0 & 0 & O & 0 & 1 \\ \hline \multicolumn{2}{c|}{O} & O & \multicolumn{2}{c}{O} \end{array}\right], \quad W_3 = \left[\begin{array}{ccc|c|ccc} 0 & 0 & 1 & & 0 & 1 & 1 \\ 0 & 0 & 0 & O & 0 & 0 & 0 \\ 0 & 0 & 0 & & 0 & 0 & 1 \\ \hline \multicolumn{3}{c|}{O} & O & \multicolumn{3}{c}{O} \end{array}\right],$$

and let $M_i = \Delta_n + W_i$ $(i = 2, 3)$. Then $A = M_2$ if $k = 2$ and $A = M_3$ if $k = 3$. Let

$$M = \Delta_n + \begin{bmatrix} O & J_2 \\ O & O \end{bmatrix}.$$

Then calculation shows that

$$n = 7: \quad \rho(M) = 4.3846\ldots, \ \rho(M_2) = 4.3504\ldots, \ \rho(M_3) = 4.3141\ldots;$$

$$n = 8: \quad \rho(M) = 4.6762\ldots, \ \rho(M_2) = 4.6133\ldots, \ \rho(M_3) = 4.5719\ldots.$$

Thus $\rho(M) > \rho(M_i)$ $(i = 2, 3)$ for $n = 7$ and $n = 8$, which is a contradiction. Hence $k \neq 2, 3$ if $n = 7$ or 8. Assume that $n \geq 9$. Let $x = (x_1, \ldots, x_n)^T$ be a positive eigenvector of A corresponding to $\rho = \rho(A)$, and again let $r = (\rho - 1)/\rho$. As before we have $x_{n-2} = r^2 x_n$ and by Lemma 2.10,

$$\rho(x_{n-2} - x_k) \geq \left(\frac{r^2(1 - r^{n-2-k})}{1-r} - 1\right)x_n.$$

Since $n \geq 9$ and $k \leq 3$, we have $n - 2 - k \geq 4$ and hence

$$\frac{r^2(1 - r^{n-2-k})}{1-r} - 1 \geq \frac{r^2(1 - r^4)}{1-r} - 1 = r^2 + r^3 + r^4 - 1 > 0$$

because $\rho > 4$. It thus follows that $x_{n-2} > x_k$ and hence $\rho(A - E_{1k} + E_{1,n-2}) > \rho(A)$, which is a contradiction. We have thus proved that $k = n-2$ and thus that $r_1(Z) = (0, \ldots, 0, 1, 1, 1)$.

To complete the proof we show that $z_{2n} = 1$. Flipping A to get $\widehat{A} = [\hat{a}_{ij}]$ we have $c_n(\widehat{A}) = (1, 1, 1, 0, \ldots, 0, 1)^T$. We need to show that $\hat{a}_{1,n-1} = 1$ or, equivalently, that $\hat{a}_{12} = \cdots = \hat{a}_{1,n-2} = 0$. Let $\widehat{x} = (\hat{x}_1, \ldots, \hat{x}_n)^T$ be a positive eigenvector of $\widehat{A}$ corresponding to ρ. Then $\hat{x}_{n-1} = r\hat{x}_n$, where $r = (\rho - 1)/\rho$, and for each $j = 2, \ldots, n-2$, we have

$$\rho(\hat{x}_{n-1} - \hat{x}_j) = \left(\frac{r(1 - r^{n-1-j})}{1-r} - \epsilon_j\right)\hat{x}_n,$$

where $\epsilon_j = 1$ if $2 \leq j \leq 3$ and $\epsilon_j = 0$ if $4 \leq j \leq n-2$. From this it can be proved that $\hat{x}_{n-1} > \hat{x}_j$ $(j = 2, \ldots, n-2)$ for all $n \geq 6$. It thus follows from Lemma 1.3 that none of $\hat{a}_{12}, \hat{a}_{13}, \ldots, \hat{a}_{1,n-2}$ can be equal to one and hence that $\hat{a}_{1,n-1} = 1$, as desired. Thus in this subcase Z equals

$$\left[\begin{array}{c|ccc} O & 1 & 1 & 1 \\ & 0 & 0 & 1 \\ \hline O & \multicolumn{3}{c}{O} \end{array}\right]$$

or the matrix obtained by flipping. Since by Lemma 2.12, the matrices

$$\Delta_n + \begin{bmatrix} O & J_2 \\ O & O \end{bmatrix} \quad \text{and} \quad \Delta_n + \left[\begin{array}{c|ccc} O & 1 & 1 & 1 \\ & 0 & 0 & 1 \\ \hline O & \multicolumn{3}{c}{O} \end{array}\right]$$

have the same spectral radius. The proof is complete. □

Let A be a matrix in $\mathcal{A}(n, d)$, and let $Z = A - \Delta_n = [z_{ij}]$. Then A has a *staircase pattern in the upper right corner*, abbreviated SPURC, provided that

$$z_{i1} \leq z_{i2} \leq \cdots \leq z_{in} \qquad (i = 1, 2, \ldots, n)$$

and

$$z_{1j} \geq z_{2j} \geq \cdots \geq z_{nj} \qquad (j = 1, 2, \ldots, n).$$

Each of the matrices in $\mathcal{G}(n, d)$ in the statement of Theorem 2.13 has a SPURC. However, it is not always the case that each matrix in $\mathcal{G}(n, d)$ has a SPURC. For example, one can check that

$$G_1 = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad \text{and} \quad G_2 = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

are matrices in $\mathcal{G}(4, 2)$ and $\mathcal{G}(5, 3)$, respectively, but neither has a SPURC. The SPURC property does hold in the following asymptotic sense.

THEOREM 2.14. *Let d be a positive integer. Then for n sufficiently large, each matrix in $\mathcal{G}(n, d)$ has a SPURC.*

Proof. Let $A = [a_{ij}]$ be a matrix in $\mathcal{G}(n, d)$, and let $\rho = \rho(A)$ and $r = (\rho - 1)/\rho$. By Corollary 2.2, by choosing n large enough we can make ρ as large as we wish and r as close to 1 as we wish. Choose n so large that

(i) $n \geq 9d + 3$,

(ii) $r + r^2 + \cdots + r^{3d} > 2d$, and

(iii) $\rho \geq 2d$.

Let $x = (x_1, \ldots, x_n)^T$ be a positive eigenvector of A corresponding to ρ, let $Z = A - \Delta_n = [z_{ij}]$, and let $y = Zx = (y_1, \ldots, y_n)^T$. For each $k = 1, 2, \ldots, n$ we have $y_k \leq dx_n$ because

$$y_k = \sum_{j=1}^{n} z_{kj} x_j \leq \sum_{j=1}^{n} z_{kj} x_n \leq dx_n.$$

Let $l = \min\{j \mid z_{1j} = 1\}$. We show that $l > n/2$. If $z_{1,l+1} = \cdots = z_{1n} = 1$, then since $d \geq \sum_{j=l}^{n} z_{1j} = n - l + 1$, we have $l \geq n + 1 - d > n/2$ (by (i)) and we are done. Assume that there is a $j > l$ such that $z_{ij} = 0$, and let m be the largest integer such that $z_{1m} = 0$. Then $0 < n - m \leq d$. We prove that $m - l \leq n/3$. Suppose to the contrary that $m - l > n/3$. By Lemma 2.10 we have

$$\begin{aligned} \rho(x_m - x_l) &\geq \frac{1 - r^{m-l}}{1 - r} x_m - y_l \\ &\geq (1 + r + \cdots + r^{m-l-1}) x_m - d x_n. \end{aligned} \tag{8}$$

From

$$x_{m+1} + \cdots + x_n - y_m = \rho(x_n - x_m)$$

we get

$$\begin{aligned} \rho x_m &= \rho x_n - x_{m+1} - \cdots - x_n + y_m \\ &\geq \rho x_n - (n - m) x_n \end{aligned}$$

so that

$$x_m \geq \frac{\rho - (n - m)}{\rho} x_n. \tag{9}$$

Combining (8) and (9) we get

$$\rho(x_m - x_l) \geq \left(\frac{\rho - (n - m)}{\rho} (1 + r + \cdots + r^{m-l-1}) - d \right) x_n.$$

From (i) we get $m - l - 1 > n/3 - 1 \geq 3d$, and then from (iii) we get

$$\frac{\rho - (n - m)}{\rho} \geq \frac{\rho - d}{\rho} \geq \frac{1}{2}.$$

Hence using (ii) we get

$$\begin{aligned} \frac{\rho - (n - m)}{\rho} (1 + r + \cdots + r^{m-l-1}) - d &> \frac{1 + r + \cdots + r^{3d}}{2} - d \\ &> \frac{1 + 2d}{2} - d > 0. \end{aligned}$$

We can now conclude that $x_m > x_l$. This fact, together with $a_{1l} = z_{1l} = 1 > 0 = z_{1m} = a_{1m}$, contradicts Lemma 1.3. This proves that $m - l \leq n/3$. Now

$$n - l = n - m + m - l \leq d + \frac{n}{3} < \frac{n}{2},$$

and hence we have $l > n/2$.

In a similar way we can prove that

$$\begin{aligned} \min\{j \mid z_{ij} = 1\} &> \frac{n}{2} \qquad (i = 1, 2, \ldots, n), \\ \max\{i \mid z_{ij} = 1\} &< \frac{n}{2} \qquad (j = 1, 2, \ldots, n). \end{aligned}$$

Hence it follows that

$$Z \leq \begin{bmatrix} O & J_{\lfloor n/2 \rfloor} \\ O & O \end{bmatrix}. \tag{10}$$

Then since $r_j(A) \leq r_{j+1}(A), r_j(A) \neq r_{j+1}(A)\,(j = \lfloor n/2 \rfloor + 1, \ldots, n-1)$, we get from Lemma 2.7 and from (10) that

$$z_{i1} \leq z_{i2} \leq \cdots \leq z_{in} \qquad (i = 1, 2, \ldots, n);$$

in a similar way we get

$$z_{1j} \geq z_{2j} \geq \cdots \geq z_{nj} \qquad (j = 1, 2, \ldots, n).$$

Thus A has a SPURC. □

3. Minimizing the spectral radius over $\mathcal{A}(n, d)$. Throughout this section n and d again denote positive integers with $d \leq n(n-1)/2$. Here we are concerned with the minimum spectral radius of matrices in $\mathcal{A}(n, d)$ and the characterization of those matrices with the minimum spectral radius. Let

$$\mu_{n,d} = \min\{\rho(A) : A \in \mathcal{A}(n, d)\},$$

and let

$$\mathcal{H}(n, d) = \{A : A \in \mathcal{A}(n, d), \rho(A) = \mu_{n,d}\}.$$

Clearly, $\mu_{n',d} \leq \mu_{n,d}$ and $\mu_{n,d'} \leq \mu_{n,d}$ if $n' \geq n$ and $d' \leq d$.

We first derive an analogue for $\mu_{n,d}$ of the results of Schwarz described in the introduction.

LEMMA 3.1. *Let P and Q be permutation matrices of order n. Then the matrix obtained from $P\Delta_n Q$ by first* (i) *moving all the* 1*'s in each row to the left and then* (ii) *moving all the* 1*'s in each column to the bottom equals Δ_n.*

Proof. The row sums of Δ_n are $1, 2, \ldots, n$, and hence the row sums of $P\Delta_n Q$ are $1, 2, \ldots, n$ in some order. The matrix B obtained from $P\Delta_n Q$ by applying (i) also has row sums $1, 2, \ldots, n$ in some order and hence there exists a permutation matrix R such that $RB = \Delta_n$. Therefore applying (ii) to B yields the same matrix as applying (ii) to Δ_n, that is, yields Δ_n. □

COROLLARY 3.2. *Let A be a matrix in $\mathcal{A}(n, d)$ and let P and Q be permutation matrices of order n. Then the matrix obtained from PAQ by first* (i) *moving all the* 1*'s in each row to the left and then* (ii) *moving all the* 1*'s in each column to the bottom is also in $\mathcal{A}(n, d)$.*

THEOREM 3.3. *There exists a matrix in $\mathcal{H}(n, d)$ such that in each row all the* 1*'s precede the* 0*'s and in each column all the* 0*'s precede the* 1*'s.*

Proof. Let A be a matrix in $\mathcal{H}(n, d)$. Since $\Delta_n \leq A$, it follows from Lemma 1.2 that for some $r \geq 1$, $A = A_1 \# \cdots \# A_r$, where $A_1, A_2, \ldots, A_r$ are the irreducible components of A. Each A_i has a positive eigenvector corresponding to its spectral radius $\rho(A_i)$. For each integer i with $1 \leq i \leq r$, let Q_i be a permutation matrix such that the components of the positive eigenvector of $Q_i A_i Q_i^T$ corresponding to $\rho(A_i)$ are in nondecreasing order. Applying Lemma 1.3 (and its analogue for columns) to $Q_i A_i Q_i^T$ we obtain a matrix B_i such that in each row all the 1's precede the 0's and in each column all the 0's precede the 1's and $\rho(B_i) \leq \rho(A_i)\,(i = 1, 2, \ldots, r)$. Let

$B = B_1 \# \cdots \# B_r$. It follows from Corollary 3.2 that $\Delta_n \leq B$ and hence $B \in \mathcal{A}(n,d)$. Since

$$\rho(B) = \min\{\rho(B_i) : i = 1,2,\ldots,r\} \leq \min\{\rho(A_i) : i = 1,2,\ldots,r\} = \mu(n,d),$$

we conclude that $\rho(B) = \mu_{n,d}$ and hence B is in $\mathcal{H}(n,d)$. □

We next investigate $\mu_{n,d}$ and $\mathcal{H}(n,d)$ for $d \leq n-1$. Throughout the remainder of this section, we let H_3 and H_4 denote the (0, 1) lower Hessenberg matrices of orders 3 and 4, respectively, and we let $B_4 = \Delta_4 + E_{14}$. Each of these matrices is given as follows along with their spectral radii:

$$H_3 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \quad \rho(H_3) = \tfrac{3+\sqrt{5}}{2};$$

$$H_4 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad \rho(H_4) = 3;$$

$$B_4 = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad \rho(B_4) = \tfrac{3+\sqrt{5}}{2}.$$

The characteristic polynomial of H_3 equals $x(x^2 - 3x + 1)$ and that of B_4 equals $(x^2 - 3x + 1)(x^2 - x + 1)$. This implies that the spectral radii of H_3 and B_4 equal $(3 + \sqrt{5})/2$ as given above.

The following lemma is a direct consequence of Lemma 1.3.

LEMMA 3.4. *Let $A = [a_{ij}]$ be a $(0,1)$-matrix of of order n having a positive eigenvector corresponding to $\rho(A)$. Suppose that for some i, j, and l we have $a_{ij} = 0$ and $a_{il} = 1$ and $r_j(A) \leq r_l(A)$ but $r_j(A) \neq r_l(A)$. If the matrix $B = A + E_{ij} - E_{il}$ is irreducible, then $\rho(B) < \rho(A)$.*

LEMMA 3.5. *Let A be a matrix in $\mathcal{H}(n,d)$.*

(a) *If $d = 1$, then $\rho(A) = 2$ and $A = \Delta_n + \sum_{i=1}^{n-1} \epsilon_i E_{i,i+1}$, where each $\epsilon_i = 0$ or 1 and $\epsilon_1 + \cdots + \epsilon_{n-1} = 1$.*

(b) *If $(n,d) = (3,2)$, then $\rho(A) = (3 + \sqrt{5})/2$ and $A = H_3$ up to permutation similarity.*[2]

(c) *If $(n,d) = (4,2)$, then $\rho(A) = 2$ and $A = J_2 \# J_2$.*

(d) *If $(n,d) = (4,3)$, then $\rho(A) = 3$ and $A = J_1 \# J_3$, $A = J_3 \# J_1$, or $A = H_4$ up to permutation similarity.*[3]

Proof. Assertion (a) is almost immediate, and assertion (b) follows from the fact that all matrices in $\mathcal{A}(3,2)$ are permutation similar and have spectral radius equal to $(3+\sqrt{5})/2$. Now assume that $(n,d) = (4,2)$. If $A \neq J_2 \# J_2$, then A has an irreducible component of order 3 or 4 with minimum row sum at least 2 (but not all row sums equal 2), and hence by (PFII), $\rho(A) > 2$. Since $\rho(J_2 \# J_2) = 2$, $A = J_2 \# J_2$, and assertion (c) follows.

[2] There are three matrices in $\mathcal{H}(3,2)$.

[3] There are four matrices in $\mathcal{H}(4,2)$.

Finally, assume that $(n, d) = (4, 3)$. The only reducible matrices in $\mathcal{A}(4, 3)$ are $J_1 \# J_3$ and $J_3 \# J_1$, each of which has spectral radius equal to 3. In particular, $\mu_{4,3} \leq 3$. Suppose that A is irreducible. It is now easy to check using (PFII) that the only irreducible matrices in $\mathcal{H}(4, 3)$ with spectral radius at most 3 are the matrices H_4 and the matrix

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

which is permutation similar to H_4. Since $\rho(H_4) = 3$, assertion (d) now follows. □

LEMMA 3.6. *For each $n \geq 4$ we have $\mu_{n,n-1} = 3$.*

Proof. Let $n = 3t + a$, where $t \geq 1$ and $1 \leq a \leq 3$, and let $B = J_3 \# \cdots \# J_3 \# K$, where the number of J_3's is t and where

$$K = \begin{cases} J_1 & \text{if } a = 1, \\ J_2 & \text{if } a = 2, \\ H_3 & \text{if } a = 3. \end{cases}$$

Then B is in $\mathcal{A}(n, n-1)$ and $\rho(B) = 3$. Thus $\mu_{n,n-1} \leq 3$. We prove $\mu_{n,n-1} \geq 3$ by induction on n. By (d) of Lemma 3.5, we have $\mu_{4,3} = 3$. Assume that $n \geq 5$ and let $A = [a_{ij}]$ be a matrix in $\mathcal{A}(n, n-1)$. Then A satisfies at least one of the following properties.

(i) $\sigma(r_1(A - \Delta_n)) \leq 1$.
(ii) $\sigma(r_2(A - \Delta_n)) = 0$.
(iii) $\sigma(r_1(A - \Delta_n)) \geq 2$ and $\sigma(r_2(A - \Delta_n)) \geq 1$.

If (i) holds, then $A(1 \mid 1) \geq F$ for some matrix F in $\mathcal{A}(n-1, n-2)$ and hence by (PFIV), (PFV), and the inductive assumption, $\rho(A) \geq \rho(F) \geq 3$. If (ii) holds, then $A(2 \mid 2) \geq G$, where G is in $\mathcal{A}(n-1, n-2)$ and again $\rho(A) \geq 3$. If (iii) holds, then each row of A contains at least three 1's and hence $\rho(A) \geq 3$ by (PFII). Therefore $\rho(A) \geq 3$ for each matrix in $\mathcal{A}(n, n-1)$ and hence $\mu_{n,n-1} \geq 3$. □

COROLLARY 3.7. *If $n \geq 4$ and $d \leq n - 1$, then $\mu_{n,d} \leq 3$.*

Assume that $n \geq 4$ and $d \leq n - 1$. In view of the above corollary, no matrix A in $\mathcal{A}(n, d)$ with $\rho(A) > 3$ belongs to $\mathcal{H}(n, d)$. To investigate matrices in $\mathcal{H}(n, d)$ we now identify some classes of matrices whose spectral radii are greater than 3. Any matrix of $\mathcal{A}(n, d)$ having an irreducible component that belongs to one of these classes cannot belong to $\mathcal{H}(n, d)$.

LEMMA 3.8. *If $A = [a_{ij}]$ is an irreducible $(0, 1)$-matrix of order 6 such that $\Delta_6 \leq A$, then $\rho(A) > 3$.*

Proof. If $a_{16} = 1$, then $\rho(A) \geq \rho(\Delta_6 + E_{16}) > 3$. We now assume that $a_{16} = 0$ and consider a number of cases.

Case (i). $a_{15} = 1$. Then $a_{k6} = 1$ for some $k \in \{2, 3, 4, 5\}$ and thus $A \geq \Delta_6 + E_{15} + E_{k6}$. If $k = 5$, then $\Delta_6 + E_{15} + E_{k6}$ is permutation similar to $\Delta_6 + E_{16} + E_{56}$, and hence $\rho(A) \geq \rho(\Delta_6 + E_{16}) > 3$. If $2 \leq k \leq 4$, then by Lemma 3.4 applied to $\widehat{A}$ we have

$$\rho(A) \geq \rho(\Delta_6 + E_{15} + E_{k6}) \geq \rho(\Delta_6 + E_{15} + E_{46}) = 3.1497\ldots > 3.$$

Case (ii). $a_{26} = 1$. We apply Case (i) to $\widehat{A}$.

Case (iii). $a_{15} = a_{26} = 0$ and $a_{14} = 1$. Then $a_{k6} = 1$ for some $k \in \{3, 4, 5\}$ and hence $A \geq \Delta_6 + E_{14} + E_{k6}$. If $k \leq 4$, then $\rho(A) > 3$ since $\rho(\Delta_6 + E_{14} + E_{36}) = 3.0907\ldots$

and $\rho(\Delta_6 + E_{14} + E_{46}) = 3.2201\ldots$. If $k = 5$, then there is an $i \in \{2,3,4\}$ such that $a_{i5} = 1$. Then permuting the last two rows and last two columns of A, we may apply one of the previous cases to conclude that $\rho(A) > 3$.

Case (iv). $a_{15} = a_{26} = 0$ and $a_{36} = 1$. We apply the previous case to $\widehat{A}$.

Case (v). $a_{15} = a_{14} = a_{26} = a_{36} = 0$ and $a_{13} = 1$. There exists a $k \in \{4,5\}$ such that $a_{k6} = 1$. Since A is irreducible, we also have $a_{st} = 1$ for some $(s,t) \in \{(2,4),(2,5),(3,4),(3,5)\}$. Then $A \geq \Delta_6 + E_{13} + E_{st} + E_{k6}$. First suppose that $k = 4$ and let $X_{st} = \Delta_6 + E_{13} + E_{st} + E_{46}$. Then $\rho(X_{24}) = 3.2092\ldots$, $\rho(X_{25}) = 3.1479\ldots$, $\rho(X_{34}) = 3.2695\ldots$, and $\rho(X_{35}) = 3.2092\ldots$ and hence $\rho(A) > 3$. Now suppose that $k = 5$. Since A is irreducible, at least one of a_{25}, a_{35}, and a_{45} equals 1. That $\rho(A) > 3$ follows by applying the previous cases to the matrix obtained from A by interchanging rows 5 and 6 and columns 5 and 6.

Case (vi). $a_{15} = a_{14} = a_{13} = a_{26} = a_{36} = 0$ and $a_{46} = 1$. We apply the previous case to $\widehat{A}$.

Case (vii). $a_{15} = a_{14} = a_{13} = a_{26} = a_{36} = a_{46} = 0$. Since A is irreducible, $a_{12} = a_{56} = 1$. In this case at least one of a_{25}, a_{35}, and a_{45} equals 1, and by an argument similar to that used in Case (v), we can show that $\rho(A) > 3$. □

THEOREM 3.9. *Let A be an irreducible $(0,1)$-matrix of order $n \geq 6$ such that $\Delta_n \leq A$. Then $\rho(A) > 3$.*

Proof. It follows by repeated application of (b) of Lemma 1.2 that A has an irreducible, principal submatrix of order 6. Hence $\rho(A) > 3$ follows from Lemma 3.8 and (PFV). □

COROLLARY 3.10. *If $1 \leq d \leq n-1$ and A is a matrix in $\mathcal{H}(n,d)$, then for some r, $A = A_1 \# \cdots \# A_r$ where each A_i is an irreducible matrix of order at most 5.*

Proof. This is a direct consequence of Corollary 3.7 and Theorem 3.9. □

THEOREM 3.11. *Let $1 \leq d \leq n-1$. Then*

$$\mu_{n,d} = \begin{cases} 2 & \text{if } 1 \leq d \leq \lfloor \frac{n}{2} \rfloor, \\ \frac{3+\sqrt{5}}{2} & \text{if } \lfloor \frac{n}{2} \rfloor + 1 \leq d \leq \lfloor \frac{2n}{3} \rfloor, \\ 3 & \text{if } \lfloor \frac{2n}{3} \rfloor + 1 \leq d \leq n-1. \end{cases} \tag{11}$$

Proof. Let A be a matrix in $\mathcal{A}(n,d)$. First suppose that $1 \leq d \leq \lfloor n/2 \rfloor$. Then A has an irreducible component of order at least 2 and hence $\mu_{n,d} \geq 2$ by (PFII). Since there exists a matrix in $\mathcal{A}(n,d)$ each of whose irreducible components has order at most 2, $\mu_{n,d} = 2$.

Now suppose that $\lfloor n/2 \rfloor + 1 \leq d \leq \lfloor 2n/3 \rfloor$. Then A has an irreducible component of order at least 3 and, by Corollary 3.10, no irreducible component of order greater than 5. There exists a matrix in $\mathcal{A}(n,d)$ each of whose irreducible components is contained in $\{H_3, J_2, J_1\}$ and hence $\mu_{n,d} \leq (3+\sqrt{5})/2$. It can be checked by computation that if X is an irreducible $(0,1)$-matrix of order 4 with $\Delta_4 \leq X$ or X is an irreducible matrix of order 5 with $\Delta_5 \leq X$, then $\rho(X) \geq (3+\sqrt{5})/2$ with equality if and only if $X = B_4$. It now follows easily that $\mu_{n,d} = (3+\sqrt{5})/2$.

Finally, suppose that $\lfloor 2n/3 \rfloor + 1 \leq d \leq n-1$. Then again A has an irreducible component of order at least 3 and no irreducible component of order greater than 5. If at least one component of A equals J_3, then $\rho(A) \geq 3$. Suppose that A has no components of order 3. Then the lower bound on d implies that either A has a component of order 5 with at least four 1's or A has a component of order 4 with at least three 1's. Since Lemma 3.6 implies that $\mu_{5,4} = \mu_{4,3} = 3$, we have $\rho(A) \geq 3$.

There exists a matrix in $\mathcal{A}(n,d)$ each of whose irreducible components belongs to $\{J_3, J_2, J_1\}$, and hence we now conclude that $\mu_{n,d} = 3$. □

To conclude we determine the minimum spectral radius $\mu_{n,n}$ of matrices in $\mathcal{A}(n,n)$. In the next theorem we shall make use of the the matrices below given with their characteristic polynomials and spectral radii:

$$T_4 = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad x^2(x^2 - 4x + 2), \quad \rho(T_4) = 2 + \sqrt{2};$$

$$T_5 = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \quad x^3(x^2 - 5x + 5), \quad \rho(T_5) = \frac{5 + \sqrt{5}}{2}.$$

THEOREM 3.12. *Let n be an integer with $n \geq 3$. Then*

$$\mu_{n,n} = \begin{cases} 3 & \text{if } n \equiv 0 \pmod 3, \\ 2 + \sqrt{2} & \text{if } n \neq 2, 5 \text{ and } n \equiv 1 \text{ or } 2 \pmod 3, \\ \frac{5+\sqrt{5}}{2} & \text{if } n = 5. \end{cases} \tag{12}$$

Proof. The matrices $J_3 \oplus \cdots \oplus J_3$, $J_3 \oplus \cdots \oplus J_3 \oplus T_4$, $J_3 \oplus \cdots \oplus J_3 \oplus T_4 \oplus T_4$, and T_5 show that the values for $\mu_{n,n}$ given in the statement of the theorem are upper bounds for $\mu_{n,n}$. Since $\mu_{n,n} \geq \mu_{n,n-1}$, Theorem 3.11 now implies that $\mu_{n,n} = 3$ if $n \equiv 0 \pmod 3$. For the remaining two cases we let $A = A_1 \# \cdots \# A_r$ be a matrix in $\mathcal{H}(n,n)$ where $A_1, \ldots, A_r$ are the irreducible components of A with orders $n_1, \ldots, n_r$, respectively.

Case $n \equiv 1 \pmod 3$. By Theorem 3.3 we may also assume that in each row of A all the 1's precede the 0's and in each column all the 0's precede the 1's. Each irreducible component of A also has this property and hence has only 1's on its superdiagonal.

First assume that $\sigma(A_i - \Delta_i) = n_i$ for all i. Suppose that $n_j \geq 5$ for some j. Then A_j is an irreducible matrix and contains a principal submatrix B such that $T_4 \leq B$. From (PFIV) and (PFV) we conclude that $\rho(A) > \rho(T_4)$, which is a contradiction. Hence $n_i \leq 4$ for all i. Since $\sigma(A - \Delta_n) = n$ and $n \equiv 1 \pmod 3$, there is a k such that n_k equals 4. Since $T_4 \leq A_k$ or $\widehat{T}_4 \leq A$, it follows from (PFIV) that $\rho(A) \geq \rho(A_k) \geq \rho(T_4)$.

Now assume that it is not the case that $\sigma(A_i - \Delta_i) = n_i$ for all i. Then $\sigma(A_k - \Delta_k) > n_k$ for some k where $n_k \geq 4$. As above it follows that $\rho(A) \geq \rho(T_4)$.

Case $n \equiv 2 \pmod 3$. First assume that $n = 5$. If A is reducible, then it is easy to check that $\rho(A) > 4$, which is a contradiction. Thus A is an ireducible matrix with $\Delta_5 \leq A$ having 1's on its superdiagonal and one additional 1. Except for T_5 and $\widehat{T}_5$, there is only one other such matrix and its spectral radius is greater than that of T_5. If $n \geq 8$, it follows as above that $\rho(A) \geq \rho(T_4)$. □

REFERENCES

[BP79] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[BH85] R. A. BRUALDI AND A. J. HOFFMAN, *On the spectral radius of* $(0,1)$*-matrices*, Linear Algebra Appl., 65 (1985), pp. 133–146.

[BR91] R. A. BRUALDI AND H. J. RYSER, *Combinatorial Matrix Theory*, Cambridge University Press, London and New York, 1991.

[BS86] R. A. BRUALDI AND E. S. SOLHEID, *On the spectral radius of complementary acyclic matrices of zeros and ones*, SIAM J. Alg. Discrete Methods, 7 (1986), pp. 265–272.

[BS87] ———, *On the minimum spectral radius of matrices of zeros and ones*, Linear Algebra Appl., 85 (1987), pp. 81–100.

[Ch90] LI CHING, *A bound on the spectral radius of matrices of zeros and ones*, Linear Algebra Appl., 132 (1990), pp. 179–183.

[Fr85] S. FRIEDLAND, *The maximal eigenvalue of* 0-1 *matrices with prescribed number of ones*, Linear Algebra. Appl., 69 (1985), pp. 33–69.

[Ga59] F. R. GANTMACHER, *The Theory of Matrices, Vol.* 2, K. A. Hirsch, transl., Chelsea, New York, 1959.

[Sc64] B. SCHWARZ, *Rearrangements of square matrices with non-negative elements*, Duke Math. J., 31 (1964), pp. 45–62.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 3, pp. 509–514, July 1996

004

ON THE CONDITION BEHAVIOUR IN THE JACOBI METHOD*

ZLATKO DRMAČ†

Abstract. The aim of this note is to show that the matrix $S(n,\alpha) = (1-\alpha)I + \alpha ee^T$, $e = (1,\ldots,1)^T$, $\alpha \in (0,1)$ is not a counterexample for the accuracy properties of the Jacobi method for computing the singular and eigenvalue decomposition, as might be understood from a recent article of Mascarenhas in this journal. In fact, the Jacobi process on $S(n,\alpha)$ is an example of the perfect behaviour of the algorithm. It is shown that Jacobi rotations preserve the optimal (with respect to diagonal scalings) spectral condition number of $S(n,\alpha)$.

Key words. Jacobi method, accuracy, condition behaviour, optimal scaling

AMS subject classifications. 65F15, 65G05

1. Introduction.

Jacobi's method is more accurate than QR. This was stated by Demmel and Veselić [2], where the claim about the accuracy of the Jacobi method was dependent on the behaviour of the condition numbers during the process. For the reader's convenience, we give a short description of the Jacobi method for computing the spectral decomposition of positive definite matrices. Let $n \in \mathbf{N}$ and let

$$H = LL^T \in \mathbf{R}^{n\times n}$$

be positive definite. The *Jacobi process* is defined by

$$H^{(0)} = H, \quad H^{(k+1)} = (U^{(k)})^T H^{(k)} U^{(k)}, \quad k = 0,1,2,\ldots,$$

where $U^{(k)}$, $k = 0,1,2,\ldots$ are chosen from the orthogonal group $\mathcal{O}(n)$. Each $U^{(k)}$ is given by *pivot position* $(p,q) = \mathcal{P}(k)$, which depends on chosen *pivot strategy* $\mathcal{P}$: $\mathbf{N} \to \{(p,q),\ 1 \le p < q \le n\}$, and by the parameter (angle) $\phi_k \in [-\pi/4, \pi/4]$, as follows:

$$\begin{bmatrix} (U^{(k)})_{pp} & (U^{(k)})_{pq} \\ (U^{(k)})_{qp} & (U^{(k)})_{qq} \end{bmatrix} = \begin{bmatrix} \cos\phi_k & \sin\phi_k \\ -\sin\phi_k & \cos\phi_k \end{bmatrix},$$
$$(U^{(k)})_{ij} = \delta_{ij}, \quad (i,j) \notin \{(p,p),(p,q),(q,p),(q,q)\},$$

where δ_{ij} denotes Kronecker's symbol. The angle ϕ_k is chosen to satisfy

$$(H^{(k+1)})_{pq} = 0,$$

i.e., for $(H^{(k)})_{pq} \neq 0$ (see [5]),

$$\cot 2\phi_k = \frac{(H^{(k)})_{qq} - (H^{(k)})_{pp}}{2(H^{(k)})_{pq}}, \quad \tan\phi_k = \frac{\operatorname{sign}\cot 2\phi_k}{|\cot 2\phi_k| + \sqrt{1+\cot^2 2\phi_k}}.$$

For suitably chosen pivot strategy $\mathcal{P}$ we have a *convergent process*—there are

$$\Lambda = \operatorname{diag}(\lambda_1,\ldots,\lambda_n) = \lim_{k\to\infty} H^{(k)}, \quad U = \prod_{k=0}^{\infty} U^{(k)} \in \mathcal{O}(n),$$

* Received by the editors January 19, 1994; accepted for publication (in revised form) by F. T. Luk May 30, 1995. This research was supported by National Science Foundation grant ACS-9357812, Department of Energy grant DE-FG03-94ER25215, and Intel Corporation.

† Department of Computer Science, University of Colorado, Boulder, CO 80309-0430 (zlatko@cs.colorado.edu).

and $H = U\Lambda U^\tau$ is the spectral decomposition of H. For $k \in \mathbf{N}\bigcup\{0\}$ let

$$H^{(k)} = \Delta^{(k)} H_S^{(k)} \Delta^{(k)}, \quad (H_S^{(k)})_{ii} = 1, \quad 1 \le i \le n.$$

According to the theory of Demmel and Veselić, the spectral condition numbers $\kappa_2(H_S^{(k)}) = \sigma_{\max}(H_S^{(k)})/\sigma_{\min}(H_S^{(k)})$, $k \in \mathbf{N}\bigcup\{0\}$ determine the relative accuracy of the computed eigenvalues. Here $\sigma_{\min}(\cdot)$ and $\sigma_{\max}(\cdot)$ denote the minimal and the maximal singular values of a matrix, respectively. If the elements of the initial matrix H have relative uncertainties (which is the case in the applications), then the moderate size of the quotient

$$\mu = \mu(H, \mathcal{P}) = \max_{k \in \mathbf{N}} \frac{\kappa_2(H_S^{(k)})}{\kappa_2(H_S^{(0)})}$$

gives sense to the first sentence of this section. In extensive numerical testing Demmel and Veselić have never measured μ above 1.82. Explaining the excellent behaviour of μ is an important open problem; see [2], [1]. Wang [7] used the matrix $S(n,\alpha)$ (see (2.1) below) and obtained μ considerably larger than 1.82. Recently, Mascarenhas [6] used $S(n,\alpha)$ for the construction of a Jacobi process with condition growth of the order $n/4$. Thus, the matrix $S(n,\alpha)$ started to play the role of a counterexample for the relative accuracy properties of the Jacobi method. *Is that really so?*

The aim of this note is to show that the Jacobi process on $S(n,\alpha)$ is an example of the perfect behaviour of the Jacobi method and that the search for a real counterexample remains open.

2. The matrix $S(n,\alpha)$. Let $n \in \mathbf{N}$, $\alpha \in (0,1)$. The matrix $S(n,\alpha)$ is defined by

$$S(n,\alpha) = (1-\alpha) I_n + \alpha e e^\tau, \quad e = (1,\dots,1)^\tau. \tag{2.1}$$

It was noted in [4] that $S(3,\alpha)$ is optimally scaled with respect to diagonal scalings, i.e., no scaling $DS(3,\alpha)D$, D diagonal and nonsingular, can decrease its spectral condition number. Because of the importance of $S(n,\alpha)$ in the theoretical investigation of the Jacobi process, we prove the optimal scaling property here. We use the technique and characterization developed in [4].

THEOREM 2.1. *For $n \in \mathbf{N}$, $\alpha \in (0,1)$ let the matrix $S(n,\alpha)$ be defined by* (2.1). *Then*

$$\kappa_2(S(n,\alpha)) = \min\{\kappa_2(DS(n,\alpha)D), \quad D \in \mathcal{D}\},$$

where $\mathcal{D}$ denotes the set of $n \times n$ diagonal nonsingular matrices.

Proof. It is easily seen that the spectrum of $S(n,\alpha)$ is given by

$$\lambda_1(S(n,\alpha)) = \cdots = \lambda_{n-1}(S(n,\alpha)) = 1-\alpha, \quad \lambda_n(S(n,\alpha)) = 1 + (n-1)\alpha.$$

The eigenspace of $\lambda_n(S(n,\alpha))$ is spanned by $e = \sum_{i=1}^n e_i$, and if we look at the intersection of this eigenspace and the Euclidean unit sphere $S(0,1) = \{x \in \mathbf{R}^n, \ \|x\|_2 \equiv \sqrt{x^\tau x} = 1\}$, we get two points, $\pm\tilde{e} = \pm\frac{1}{\sqrt{n}} e$. Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the minimal and the maximal eigenvalues of a matrix, respectively. The subspace belonging to $\lambda_{\min}(S(n,\alpha))$ is obviously the hyperplane $e^\perp$, and therefore the normed eigenvectors from $e^\perp$ build the sphere $S^{n-2}(0,1) = e^\perp \bigcap S(0,1)$. Let $D = \operatorname{diag}(d_{11},\dots,d_{nn}) \in \mathcal{D}$ and let $S_D(n,\alpha) = DS(n,\alpha)D$ be the scaled matrix. Since for $x \neq \mathbf{0}$

$$S_D(n,\alpha)x = \lambda x \iff S(n,\alpha)(Dx) = \lambda D^{-2}(Dx),$$

we have

$$\lambda_{\min}(S_D(n,\alpha)) = \min_{\|x\|_2=1} \frac{x^T S(n,\alpha)x}{x^T D^{-2}x}$$
$$\leq \min_{x\in S^{n-2}(0,1)} \frac{x^T S(n,\alpha)x}{x^T D^{-2}x} = \frac{\lambda_{\min}(S(n,\alpha))}{\max\limits_{x\in S^{n-2}(0,1)} x^T D^{-2}x},$$
$$\lambda_{\max}(S_D(n,\alpha)) = \max_{\|x\|_2=1} \frac{x^T S(n,\alpha)x}{x^T D^{-2}x}$$
$$\geq \max_{x\in\{\pm\tilde{e}\}} \frac{x^T S(n,\alpha)x}{x^T D^{-2}x} = \frac{\lambda_{\max}(S(n,\alpha))}{\min\limits_{x\in\{\pm\tilde{e}\}} x^T D^{-2}x}.$$

Therefore,

$$\frac{\lambda_{\max}(S_D(n,\alpha))}{\lambda_{\min}(S_D(n,\alpha))} \geq \frac{\lambda_{\max}(S(n,\alpha))}{\lambda_{\min}(S(n,\alpha))} \frac{\max\limits_{x\in S^{n-2}(0,1)} x^T D^{-2}x}{\min\limits_{x\in\{\pm\tilde{e}\}} x^T D^{-2}x}.$$

Note that

$$x \in \{\pm\tilde{e}\} \Longrightarrow x^T D^{-2}x = \sum_{i=1}^{n} \frac{x_i^2}{d_{ii}^2} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{d_{ii}^2}.$$

On the other hand, if we choose n vectors in $S^{n-2}(0,1)$,

$$x_i = \frac{1}{\sqrt{2}}(e_i - e_{i+1}), \quad 1 \leq i \leq n-1, \quad x_n = \frac{1}{\sqrt{2}}(e_1 - e_n)$$

and compute

$$x_i^T D^{-2} x_i = \frac{1}{2}\left(\frac{1}{d_{ii}^2} + \frac{1}{d_{i+1,i+1}^2}\right), \quad 1 \leq i \leq n-1,$$
$$x_n^T D^{-2} x_n = \frac{1}{2}\left(\frac{1}{d_{11}^2} + \frac{1}{d_{n,n}^2}\right),$$

we immediately see that

$$\max_{x\in S^{n-2}(0,1)} x^T D^{-2}x \geq \min_{x\in\{\pm\tilde{e}\}} x^T D^{-2}x,$$

i.e.,

$$\kappa_2(S_D(n,\alpha)) \geq \kappa_2(S(n,\alpha)).$$

Thus, $\kappa_2(S(n,\alpha)) = \min\{\kappa_2(DS(n,\alpha)D),\ D \in \mathcal{D}\}$. □

Remark 2.2. Since $S(n,\alpha)^T S(n,\alpha) = (1+(n-1)\alpha^2)S(n, \frac{(n-2)\alpha^2+2\alpha}{1+(n-1)\alpha^2})$, it also holds that

$$\kappa_2(S(n,\alpha)) = \min\{\kappa_2(S(n,\alpha)D),\ D \in \mathcal{D}\},$$

due to the homogeneity of $\kappa_2(\cdot)$ and the properties of the spectral norm.

Wang [7] used $S(n,\alpha)$ as an example of the matrix for which the scaled condition during the Jacobi process grows more than Demmel and Veselić's 1.82 growth factor. Recently, Mascarenhas [6] described a strategy with the condition[1] growth factor $n/4$. *Although used as a counterexample for the good behaviour of the Jacobi rotations, the matrix $S(n,\alpha)$ is actually an example of the perfect behaviour of Jacobi rotations.* We shall explain our claim. For the reader's convenience we restrict ourselves to the technically simple part of the analysis. We consider two pivot strategies: the well-known row-cyclic strategy $((1,2),\dots,(1,n),(2,3),\dots,(2,n),\dots,(n-1,n))$ and the strategy defined by Mascarenhas [6].

For the sake of simplicity, let $n>2$ be even. In that case the optimal scaling of $S(n,\alpha)$ is obvious, at least in the theory of Forsythe and Straus. Indeed,

$$x_M=(1,1,\dots,1)^T,\quad x_m=(1,-1,\dots,1,-1)^T \tag{2.2}$$

are eigenvectors corresponding to $\lambda_{\max}(S(n,\alpha))$ and $\lambda_{\min}(S(n,\alpha))$, respectively, and the corresponding eigenspaces are *not separable by the set of nonsingular diagonal matrices.* (See Definition 1 and Theorems 1 and 3 in [4].)

THEOREM 2.3. *Let the Jacobi method with the row-cyclic strategy be applied on $S=S(n,\alpha)$ and let the Jacobi rotations be followed by the swapping of pivot rows and columns (see below). Then only $n-1$ rotations are needed to finish the diagonalization process. Furthermore, if $S^{(k)}$, $1\le k\le n-1$ denote corresponding matrices generated by the method, then for $1\le k\le (n-2)/2$*

$$\kappa_2(S^{(k)})=\min\{\kappa_2(DS^{(k)}D),\ \ D\in\mathcal{D}\}.$$

Proof. After k steps of the row-cyclic strategy (with certain permutations), the obtained matrix has form $S^{(k)}=(1-\alpha)I_k\oplus\hat{S}^{(k)}$, where

$$\hat{S}^{(k)}=\begin{bmatrix} 1+k\alpha & \sqrt{1+k}\alpha & \cdots & \cdots & \sqrt{1+k}\alpha \\ \sqrt{1+k}\alpha & 1 & \alpha & \cdots & \alpha \\ \vdots & \alpha & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & 1 & \alpha \\ \sqrt{1+k}\alpha & \alpha & \cdots & \alpha & 1 \end{bmatrix}.$$

This will be clear if we consider rotation, which annihilates the $(1,2)$ position of $\hat{S}^{(k)}$. Let

$$U^{(k)}=\begin{bmatrix} c_k & s_k \\ -s_k & c_k \end{bmatrix}\oplus I_{n-k-2},\quad c_k=\frac{\sqrt{1+k}}{\sqrt{2+k}},\quad s_k=-\frac{1}{\sqrt{2+k}}.$$

Then

$$(U^{(k)})^T\hat{S}^{(k)}U^{(k)}=\begin{bmatrix} 1+(k+1)\alpha & 0 & \sqrt{2+k}\alpha & \cdots & \sqrt{2+k}\alpha \\ 0 & 1-\alpha & 0 & \cdots & 0 \\ \sqrt{2+k}\alpha & 0 & 1 & \alpha & \alpha \\ \vdots & \vdots & \alpha & 1 & \alpha \\ \sqrt{2+k}\alpha & 0 & \alpha & \alpha & 1 \end{bmatrix},$$

[1] Mascarenhas used $1/\sigma_{\min}(\cdot)$ as a condition number. This can be misleading because n overestimates the spectral norm of the scaled matrix up to n times (roughly).

and if we swap the first two columns and rows we obtain the matrix $(1-\alpha)\oplus\hat{S}^{(k+1)}$. Now we prove that for $k \le (n-2)/2$ the matrices $S^{(k)}$ are optimally scaled; i.e., any diagonal symmetric scaling would increase the spectral condition number of $S^{(k)}$. Obviously, the problem reduces to the optimal scaling of $\hat{S}^{(k)}$. Consider the vectors

$$x_{\max} = (\sqrt{1+k}, 1, 1, \ldots, 1)^\tau \in \mathbf{R}^{n-k},$$
$$x_{\min} = (\sqrt{1+k}, \underbrace{-1, -1, \ldots, -1}_{1+k}, \underbrace{1, -1}, \ldots, \underbrace{1, -1})^\tau \in \mathbf{R}^{n-k}.$$

They belong to the eigenspaces of $\lambda_{\max}(\hat{S}^{(k)})$ and $\lambda_{\min}(\hat{S}^{(k)})$, respectively, and are reflections of each other; i.e., $|(x_{\max})_j| = |(x_{\min})_j|$, $1 \le j \le n-k$. Now by Theorem 3 in [4] the matrix $\hat{S}^{(k)}$ is optimally scaled. Thus, Jacobi rotations preserve the optimal scaling property and, because of orthogonality, the value of the optimal spectral condition number. □

The case $(n-2)/2 < k < n-1$ is technically not easy to handle. As an illustration, we shortly analyse the case $S = S(4,\alpha)$. After rotating at pivot positions $(1,2)$ and $(2,3)$, respectively, we have (up to a certain permutation)

$$S^{(1)} = \begin{bmatrix} 1-\alpha & 0 & 0 & 0 \\ 0 & 1+\alpha & \alpha\sqrt{2} & \alpha\sqrt{2} \\ 0 & \alpha\sqrt{2} & 1 & \alpha \\ 0 & \alpha\sqrt{2} & \alpha & 1 \end{bmatrix}, \quad S^{(2)} = \begin{bmatrix} 1-\alpha & 0 & 0 & 0 \\ 0 & 1-\alpha & 0 & 0 \\ 0 & 0 & 1+2\alpha & \alpha\sqrt{3} \\ 0 & 0 & \alpha\sqrt{3} & 1 \end{bmatrix}.$$

The lower right 3×3 submatrix of $S^{(1)}$ is optimally scaled. On the other hand, the lower right 2×2 submatrix of $S^{(2)}$ is not optimally scaled, but its standard scaling (diagonals to one) is the optimal choice. Furthermore, rotating at $(3,4)$ in $S^{(1)}$ gives

$$S'^{(2)} = \begin{bmatrix} 1-\alpha & 0 & 0 & 0 \\ 0 & 1+\alpha & 2\alpha & 0 \\ 0 & 2\alpha & 1+\alpha & 0 \\ 0 & 0 & 0 & 1-\alpha \end{bmatrix},$$

which is an optimally scaled matrix.

In the case of the strategy defined by Mascarenhas, the situation is much simpler. Mascarenhas uses $n = 2^l$ and the following strategy, which we will call $\mathcal{P}_M$.

(i) Partition $H = H^\tau \in \mathbf{R}^{n\times n}$ by

$$H = \begin{bmatrix} H_{1,1} & H_{1,2} \\ (H_{1,2})^\tau & H_{2,2} \end{bmatrix}, \quad H_{1,1}, H_{2,2} \in \mathbf{R}^{n/2\times n/2}, \tag{2.3}$$

and first choose $n/2$ pivot pairs from the main diagonal of the $(1,2)$ block in (2.3), i.e., $(k, n/2+k)$, $1 \le k \le n/2$. Next, the remaining pivot positions in the blocks $(1,2)$ and $(1,1)$ in (2.3) are chosen in any order.

(ii) If $n > 2$, then apply (i) recursively on the $(2,2)$ block in (2.3). (In the case $H = S(n,\alpha)$, (i) reduces to pivoting at $(k, n/2+k)$, $1 \le k \le n/2$.)

THEOREM 2.4. *Let $S^{(k)}$, $1 \le k \le n-1$ denote matrices from the Jacobi process with pivot strategy $\mathcal{P}_M$ and $S^{(0)} = S(n,\alpha)$, $n = 2^l > 2$. Then $S^{(n-1)}$ is diagonal and for $1 \le k \le n-2$*

$$\kappa_2(S^{(k)}) = \min\{\kappa_2(DS^{(k)}D), \quad D \in \mathcal{D}\}.$$

Proof. Recall that $S^{(k)}$ is obtained after rotating by $U^{(k-1)}$ at the pivot position $(k, n/2+k)$ and with the angle $\pi/4$. It is easily seen that $S^{(k)} = (1-\alpha)I_k \oplus \tilde{S}^{(k)}$ with some $\tilde{S}^{(k)}$ and that $\tilde{S}^{(n/2)} = (1+\alpha)S(n/2, 2\alpha/(1+\alpha))$. Hence it is sufficient to consider the case $1 \le k \le (n-2)/2$. If x_m and x_M are as in (2.2) then

$$x_m^{(k)} = (U^{(k-1)})^T \cdots (U^{(0)})^T x_m, \quad x_M^{(k)} = (U^{(k-1)})^T \cdots (U^{(0)})^T x_M$$

are eigenvectors corresponding to $\lambda_{\min}(S^{(k)})$ and $\lambda_{\max}(S^{(k)})$, respectively. Furthermore, it can be shown that $|(x_m^{(k)})_j| = |(x_M^{(k)})_j|$, $1 \le j \le n$, and the application of Theorem 3 from [4] completes the proof. □

The spectral condition number remains optimal the entire time with respect to diagonal scalings, and in the last step it drops to one.

Acknowledgement. The material presented in this paper is a part of the author's Ph.D. thesis [3], written under the supervision of Professor K. Veselić at the Department of Mathematics, University of Hagen. The author thanks Professor J. Barlow, Professor J. Demmel, and Professor K. Veselić for their comments. He also thanks anonymous referees for their constructive reports.

REFERENCES

[1] J. DEMMEL, *Open Problems in Numerical Linear Algebra*, LAPACK Working Note 47, Computer Science Division and Mathematics Department, University of California, Berkeley, CA, 1992.

[2] J. DEMMEL AND K. VESELIĆ, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.

[3] Z. DRMAČ, *Computing the Singular and the Generalized Singular Values*, Ph.D. thesis, Lehrgebiet Mathematische Physik, University of Hagen, D-58084 Hagen, Germany, 1994.

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

[5] C. G. J. JACOBI, *Über ein leichtes Verfahren die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen*, Crelle's J. für Reine und Angew. Math., 30 (1846), pp. 51–95.

[6] W. F. MASCARENHAS, *A note on Jacobi being more accurate than QR*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 215–218.

[7] X. WANG, *A few propositions on upper bounds for* $\max_{0 \le m \le M} \kappa(A_m)/\kappa(A_0)$, University of Hagen, D-58084 Hagen, Germany, 1991, unpublished manuscript.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 3, pp. 515–529, July 1996

005

AN OPERATOR RELATION OF THE USSOR AND THE JACOBI ITERATION MATRICES OF A p-CYCLIC MATRIX *

DIMITRIOS NOUTSOS†

Abstract. Let the Jacobi matrix B associated with the linear system $Ax = b$ be a weakly cyclic matrix, generated by the cyclic permutation $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_p)$ as this is defined by Li and Varga. The same authors derived the corresponding functional equation connecting the eigenvalues λ of the unsymmetric successive overrelaxation (USSOR) iteration matrix $T_{\omega\hat{\omega}}$ and the eigenvalues μ of the Jacobi matrix B extending previous results by Gong and Cai. In this paper, the validity of an analogous matrix relationship connecting the operators $T_{\omega\hat{\omega}}$ and B is proved. Moreover, the "equivalence" of the USSOR method and a certain two-parametric p-step method for the solution of the initial system is established. The tool for the proof of our main result is elementary graph theory.

Key words. USSOR method, p-cyclic matrices, graph theory, matrix relationship

AMS subject classification. 65F10

1. Introduction.

Let us consider the matrix $A \in \mathbb{C}^{n,n}$ and let us suppose that it is partitioned into $p \times p$ blocks where its diagonal blocks are square and nonsingular. For the solution of the linear system

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

we consider the unsymmetric successive overrelaxation (USSOR) iterative method

$$x^{(m+1)} = T_{\omega\hat{\omega}} x^{(m)} + c, \qquad m = 0, 1, 2, \ldots, \tag{1.2}$$

where $x^{(0)} \in \mathbb{C}^n$ is arbitrary, and ω and $\hat{\omega}$ are the overrelaxation parameters. The iteration matrix $T_{\omega\hat{\omega}}$ is given by

$$T_{\omega\hat{\omega}} = (I - \hat{\omega}U)^{-1}[(1-\hat{\omega})I + \hat{\omega}L](I - \omega L)^{-1}[(1-\omega)I + \omega U], \tag{1.3}$$

where L and U are, respectively, the strictly lower and the strictly upper block triangular parts of the block Jacobi matrix B and the vector c is given by

$$c = (\omega + \hat{\omega} - \omega\hat{\omega})(I - \hat{\omega}U)^{-1}(I - \omega L)^{-1} b. \tag{1.4}$$

Let the associated block Jacobi matrix B be a weakly cyclic matrix generated by the cyclic permutation $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_p)$. This definition given by Li and Varga [9] is as follows.

DEFINITION. *The $p \times p$ block matrix B is a weakly cyclic matrix, generated by the cyclic permutation $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_p)$, if there exists a permutation $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_p)$ of the integers $\{1, 2, \ldots, p\}$ such that*

$$B_{\sigma_j \sigma_{j+1}} \not\equiv 0,\ j = 1(1)p, \text{ and } B_{ij} \equiv 0 \text{ otherwise}, \tag{1.5}$$

* Received by the editors May 13, 1994; accepted for publication (in revised form) by R. Freund June 13, 1995.

† Department of Mathematics, University of Ioannina, GR-451 10, Ioannina, Greece (dnoutsos@cc.uoi.gr).

where $\sigma_{p+1} = \sigma_1$.

We remark here that the well-known definition for the consistently ordered matrix ([16] and [21]) is derived from the one above with $\sigma = (p, p-1, p-2, \ldots, 1)$, while that of the $(q, p-q)$-generalized consistently ordered $(q, p-q)$-GCO matrix ([2], [7], and [4]) is derived from the permutation $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_p)$, where $\sigma_{j+1} = p - q + \sigma_j$ or $\sigma_{j+1} = \sigma_j - q$ such that $1 \leq \sigma_j \leq p, j = 1(1)p$. So, the definition (1.5) is the most general for the family of p-cyclic matrices. It is obvious that the graph of the block matrix B is a cycle as this is also noted in [9].

Li and Varga [9] derived the functional equation

$$
\begin{aligned}
&[\lambda - (1-\omega)(1-\hat{\omega})]^p \\
&\quad = (\omega + \hat{\omega} - \omega\hat{\omega})^{2k} \lambda^k [\lambda\omega + \hat{\omega} - \omega\hat{\omega}]^{|\zeta_L| - k} [\lambda\hat{\omega} + \omega - \omega\hat{\omega}]^{|\zeta_U| - k} \mu^p,
\end{aligned} \tag{1.6}
$$

which couples the nonzero eigenvalues λ of the USSOR iteration matrix $T_{\omega\hat{\omega}}$ with the eigenvalues μ of the Jacobi matrix B. In (1.6) $|\zeta_L|$ and $|\zeta_U|$ are the cardinalities of the sets ζ_L and ζ_U, which are the two disjoint subsets of $P \equiv \{1, 2, \ldots, p\}$ associated with the cyclic permutation $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_p)$ as these are defined in [9], i.e.,

$$
\zeta_L = \{\sigma_j : \sigma_j > \sigma_{j+1}\}, \qquad \zeta_U = \{\sigma_j : \sigma_j < \sigma_{j+1}\}. \tag{1.7}
$$

The integer k is well defined in [9] as is the number of nonzero block elements of the matrix product LU. Li and Varga gave also the directed graph interpretation of the number k. It is obvious that $\zeta_L \cup \zeta_U = \{1, 2, \ldots, p\}$ and $\zeta_L \cap \zeta_U = \varnothing$, consequently, $|\zeta_L| + |\zeta_U| = p$. In other words $|\zeta_L|$ and $|\zeta_U|$ are the numbers of the nonzero block elements of the matrices L and U, respectively.

Equation (1.6) generalizes the following previous works: (i) The results of Saridakis [12] on the USSOR iteration matrix for consistently ordered weakly p-cyclic matrices; (ii) the ones of Gong and Cai [5] and of Varga, Niethammer, and Cai [17] on the SSOR iteration matrix for p-cyclic matrices; (iii) the well-known results of Young [19, 21] on the SOR matrix for the two-cyclic case; (iv) the well-known results of Varga [15, 16] on the SOR iteration matrix for the consistently ordered weakly p-cyclic Jacobi matrix; and (v) the results of Verner and Bernal [18] on the SOR matrix for the $(q, p-q)$-GCO case. It should be noted that the result in the last case was mentioned for the first time by Varga in [16]. Finally, a relationship similar in character on the modified SOR (MSOR) matrix for the $(q, p-q)$-GCO case, was derived by Taylor [14].

Our main objective in this work is to derive the matrix analogue of the functional equation (1.6). More specifically, we show that the identity

$$
\begin{aligned}
&[T_{\omega\hat{\omega}} - (1-\omega)(1-\hat{\omega})I]^p \\
&\quad = (\omega + \hat{\omega} - \omega\hat{\omega})^{2k} T_{\omega\hat{\omega}}^k [\omega T_{\omega\hat{\omega}} + (\hat{\omega} - \omega\hat{\omega})I]^{|\zeta_L| - k} [\hat{\omega} T_{\omega\hat{\omega}} + (\omega - \omega\hat{\omega})I]^{|\zeta_U| - k} B^p
\end{aligned} \tag{1.8}
$$

always holds.

It is interesting to mention that the matrix analogues of the functional equations of cases (ii)–(v) were derived by Galanis, Hadjidimos, and Noutsos (see [1–3]), by using elementary graph theory (Harary [8], Varga [16]). The matrix analogue of the equation corresponding to the MSOR case was derived by Young and Kincaid [20] for the special case $(p, q) = (2, 1)$, by Hadjidimos and Yeyios [6] for the cases $(p, q) = (3, 1), (3, 2)$ by the straightforward analytic calculations and by Hadjidimos and Noutsos [7] for all values of p and q, by elementary graph theory.

The proof of (1.8) is given in §2. As will be seen, the main tool will be combinatorics and to guide intuition elementary graph theory will be used. Also by considering special cases of (1.8) with $\hat{\omega} = 0$ or $\omega = 0$, known, other results for the SOR as well as the backward SOR methods will be obtained. In §3 the "equivalence" of the USSOR method and a certain two-parametric p-step method in the sense of Niethammer and Varga [10] is established. Apart from the theoretical interest presented by the identity (1.8), it is also of practical importance, since the problem of determination of "good" or "optimal" parameters ω and $\hat{\omega}$ for the solution of the linear system (1.1), using the USSOR method, is equivalent to that of the determination of the same parameters of a two-parametric p-step iterative method. This problem, however, still remains an open one.

2. Main result and preliminary analysis. The statement of our main result is given in the following theorem.

THEOREM 2.1. *Let B be the weakly cyclic block Jacobi matrix, generated by the cyclic permutation $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_p)$, and $T_{\omega\hat{\omega}}$ in* (1.3) *be the block* USSOR *iteration matrix associated with A in* (1.1). *Then the matrix relationship* (1.8) *holds.*

The proof of Theorem 2.1 will be given later, where a number of other auxiliary statements will be stated and proved. First, the background material on which these proofs are based is developed.

It is noted that (1.8) trivially holds if $\omega = \hat{\omega} = 0$. So we assume that $\omega \neq 0$ and $\hat{\omega} \neq 0$. We will see that this assumption can be made without any loss of generality.

To simplify the proof of Theorem 2.1, we will prove the validity of another simpler relationship which is produced from (1.8) by setting

$$\tilde{T}_{\omega\hat{\omega}} = (I - \hat{\omega}U)T_{\omega\hat{\omega}}(I - \hat{\omega}U)^{-1} \tag{2.1}$$

in the place of $T_{\omega\hat{\omega}}$. We then begin our analysis by introducing the directed graphs.

The directed graph G is a pair (V, E) where $E \subseteq V \times V$ (see [16] or [8]). In our analysis the vertex set $V \equiv P$, following [13] or [7], we identify G with the edge set E. Also for a block partitioned matrix A, the graph of A is defined to be $G(A) = \{(i,j) : A_{ij} \neq 0\}$. So the directed graph $G(B)$ of the Jacobi matrix B will be

$$G(B) = \bigcup_{i=1}^{p} \{(\sigma_i, \sigma_{i+1})\}, \tag{2.2}$$

where $\sigma_{p+1} = \sigma_1$. (In the sequel the node σ_{p+1} will be denoted as σ_1.)

An example for $p = 5$ is given now to demonstrate the analysis. Let

$$B = \begin{pmatrix} 0 & 0 & 0 & B_{14} & 0 \\ 0 & 0 & 0 & 0 & B_{25} \\ 0 & B_{32} & 0 & 0 & 0 \\ 0 & 0 & B_{43} & 0 & 0 \\ B_{51} & 0 & 0 & 0 & 0 \end{pmatrix}$$

be the Jacobi matrix. From the definition we have $\sigma = (2,5,1,4,3), \zeta_L = \{5,4,3\}$, and $\zeta_U = \{2,1\}$. The graph $G(B)$ is shown in Fig. 1.

From Fig. 1 it is easily seen that $G(B)$ is a cyclic graph. It is also noted that (i) there are exactly k paths which go from a node of ζ_L to a node of ζ_U corresponding to the nonzero blocks of LU. We call these paths "backward" paths. (ii) There are exactly k paths which go from a node of ζ_U to a node of ζ_L corresponding to the

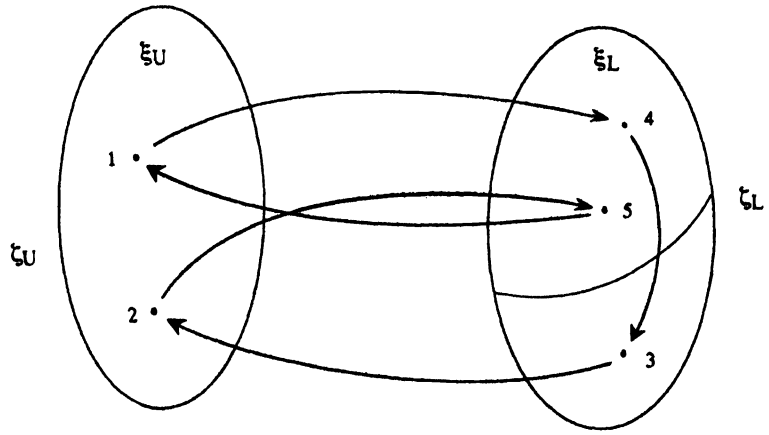


FIG. 1.

nonzero blocks of UL. We call these paths "forward" paths. In our example, $k = 2$ corresponds to the two backward paths $(3,2)$ and $(5,1)$ or to the two forward paths $(2,5)$ and $(1,4)$.

To derive the graph of the matrix B^p we can observe that starting from the node σ_i, we return to σ_i after p paths of B passing through all the nodes $\sigma_{i+1}, \sigma_{i+2}, \ldots, \sigma_{i-1}$. So B^p is a block diagonal matrix which comes from a sum of products of powers of L's and U's. Each product contains totally a number of $|\zeta_L|$, L's and $|\zeta_U|$, U's.

The graph $G(\tilde{T}_{\omega\hat{\omega}})$ of the matrix $\tilde{T}_{\omega\hat{\omega}}$ is now studied. From (1.3) and (2.1) we get that

$$\tilde{T}_{\omega\hat{\omega}} = [(1-\hat{\omega})I + \hat{\omega}L](I-\omega L)^{-1}[(1-\omega)I + \omega U](I-\hat{\omega}U)^{-1}. \tag{2.3}$$

Obviously the following relations hold:

$$(I-\omega L)^{-1} = \sum_{i=0}^{q_L} (\omega L)^i \text{ and } (I-\hat{\omega}U)^{-1} = \sum_{i=0}^{q_U} (\hat{\omega}U)^i, \tag{2.4}$$

where q_L and q_U are the largest integers such that $L^{q_L} \neq 0$ and $U^{q_U} \neq 0$, respectively, (in the above example, $q_L = 2$ and $q_U = 1$). By substituting (2.4) in (2.3) and after simple operations, we obtain that

$$\begin{aligned} \tilde{T}_{\omega\hat{\omega}} &= (1-\omega)(1-\hat{\omega})I + (\omega + \hat{\omega} - \omega\hat{\omega}) \\ &\times \left\{ (1-\omega)\sum_{i=1}^{q_L} \omega^{i-1}L^i + (1-\hat{\omega})\sum_{i=1}^{q_U} \hat{\omega}^{i-1}U^i \right. \\ &\left. + (\omega + \hat{\omega} - \omega\hat{\omega})\sum_{i=1}^{q_L}\sum_{j=1}^{q_U} \omega^{i-1}L^i\hat{\omega}^{j-1}U^j \right\}. \end{aligned} \tag{2.5}$$

It is noted that ωL and $\hat{\omega}U$ are of exactly the same form as L and U. So ωL and $\hat{\omega}U$ will be denoted from now on by L and U. Thus, after this convention (2.5) can be written as

$$\begin{aligned} \tilde{T}_{\omega\hat{\omega}} &= (1-\omega)(1-\hat{\omega})I + (\omega + \hat{\omega} - \omega\hat{\omega}) \\ &\times \left\{ \frac{1-\omega}{\omega}\sum_{i=1}^{q_L} L^i + \frac{1-\hat{\omega}}{\hat{\omega}}\sum_{i=1}^{q_U} U^i + \frac{\omega + \hat{\omega} - \omega\hat{\omega}}{\omega\hat{\omega}}\sum_{i=1}^{q_L}\sum_{j=1}^{q_U} L^iU^j \right\}. \end{aligned} \tag{2.6}$$

Since in relation (2.6) we have different scalar coefficients for the matrices I, L^i, U^i, and L^iU^j, we introduce the weighted graph of $\tilde{T}_{\omega\hat{\omega}}$. Thus we define (i) the paths weighted by $(\omega + \hat{\omega} - \omega\hat{\omega})\ \frac{1-\omega}{\omega}$ as single-arrowed paths; (ii) the paths weighted by $(\omega + \hat{\omega} - \omega\hat{\omega})\ \frac{1-\hat{\omega}}{\hat{\omega}}$ as double-arrowed paths; (iii) the paths weighted by $\frac{(\omega+\hat{\omega}-\omega\hat{\omega})^2}{\omega\hat{\omega}}$ as triple-arrowed paths; and (iv) the paths weighted by $(1-\omega)(1-\hat{\omega})$ as four-arrowed paths. So from the right-hand side of (2.6) we have the following. The first term of (2.6) gives the four-arrowed identity paths

$$(\overrightarrow{\sigma_i, \sigma_i}), \quad i = 1(1)p. \tag{2.7}$$

The second term, which contains a sum of powers of L, gives the single-arrowed paths

$$(\sigma_i, \overrightarrow{\sigma_{i+j}}), \quad j = 1(1)q_{L,i}, \quad \sigma_i \in \zeta_L, \tag{2.8}$$

where $q_{L,i}$ is an integer such that all the successive nodes $\sigma_i, \sigma_{i+1}, \ldots, \sigma_{i+q_{L,i}-1}$ belong to ζ_L and $\sigma_{i+q_{L,i}} \in \zeta_U$. (From Fig. 1 we can see that if $\sigma_i = 5$ or $\sigma_i = 3$ then $q_{L,i} = 1$ while if $\sigma_i = 4$ then $q_{L,i} = 2$.) The third term, which contains a sum of powers of U, gives the double-arrowed paths

$$(\sigma_i, \overrightarrow{\sigma_{i+j}}), \quad j = 1(1)q_{U,i}, \quad \sigma_i \in \zeta_U, \tag{2.9}$$

where $q_{U,i}$ is an integer such that all the successive nodes $\sigma_i, \sigma_{i+1}, \ldots, \sigma_{i+q_{U,i}-1}$ belong to ζ_U and $\sigma_{i+q_{U,i}} \in \zeta_L$. (Figure 1 gives that $q_{U,i} = 1$ for both cases $\sigma_i = 1$ or 2.) Finally the last term, which contains a double sum of products of powers of L and U, gives the triple-arrowed paths

$$(\overrightarrow{\sigma_i, \sigma_{i+j}}), \quad j = q_{L,i} + 1(1)q_{LU,i}, \quad \sigma_i \in \zeta_L, \tag{2.10}$$

where $q_{LU,i} = q_{L,i} + q_{U,i+q_{L,i}}$ (in our example $q_{LU,i} = 3$ for $\sigma_i = 4$, which corresponds to the three successive paths $(4,3), (3,2)$, and $(2,5)$). It is noted here that $\sigma_s := \sigma_{s-p}$ if $s > p$ in (2.8), (2.9), and (2.10). The union of all the paths in (2.7), (2.8), (2.9), and (2.10) gives the graph of $\tilde{T}_{\omega\hat{\omega}}$.

$$\begin{aligned} G(\tilde{T}_{\omega\hat{\omega}}) = & \left(\bigcup_{i=1}^{p} \{ \overrightarrow{(\sigma_i, \sigma_i)} \} \right) \\ & \cup \left(\bigcup_{\sigma_i \in \zeta_L} \left[\bigcup_{j=1}^{q_{L,i}} \{ (\sigma_i, \overrightarrow{\sigma_{i+j}}) \} \bigcup_{j=q_{L,i}+1}^{q_{LU,i}} \{ \overrightarrow{(\sigma_i, \sigma_{i+j})} \} \right] \right) \\ & \cup \left(\bigcup_{\sigma_i \in \zeta_U} \left[\bigcup_{j=1}^{q_{U,i}} \{ (\sigma_i, \overrightarrow{\sigma_{i+j}}) \} \right] \right). \end{aligned} \tag{2.11}$$

The subgraphs of $G(\tilde{T}_{\omega\hat{\omega}})$ of our example that contain only the paths that have the origin node $4 \in \zeta_L$ or $1 \in \zeta_U$ are illustrated in Fig. 2(a) and Fig. 2(b), respectively.

We distinguish the subset ξ_L of ζ_L which contains the nodes σ_j such that $\sigma_{j-1} \in \zeta_U$ and the subset ξ_U of ζ_U which contains the nodes σ_j such that $\sigma_{j-1} \in \zeta_L$. It is easily seen from Fig. 1 that both ξ_L and ξ_U contain exactly k nodes. In our example we have $\xi_L = \{4, 5\}$ and $\xi_U = \{1, 2\}$.

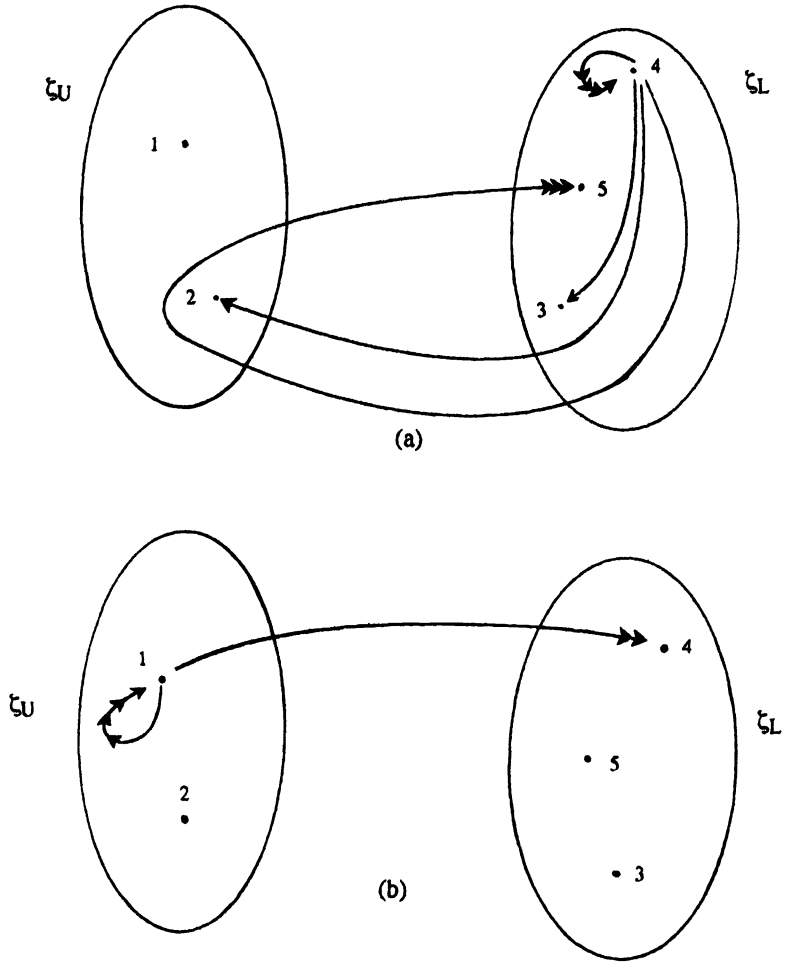


FIG. 2.

After replacing ωL and $\hat{\omega}U$ by L and U, the matrix relationship (1.8) will be equivalent to

(2.12)

$$[\tilde{T}_{\omega\hat{\omega}} - (1-\omega)(1-\hat{\omega})I]^p = \left[\frac{(\omega+\hat{\omega}-\omega\hat{\omega})^2}{\omega\hat{\omega}}\right]^k \tilde{T}^k_{\omega\hat{\omega}} \left[\tilde{T}_{\omega\hat{\omega}} + \frac{\hat{\omega}(1-\omega)}{\omega} I\right]^{|\zeta_L|-k} \times \left[\tilde{T}_{\omega\hat{\omega}} + \frac{\omega(1-\hat{\omega})}{\hat{\omega}} I\right]^{|\zeta_U|-k} B^p.$$

It is noted that in (2.12) we have put B^p for $\omega^{|\zeta_L|}\hat{\omega}^{|\zeta_U|}B^p$, since B^p constitutes the sum of products of $|\zeta_L|, L$'s and $|\zeta_U|, U$'s.

From (2.6) we can see that the graph of the matrix $\tilde{T}_{\omega\hat{\omega}} - (1-\omega)(1-\hat{\omega})I$ con-

tains no identity paths. So, from (2.11), we have

(2.13)

$$G(\tilde{T}_{\omega\hat{\omega}} - (1-\omega)(1-\hat{\omega})I) = \left(\bigcup_{\sigma_i \in \zeta_L} \left[\bigcup_{j=1}^{q_{L,i}} \{(\sigma_i, \overrightarrow{\sigma_{i+j}})\} \bigcup_{j=q_{L,i}+1}^{q_{LU,i}} \{(\sigma_i, \overrightarrow{\sigma_{i+j}})\} \right] \right) \cup \left(\bigcup_{\sigma_i \in \zeta_U} \left[\bigcup_{j=1}^{q_{U,i}} \{(\sigma_i, \overrightarrow{\sigma_{i+j}})\} \right] \right).$$

The graph in (2.13) is derived from $G(\tilde{T}_{\omega\hat{\omega}})$ by simply omitting the identity paths.

It is easily checked from (2.6) that the matrix $\tilde{T}_{\omega\hat{\omega}} + (\hat{\omega}(1-\omega)/\omega)I$ is given by

(2.14)

$$\tilde{T}_{\omega\hat{\omega}} + \frac{\hat{\omega}(1-\omega)}{\omega} I = (\omega + \hat{\omega} - \omega\hat{\omega}) \times \left\{ \frac{1-\omega}{\omega} \sum_{i=0}^{q_L} L^i + \frac{1-\hat{\omega}}{\hat{\omega}} \sum_{i=1}^{q_U} U^i + \frac{\omega + \hat{\omega} - \omega\hat{\omega}}{\omega\hat{\omega}} \sum_{i=1}^{q_L} \sum_{j=1}^{q_U} L^i U^j \right\}.$$

So the identity paths now become single-arrowed paths and the graph of the matrix is given by

(2.15)

$$G\left(\tilde{T}_{\omega\hat{\omega}} + \frac{\hat{\omega}(1-\omega)}{\omega} I\right) = \left(\bigcup_{i=1}^{p} \{(\overrightarrow{\sigma_i, \sigma_i})\} \right) \cup \left(\bigcup_{\sigma_i \in \zeta_L} \left[\bigcup_{j=1}^{q_{L,i}} \{(\sigma_i, \overrightarrow{\sigma_{i+j}})\} \bigcup_{j=q_{L,i}+1}^{q_{LU,i}} \{(\sigma_i, \overrightarrow{\sigma_{i+j}})\} \right] \right) \cup \left(\bigcup_{\sigma_i \in \zeta_U} \left[\bigcup_{j=1}^{q_{U,i}} \{(\sigma_i, \overrightarrow{\sigma_{i+j}})\} \right] \right).$$

This graph is derived from $G(\tilde{T}_{\omega\hat{\omega}})$ by simply replacing the four-arrowed identity paths with single-arrowed paths. Similarly, the matrix $\tilde{T}_{\omega\hat{\omega}} + (\omega(1-\hat{\omega})/\hat{\omega})I$ is given by

(2.16)

$$\tilde{T}_{\omega\hat{\omega}} + \frac{\omega(1-\hat{\omega})}{\hat{\omega}} I = (\omega + \hat{\omega} - \omega\hat{\omega}) \times \left\{ \frac{1-\omega}{\omega} \sum_{i=1}^{q_L} L^i + \frac{1-\hat{\omega}}{\hat{\omega}} \sum_{i=0}^{q_U} U^i + \frac{\omega + \hat{\omega} - \omega\hat{\omega}}{\omega\hat{\omega}} \sum_{i=1}^{q_L} \sum_{j=1}^{q_U} L^i U^j \right\}$$

and its graph by

(2.17)

$$G\left(\tilde{T}_{\omega\hat{\omega}} + \frac{\omega(1-\hat{\omega})}{\hat{\omega}} I\right) = \left(\bigcup_{i=1}^{p} \{(\overrightarrow{\sigma_i, \sigma_i})\} \right) \cup \left(\bigcup_{\sigma_i \in \zeta_L} \left[\bigcup_{j=1}^{q_{L,i}} \{(\sigma_i, \overrightarrow{\sigma_{i+j}})\} \bigcup_{j=q_{L,i}+1}^{q_{LU,i}} \{(\sigma_i, \overrightarrow{\sigma_{i+j}})\} \right] \right) \cup \left(\bigcup_{\sigma_i \in \zeta_U} \left[\bigcup_{j=1}^{q_{U,i}} \{(\sigma_i, \overrightarrow{\sigma_{i+j}})\} \right] \right),$$

which is derived from $G(\tilde{T}_{\omega\hat{\omega}})$ by simply replacing the four-arrowed identity paths with double-arrowed identity paths.

A lemma is now stated and proved that shows the equivalence of (1.8) and (2.12).

LEMMA 2.2. *If the matrix relationship* (2.12) *holds then so does* (1.8) *and vice versa.*

Proof. We prove the validity of the matrix relationship (2.12) from that of (1.8) by replacing at the same time the ωL's and $\hat{\omega}U$'s by L's and U's, respectively.

By taking the inverse similarity transformation of (2.1) on both sides of (2.12), we have

$$\begin{aligned}(I-U)^{-1}&[\tilde{T}_{\omega\hat{\omega}}-(1-\omega)(1-\hat{\omega})I]^p(I-U)\\ &=(I-U)^{-1}\left[\frac{(\omega+\hat{\omega}-\omega\hat{\omega})^2}{\omega\hat{\omega}}\right]^k\tilde{T}^k_{\omega\hat{\omega}}\left[\tilde{T}_{\omega\hat{\omega}}+\frac{\hat{\omega}(1-\omega)}{\omega}\,I\right]^{|\zeta_L|-k}\\ &\quad\times\left[\tilde{T}_{\omega\hat{\omega}}+\frac{\omega(1-\hat{\omega})}{\hat{\omega}}\,I\right]^{|\zeta_U|-k}B^p(I-U)\end{aligned}$$

or from (2.1)

(2.18)

$$\begin{aligned}[T_{\omega\hat{\omega}}-(1-\omega)(1-\hat{\omega})I]^p&=\left[\frac{(\omega+\hat{\omega}-\omega\hat{\omega})^2}{\omega\hat{\omega}}\right]^k T^k_{\omega\hat{\omega}}\left[T_{\omega\hat{\omega}}+\frac{\hat{\omega}(1-\omega)}{\omega}\,I\right]^{|\zeta_L|-k}\\ &\quad\times\left[T_{\omega\hat{\omega}}+\frac{\omega(1-\hat{\omega})}{\hat{\omega}}\,I\right]^{|\zeta_U|-k}(I-U)^{-1}B^p(I-U).\end{aligned}$$

For (2.18) to hold it must be proved that $(I-U)^{-1}B^p(I-U)=B^p$ or $B^p(I-U)=(I-U)B^p$ or simply that

$$B^pU=UB^p. \tag{2.19}$$

The proof of (2.19) is given by elementary graph theory. Since the graph $G(B^p)$ contains the identity paths $(\sigma_i,\sigma_i), i=1(1)p$ which constitute p successive simple paths of $G(B)$ and the graph $G(U)$ contains the simple paths $(\sigma_i,\sigma_{i+1}),\sigma_i\in\zeta_U$, the graph $G(B^pU)$ contains the paths $(\sigma_i,\sigma_{i+1}),\sigma_i\in\zeta_U$ which constitute $p+1$ successive simple paths of $G(B)$. Similarly, the graph $G(UB^p)$ contains the same paths. So these two graphs describe the graphs of the same matrices and the proof is complete. Moreover, it is noted here that an analogous proof gives that the matrices B^p and L also commute. □

Now we have all the necessary tools to prove our main theorem.

Proof of Theorem 2.1. Let C and D be the matrices denoting the left- and right-hand sides of (2.12), respectively. The proof is due to the following simple idea: Since C and D have been expanded in sums of terms of products of L's and U's, we must prove that if there exists a term of the expansion of C then there exists also such a term of the expansion of D with the same coefficient and vice versa. This means, in graph analogue, that if there exists a path (σ_i,σ_j) of $G(C)$ then there exists also such a path of $G(D)$ weighted with the same weight, for all the pairs σ_i,σ_j and vice versa. Each of these paths consists of consecutive subpaths and represents the graph of a nonidentically zero block of the term in question. Our objective will be accomplished if we show that all paths in $G(C)$ and $G(D)$ from σ_i to σ_j with m backward subpaths $(0\le m\le p)$ coincide and are associated with equal overall weights. It is obvious that any two paths (σ_i,σ_j) of $G(C)$ and (σ_i,σ_j) of $G(D)$ with a particular number m

of backward edges correspond to the same expansion in terms of nonidentically zero products of L's and U's. They differ from each other only because of the different weights of the single-, double-, triple-, or four-arrowed subpaths as they are described above. For example, let $\sigma_i = 4$ and $\sigma_j = 5$. Then Figs. 1 and 2 give that the path $(4, 5)$ of $G(C)$ constitutes paths associated with three different numbers m of backward edges. $m = 1$ corresponds to the matrix product $LLU, m = 3$ corresponds to $LLULULLU$, and $m = 5$ corresponds to $LLULULLULULLU$. The union of all the above paths from σ_i to σ_j with m backward edges will be considered as one path, with which an overall weight will be associated. This overall weight will be equal to the sum of all the weights associated with each individual path. The determination of this weight constitutes the basic key to the proof of our main result.

We try to find the overall weight of $G(C)$ with $k + m$ backward subpaths ($0 \leq m \leq p - k$). (The number of $k + m$ backward subpaths is taken since the smallest number of backward subpaths of the matrix $C = [\tilde{T}_{\omega\hat{\omega}} - (1 - \omega)(1 - \hat{\omega})I]^p$ is k. This is obtained by considering the path of the smallest possible way, which contains p consecutive subpaths of the form (σ_l, σ_{l+1}) with their weights.) From the graph expression (2.13) of the matrix $\tilde{T}_{\omega\hat{\omega}} - (1 - \omega)(1 - \hat{\omega})I$, from Fig. 2, and from elementary graph theory we can see that this path consists of the union of all possible combinations of p consecutive subpaths of $G(\tilde{T}_{\omega\hat{\omega}} - (1 - \omega)(1 - \hat{\omega})I)$ that go from σ_i to σ_j with $k + m$ backward edges. This remark leads us to the conclusion that to analyze and study the problem at hand, the use of combinatorics theory together with elementary graph theory must be made.

The analysis requires that we distinguish four cases

(i) $\sigma_i, \sigma_j \in \zeta_L$,

(ii) $\sigma_i, \sigma_j \in \zeta_U$,

(iii) $\sigma_i \in \zeta_L$ and $\sigma_j \in \zeta_U$,

(iv) $\sigma_i \in \zeta_U$ and $\sigma_j \in \zeta_L$.

Since the argumentation is quite similar in all the four cases, only the first case is presented in detail. The others can be found in [11].

From Fig. 2(a) we see that there are two types of backward edges: the single-arrowed path with ending node belonging to ξ_U (see path $(\overrightarrow{4, 2})$) and the triple-arrowed path with ending node belonging to $(\zeta_U \backslash \xi_U) \cup \xi_L$ (see path $(\overrightarrow{4, 5})$). If we take r ending nodes of the first type and $k + m - r$ of the second type we have $\binom{k+m}{r}$ cases to consider. Then let t be the number of consecutive nodes in the way from σ_i to σ_j with $k + m$ backward edges, with σ_j being included and t_L and t_U being the number of those nodes, respectively, which belong to ζ_L and $\zeta_U (t_L + t_U = t)$. We consider all possible combinations of t nodes by taking p of them as ending nodes of $G(\tilde{T}_{\omega\hat{\omega}} - (1 - \omega)(1 - \hat{\omega})I)$. In our example, from node 4 to node 5 with three backward edges, we have the consecutive nodes: 3, 2, 5, 1, 4, 3, 2, and 5 as we can see in Fig. 1. So, $t = 8$. Five of these nodes (3, 5, 4, 3, and 5) are taken from ζ_L and three nodes (2, 1, and 2) from ζ_U. So, $t_L = 5$ and $t_U = 3$. This way corresponds to the product of blocks $B_{43}B_{32}B_{25}B_{51}B_{14}B_{43}B_{32}B_{25}$.

Let $\sigma_j \in \xi_L$, as in our example $\sigma_j = 5$. The analysis of the case $\sigma_j \in \zeta_L \backslash \xi_L$ is similar. In the sequel we will see that the way of going from one node of the set ζ_U to one of ζ_L can be given by means of the nodes of ξ_L only. So the $k + m$ nodes of ξ_L will all be taken (the three nodes 5, 4, and 5 of our example). The number of nodes that have been taken so far is $r + k + m (r$ of ξ_U and $k + m$ of $\xi_L)$. From the remaining nodes we take q nodes belonging to $\zeta_U \backslash \xi_U$ and s nodes belonging to $\zeta_L \backslash \xi_L$. So, $r + q + s = p - k - m$. The q nodes are taken from t_U nodes of ζ_U except the

$k+m$ nodes of ξ_U corresponding to the $k+m$ backward edges which were taken before. This gives $\binom{t_U-k-m}{q}$ different ways to consider. In our example we have $t_U = k+m = 3$ since $\zeta_U = \xi_U = \{1,2\}$. This means that we have only one possible way. The s nodes are taken from t_L nodes of ζ_L except the $k+m$ nodes of ξ_L. Similarly this gives a number of $\binom{t_L-k-m}{s}$ different ways to consider. Totally, we have

$$\binom{k+m}{r}\binom{t_U-k-m}{q}\binom{t_L-k-m}{s} \tag{2.20}$$

different ways to consider. The associated weight comes from $k+m-r$ triple-arrowed subpaths, from $q+k+m-(k+m-r) = q+r$ double-arrowed subpaths (q nodes of $\zeta_U \backslash \xi_U$ plus $k+m$ nodes of ξ_L except the $k+m-r$ triple-arrowed subpaths), and from the remaining $r+s$ single-arrowed subpaths. So this weight is

(2.21)

$$\left[(\omega+\hat{\omega}-\omega\hat{\omega})\,\frac{1-\hat{\omega}}{\hat{\omega}}\right]^{q+r}\left[(\omega+\hat{\omega}-\omega\hat{\omega})\,\frac{1-\omega}{\omega}\right]^{r+s}\left[\frac{(\omega+\hat{\omega}-\omega\hat{\omega})^2}{\omega\hat{\omega}}\right]^{k+m-r}.$$

By considering all possible values of r, q, and s such that $r+q+s = p-k-m$, we get the total overall weight equal to

$$\begin{aligned} N_C =& (\omega+\hat{\omega}-\omega\hat{\omega})^p \sum_{r+q+s=p-k-m}\binom{k+m}{r}\binom{t_U-k-m}{q}\binom{t_L-k-m}{s} \\ &\times\left[\frac{1-\hat{\omega}}{\hat{\omega}}\right]^{q+r}\left[\frac{1-\omega}{\omega}\right]^{r+s}\left[\frac{\omega+\hat{\omega}-\omega\hat{\omega}}{\omega\hat{\omega}}\right]^{k+m-r}. \end{aligned} \tag{2.22}$$

In our example:

$$N_C = (\omega+\hat{\omega}-\omega\hat{\omega})^5\sum_{r=0}^{2}\binom{3}{r}\binom{2}{2-r}\left[\frac{1-\hat{\omega}}{\hat{\omega}}\right]^r\left[\frac{1-\omega}{\omega}\right]^2\left[\frac{\omega+\hat{\omega}-\omega\hat{\omega}}{\omega\hat{\omega}}\right]^{3-r}.$$

Now we try to show that there exists the same path in $G(D)$ with the same weight. Only the case where $\sigma_i \in \zeta_L$ and $\sigma_j \in \xi_L$ is studied here.

From (2.12) we note that B^p is the last factor of the matrix D. The graph $G(B^p)$ consists of the identity paths (σ_i, σ_i) containing k backward edges. This means that in the graph of D, the last path (σ_j, σ_j) containing k backward edges belongs to the graph of B^p and has no weight. So, we must find the overall weight of the path from σ_i to σ_j with m backward edges of the graph

$$G\left(\tilde{T}^k_{\omega\hat{\omega}}\left[\tilde{T}_{\omega\hat{\omega}}+\frac{\hat{\omega}(1-\omega)}{\omega}\,I\right]^{|\zeta_L|-k}\left[\tilde{T}_{\omega\hat{\omega}}+\frac{\omega(1-\hat{\omega})}{\hat{\omega}}\,I\right]^{|\zeta_U|-k}\right). \tag{2.23}$$

From the graph expressions (2.11), (2.15), (2.17), and Fig. 2 it is easily seen that this path exists since the same nodes are used in the way from σ_i to σ_j. The total number of nodes are $t-p$ (p nodes belong to the graph of B^p). In our example we have the nodes 3, 2, and 5. However, $t_U - |\zeta_U|$ of them belong to ζ_U and $t_L - |\zeta_L|$ belong to ζ_L. The main difference from the previous case is that now there are identity paths involved. We can also see that the graph expressions (2.11), (2.15), and (2.17) have the same paths which differ only in the weight of the identity paths. We then must

find all the possible combinations by taking the first k consecutive paths from (2.11), the second $|\zeta_L| - k$ consecutive paths from (2.15), and the last $|\zeta_U| - k$ consecutive paths from (2.17).

Let us consider r_1 nodes from ξ_U, q_1 nodes from $\zeta_U \backslash \xi_U$ and q_2 nodes from $\zeta_L \backslash \xi_L$ of the path in the way from σ_i to σ_j with m backward edges. This gives a number of

$$\binom{m}{r_1}\binom{t_U - |\zeta_U| - m}{q_1}\binom{t_L - |\zeta_L| - m}{q_2} \tag{2.24}$$

different ways. Let us also take s_1 four-arrowed identity paths from the k paths of the graph (2.11), s_2 single-arrowed identity paths from the $|\zeta_L| - k$ paths of the graph (2.15), and s_3 double-arrowed identity paths from the $|\zeta_U| - k$ paths of the graph (2.17). So we must first distribute the number of times of the above s_1 identity paths to the $k - s_1 + 1$ nodes (the first σ_i node being included). This gives the number of combinations with repetitions of $k - s_1 + 1$ chosen s_1, that is

$$\binom{k - s_1 + 1 + s_1 - 1}{k - s_1 + 1 - 1} = \binom{k}{k - s_1} = \binom{k}{s_1}. \tag{2.25}$$

Similarly we obtain a number of $\binom{|\zeta_L| - k}{s_2}$ different cases because of the identity paths of (2.15) and a number of $\binom{|\zeta_U| - k}{s_3}$ different cases because of the identity paths of (2.17). After these considerations are made it is obvious that there is a number of

$$\binom{m}{r_1}\binom{t_U - |\zeta_U| - m}{q_1}\binom{t_L - |\zeta_L| - m}{q_2}\binom{k}{s_1}\binom{|\zeta_L| - k}{s_2}\binom{|\zeta_U| - k}{s_3} \tag{2.26}$$

different ways. The associated weight consists of $r_1 + q_2 + s_2$ single-arrowed paths of $r_1 + q_1 + s_3$ double-arrowed paths, of $m - r_1$ triple-arrowed paths, and of s_1 four-arrowed paths. This gives a weight of

$$\begin{aligned} &\left[(\omega + \hat{\omega} - \omega\hat{\omega})\,\frac{1 - \hat{\omega}}{\hat{\omega}}\right]^{r_1 + q_1 + s_3}\left[(\omega + \hat{\omega} - \omega\hat{\omega})\,\frac{1 - \omega}{\omega}\right]^{r_1 + q_2 + s_2} \\ &\quad \times \left[\frac{(\omega + \hat{\omega} - \omega\hat{\omega})^2}{\omega\hat{\omega}}\right]^{m - r_1}[(1 - \omega)(1 - \hat{\omega})]^{s_1}. \end{aligned} \tag{2.27}$$

The total number of subpaths of (2.23) is $k + (|\zeta_L| - k) + (|\zeta_U| - k) = p - k$. Since the m subpaths with ending nodes in ξ_L must be taken, the integers r_1, q_1, q_2, s_1, s_2, and s_3 vary but satisfy the relationship $r_1 + q_1 + q_2 + s_1 + s_2 + s_3 = p - k - m$. From (2.26), and (2.27) we have the total weight of the path of the graph (2.23) from σ_i to σ_j with m backward edges, which is

$$\begin{aligned} &\sum_{r_1 + q_1 + q_2 + s_1 + s_2 + s_3 = p - k - m}\binom{m}{r_1}\binom{t_U - |\zeta_U| - m}{q_1}\binom{t_L - |\zeta_L| - m}{q_2}\binom{k}{s_1} \\ &\quad \times \binom{|\zeta_L| - k}{s_2}\binom{|\zeta_U| - k}{s_3}(\omega + \hat{\omega} - \omega\hat{\omega})^{p - k}\left[\frac{1 - \hat{\omega}}{\hat{\omega}}\right]^{q_1 + r_1 + s_3 + s_1} \\ &\quad \times \left[\frac{1 - \omega}{\omega}\right]^{r_1 + q_2 + s_2 + s_1}\left[\frac{\omega + \hat{\omega} - \omega\hat{\omega}}{\omega\hat{\omega}}\right]^{m - r_1 - s_1}. \end{aligned} \tag{2.28}$$

By considering $q_1+s_3=q, q_2+s_2=s$, and $r_1+s_1=r$, the sum (2.28) takes the form

(2.29)

$$\sum_{r+q+s=p-k-m}\left(\sum_{r_1+s_1=r}\binom{m}{r_1}\binom{k}{s_1}\left[\frac{1-\hat{\omega}}{\hat{\omega}}\right]^{r_1+s_1}\left[\frac{1-\omega}{\omega}\right]^{r_1+s_1}\left[\frac{\omega+\hat{\omega}-\omega\hat{\omega}}{\omega\hat{\omega}}\right]^{m-r_1-s_1}\right)$$
$$\times\left(\sum_{q_1+s_3=q}\binom{t_U-|\zeta_U|-m}{q_1}\binom{|\zeta_U|-k}{s_3}\left[\frac{1-\hat{\omega}}{\hat{\omega}}\right]^{q_1+s_3}\right)$$
$$\times\left(\sum_{q_2+s_2=s}\binom{t_L-|\zeta_L|-m}{q_2}\binom{|\zeta_L|-k}{s_2}\left[\frac{1-\omega}{\omega}\right]^{q_2+s_2}\right)(\omega+\hat{\omega}-\omega\hat{\omega})^{p-k}.$$

By applying combinatorics theory, (2.29) gives

(2.30)

$$(\omega+\hat{\omega}-\omega\hat{\omega})^{p-k}\sum_{r+q+s=p-k-m}\binom{k+m}{r}\binom{t_U-k-m}{q}\binom{t_L-k-m}{s}$$
$$\times\left[\frac{1-\hat{\omega}}{\hat{\omega}}\right]^{q+r}\left[\frac{1-\omega}{\omega}\right]^{r+s}\left[\frac{\omega+\hat{\omega}-\omega\hat{\omega}}{\omega\hat{\omega}}\right]^{m-r}.$$

The total weight N_D of the path from σ_i to σ_j with $k+m$ backward edges of $G(D)$ is given by multiplying (2.30) with the coefficient

$$\left[\frac{(\omega+\hat{\omega}-\omega\hat{\omega})^2}{\omega\hat{\omega}}\right]^k$$

of the right-hand side of (2.12). This gives exactly the quantity N_C of (2.22). So

$$N_C\equiv N_D. \tag{2.31}$$

Obviously (2.31) is satisfied for all pairs $(\sigma_i,\sigma_j), i,j=1(1)p$ and the proof of our theorem is complete. □

Based on the analysis so far, it is easy to prove the following statement.

THEOREM 2.3. *Under the assumptions of Theorem* 2.1 *there holds*

$$B^pT_{\omega\hat{\omega}}=T_{\omega\hat{\omega}}B^p, \tag{2.32}$$

that is, the matrices B^p *and* $T_{\omega\hat{\omega}}$ *commute.*

Proof. The proof is obvious from Lemma 2.2, since the matrix B^p is commutative with the matrices U and L. □

The above result gives a more general matrix relationship than (1.8). In fact, it is not necessary that the factor B^p of the right-hand side be put as the last factor of the product. It can be put as its first factor or as any intermediate one.

Based on the main result already obtained, we can obtain some similar results for the SSOR, the SOR, and the backward SOR methods. These are presented in the following corollary.

COROLLARY 2.4. *Let* B *be the weakly cyclic block Jacobi matrix, generated by the cyclic permutation* $\sigma=(\sigma_1,\sigma_2,\ldots,\sigma_p)$. *Let also* S_ω *be the block* SSOR, L_ω *be the block*

SOR, and $\cup_\omega$ *be the block backward* SOR *iteration matrices, respectively, associated with* A *in* (1.1). *Then the following matrix relationships*

$$[S_\omega - (1-\omega)^2 I]^p = \omega^p (2-\omega)^{2k} S_\omega^k [S_\omega + (1-\omega)I]^{p-2k} B^p, \tag{2.33}$$

$$[L_\omega - (1-\omega)I]^p = \omega^p L_\omega^{|\zeta_L|} B^p, \tag{2.34}$$

and

$$[\cup_\omega - (1-\omega)I]^p = \omega^p \cup_\omega^{|\zeta_U|} B^p \tag{2.35}$$

hold.

Proof. It is easily proved that the analysis of the proof of our main result above holds when $\omega = \hat{\omega}$ or $\omega = 0$ or $\hat{\omega} = 0$; see also [11]. Putting $\omega = \hat{\omega}$ or $\hat{\omega} = 0$ or $\omega = 0$ (and using ω instead of $\hat{\omega}$) in $T_{\omega\hat{\omega}}$ in (1.3) reduces this matrix to the SSOR matrix S_ω, the SOR matrix L_ω, or to the backward SOR matrix $\cup_\omega$, respectively. Consequently, putting $\omega = \hat{\omega}$ or $\hat{\omega} = 0$ or $\omega = 0$ (and using ω instead of $\hat{\omega}$) in (1.8) reduces the relationship in question to the matrix relationships (2.33), (2.34), or (2.35), respectively. □

The first result generalizes the previous result by Galanis, Hadjidimos, and Noutsos [3] for the p-cyclic consistently ordered case and the second result generalizes the previous one by Galanis, Hadjidimos, and Noutsos [2] for the $(q, p-q)$-generalized consistently ordered case. It is noted here that the proof of Corollary 2.4 can be obtained independently of the result (1.8) by using an analogous analysis and elementary graph theory in each particular case.

3. Equivalence of the USSOR and a two-parametric p-step method. To show that the USSOR method, used for the solution of (1.1), is equivalent to a certain two-parametric p-step method we proceed in a way analogous to that in [1–3]. For this let $x^{(m-p)}$ be the $(m-p)$th iteration of (1.2) with $m = p, p+1, p+2, \ldots$. From (1.8) we have

(3.1)

$$[T_{\omega\hat{\omega}} - (1-\omega)(1-\hat{\omega}))I]^p x^{(m-p)} = (\omega + \hat{\omega} - \omega\hat{\omega})^{2k} T_{\omega\hat{\omega}}^k [\omega T_{\omega\hat{\omega}} + (\hat{\omega} - \omega\hat{\omega})I]^{|\zeta_L|-k} \times [\hat{\omega} T_{\omega\hat{\omega}} + (\omega - \omega\hat{\omega})I]^{|\zeta_U|-k} B^p x^{(m-p)}.$$

By expanding both sides of (3.1) in terms of $T_{\omega\hat{\omega}}$ and by successively applying (1.2), after some modest amount of algebra takes place (see [11]), we get the following two-parametric p-step iterative scheme:

$$\begin{aligned} x^{(m)} = & -\sum_{j=1}^{p} (-1)^j (1-\omega)^j (1-\hat{\omega})^j \binom{p}{j} x^{(m-j)} \\ & + (\omega + \hat{\omega} - \omega\hat{\omega})^{2k} B^p \sum_{i=0}^{|\zeta_L|-k} \sum_{j=0}^{|\zeta_U|-k} \binom{|\zeta_L|-k}{i} \\ & \times \binom{|\zeta_U|-k}{j} (\hat{\omega} - \omega\hat{\omega})^i (\omega - \omega\hat{\omega})^j \omega^{|\zeta_L|-k-i} \hat{\omega}^{|\zeta_U|-k-j} x^{(m-k-i-j)} \\ & + (\omega + \hat{\omega} - \omega\hat{\omega})^p \left(\sum_{i=0}^{p-1} B^i \right) b, \end{aligned} \tag{3.2}$$

where $x^{(j)} \in \mathbb{C}^n, j = 0(1)p-1$ are arbitrary.

In the sense explained above, the USSOR method (1.2) and (3.2) are equivalent and the study of (1.2) can be made by studying (3.2) and vice versa.

We must remark here that by putting $\omega = \hat{\omega}$ or $\hat{\omega} = 0$ or $\omega = 0$ in (3.2), we recover the monoparametric p-step schemes related to the SSOR, SOR, or backward SOR iterative methods, respectively. These schemes can also be obtained from the matrix relationships (2.33), (2.34), or (2.35), respectively.

One may also observe that because of the special cyclic nature of B, scheme (3.2) can be split into p simpler and smaller-dimension p-step iterative methods provided that all the vectors involved are partitioned in accordance with B. Each of these p simpler p-step methods has the same convergence rate, in the way considered in [10], as that of (3.2). So the solution of any one of these simpler methods provides us with the corresponding vector component of the solution x of (1.1), and from (1.1) all the other components of x. Therefore x itself can be readily obtained.

Acknowledgments. The author wishes to thank professor Apostolos Hadjidimos for various comments and suggestions on an early version of this manuscript. He is also indebted to the referees for their constructive criticism.

REFERENCES

[1] S. Galanis, A. Hadjidimos, and D. Noutsos, *On the equivalence of the k-step iterative Euler methods and successive overrelaxation* (SOR) *methods for k-cyclic matrices*, Math. Comput. Simulation, 30 (1988), pp. 213–230.

[2] ———, *The relationship between the Jacobi and the successive overrelaxation* (SOR) *matrices of a k-cyclic matrix*, Comput. Math. Appl., 17 (1989), pp. 1351–1357.

[3] ———, *On an* SSOR *matrix relationship and its consequences*, Internat. J. Numer. Math. Engrg., 27 (1989), pp. 559–570.

[4] S. Galanis, A. Hadjidimos, D. Noutsos, and M. Tzoumas, *On the optimum factor associated with p-cyclic matrices*, Linear Algebra Appl., 162–164 (1992), pp. 433–445.

[5] L. Gong and D. Y. Cai, *Relationship between eigenvalues of Jacobi and* SSOR *iterative matrix with p-weak cyclic matrix*, J. Comput. Math. Colleges and Universities, 1 (1985), pp. 79–84. (In Chinese.)

[6] A. Hadjidimos and A. K. Yeyios, *Some recent results on the modified* SOR *theory*, Linear Algebra Appl., 154–156 (1991), pp. 5–21.

[7] A. Hadjidimos and D. Noutsos, *On a matrix identity connecting iteration operators associated with a p-cyclic matrix*, Linear Algebra Appl., 182 (1993), pp. 157–178.

[8] F. Harary, *Graph Theory*, Addison-Wesley, Reading, MA, 1969.

[9] X. Li and R. S. Varga, *A note on the* SSOR *and* USSOR *iterative methods applied to p-cyclic matrices*, Numer. Math., 56 (1989), pp. 109–121.

[10] W. Niethammer and R. S. Varga, *The analysis of k-step iterative methods for linear systems from summability theory*, Numer. Math., 41 (1983), pp. 177–206.

[11] D. Noutsos, *The operator relation of the* USSOR *and the Jacobi iteration matrices of a p-cyclic matrix*, TR 234, Department of Mathematics, University of Ioannina, Greece, 1993.

[12] Y. G. Saridakis, *On the analysis of the unsymmetric overrelaxation method when applied to p-cyclic matrices*, Numer. Math., 49 (1986) pp. 461–473.

[13] H. Schneider, *Theorems on M-splittings of a singular M-matrix which depend on graph structure*, Linear Algebra Appl., 58 (1984), pp. 407–424.

[14] P. J. Taylor, *A generalization of symmetric relaxation methods for consistently ordered matrices*, Numer. Math., 13 (1969), pp. 377–395.

[15] R. S. Varga, *p-cyclic matrices: A generalization of the Young–Frankel successive overrelaxation scheme*, Pacific J. Math., 9 (1959), pp. 617–628.

[16] ———, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[17] R. S. Varga, W. Niethammer, and D. Y. Cai, *p-cyclic matrices and the symmetric successive overrelaxation method*, Linear Algebra Appl., 58 (1984), pp. 425–439.

[18] J. H. VERNER AND M. J. M. BERNAL, *On generalizations of the theory of consisting orderings for successive over-relaxation methods*, Numer. Math., 12 (1968), pp. 215–222.

[19] D. M. YOUNG, *Iterative methods for solving partial difference equations of elliptic type*, Trans. Amer. Math. Soc., 76 (1954), pp. 92–111.

[20] D. M. YOUNG AND D. R. KINCAID, *Norms of the Successive Overrelaxation Method and Related Methods*, Report, TNN-94, Computation Center, University of Texas, Austin, 1969.

[21] D. M. YOUNG, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 3, pp. 530–547, July 1996

006

ON THE FACIAL STRUCTURE OF THE SET OF CORRELATION MATRICES*

MONIQUE LAURENT† AND SVATOPLUK POLJAK‡

Abstract. We study the facial structure of the set $\mathcal{E}_{n\times n}$ of correlation matrices (i.e., the positive semidefinite matrices with diagonal entries equal to 1). In particular, we determine the possible dimensions for a face, as well as for a polyhedral face, of $\mathcal{E}_{n\times n}$. It turns out that the spectrum of face dimensions is lacunary and that $\mathcal{E}_{n\times n}$ has polyhedral faces of dimension up to $\approx \sqrt{2n}$. As an application, we describe in detail the faces of $\mathcal{E}_{4\times 4}$. We also discuss results related to optimization over $\mathcal{E}_{n\times n}$.

Key words. correlation matrix, face, max-cut problem

AMS subject classifications. 15A48, 52A20, 90C27

1. Introduction.

A positive semidefinite matrix whose diagonal entries are equal to 1 is called a *correlation matrix*. Let $\mathcal{E}_{n\times n}$ denote the set of $n \times n$ correlation matrices, i.e.,

$$\mathcal{E}_{n\times n} := \{X \in \mathbb{R}^{n\times n} \mid X \succeq 0, x_{ii} = 1 \text{ for all } i = 1, \ldots, n\}.$$

The notation $X \succeq 0$ means that X is a symmetric positive semidefinite matrix. The convex set $\mathcal{E}_{n\times n}$ is called the *elliptope*. Let us recall two previously known results that are also crucial for this paper.

THEOREM 1.1 (see [LT94]). *Let $A \in \mathcal{E}_{n\times n}$ be a correlation matrix of rank r and let $F(A)$ be the smallest face of $\mathcal{E}_{n\times n}$ containing A. Then*

$$\dim F(A) = \binom{r+1}{2} - \operatorname{rank}(v_i v_i^T \mid 1 \le i \le n), \tag{1.1}$$

where $v_1, \ldots, v_n \in \mathbb{R}^r$ is a collection of vectors such that $A = \operatorname{Gram}(v_1, \ldots, v_n)$.

Theorem 1.1 generalizes results of [CM79, Loe80, GPW90], which mainly considered the question of determining the possible ranks for extreme elements of $\mathcal{E}_{n\times n}$. The elliptope is a nonpolyhedral convex set and it has a nonsmooth boundary. The points $X \in \mathcal{E}_{n\times n}$ with full-dimensional normal cone are called *vertices*.

THEOREM 1.2 (see [LP93]). *The elliptope $\mathcal{E}_{n\times n}$ has precisely 2^{n-1} vertices, each of the form aa^T for $a \in \{-1, 1\}^n$.*

Theorem 1.2 was motivated by the fact that $\mathcal{E}_{n\times n}$ is a relaxation of a hard combinatorial optimization problem, namely, the max-cut problem. Indeed, the rank-one matrices of $\mathcal{E}_{n\times n}$ are of the form aa^T for $a \in \{-1, 1\}^n$; they are called *cut matrices* because they correspond to the cuts of the complete graph. The convex hull of the cut matrices defines a polytope, called the *cut polytope* and denoted by $\mathrm{CUT}_{n\times n}$. Then the max-cut problem is the problem of optimizing a linear objective function over the cut polytope. Hence, $\mathcal{E}_{n\times n}$ can be seen as a (nonpolyhedral) relaxation of the cut polytope (see [LP93, La94]). Moreover, a recent result of [GW94] shows that by

* Received by the editors January 3, 1995; accepted for publication (in revised form) by R. Brualdi June 23, 1995.

† LIENS, Ecole Normale Supérieure, 45 rue d'Ulm, 75230 Paris cedex 05, France. This research was done while the author was visiting CWI, Amsterdam.

‡ Fakultät für Mathematik und Informatik, Universität Passau, Innstrasse 33, 94030 Passau, Germany, until his death in April 1995. This research was partially done while the author was visiting CWI, Amsterdam, with a grant from the Stieltjes Institute, whose support is gratefully acknowledged. Also partially supported by Grantova Agentura University Karlovi.

optimizing over the elliptope one obtains a very good approximation for the max-cut problem.

Some other papers [GJSW84, BJT93, BJL, La94] study the projection $\mathcal{E}(G)$ of $\mathcal{E}_{n\times n}$ on the edge set of a graph G; this corresponds to the question of determining what partial matrices can be completed to a positive semidefinite matrix.

The subject of this paper is the facial structure of the elliptope $\mathcal{E}_{n\times n}$. Section 2 contains several old and new preliminary results. In §3, we describe all possible values for the dimension of a face of $\mathcal{E}_{n\times n}$. We show that for all "admissible" values k within the range of (1.1), there exists a face of dimension k. Our further results from §4 concern the polyhedral faces of $\mathcal{E}_{n\times n}$. A polyhedral face is, in some sense, the most "nonsmooth part" of the boundary of $\mathcal{E}_{n\times n}$. We determine the largest possible dimension for a polyhedral face and show that it can be realized by a simplex face whose vertices are cut matrices. In §5, we group some results related to optimization over the elliptope. In particular, we present a link between the faces of the elliptope and the dimension of the optimized eigenspace in the dual problem. Finally, we treat in detail in §6 the elliptope $\mathcal{E}_{4\times 4}$, the elliptope $\mathcal{E}_{3\times 3}$ having been described in [LP93]. We describe the proper faces of $\mathcal{E}_{4\times 4}$, whose possible dimensions are 0, 1, 2, and 3; faces of dimension 1 are edges between two cut matrices and faces of dimension 3 are ismorphic to $\mathcal{E}_{3\times 3}$. The highest dimension for a polyhedral face of $\mathcal{E}_{4\times 4}$ is 2.

2. Old and new basic facts. Throughout the paper, when dealing with matrices, we take as ambient space the set of symmetric matrices equipped with the inner product

$$\langle A, B\rangle := \mathrm{Tr}(AB) = \sum_{1\le i,j\le n} a_{ij}b_{ij}.$$

We start with some well-known facts, formulated in the following two lemmas.

LEMMA 2.1. *Let $x_1,\dots,x_n$ be n linearly independent vectors in $\mathbb{R}^n$. Then the system*

$$\mathcal{S} := \{x_i x_i^T \mid 1\le i\le n\} \cup \{(x_i - x_j)(x_i - x_j)^T \mid 1\le i < j \le n\}$$

is linearly independent.

Proof. Since $\mathcal{S}$ consists of $n + \binom{n}{2} = \binom{n+1}{2}$ elements, it suffices to show that, if X is a symmetric $n\times n$ matrix orthogonal to all members of $\mathcal{S}$, then X is the zero matrix. By assumption, $\langle X, x_i x_i^T\rangle = x_i^T X x_i = 0$ for $i = 1,\dots,n$ and $\langle X, (x_i - x_j)(x_i - x_j)^T\rangle = (x_i - x_j)^T X (x_i - x_j) = 0$, implying that $x_i^T X x_j + x_j^T X x_i = 0$ for $1\le i < j \le n$. We check that $x^T X x = 0$ for all $x\in\mathbb{R}^n$. Indeed, let $x = \sum_{1\le i\le n}\alpha_i x_i$ for some scalars α_i. Then $x^T X x = \sum_{1\le i\le n}\alpha_i^2 x_i^T X x_i + \sum_{1\le i<j\le n}\alpha_i\alpha_j(x_i^T X x_j + x_j^T X x_i) = 0$. This implies that $X = 0$; indeed, if x is an eigenvector of X for the eigenvalue λ, then $0 = x^T X x = \lambda \parallel x \parallel^2$, yielding $\lambda = 0$. □

The *Gram matrix* $\mathrm{Gram}(v_1,\dots,v_k)$ of a collection of vectors $v_1,\dots,v_k$ is the $k\times k$ symmetric matrix whose (i,j)th entry is equal to $v_i^t v_j$. The linear subspace spanned by vectors $v_1,\dots,v_k$ is denoted $\langle v_1,\dots,v_k\rangle$.

LEMMA 2.2. *Let $v_1,\dots,v_k\in\mathbb{R}^n$. Then*

$$\dim\left(\langle v_1,\dots,v_k\rangle\right) = \mathrm{rank}\ \left(\mathrm{Gram}(v_1,\dots,v_k)\right) = \mathrm{rank}\left(\sum_{i=1}^{k} v_i v_i^t\right). \qquad \square$$

2.1. The kernel of a correlation matrix. It is easy to see Lemma 2.3.

LEMMA 2.3. *The relative interior of $\mathcal{E}_{n\times n}$ consists of the positive definite correlation matrices and its relative boundary of the correlation matrices X with* $\operatorname{rank}(X) < n$. □

Let $X \in \mathcal{E}_{n\times n}$. Clearly, each nonzero vector of $\ker(X)$ has at least two nonzero coordinates. It is shown in [DP93b] that every vector $v \in \ker(X)$ is *balanced*, i.e., satisfies

$$|v_i| \le \sum_{1\le j\le n,\ j\ne i} |v_j| \quad \text{for all } i = 1,\dots,n.$$

THEOREM 2.4 (see [DP93b]). *Given a vector $v \in \mathbb{R}^n$, there exists a correlation matrix $X \in \mathcal{E}_{n\times n}$ such that $Xv = 0$ if and only if v is balanced.* □

Note that there exist balanced vectors $v \in \mathbb{R}^n$ for which there exists no matrix $X \in \mathcal{E}_{n\times n}$ for which equality $\ker(X) = \langle v\rangle$ holds. This is the case, for instance, for the vector $v = (n-1, 1, \dots, 1)$; see Theorem 2.6. Call a vector $v \in \mathbb{R}^n$ *strictly balanced* if it satisfies

$$|v_i| < \sum_{1\le j\le n,\ j\ne i} |v_j| \quad \text{for all } i = 1,\dots,n.$$

LEMMA 2.5. *Let $X \in \mathcal{E}_{n\times n}$ with $|x_{ij}| < 1$ for all $i \ne j$. Then every nonzero vector $v \in \ker(X)$ is strictly balanced.*

Proof. Suppose that $|v_1| = |v_2| + \cdots + |v_n|$. From $Xv = 0$, we obtain that $\sum_{2\le i\le n} x_{1i}v_i = -v_1$. Therefore,

$$|v_1| = \left|\sum_{2\le i\le n} x_{1i}v_i\right| \le \sum_{2\le i\le n} |x_{1i}||v_i| \le \sum_{2\le i\le n} |v_i| = |v_1|.$$

Hence, equality holds throughout, which implies that $\sum_{2\le i\le n}(|x_{1i}| - 1)|v_i| = 0$. Therefore, $v_2 = \cdots = v_n = 0$, which is a contradiction. □

THEOREM 2.6. *Let $v \in \mathbb{R}^n$ such that $v_i \ne 0$ for all i. The following statements are equivalent.*

(i) *There exists $X \in \mathcal{E}_{n\times n}$ such that* $\ker(X) = \langle v\rangle$.

(ii) *The vector v is strictly balanced.*

Proof. (i) $\Longrightarrow$ (ii). Let $X \in \mathcal{E}_{n\times n}$ such that $\ker(X) = \langle v\rangle$. Then $|x_{ij}| < 1$ for all $i \ne j$. (If, say, $x_{12} = 1$, then the vector $(1, -1, 0, \dots, 0)$ belongs to $\ker(X)$; hence, it coincides with v, which contradicts the fact that all entries of v are nonzero.) Therefore, v is strictly balanced by Lemma 2.5.

(ii) $\Longrightarrow$ (i). We partly follow the proof of Theorem 3.2 in [DP93b]. We can suppose without loss of generality that $v_1, \dots, v_n > 0$. For $h = 1, \dots, n$, set

$$1 + \epsilon_h := \left(\frac{\sum_{i\ne h} v_i}{v_h}\right)^2;$$

then $\epsilon_h > 0$. Define the vector

$$x_h := (1, \dots, 1, -\sqrt{1+\epsilon_h}, 1, \dots, 1) \in \mathbb{R}^n,$$

where $\sqrt{1+\epsilon_h}$ stands at the hth position. Also set

$$t := \frac{\sum_{1\le h\le n} \frac{1}{\epsilon_h}}{1 + \sum_h \frac{1}{\epsilon_h}}, \quad \alpha_h := \frac{1-t}{\epsilon_h} \quad \text{for } h = 1, \dots, n.$$

Finally, let

$$X := \sum_{1 \le h \le n} \alpha_h x_h x_h^T.$$

Clearly, $X \succeq 0$ as $\alpha_h > 0$ since $0 < t < 1$. One can check that the diagonal entries of X are equal to 1. Moreover, $Xv = 0$ since v is orthogonal to $x_1, \ldots, x_n$ and $\ker(X) = \langle v \rangle$ as the rank of X is equal to the rank of $\{x_1, \ldots, x_n\}$, i.e., to $n-1$ (see Lemma 2.2). □

Note that Theorem 2.6 does not hold if some entries of v are equal to 0. For instance, the vector $v = (0, 1, 1)$ is not strictly balanced but the kernel of the matrix

$$\begin{pmatrix} 1 & 1/2 & -1/2 \\ 1/2 & 1 & -1 \\ -1/2 & -1 & 1 \end{pmatrix}$$

is spanned by v.

2.2. Faces. A convex subset F of a convex set K is called a *face* (or *extreme set*) of K if, for all $x \in F, y, z \in K$, $0 \le \alpha \le 1$, $x = \alpha y + (1-\alpha)z$ implies that $y, z \in F$. We recall some facts, taken from [LP93], on the faces of $\mathcal{E}_{n\times n}$.

THEOREM 2.7 (see [LP93]). *For every subspace V of $\mathbb{R}^n$, the set*

$$F_V := \{X \in \mathcal{E}_{n\times n} \mid \ker(X) \supseteq V\}$$

is a face of $\mathcal{E}_{n\times n}$. Conversely, every face F of $\mathcal{E}_{n\times n}$ is of the form F_V, where $V = \bigcap_{X\in F} \ker(X)$. In particular, given $X_0 \in \mathcal{E}_{n\times n}$, let $F(X_0)$ denote the smallest face of $\mathcal{E}_{n\times n}$ that contains X_0. Then

$$F(X_0) = \{X \in \mathcal{E}_{n\times n} \mid \ker(X) \supseteq \ker(X_0)\}. \qquad \square$$

Faces of $\mathcal{E}_{n\times n}$ can be "lifted" to faces of $\mathcal{E}_{(n+1)\times(n+1)}$ (of the same dimension) in the following way. Let X be a symmetric $n \times n$ matrix with diagonal entries equal to 1, of the form

$$X = \left(\begin{array}{c|c} Y & a \\ \hline a^T & 1 \end{array}\right),$$

where $a \in \mathbb{R}^{n-1}$ and Y is a symmetric $(n-1) \times (n-1)$ matrix. Consider the $(n+1) \times (n+1)$ symmetric matrices X' and X'' defined by

$$X' = \left(\begin{array}{c|c|c} Y & a & a \\ \hline a^T & 1 & 1 \\ \hline a^T & 1 & 1 \end{array}\right), \quad X'' = \left(\begin{array}{c|c|c} Y & a & -a \\ \hline a^T & 1 & -1 \\ \hline -a^T & -1 & 1 \end{array}\right).$$

For a subset F of $\mathcal{E}_n$, set $F' := \{X' \mid X \in F\}$ and $F'' := \{X'' \mid X \in F\}$. Then

$$X \in \mathcal{E}_n \iff X' \in \mathcal{E}_{n+1} \iff X'' \in \mathcal{E}_{n+1},$$

$$F \text{ is a face of } \mathcal{E}_n \iff F' \text{ is a face of } \mathcal{E}_{n+1} \iff F'' \text{ is a face of } \mathcal{E}_{n+1}.$$

Clearly, F, F', and F'' all have the same dimension. We say that F', F'' are *liftings* of the face F. Moreover, if F is a face of $\mathcal{E}_{n\times n}$ and $V = \bigcap_{X\in F} \ker(X)$, then the subspace

$V' := \bigcap_{Y \in F'} \ker(Y)$ is generated by the vectors $(v,0)$ $(v \in V)$ and $(0,\ldots,0,1,-1)$, while the subspace $V'' := \bigcap_{Y \in F''} \ker(Y)$ is generated by the vectors $(v,0)$ $(v \in V)$ and $(0,\ldots,0,1,1)$. The following result permits to recognize if a face arises as a lifting of another face.

LEMMA 2.8. *Let F be a face of $\mathcal{E}_{(n+1)\times(n+1)}$ and $V = \bigcap_{X \in F} \ker(X)$. Then F is a lifting of a face of $\mathcal{E}_{n\times n}$ if and only if there exists a vector $v \in V$ having exactly two nonzero coordinates.*

Proof. Necessity is clear. Conversely, suppose that $v \in V$ with $v = (0,\ldots,0,\alpha,\beta)$. Since v is balanced, we deduce that $|\alpha| = |\beta|$, i.e., $\alpha = \pm\beta$. This implies easily that F is a lifting of a face of $\mathcal{E}_{n\times n}$. □

2.3. The normal cone. Given a convex set K in a space V with inner product $\langle , \rangle$ and a boundary point x_0 of K, the *normal cone* $\mathcal{N}(K,x_0)$ at x_0 is defined by

$$\mathcal{N}(K,x_0) = \{c \in V \mid \langle c,x\rangle \leq \langle c,x_0\rangle \text{ for all } x \in K\}.$$

The normal cone $\mathcal{N}(\mathcal{E}_{n\times n},A)$ of a matrix $A \in \mathcal{E}_{n\times n}$ will be denoted as $\mathcal{N}(A)$. It can be characterized as follows.

THEOREM 2.9 (see [LP93]). *We have*

$$\mathcal{N}(A) = \{D - M \mid D \text{ is a diagonal matrix}, M \succeq 0, \langle M,A\rangle = 0\}. \quad \square$$

In fact, we can compute the exact dimension of the normal cone at a correlation matrix A in terms of the rank of A.

THEOREM 2.10. *Let $A \in \mathcal{E}_{n\times n}$ with $q := \dim \ker(A)$. Then*

$$\dim \mathcal{N}(A) = \binom{q+1}{2} + n.$$

Proof. Let $b_1,\ldots,b_q$ be linearly independent vectors in $\ker(A)$. Then the matrices $-(b_i+b_j)(b_i+b_j)^T$ $(1 \leq i \leq j \leq q)$ belong to $\mathcal{N}(A)$. The elementary diagonal matrix E_{ii} $(1 \leq i \leq n)$ is defined as the matrix with all entries 0 but the (i,i)th entry which equals 1. All the n matrices E_{ii} also belong to $\mathcal{N}(A)$. We show that the system $\{(b_i+b_j)(b_i+b_j)^T \mid 1 \leq i \leq j \leq q\} \cup \{E_{ii} \mid 1 \leq i \leq n\}$ is linearly independent. For this, let λ_{ij}, μ_i be scalars such that

$$\sum_{1\leq i\leq j\leq q} \lambda_{ij}(b_i+b_j)(b_i+b_j)^T + \sum_{1\leq i\leq n} \mu_i E_{ii} = 0.$$

We show that all λ_{ij}'s and μ_i's are equal to 0. Let $u \in (\ker(A))^\perp$. Applying the above relation to u, we obtain that $\sum_{1\leq i\leq n} \mu_i E_{ii} u = 0$, i.e., $\mu_i u_i = 0$ for all $i = 1,\ldots,n$.

CLAIM 2.11. *For all $i \in \{1,\ldots,n\}$, there exists $u \in (\ker(A))^\perp$ such that $u_i \neq 0$.*

Proof. Suppose that $u_i = 0$ for all $u \in (\ker(A))^\perp$. Then $(\ker(A))^\perp \subseteq \{u \in \mathbb{R}^n \mid u_i = 0\}$. Therefore, $\ker(A) \supseteq \{u \in \mathbb{R}^n \mid u_i = 0\}^\perp$. This implies that the ith unit vector belongs to $\ker(A)$, which is a contradiction with Theorem 2.4. □

Therefore, $\mu_i = 0$ for all $i = 1,\ldots,n$. Using Lemma 2.1, we obtain that $\lambda_{ij} = 0$ for all $1 \leq i \leq j \leq q$. Hence, we have found a system of $\binom{q+1}{2}+n$ linearly independent members of $\mathcal{N}(A)$. This shows that

$$\dim \mathcal{N}(A) \geq \binom{q+1}{2} + n.$$

We now show the converse inequality. Let $\mathcal{B}$ be a system of linearly independent members of $\mathcal{N}(A)$ of maximum cardinality. Since all diagonal matrices belong to

$\mathcal{N}(A)$, we can suppose without loss of generality that $\mathcal{B}$ is composed of the elementary diagonal matrices $E_{11}, \ldots, E_{nn}$ together with some matrices $-M_1, \ldots, -M_k$, where each M_i is positive semidefinite and satisfies $\langle M_i, A \rangle = 0$. By the latter condition, all matrices M_i belong to the set $F := \{M \succeq 0 \mid \ker(M) \supseteq (\ker(A))^{\perp}\}$. One can check that the set F has dimension $\binom{q+1}{2}$ (see also [HW87]). This implies that $k \leq \binom{q+1}{2}$. Therefore, $\dim \mathcal{N}(A) \leq \binom{q+1}{2} + n$. This concludes the proof. □

Note that Theorem 2.10 implies the characterization of the vertices of $\mathcal{E}_{n\times n}$ from Theorem 1.2. Let $A \in \mathcal{E}_{n\times n}$. Suppose that A has rank r and is the Gram matrix of the vectors $v_1, \ldots, v_n \in \mathbb{R}^r$. Set $g := \dim (v_1 v_1^T, \ldots, v_n v_n^T)$. Then the dimension of the face $F(A)$ and of the normal cone of A are linked by

$$\dim F(A) + \dim \mathcal{N}(A) = \binom{n+1}{2} + n - r(n-r) - g. \tag{2.1}$$

(This follows from Theorems 1.1 and 2.10.) It implies the following corollary.

COROLLARY 2.12.

$$\binom{n+1}{2} - r(n-r) \leq \dim F(A) + \dim \mathcal{N}(A) \leq \binom{n+1}{2} - (r-1)(n-r). \qquad \square$$

Note that equality holds in the upper bound if, for instance, A is a cut matrix or A lies in the relative interior of $\mathcal{E}_{n\times n}$.

3. The dimension of the faces of $\mathcal{E}_{n\times n}$. We group in this section several results on the faces of the elliptope $\mathcal{E}_{n\times n}$. Using a result of [LT94] recalled in Theorem 1.1 above, we describe all the possible values that can take the dimension of a face of $\mathcal{E}_{n\times n}$; it turns out that the spectrum of feasible dimensions for proper faces is a union of intervals that ranges from 0 to $\binom{n-1}{2}$.

Suppose $A \in \mathcal{E}_{n\times n}$ has rank r. Then A is the Gram matrix of a set of vectors $v_1, \ldots, v_n \in \mathbb{R}^r$ of rank r, i.e.,

$$A_{ij} = v_i^T v_j \quad \text{for } 1 \leq i, j \leq n.$$

A *perturbation* of A is any symmetric matrix B such that $A \pm tB \in \mathcal{E}_{n\times n}$ for some small $t > 0$. Then the dimension of the face $F(A)$ (the smallest face of $\mathcal{E}_{n\times n}$ containing A) is defined as the dimension of the space of perturbations of A. Let Z denote the $r \times n$ matrix whose columns are $v_1, \ldots, v_n$; thus, $A = Z^T Z$. Li and Tam [LT94] show that B is a perturbation of A if and only if

$$B = Z^T R Z, \tag{3.1}$$

where R belongs to the orthogonal complement of $\langle v_1 v_1^T, \ldots, v_n v_n^T \rangle$ in the space of symmetric $r \times r$ matrices (this latter condition ensures that the diagonal entries of B are equal to 0). This implies that the dimension of $F(A)$ can be expressed as in (1.1). More generally, we have the following result.

THEOREM 3.1.

(i) *Let $A \in \mathcal{E}_{n\times n}$ of rank r and let k denote the dimension of $F(A)$. Then $\binom{r+1}{2} - n \leq k \leq \binom{r}{2}$.*

(ii) *Let $r, k \geq 0$ be integers such that $1 \leq r \leq n$ and $\max(0, \binom{r+1}{2} - n) \leq k \leq \binom{r}{2}$. Then there exists a matrix $A \in \mathcal{E}_{n\times n}$ of rank r and for which $\dim(F(A)) = k$.*

Proof. (i) follows from the inequalities $r \leq \operatorname{rank}(v_i v_i^T \mid 1 \leq i \leq n) \leq n$. (The upper bound is obvious. For the lower bound, observe that the set $(v_1, \ldots, v_n)$ has

rank r and that if, say, $v_1, \ldots, v_r$ are linearly independent, then $v_1 v_1^T, \ldots, v_r v_r^T$ are also linearly independent by Lemma 2.1.)

For (ii), we use a construction proposed in [LT94] (also in [GPW90]). Let $e_1, \ldots, e_r \in \mathbb{R}^r$ denote the unit vectors in $\mathbb{R}^r$ and set

$$w_{ij} := \frac{1}{\sqrt{2}}(e_i + e_j) \text{ for } 1 \leq i < j \leq r.$$

One can easily check that the $\binom{r+1}{2}$ matrices $\{e_i e_i^T \mid 1 \leq i \leq r\} \cup \{w_{ij} w_{ij}^T \mid 1 \leq i < j \leq r\}$ are linearly independent.

Let k be such that $0 \leq k \leq \binom{r}{2}$. Suppose first that $n = \binom{r+1}{2} - k$. Then $r \leq n \leq \binom{r+1}{2}$. Define A as the Gram matrix of the n vectors $e_1, \ldots, e_r$ together with $n - r$ of the vectors w_{ij}. By construction, A has rank r. Using relation (1.1), one obtains that $\dim(F(A)) = \binom{r+1}{2} - n = k$. This shows (ii) in the case when $n = \binom{r+1}{2} - k$. Suppose now that $n > \binom{r+1}{2} - k$. Then we choose for A a lifting of the matrix defined above; for instance, we can take for A the Gram matrix of the n vectors e_1 (repeated $n - \binom{r+1}{2} + k + 1$ times), $e_2, \ldots, e_r$, together with $\binom{r}{2} - k$ of the vectors w_{ij}. □

A correlation matrix X is called *extreme* if the set $F := \{X\}$ is a zero-dimensional face of $\mathcal{E}_{n\times n}$. Thus, as a special case of Theorem 3.1, we obtain the result of Li and Tam.

COROLLARY 3.2 (see [LT94]). *Let $r_{\max}$ be the largest integer r such that $\binom{r+1}{2} \leq n$. Then*

(i) $1 \leq \operatorname{rank}(X) \leq r_{\max}$ *for every extreme correlation matrix $X \in \mathcal{E}_{n\times n}$.*
(ii) *For every r, $1 \leq r \leq r_{\max}$, there is an extreme correlation matrix $X \in \mathcal{E}_{n\times n}$ of rank r.* □

As shown in [LP93], any two cut matrices of $\mathcal{E}_{n\times n}$ form an edge (one-dimensional face) of $\mathcal{E}_{n\times n}$. For $n = 3, 4$, these are the only edges of $\mathcal{E}_{n\times n}$ (see §6). However, for $n \geq 5$, $\mathcal{E}_{n\times n}$ has edges whose extremities are *not* cut matrices. A construction for such an edge is given in Example 3.3.

Example 3.3. We apply the construction from the proof of Theorem 3.1 (ii) in the case $n = 5$, $r = 3$, $k = 1$. Let $A \in \mathcal{E}_{5\times 5}$ be the Gram matrix of the vectors $e_1 = (1,0,0)$, $e_2 = (0,1,0)$, $e_3 = (0,0,1)$, $w_{12} = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0)$, and $w_{13} = (\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}})$, i.e.,

$$A = \begin{pmatrix} 1 & 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 1 & \frac{1}{2} \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} & \frac{1}{2} & 1 \end{pmatrix}.$$

Hence, $F(A)$ is an edge of $\mathcal{E}_{5\times 5}$. In order to describe this edge, we note that $\ker(A)$ is spanned by the vectors

$$a := (-1, -1, 0, \sqrt{2}, 0), \; b := (-1, 0, -1, 0, \sqrt{2}).$$

Then $X \in \mathcal{E}_{5\times 5}$ belongs to $F(A)$ if and only if $Xa = 0$ and $Xb = 0$. One can check

that X must be of the following form:

$$X(\alpha) := \begin{pmatrix} 1 & 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 1 & \alpha & \frac{1}{\sqrt{2}} & \frac{\alpha}{\sqrt{2}} \\ 0 & \alpha & 1 & \frac{\alpha}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{\alpha}{\sqrt{2}} & 1 & \frac{1+\alpha}{2} \\ \frac{1}{\sqrt{2}} & \frac{\alpha}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1+\alpha}{2} & 1 \end{pmatrix},$$

where $-1 \leq \alpha \leq 1$. Hence the edge $F(A)$ has the matrices $X(-1)$ and $X(1)$ as extremities, where $X(-1)$ and $X(1)$ are of the above form for $\alpha = -1, 1$. □

As an application of Theorem 3.1, we can describe the range $\mathcal{D}_n$ of the values taken by the dimension of the faces of $\mathcal{E}_{n\times n}$. Let s_n denote the smallest integer such that $\binom{s_n+2}{2} - n > \binom{s_n}{2} + 1$, i.e., $2s_n > n$ or, equivalently,

$$s_n = \left\lfloor \frac{n}{2} \right\rfloor + 1.$$

Then

$$\mathcal{D}_n = \left[0, \binom{s_n}{2}\right] \cup \bigcup_{r=s_n+1}^{n} \left[\binom{r+1}{2} - n, \binom{r}{2}\right]. \tag{3.2}$$

(Given two integers a, b, $[a, b]$ denotes the set of integers x lying between a and b.) For instance,

$$s_3 = 2,\ \mathcal{D}_3 = [0,1] \cup \{3\},$$

$$s_4 = 3,\ \mathcal{D}_4 = [0,3] \cup \{6\},$$

$$s_5 = 3,\ \mathcal{D}_5 = [0,3] \cup [5,6] \cup \{10\},$$

$$s_6 = 4,\ \mathcal{D}_6 = [0,6] \cup [9,10] \cup \{15\},$$

$$s_7 = 4,\ \mathcal{D}_7 = [0,6] \cup [8,10] \cup [14,15] \cup \{21\}.$$

In particular, the largest dimension of a proper face of $\mathcal{E}_{n\times n}$ is $\binom{n-1}{2}$. We give below a direct simple proof of this fact which permits us, moreover, to show that every face of $\mathcal{E}_{n\times n}$ of dimension $\binom{n-1}{2}$ is a lifting of $\mathcal{E}_{(n-1)\times(n-1)}$.

PROPOSITION 3.4. *Let F be a proper face of $\mathcal{E}_{n\times n}$. Then* $\dim(F) \leq \binom{n-1}{2}$, *with equality if and only if F is a lifting of $\mathcal{E}_{(n-1)\times(n-1)}$.*

Proof. Let F be a proper face of $\mathcal{E}_{n\times n}$. Then $F = F_V$ for some subspace V of $\mathbb{R}^n$, $V \neq \{0\}$. Let $v \in V, v \neq 0$. We can suppose that $v_1 \neq 0$. Then $Xv = 0$ for all $X \in F$. The equation $Xv = 0$ can be written as the following system of n equations in the $\binom{n}{2}$ variables x_{ij} $(1 \leq i < j \leq n)$:

$$\begin{cases} x_{12}v_2 + x_{13}v_3 + \ldots + x_{1n}v_n & = -v_1, \\ x_{12}v_1 + x_{23}v_3 + \cdots & = -v_2, \\ x_{13}v_1 + x_{23}v_2 + \cdots & = -v_3, \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots & \\ x_{1n}v_1 + \cdots & = -v_n. \end{cases}$$

Since $v_1 \neq 0$, the matrix of the system obviously has rank $\geq n-1$. This implies that $\dim(F) \leq \binom{n}{2} - (n-1) = \binom{n-1}{2}$. Moreover, the equality $\dim(F) = \binom{n-1}{2}$ holds if and only if the matrix of the system has rank equal to $n-1$. It is not difficult to check that this holds only if $v_i v_j = 0$ for all $2 \leq i < j \leq n$. Hence, we may suppose, for instance, that $v_3 = v_4 = \cdots = v_n = 0$. Hence, v has only two nonzero components. Using Lemma 2.8, we obtain that F is a lifting of a face (of the same dimension $\binom{n-1}{2}$) of $\mathcal{E}_{(n-1)\times(n-1)}$. Therefore, F is a lifting of $\mathcal{E}_{(n-1)\times(n-1)}$. □

We conclude with an example of a face of the next smaller dimension $\binom{n-1}{2} - 1$.

Example 3.5. Consider the face

$$F := \{X \in \mathcal{E}_{n\times n} \mid Xe = 0\},$$

where e is the all-ones vector. Then $\dim F = \binom{n-1}{2} - 1$. (To see it, one can proceed in the same way as in the proof of Proposition 3.4. Namely, the condition $Xe = 0$ can be rewritten as the system

$$\sum_{j=1,\ldots,n,\ j\neq i} x_{ij} = -1 \quad \text{for all } i = 1,\ldots,n.$$

Since the matrix of this system has rank n, we deduce that $\dim F = \binom{n}{2} - n = \binom{n-1}{2} - 1$.) Let X_0 denote the matrix with ones on the diagonal and $-\frac{1}{n-1}$ on the off-diagonal positions. Then X_0 belongs to the relative interior of F since $\ker(X_0) = \langle e \rangle$. Hence, $F = F(X_0)$.

Suppose that n is even. Then F contains the cut matrices ff^T for all vectors $f \in \{-1,1\}^n$ having exactly $\frac{n}{2}$ entries 1 and $\frac{n}{2}$ entries -1. Hence, F contains $\frac{1}{2}\binom{n}{n/2}$ cut matrices.

Let us look in more detail at the case $n = 4$. Then one can easily check that a matrix $X \in \mathcal{E}_{4\times 4}$ belongs to the face F if and only if it is of the form

$$X = \begin{pmatrix} 1 & x & y & -1-x-y \\ x & 1 & -1-x-y & y \\ y & -1-x-y & 1 & x \\ -1-x-y & y & x & 1 \end{pmatrix},$$

with the conditions $-1 \leq x, y \leq 1$, and $x + y \leq 0$. Therefore, F is a polyhedral face of $\mathcal{E}_{4\times 4}$, whose vertices are the three cut matrices ff^T for $f = (1,1,-1,-1)$, $(1,-1,-1,1)$, and $(1,-1,1,-1)$.

Finally, note that, for any $n \geq 5$, the face F cannot be a polyhedral face because its dimension is too large; see Theorem 4.1. □

4. Polyhedral faces of $\mathcal{E}_{n\times n}$. We consider here the polyhedral faces of the elliptope $\mathcal{E}_{n\times n}$. In particular, we describe the range of their feasible dimensions.

As was mentioned in Proposition 3.4, every face of $\mathcal{E}_{n\times n}$ of dimension $\binom{n-1}{2}$ is isomorphic to $\mathcal{E}_{(n-1)\times(n-1)}$. Hence, $\mathcal{E}_{n\times n}$ has no polyhedral face of dimension $\binom{n-1}{2}$. In fact, we can show that the feasible dimensions for polyhedral faces of $\mathcal{E}_{n\times n}$ range from 0 to k_n, where k_n is the largest integer such that $\binom{k_n+1}{2} \leq n-1$. We also consider the polyhedral faces of $\mathcal{E}_{n\times n}$ having only cut matrices as vertices, i.e., the faces of $\mathcal{E}_{n\times n}$ that are inherited from the cut polytope. It turns out that such a face is necessarily a simplex. In fact, a simplex face of dimension k can be constructed for any $k \leq k_n$.

THEOREM 4.1. *Let F be a polyhedral face of $\mathcal{E}_{n\times n}$ of dimension $k-1$. Then $\binom{k}{2} \leq n-1$. Moreover, if all vertices of F are cut matrices, then F is a simplex.*

Proof. Let $F_0 \subset F_1 \subset \cdots \subset F_i \subset F_{i+1} \subset \cdots \subset F_{k-1} := F$ be a chain of faces of F, where F_i has dimension i for each $i = 0, 1, \ldots, k-1$. Using Theorem 2.7, each F_i is of the form $F_{V_i} = \{X \in \mathcal{E}_{n\times n} \mid V_i \subseteq \ker(X)\}$, where the V_i are subspaces of $\mathbb{R}^n$ forming a strict chain:

$$V_0 \supset V_1 \supset \cdots \supset V_i \supset V_{i+1} \supset \cdots \supset V_{k-1}.$$

Then $\dim(V_{k-1}) \leq \dim(V_0) - k + 1 \leq n - 1 - k + 1 = n - k$. Let X be an interior point of F and let r denote the rank of X. Then $r = n - \dim(V_{k-1}) \geq k$. Using the dimension formula (1.1), we deduce that $k - 1 = \dim(F) \geq \binom{r+1}{2} - n \geq \binom{k+1}{2} - n$. This implies that $n \geq \binom{k}{2} + 1$.

Suppose now that all the vertices of F are cut matrices, say, of the form $f_h f_h$ for $h \in H$, where $f_h \in \{-1,1\}^n$ for all $h \in H$. Then

$$V_{k-1} = \bigcap_{h\in H} \ker(f_h f_h^T) = \langle f_h \mid h \in H\rangle^{\perp}.$$

Hence, $\dim(V_{k-1}) = n - \dim(\langle f_h \mid h \in H\rangle) \leq n - k$, which implies that

$$\dim(\langle f_h \mid h \in H\rangle) \geq k.$$

Let $f_0, f_1, \ldots, f_{k-1}$ be k linearly independent vectors in the set $\{f_h \mid h \in H\}$. Then the vertices $f_i f_i^T$ $(i = 0, 1, \ldots, k-1)$ affinely span the polyhedron F. We show that they are the only vertices of F. For this, let X be another vertex of F. Then $X = \sum_{0\leq i\leq k-1} \alpha_i f_i f_i^T$ with $\sum_{0\leq i\leq k-1} \alpha_i = 1$. We show that each α_i is nonnegative. Indeed, let

$$u \in \langle f_j \mid j = 0, 1, \ldots, k-1, j \neq i\rangle^{\perp} \cap \langle f_0, f_1, \ldots, f_{k-1}\rangle$$

such that $u \neq 0$. Then $u^T X u = \alpha_i (u^T f_i)^2 \geq 0$ with $u^T f_i \neq 0$, yielding $\alpha_i \geq 0$. Hence, X is a vertex of F which can be written as a convex combination of other vertices of F. This shows that $f_0 f_0^T, \ldots, f_{k-1} f_{k-1}^T$ are the only vertices of F. Therefore, F is a simplex. □

We propose in Proposition 4.7 a construction for polyhedral faces of dimension $k-1$ for each integer k such that $\binom{k}{2} \leq n - 1$. For this, we state an intermediate result.

We recall the following notation. Given two vectors $x, y \in \mathbb{R}^n$, their *Hadamard product* is the vector $z := x \circ y \in \mathbb{R}^n$ with entries $z_i := x_i y_i$.

THEOREM 4.2. *Let $f_1, \ldots, f_k \in \{-1,1\}^n$ and set $e := (1, \ldots, 1) \in \{-1,1\}^n$. Suppose that the following assertions hold.*

(i) *The vectors $\{f_1, \ldots, f_k\}$ are linearly independent.*

(ii) *The vectors $\{f_h \circ f_{h'} \mid 1 \leq h < h' \leq k\} \cup \{e\}$ are linearly independent.*

Then the set $F := \mathrm{Conv}(f_h f_h^T \mid h = 1, \ldots, k)$ is a face of $\mathcal{E}_{n\times n}$ of dimension $k-1$.

(Here "Conv" denotes the operation of taking the convex hull.) Note that the face F constructed in the previous theorem is a simplex face with cut matrices as vertices.

Proof. Set

$$X_0 := \frac{1}{k}\left(\sum_{1\leq h\leq k} f_h f_h^T\right).$$

Then $\ker(X_0) = \langle f_1, \ldots, f_k \rangle^\perp$. Therefore, by (i), X_0 has rank k. Let $F(X_0)$ denote the smallest face of $\mathcal{E}_{n\times n}$ containing X_0. Clearly, $F(X_0)$ contains F. Our goal is to show that $F(X_0) = F$.

Consider the $k \times n$ matrix M whose rows are the vectors $f_1, \ldots, f_k$. Denote by $v^1, \ldots, v^n \in \mathbb{R}^k$ its columns. Set $w^i := \frac{1}{\sqrt{k}} v^i$ for $i = 1, \ldots, n$. It is easy to see that X_0 is equal to the Gram matrix of $w^1, \ldots, w^n$. Therefore, by the dimension formula (1.1),

$$\dim\ F(X_0) = \binom{k+1}{2} - \operatorname{rank}\{w^1(w^1)^T, \ldots, w^n(w^n)^T\}.$$

CLAIM 4.3. $\operatorname{rank}(w^1(w^1)^T, \ldots, w^n(w^n)^T) \geq \binom{k}{2} + 1$.

Proof. By assumption (ii), the vectors $\{f_h \circ f_{h'} \mid 1 \leq h < h' \leq k\} \cup \{e\}$ are linearly independent in $\mathbb{R}^n$. Let I be a subset of $\{1, \ldots, n\}$ of size $\binom{k}{2} + 1$ corresponding to the positions of independent coordinates. We show that the set $\{w^i(w^i)^T \mid i \in I\}$ is linearly independent. For this suppose that

$$\sum_{i\in I} \lambda_i w^i (w^i)^T = 0.$$

Note that $w^i(w^i)^T(h,h') = \frac{1}{k} f_h(i) f_{h'}(i)$, which is equal to $\frac{1}{k}(f_h \circ f_{h'})(i)$ if $h \neq h'$ and to $\frac{1}{k} e(i)$ if $h = h'$. This implies that all λ_i's are zero. □

As a consequence of the previous claim, we deduce that

$$\dim\ F(X_0) \leq \binom{k+1}{2} - \binom{k}{2} - 1 = k - 1.$$

On the other hand, $\dim\ F(X_0) \geq \dim(F) = k - 1$. Therefore, $\dim\ F(X_0) = \dim(F) = k - 1$. This implies, in particular, that $F(X_0)$ is contained in the affine hull of $\{f_1 f_1^T, \ldots, f_k f_k^T\}$. Now, by an argument similar to the one used in the proof of Theorem 4.1, we show that $F(X_0) \subseteq F$. For this, let $X \in F(X_0)$; then $X = \sum_{1\leq h\leq k} \mu_h f_h f_h^T$ with $\sum_{1\leq h\leq k} \mu_h = 1$. We claim that $\mu_h \geq 0$ for all h, which will imply that $X \in F$. Indeed, take a nonzero vector u in the intersection of the spaces $\langle f_1, \ldots, f_k \rangle$ and $\langle f_1, \ldots, f_{k-1} \rangle^\perp$. Then $u^T X u = \mu_k (u^T f_k)^2 \geq 0$, implying that $\mu_k \geq 0$. The same argument shows that all μ_h's are nonnegative. □

Remark 4.4. We can suppose without loss of generality in Theorem 4.2 that the vectors $f_1, \ldots, f_k$ have a common entry equal to 1, say, $f_h(n) = 1$ for $h = 1, \ldots, k$. Set $S_h := \{i \mid f_h(i) = 1\}$ for $h = 1, \ldots, k$. It is easy to check that assumption (ii) of Theorem 4.2 can be reformulated as the following.

(iii) *The* $\binom{k}{2}$ *vectors* $\chi^{S_h \triangle S_{h'}}$ $(1 \leq h < h' \leq k)$ *are linearly independent.* (Here χ^A denotes the 0,1-incidence vector of the set A and $A \triangle B := (A \setminus B) \cup (B \setminus A)$ denotes the symmetric difference of the sets A and B.) □

We say that the sets $S_1, \ldots, S_k \subseteq V := \{1, \ldots, n\}$ are in *general position* if each of the 2^k sets $\bigcap_{h\in H} S_h \cap \bigcap_{h\notin H} (V \setminus S_h)$ is nonempty for $H \subseteq \{1, \ldots, k\}$. Then $k \leq \log_2 n$. We say that the vectors $f_1, \ldots, f_k \in \{-1, 1\}^n$ are in *general position* if the sets $S_h := \{i \mid f_h(i) = 1\}$ $(h = 1, \ldots, k)$ are in general position.

COROLLARY 4.5. *Let* $f_1, \ldots, f_k \in \{-1, 1\}^n$ *be in general position. Then the set* $\operatorname{Conv}(f_1 f_1^T, \ldots, f_k f_k^T)$ *is a face of* $\mathcal{E}_{n\times n}$.

Proof. By Theorem 4.2 and Remark 4.4, it suffices to verify that conditions (i) and (iii) hold, which can be easily done. □

Example 4.6. Let $n = 4$, $f_1 = (1, -1, -1, -1)$, $f_2 = (1, -1, 1, 1)$, $f_3 = (1, 1, -1, 1)$. The sets $S_1 := \{1\}$, $S_2 := \{1, 3, 4\}$, $S_3 := \{1, 2, 4\}$ are not in general position but nevertheless satisfy assumption (iii). Also (i) holds. Hence, the set $\operatorname{Conv}(f_1 f_1^T, f_2 f_2^T, f_3 f_3^T)$

is a polyhedral face of $\mathcal{E}_{4\times 4}$ of dimension 2. Note that this face falls into the category of the so-called *elliptic* faces of $\mathcal{E}_{4\times 4}$ (see §6). Also, $F = F_V$, where $V = \langle f_1, f_2, f_3\rangle^{\perp} = \langle (1,1,1,-1)\rangle$. □

PROPOSITION 4.7. *For each integer k such that $\binom{k}{2} + 1 \leq n$, the elliptope $\mathcal{E}_{n\times n}$ has a polyhedral face of dimension $k-1$ (which is a simplex with cut matrices as vertices).*

Proof. It is enough to show the result for $n = \binom{k}{2} + 1$ (for larger values of n, apply lifting). Let G denote the graph with node set $\{1,\ldots,k,k+1\}$, obtained from the complete graph K_k on $\{1,\ldots,k\}$ by adding an edge e, say, $e = (1, k+1)$. We consider the edge set of G as our groundset of n elements. For $h = 1,\ldots,k$, let S_h denote the set of edges in the star of the node h plus the edge e, i.e., S_h consists of the edges (h,i) $(i \in \{1,\ldots,k\} \setminus \{h\})$ together with the edge e. Let f_h denote the ± 1-incidence vector of S_h. Then $\mathrm{Conv}(f_1 f_1^T, \ldots, f_k f_k^T)$ is a face of $\mathcal{E}_{n\times n}$ (since assumptions (i), (iii) can be easily checked to hold). □

As an application of Theorem 4.1 and Proposition 4.7, we obtain that the largest dimension of a polyhedral face of $\mathcal{E}_{n\times n}$ is equal to k_n, the largest integer such that $\binom{k_n+1}{2} \leq n-1$, i.e.,

$$k_n = \left\lfloor \frac{\sqrt{8n-7}-1}{2} \right\rfloor.$$

COROLLARY 4.8. *The maximum dimension of a polyhedral face of the elliptope $\mathcal{E}_{n\times n}$ is*

$$\left\lfloor \frac{\sqrt{8n-7}-1}{2} \right\rfloor. \qquad \square$$

Remark 4.9. It was shown in [DLP92] that, if the vectors $f_1,\ldots,f_k$ are in general position, then the set $F := \mathrm{Conv}(f_1 f_1^T, \ldots, f_k f_k^T)$ is a face of the metric polytope and, thus, of the cut polytope $\mathrm{CUT}_{n\times n}$. We recall that the *metric polytope* $\mathrm{MET}_{n\times n}$ is defined as

$$\begin{aligned} \mathrm{MET}_{n\times n} := \{X \in \mathrm{SYM}_{n\times n} \mid\ & X_{ii} = 1 && \text{for } i = 1,\ldots,n,\\ & X_{ij} - X_{ik} - X_{jk} \geq -1 && \text{for } 1 \leq i,j,k \leq n,\\ & X_{ij} + X_{ik} + X_{jk} \geq -1 && \text{for } 1 \leq i,j,k \leq n\}. \end{aligned}$$

Hence, the metric polytope is a linear relaxation of the cut polytope; see [LPR95] for more details. Corollary 4.5 shows that the set F is also a face of the elliptope $\mathcal{E}_{n\times n}$. Therefore, the elliptope $\mathcal{E}_{n\times n}$ and the metric polytope $\mathrm{MET}_{n\times n}$ are two distinct relaxations of the cut polytope $\mathrm{CUT}_{n\times n}$ that share many common faces, at least up to dimension $\log_2 n$. □

5. Optimization aspects. Let us consider the optimization problem

$$\begin{aligned} &\min && \langle C, X\rangle \\ &\text{s.t.} && X \in \mathcal{E}_{n\times n}, \end{aligned} \tag{5.1}$$

where C is a symmetric $n \times n$ matrix. Recall that

$$\langle C, X\rangle = \mathrm{Tr}(CX) = \sum_{i,j=1,\ldots,n} c_{ij} x_{ij}$$

and let J denote the all-ones matrix. The problem (5.1) is of interest because it is related to the max-cut problem. To be more precise, the problem

$$\begin{array}{lll} \max & \frac{1}{2}\sum_{1\le i<j\le n} c_{ij}(1-x_{ij}) & = \frac{1}{4}\langle C, J\rangle - \\ \text{s.t.} & X \in \mathcal{E}_{n\times n} & \end{array} \qquad \begin{array}{ll} \min & \frac{1}{4}\langle C, X\rangle \\ \text{s.t.} & X \in \mathcal{E}_{n\times n} \end{array} \tag{5.2}$$

provides a good approximation of the max-cut problem

$$\begin{array}{ll} \max & \frac{1}{2}\sum_{1\le i<j\le n} c_{ij}(1-a_i a_j) \\ \text{s.t.} & a \in \{-1,1\}^n. \end{array} \tag{5.3}$$

(For various results concerning the approximation of (5.3) by (5.2) we refer to the following papers: worst case bound of the approximation [GW94], asymptotic optimality of the approximation [DP93a], complexity and further aspects [DP93b, LP93].)

Let F_C denote the set of optimum solutions to the problem (5.2), i.e.,

$$F_C = \{A \in \mathcal{E}_{n\times n} \mid \langle C, A\rangle \le \langle C, X\rangle \text{ for all } X \in \mathcal{E}_{n\times n}\}.$$

The set F_C is exposed. Let us recall that a set F is called an *exposed set* of a convex set K if $F = K \cap H$ for some supporting hyperplane H for K. Clearly, each exposed set is a face of K. For a general convex set K, the converse is not true. However, for the elliptope $\mathcal{E}_{n\times n}$, both notions coincide.

LEMMA 5.1 (see [LP93]). *Every face of $\mathcal{E}_{n\times n}$ is exposed.* □

If F_C contains a rank-one matrix, then (5.2) provides an exact solution of the max-cut problem. Hence we are interested in finding low-rank matrices in F_C, since they (intuitively) provide a tighter approximation of the max-cut problem.

QUESTION 5.2. *Given a face F of $\mathcal{E}_{n\times n}$, what is the minimum rank of a matrix $X \in F$?*

Since there exist extreme correlation matrices of any rank r up to the bound $r_{\max}$ given in Corollary 3.2, we cannot ensure, in general, the existence of matrices with rank smaller than $r_{\max} \approx \sqrt{2n}$. However, we are able to establish the existence of a low-rank matrix under some additional constraints.

LEMMA 5.3. *For every balanced vector $c \in \mathbb{R}^n$, there is a matrix $X \in \mathcal{E}_{n\times n}$ such that $c \in \ker(X)$ and rank $(X) \le 2$.*

Proof. Without loss of generality, we may assume that $c_1 \ge c_2 \ge \cdots c_n \ge 0$. Let i_0 be such that

$$\sum_{j<i_0} c_j \le \sum_{j\ge i_0} c_j \quad \text{and} \quad \sum_{j\le i_0} c_j \ge \sum_{j>i_0} c_j.$$

Set $\bar{c}_1 := \sum_{j<i_0} c_j$, $\bar{c}_2 = c_{i_0}$, and $\bar{c}_3 := \sum_{j>i_0} c_j$. Then it easily follows that $\bar{c} = (\bar{c}_1, \bar{c}_2, \bar{c}_3)$ is balanced, since $\bar{c}_1 + \bar{c}_2 \ge \bar{c}_3$, $\bar{c}_1 \le \bar{c}_2 + \bar{c}_3$ by the choice of i_0 and $\bar{c}_2 = c_{i_0} \le c_1 \le \bar{c}_1 + \bar{c}_3$ by the nonnegativity of c. Since $\bar{c} \in \mathbb{R}^3$ is balanced, there exists a matrix $\bar{X} \in \mathcal{E}_{3\times 3}$ with $\bar{c} \in \ker(\bar{X})$ (by Theorem 2.4). Set

$$\bar{X} = \begin{pmatrix} 1 & a & b \\ a & 1 & c \\ b & c & 1 \end{pmatrix} \quad \text{and} \quad X = \left(\begin{array}{c|c|c} & a & \\ J & \vdots & bJ \\ & a & \\ \hline a \ldots a & 1 & c \ldots c \\ \hline & c & \\ bJ & \vdots & J \\ & c & \end{array}\right),$$

where we have specified the i_0th row and i_0th column in X and J denotes the all-ones matrix (of appropriate sizes). Then we see that $\operatorname{rank}(X) = \operatorname{rank}(\bar{X}) \leq 2$ and $c \in \ker(X)$. □

THEOREM 5.4. *If a face F of $\mathcal{E}_{n\times n}$ contains a matrix of rank $n-1$, then it also contains a matrix of rank at most 2.*

Proof. The statement holds trivially if $F = \mathcal{E}_{n\times n}$. Suppose now that $F = F(A)$, where A has rank $n-1$. By Lemma 5.3, there exists $B \in \mathcal{E}_{n\times n}$ of rank ≤ 2 such that $\ker(A) \subseteq \ker(B)$, i.e., $B \in F$. □

Note that, under the assumption of Theorem 5.4, $\dim(F) = \binom{n-1}{2} - 1, \binom{n-1}{2}$, or $\binom{n}{2}$.

Example 5.5. The construction from the proof of Theorem 3.1 (which was already applied in Example 3.3) for the parameters $n = 9$, $r = 4$, $k = 1$ provides a matrix A of rank 4 whose face is an edge. One can determine the extremities of this edge (as was done in Example 3.3) and check that their ranks are equal to 3. So this gives a face containing only matrices of ranks 3 and 4. □

The dual problem of (5.2) is also of interest. The dual problem reads

$$\begin{array}{ll} \frac{n}{4} & \min \quad \lambda_{\max}(L_C + \operatorname{diag}(u)) \\ & \text{s.t.} \quad u_1 + \cdots + u_n = 0. \end{array} \tag{5.4}$$

We recall that L_C denotes the *Laplacian matrix*; it is the $n \times n$ symmetric matrix with (i,i)th diagonal entry $\sum_{j=1,\ldots,n,\ j\neq i} c_{ij}$ and (i,j)th entry $-c_{ij}$ for $i \neq j$. (Note that L_C does not depend on the diagonal entries of C.) Let u denote the optimum vector for the program (5.4), set $\lambda := \lambda_{\max}(L_C + \operatorname{diag}(u))$, and let V_{eig} denote the eigenspace corresponding to this eigenvalue for the matrix $L_C + \operatorname{diag}(u)$. It has been shown that strong duality holds, i.e., that both programs (5.2) and (5.4) have the same optimum solutions. Since the maximum eigenvalue in the optimum is typically multiple (unless the corresponding eigenvector is a ± 1 vector, in which case (5.1) provides an exact solution of the max-cut problem), the following question was asked in [DP93b] and in a more general setting in [Ov88].

QUESTION 5.6. *What is the possible dimension of the space V_{eig}?*

The next result establishes a link between the eigenspace V_{eig} and the face F_C and implies a lower bound for the dimension of V_{eig}.

PROPOSITION 5.7. *We have*

$$F_C = \{X \in \mathcal{E}_{n\times n} \mid \ker(X) \supseteq (V_{eig})^{\perp}\}.$$

Proof. Set $M := \lambda I - L_C - \operatorname{diag}(u)$. By construction, M is a positive semidefinite matrix and its kernel is $\ker(M) = V_{eig}$. For $X \in \mathcal{E}_{n\times n}$, we have

$$\langle L_C, X\rangle = \sum_{i,j} L_C(i,j)x_{ij} = 2 \sum_{1\leq i<j\leq n} c_{ij}(1 - x_{ij}),$$

which implies

$$\begin{aligned} \langle M, X\rangle &= \langle \lambda I, X\rangle - \langle L_C, X\rangle - \langle \operatorname{diag}(u), X\rangle \\ &= \lambda n - 4\left(\sum_{1\leq i<j\leq n} \frac{c_{ij}}{2}(1 - x_{ij})\right) \geq 0. \end{aligned}$$

Therefore, we see that $\langle M, X\rangle = 0$ if and only if X is an optimum solution to the program (5.2), i.e., if $X \in F_C$. Suppose $M = \sum_{1\leq i\leq k} f_i f_i^T$, where $f_1, \ldots, f_k$ span the

space $(\ker(M))^\perp$. Then $\langle M, X\rangle = 0$ holds if and only if $Xf_i = 0$ for all $i = 1, \ldots, k$, i.e., if $(\ker(M))^\perp \subseteq \ker(X)$. This shows the result. □

COROLLARY 5.8. *For every matrix $X \in F_C$, $\operatorname{rank}(X) \leq \dim(V_{eig})$.* □

An alternative proof of Corollary 5.8 can be given as follows. Since $X \succeq 0$, we have $X = Z^T Z$ for a matrix Z of the same rank as X. It can be checked that the rows of Z are eigenvectors from the space V_{eig}. Hence $\operatorname{rank}(X) = \operatorname{rank}(Z) \leq \dim(V_{eig})$.

Example 5.9. Consider the cost matrix $C := J$. Then the Laplacian matrix is $L_C = nI - J$. Then $\min_{u^T e=0} \lambda_{\max}(L_C + \operatorname{diag}(u))$ is attained for $u = 0$ (by symmetry; see [DP93a]) and is equal to $\lambda_{\max}(L_C) = n$. The optimized eigenspace is $V_{eig} = \{x \in \mathbb{R}^n \mid \sum_{1 \leq i \leq n} x_i = 0\}$, with dimension $n-1$. Hence, by Proposition 5.7, the face F_C is $\{X \in \mathcal{E}_{n\times n} \mid Xe = 0\}$. Note that it coincides with the face considered in Example 3.5. In particular, $(V_{eig})^\perp = \ker(X)$ for every matrix X lying in the relative interior of F_C. □

By Corollary 5.8, $\operatorname{rank}(X) \leq \dim(V_{eig})$ for each matrix X lying in the relative interior of F_C. In the previous example, we have equality: $\operatorname{rank}(X) = \dim(V_{eig})$. However, as is shown in the following example, strict inequality may hold and, in fact, the gap can be made as large as possible.

Example 5.10. Consider the cost matrix C defined by $c_{1j} = 1$ for all $j = 2, \ldots, n$ and $c_{ij} = \frac{1}{n-1}$ for all $2 \leq i < j \leq n$. Then the Laplacian matrix has the form

$$L_C = \begin{pmatrix} n-1 & -1 & \cdots & \cdots & -1 \\ -1 & \frac{2n-3}{n-1} & & & \\ \vdots & & \ddots & -\frac{1}{n-1} & \\ \vdots & & -\frac{1}{n-1} & \ddots & \\ -1 & & & & \frac{2n-3}{n-1} \end{pmatrix}.$$

Then the optimizing vector u for $\min_{u^T e=0} \lambda_{\max}(L_C + \operatorname{diag}(u))$ satisfies $u_2 = \cdots = u_n$ (by symmetry; see [DP93b]). Using this fact, it is not difficult to check that the optimum vector u is $(-(n-1)a, a, \ldots, a)$ for $a = \frac{2(n-2)}{n}$. Then the optimum value is $\lambda_{\max}(L_C + \operatorname{diag}(u)) = \frac{4(n-1)}{n}$. Moreover, the optimized eigenspace is $V_{eig} = \{x \in \mathbb{R}^n \mid (n-1)x_1 + \sum_{2 \leq i \leq n} x_i = 0\}$, with dimension $n-1$. Hence, $(V_{eig})^\perp$ is spanned by the vector $v = (n-1, 1, \ldots, 1)$. Therefore, by Proposition 5.7, the face F_C is given by

$$F_C = \{X \in \mathcal{E}_{n\times n} \mid Xv = 0\}.$$

Since v is not strictly balanced, we know from Theorem 2.6 that there cannot exist a matrix in F_C whose kernel is spanned by v. In fact, one can check that the only matrix of $\mathcal{E}_{n\times n}$ satisfying $Xv = 0$ is the cut matrix

$$X_0 := \left(\begin{array}{c|c} 1 & -1 \ldots -1 \\ \hline -1 & \\ \vdots & J \\ -1 & \end{array}\right).$$

Hence, the rank of X_0 is 1 while the dimension of V_{eig} is $n-1$, which is the largest possible gap. □

From the characterization of the normal cone (of Theorem 2.9) can be derived the following alternative description of the face F_C:

$$\begin{aligned} A \in F_C &\iff -C \in \mathcal{N}(A) \\ &\iff \exists D \text{ diagonal matrix such that} \\ &\qquad C + D \succeq 0,\ \ker(C+D) \supseteq (\ker A)^\perp. \end{aligned}$$

Therefore,

$$F_C = \{X \in \mathcal{E}_{n\times n} \mid \ker X \supseteq (\ker(C+D))^{\perp} \quad \text{for some diagonal matrix } D \text{ s.t. } C+D \succeq 0\}.$$

An interesting question is whether it is possible, given a cost matrix C, to find an element of F_C (of smallest possible rank) not using some classical optimization algorithm, but using rather some algebraic techniques based, for instance, on the above description of F_C.

6. The elliptope $\mathcal{E}_{4\times4}$. In this section, we give a description of the faces of the set $\mathcal{E}_{4\times4}$ of 4×4 correlation matrices. This question was raised by W. Barrett (private communication, 1994). Note that $\mathcal{E}_{4\times4}$ is a convex set of dimension 6.

THEOREM 6.1. *Let F be a proper face of $\mathcal{E}_{4\times4}$. Then one of the following holds.*

(i) $\dim(F) = 0$, *i.e., F consists of a unique matrix (which is an extreme element of $\mathcal{E}_{4\times4}$).*

(ii) *F is an edge joining two cut matrices, so* $\dim(F) = 1$. *There are* $\binom{8}{2} = 28$ *such faces.*

(iii) $\dim(F) = 2$; *F is called an elliptic face.*

(iv) *F is isomorphic to $\mathcal{E}_{3\times3}$ (more precisely, F is a lifting of $\mathcal{E}_{3\times3}$), so* $\dim(F) = 3$. *There are* $2\binom{4}{3} = 12$ *such faces.*

Hence, we find again that the range of feasible dimensions for the faces of $\mathcal{E}_{4\times4}$ is $[0,3] \cup \{6\}$; recall (3.2). According to Corollary 4.5, the highest dimension of a polyhedral face of $\mathcal{E}_{4\times4}$ is 2; recall the construction of such a face from Example 4.6. The elliptope $\mathcal{E}_{4\times4}$ also has nonpolyhedral faces of dimension 2; see Examples 6.2 and 6.3.

We call a face of dimension 2 of $\mathcal{E}_{4\times4}$ an *elliptic face* because, as will be seen in the proof, it is described by a set of inequalities $f(x,y) \geq 0$, where f is a polynomial of degree less than or equal to 2 in the variables x, y.

Proof of Theorem 6.1. Let F be a face of $\mathcal{E}_{4\times4}$. Suppose first that F arises as a lifting of a face G of $\mathcal{E}_{3\times3}$. We use the description of the faces of $\mathcal{E}_{3\times3}$ given in Proposition 2.10 from [LP93]. Either $G = \mathcal{E}_{3\times3}$, in which case F is one of the faces from Theorem 6.1 (iv), or G is an edge between two cut matrices, in which case F is one of the faces from (ii). It may also be that G is reduced to a single element, in which case F is also reduced to a single element; then we are in situation (i). From now on we suppose that F is not a lifting of a face of $\mathcal{E}_{3\times3}$. Set $V := \bigcap_{X\in F} \ker(X)$. By Lemma 2.8, every vector of V has at least three nonzero components. We distinguish several cases depending on the dimension of V.

Case 1. $\dim(V) = 1$. Let $v \in V, v \neq 0$. We can suppose that $v = (1, a, b, c)$, where at least two of a, b, c are nonzero. We can suppose that $a, b \neq 0$. Let

$$X = \begin{pmatrix} 1 & x & y & z \\ x & 1 & x' & y' \\ y & x' & 1 & z' \\ z & y' & z' & 1 \end{pmatrix} \tag{6.1}$$

be a matrix in F. The condition $Xv = 0$ can be rewritten as the system

$$\begin{cases} ax + by + cz = -1, \\ x + bx' + cy' = -a, \\ y + ax' + cz' = -b, \\ z + ay' + bz' = -c \end{cases}$$

in the variables x, y, z, x', y', z'. Since $a, b \neq 0$, the variables x, y, z, x' can be uniquely expressed in terms of a, b, c, y', z', namely,

$$\text{(6.2)} \qquad \begin{cases} x = & \frac{1}{2a}(-1 - a^2 + b^2 + c^2 + 2bcz'), \\ y = & \frac{1}{2b}(-1 + a^2 - b^2 + c^2 + 2acy'), \\ z = & -ay' - bz' - c, \\ x' = & \frac{1}{2ab}(1 - a^2 - b^2 - c^2 - 2acy' - 2bcz'). \end{cases}$$

The condition $X \succeq 0$ can be expressed by asking that all 2×2 and 3×3 principal subdeterminants of X be nonnegative, i.e.,

$$\text{(6.3)} \qquad \begin{cases} -1 \leq x, y, z, x', y', z' \leq 1, \\ 1 - x^2 - y^2 - (x')^2 + 2xyx' \geq 0, \\ 1 - x^2 - z^2 - (y')^2 + 2xzy' \geq 0, \\ 1 - y^2 - z^2 - (z')^2 + 2yzz' \geq 0, \\ 1 - (x')^2 - (y')^2 - (z')^2 + 2x'y'z' \geq 0. \end{cases}$$

Hence, F is a face of dimension 2, which is determined by the systems (6.2) and (6.3). So the boundary of F is described by polynomial equations in the variables y', z' of degree less than or equal to 2. Therefore, F is an elliptic face as in Theorem 6.1 (iii).

Case 2. $\dim(V) = 2$. Let $X \in F$ that is not a cut matrix. Then $\ker(X) = V$ (else $\ker(X)$ has dimension 3, which implies that X is a cut matrix). This shows that, if F is not reduced to a single element, then its relative boundary consists only of cut matrices and, thus, F is an edge between two cut matrices. However, we have already ruled out this possibility (since we assume that F is not a lifting of a face of $\mathcal{E}_{3\times 3}$). Therefore, F is reduced to a single element, i.e., we are in the situation of Theorem 6.1 (i).

Case 3. $\dim(V) = 3$. Then F is reduced to one element which is a cut matrix. So we are in the situation of Theorem 6.1 (i). □

We recall Example 4.6, where we described a polyhedral elliptic face of $\mathcal{E}_{4\times 4}$, namely, the face $\{X \in \mathcal{E}_{4\times 4} \mid Xv = 0\}$, where $v = (1, 1, 1, -1)^T$. In Example 3.5 we also described the polyhedral face of $\mathcal{E}_{4\times 4}$ corresponding to the vector $v = (1, 1, 1, 1)^T$.

We now present two examples of nonpolyhedral elliptic faces of $\mathcal{E}_{4\times 4}$. They are of the form $F = \{X \in \mathcal{E}_{4\times 4} \mid Xv = 0\}$, where $v \in \mathbb{R}^4$ is a balanced vector.

Example 6.2. Take $v = (1, 1, 1, 0)^T$. Then F consists of the matrices

$$\begin{pmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} & x \\ -\frac{1}{2} & 1 & -\frac{1}{2} & y \\ -\frac{1}{2} & -\frac{1}{2} & 1 & -x-y \\ x & y & -x-y & 1 \end{pmatrix},$$

where x, y satisfy the condition $x^2 + xy + y^2 \leq \frac{3}{4}$. Hence, F really has the shape of an ellipse. □

Example 6.3. Let $v = (1, 1, 2, 1)^T$. Then F consists of the matrices (6.1) satisfying (6.2) and (6.3), where (6.2) reads

$$x = \frac{1}{2}(3 + 4z'),\ y = \frac{1}{4}(-3 + 2y'), z = \frac{1}{4}(-5 - 2y' - 4z'),\ x' = -1 - y' - 2z'. \qquad \square$$

Acknowledgments. We thank a referee for his careful reading of the paper.

REFERENCES

[BJT93] W. W. BARRETT, C. R. JOHNSON, AND P. TARAZAGA, *The real positive definite completion problem for a simple cycle*, Linear Algebra Appl., 192 (1993), pp. 3–31.

[BJL] W. W. BARRETT, C. R. JOHNSON, AND R. LOEWY, *The real positive definite completion problem: Cycle completability*, to appear in Memoirs of the American Mathematical Society.

[CM79] J. P. R. CHRISTENSEN AND J. VESTERSTRØM, *A note on extreme positive definite matrices*, Math. Ann., 244 (1979), pp. 65–68.

[DP93a] C. DELORME AND S. POLJAK, *Laplacian eigenvalues and the maximum cut problem*, Math. Programming, 62 (1993), pp. 557–574.

[DP93b] ———, *Combinatorial properties and the complexity of an eigenvalue approximation of the max-cut problem*, European J. Combin., 14 (1993), pp. 313–333.

[DLP92] M. DEZA, M. LAURENT, AND S. POLJAK, *The cut cone* III: *On the role of triangle facets*, Graphs Combin., 8 (1992), pp. 125–142. Updated version in Graphs Combin., 9 (1993), pp. 135–152.

[DG81] H. DYM AND I. GOHBERG, *Extensions of band matrices with band inverses*, Linear Algebra Appl., 36 (1981), pp. 1–24.

[GW94] M. X. GOEMANS AND D. P. WILLIAMSON, *.878-approximation algorithms for MAX CUT and MAX 2SAT*, in Proceedings of the 26th Annual Symposium on Theory of Computing, Montréal, Canada, 1994, pp. 422–431.

[GJSW84] R. GRONE, C. R. JOHNSON, E. M. SÁ, AND H. WOLKOWICZ, *Positive definite completions of partial hermitian matrices*, Linear Algebra Appl., 58 (1984), pp. 109–124.

[GPW90] R. GRONE, S. PIERCE, AND W. WATKINS, *Extremal correlation matrices*, Linear Algebra Appl., 134 (1990), pp. 63–70.

[HW87] R. D. HILL AND S. R. WATERS, *On the cone of positive semidefinite matrices*, Linear Algebra Appl., 90 (1987), pp. 81–88.

[La94] M. LAURENT, *The real positive semidefinite completion problem for series-parallel graphs*, Report BS-R9439, Centrum voor Wiskunde en Informatica, Amsterdam, 1994, to appear in Linear Algebra Appl..

[LP92] M. LAURENT AND S. POLJAK, *One-third-integrality in the metric polytope*, Math. Programming, 71 (1995), pp. 29–50.

[LP93] ———, *On a positive semidefinite relaxation of the cut polytope*, Linear Algebra Appl., 223/224 (1995), pp. 439–461.

[LPR95] M. LAURENT, S. POLJAK, AND F. RENDL, *Connections between semidefinite relaxations of the max-cut and stable-set problems*, Report BS-R9502, Centrum voor Wiskunde en Informatica, Amsterdam, 1995, to appear in Math. Programming.

[LRT76] G.S. LEUKER, D. J. ROSE, AND R. E. TARJAN, *Algorithmic aspects of vertex eliminations on graphs*, SIAM J. Comput., 5 (1976), pp. 266–283.

[LT94] CHI-KWONG LI AND BIT-SHUN TAM, *A note on extreme correlation matrices*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 903–908.

[Loe80] R. LOEWY, *Extreme points of a convex subset of the cone of positive semidefinite matrices*, Math. Ann., 253 (1980), pp. 227–232.

[Ov88] M. L. OVERTON, *On minimizing the maximum eigenvalue of a symmetric matrix*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 256–268.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 3, pp. 548–562, July 1996

007

INTERLACING PROPERTIES OF TRIDIAGONAL SYMMETRIC MATRICES WITH APPLICATIONS TO PARALLEL COMPUTING*

ILAN BAR-ON†

Abstract. In this paper we present new interlacing properties for the eigenvalues of an unreduced tridiagonal symmetric matrix in terms of its leading and trailing submatrices. The results stated in Hill and Parlett [*SIAM J. Matrix Anal. Appl.*, 13 (1992), pp. 239–247] are hereby improved. We further extend our results to reduced symmetric tridiagonal matrices and to specially structured full symmetric matrices. We then present new fast and efficient parallel algorithms for computing a few eigenvalues of symmetric tridiagonal matrices of very large order.

Key words. symmetric, tridiagonal, eigenvalues, parallel algorithms

AMS subject classifications. 65F15, 65Y05

1. Introduction.

Tridiagonal matrices appear in a variety of algorithms for the diagonalization of general real symmetric matrices. Dense symmetric matrices are reduced to a tridiagonal form by the Householder transformation [7] and sparse symmetric matrices by the Lanczos process [4]. There are several efficient sequential algorithms for computing the eigenvalues of tridiagonal symmetric matrices such as bisection [10], LR [13], and the QR algorithm [7], [12], [15]. However, for matrices of large order, faster parallel methods are required. Cuppen's [5], [6], [11] divide and conquer method is a useful parallel method for computing the whole spectrum of the matrix, but usually we require only a few eigenvalues of the matrix. We will present in this paper new interlacing properties for the eigenvalues of a tridiagonal symmetric matrix and show how to use them to design fast and efficient parallel algorithms.

Recently, Hill and Parlett [8] have presented some refined interlacing properties for the eigenvalues of an unreduced tridiagonal symmetric matrix in terms of the eigenvalues of its leading submatrices. In this paper we generalize their results by providing interlacing properties for the eigenvalues of the matrix in terms of the eigenvalues of the leading and trailing submatrices, and we show that their results are a special case of ours. Moreover, our exposition is elementary and easy to follow. We further extend our results to general symmetric tridiagonal matrices, and to specially structured full symmetric matrices as well. Finally, we show how to apply these ideas to the design of efficient parallel algorithms for computing a few eigenvalues of tridiagonal symmetric matrices of large order.

This paper is organized as follows. In §2, we present some definitions and notation. In §§3 and 4, we present our main theorems for unreduced symmetric tridiagonal matrices. In §5, we extend them to general symmetric and other specially structured symmetric matrices. In §6, we apply them to the parallel computation of the eigenvalues. Finally, in the conclusion we mention some open related problems.

2. Basic definitions and notation.

We denote by $\mathcal{R}^n$ the set of real vectors of order n and the standard basis for this induced vector space by

$$e_i, \quad i = 1, \ldots, n,$$

* Received by the editors July 14, 1993; accepted for publication (in revised form) by F. T. Luk June 27, 1995.

† Department of Computer Science, Technion, Haifa 32 000, Israel (baron@cs.technion.ac.il). Current address: Math and Sciences, The University of Texas of the Permian Basin, Odessa, TX 79762 (baron_i@utpb.edu).

where e_i is all zeros except the ith coordinate, which is one. When needed we emphasize that a vector is in $\mathcal{R}^n$ by writing, for example, $e_i^{(n)}$. We denote by $\mathcal{M}(n)$ the set of real matrices of order n and by A^* the transpose of the matrix $A \in \mathcal{M}(n)$. Here, we adopt the convention used in the complex case since our results can be generalized to Hermitian matrices as well. However, for the sake of simplicity we will focus our attention on real symmetric matrices.

We denote a tridiagonal symmetric matrix $T \in \mathcal{M}(n)$ by

$$T = \begin{pmatrix} a_1 & b_1 & & & & \\ b_1 & a_2 & b_2 & & & \\ & b_2 & & \ddots & & \\ & & \ddots & & & \\ & & & & b_{n-1} & \\ & & & b_{n-1} & a_n & \end{pmatrix} = \begin{pmatrix} T_k & b_k & & \\ b_k & a_{k+1} & b_{k+1} & \\ & b_{k+1} & H_{k+2} & \end{pmatrix}, \tag{1}$$

where T_k denotes the leading submatrix of order k and H_{k+2} the trailing submatrix of order $n-(k+1)$. We say that T is unreduced if $b_i \neq 0$, $i = 1, \ldots, (n-1)$, in which case the eigenvalues of T, as well as those of its leading and trailing submatrices, are all simple, that is, of multiplicity one [12]. We then denote the characteristic polynomial of T_k by

$$p_k(x) = \det(xI - T_k), \quad k = 1, \ldots, n,$$

where we also write $p(x)$ for $k = n$. We denote the extended set of eigenvalues by

$$\theta^{(k)} = \{\theta_0^{(k)} = -\infty < \theta_1^{(k)} < \theta_2^{(k)} < \cdots < \theta_{k-1}^{(k)} < \theta_k^{(k)} < \theta_{k+1}^{(k)} = \infty\}, \tag{2}$$

where we will omit the subscript (k) when possible. We also denote the eigenvalues of T by $\lambda_i, i = 1, \ldots, n$. We now give a simple proof of the Cauchy interlacing theorem [12].

THEOREM 2.1. *The eigenvalues of T_{k-1} interlace the eigenvalues of T_k.*

Proof. For $k = 2$ the proof is simple and in general,

$$p_k(x) = (x - a_k)p_{k-1}(x) - b_{k-1}^2 p_{k-2}(x).$$

By induction, the eigenvalues of T_{k-2} interlace those of T_{k-1}. Hence,

$$sign(p_{k-2}(\theta_i^{(k-1)}+)) \neq sign(p_{k-2}(\theta_{i+1}^{(k-1)}-)),$$

where by $x\pm$ we mean a real number close enough to x from the right, left, respectively. Therefore, $p_k(x)$ changes sign inside each subinterval

$$(\theta_i^{(k-1)}, \theta_{i+1}^{(k-1)}), \quad 1 \leq i < k-1,$$

and being continuous it possesses a root there. A similar approach shows that $p_k(x)$ changes sign in the extreme subintervals as well. □

We proceed in the next section to give a refinement of the Cauchy interlacing theorem, and for that purpose we denote by

$$q_l(x) = \det(xI - H_l), \quad l = 1, \ldots, n,$$

the characteristic polynomial of H_l and note that

$$\begin{aligned} p(x) &= -b_{k-1}^2 p_{k-2}(x)q_{k+1}(x) + (x-a_k)p_{k-1}(x)q_{k+1}(x) - b_k^2 p_{k-1}(x)q_{k+2}(x) \\ &= q_{k+1}(x)[(x-a_k)p_{k-1}(x) - b_{k-1}^2 p_{k-2}(x)] - b_k^2 p_{k-1}(x)q_{k+2}(x) \\ &= q_{k+1}(x)p_k(x) - b_k^2 p_{k-1}(x)q_{k+2}(x) \end{aligned}$$

for $1 \le k \le n$, with $p_0(x) = q_{n+1}(x) \equiv 1$ and $q_{n+2}(x) \equiv 0$.

3. Interlacing properties for the eigenvalues of an unreduced symmetric tridiagonal matrix. We present in this section our first main result relating the spread of the eigenvalues of the matrix to the spread of the eigenvalues of its leading and trailing submatrices.

THEOREM 3.1. *Let $T \in \mathcal{M}(n)$ be an unreduced symmetric tridiagonal matrix as in* (1). *Let θ denote the extended set of eigenvalues of T_k as in* (2), *and let*

$$\beta = \{\beta_0 = -\infty < \beta_1 < \cdots < \beta_m < \infty = \beta_{m+1}\}, \quad m = n-(k+1),$$

denote the extended set of eigenvalues of H_{k+2}. We further let

$$\gamma = \{\gamma_0 = -\infty < \gamma_1 \le \cdots \le \gamma_{n-1} < \infty = \gamma_n\} \tag{3}$$

denote the union of θ and β, where by union we distinguish between the same eigenvalues from the two different sets. Then in each interval

$$(\gamma_{i-1}, \gamma_i), \qquad i = 1, \ldots, n,$$

there is exactly one different eigenvalue of T.

Remarks.

1. In case $(\gamma_{i-1} = \gamma_i)$ we assume that the interval contains that one point alone.
2. By a different eigenvalue we mean algebraically different, that is, in case the multiplicity of an eigenvalue λ is r, each is considered to be a different one.
3. The second remark is of no significance here because the eigenvalues differ by the usual convention. However, when we later consider the more general case of reduced symmetric tridiagonal matrices, an eigenvalue may have any multiplicity up to n. For the sake of consistency we prefer to clarify this definition right at this point.

We present some lemmas before we prove Theorem 3.1.

LEMMA 3.2. *Suppose there exists an index i such that*

$$\lambda = \gamma_{i-1} = \gamma_i, \quad 2 \le i \le (n-1).$$

Then λ is an eigenvalue of T. We note that in this case, γ_{i-1}, γ_i each belongs to a different set, or more formally, there exist indices r and s such that

$$\theta_r = \gamma_{i-1} = \gamma_i = \beta_s,$$

where $1 \le r \le k$ and $1 \le s \le m$.

Proof. Consider the characteristic polynomial of T:

$$p(x) = q_{k+1}(x)p_k(x) - b_k^2 p_{k-1}(x)q_{k+2}(x).$$

Then, for r, s as above,

$$p(\lambda) = q_{k+1}(\lambda)p_k(\theta_r) - b_k^2 p_{k-1}(\lambda)q_{k+2}(\beta_s) = 0.$$

Hence, λ is an eigenvalue of T. □

LEMMA 3.3. *There exists an index r such that*

$$\lambda = \theta_r, \quad 1 \le r \le k,$$

is an eigenvalue of T if and only if there exists an index s such that

$$\lambda = \beta_s, \quad 1 \le s \le m.$$

Proof. We prove one way; the other is proved similarly. Let $\lambda = \theta_r$ be an eigenvalue of T. Then

$$0 = p(\lambda) = q_{k+1}(\lambda)p_k(\theta_r) - b_k^2 p_{k-1}(\theta_r)q_{k+2}(\lambda) = -b_k^2 p_{k-1}(\theta_r)q_{k+2}(\lambda).$$

However, since T is unreduced $b_k \neq 0$ and since the eigenvalues of T_k strictly interlace those of T_{k-1} we have $p_{k-1}(\theta_r) \neq 0$. Hence, $q_{k+2}(\lambda) = 0$ and λ is an eigenvalue of H_{k+2}. □

COROLLARY 3.4. *For the proof of Theorem* 3.1 *it is sufficient to show that there is at most one eigenvalue in each nonredundant interval:*

$$(\gamma_{i-1}, \gamma_i), \quad \gamma_{i-1} < \gamma_i, \qquad 1 \le i \le n.$$

Proof. Consider the case of a redundant interval. Then

$$\gamma_{i-2} < \gamma_{i-1} = \gamma_i < \gamma_{i+1}$$

since the eigenvalues of T_k and H_{k+2} are all different. By Lemma 3.2 there is exactly one eigenvalue in each such redundant interval because the eigenvalues of T are all different. By Lemma 3.3 if the endpoint of a nonredundant interval is an eigenvalue of T it is accounted for in its respective redundant interval. Hence, if there is at most one eigenvalue in each of the remaining nonredundant intervals, there must be exactly one there, for the total number of eigenvalues is exactly n. □

We will now prove Theorem 3.1 in the following.

Proof. By the last corollary we need only prove that there is at most one eigenvalue in each nonredundant interval. For a given index $s, 1 \le s \le m+1$, we consider the sth interval of β, that is, (β_{s-1}, β_s). Then there is an index $i, s \le i \le (n - (m + 1 - s))$, and indices l and r such that

$$\theta_{r-1} \le \gamma_{i-1} = \beta_{s-1} < \gamma_i = \theta_r < \cdots < \theta_{r+l-1} = \gamma_{i+l-1} < \beta_s = \gamma_{i+l} \le \theta_{r+l},$$

where $1 \le r \le k$ and $0 \le l \le (k+1) - r$. We will then show that in each interval

$$(\gamma_{i+j-1}, \gamma_{i+j}), \quad j = 0, \ldots, l, \tag{4}$$

there exists at most one eigenvalue of T. Hence, since for $s = 1, \ldots, (m+1)$ these sets of intervals correspond to the complete set of nonredundant intervals of γ, we are done. For the proof we make use of the Sylvester theorem, which states that the inertia of T and F^*TF is the same provided F is nonsingular; see Horn and Johnson [9]. We denote the number of positive eigenvalues of T by $\pi(T)$. Given a real number x which is neither an eigenvalue of T_k nor of H_{k+2}, we construct the matrix F as follows:

$$F^* = \begin{pmatrix} I_k & 0 & \\ -b_k(e_k^{(k)})^*(T_k - xI)^{-1} & 1 & -b_{k+1}(e_1^{(m)})^*(H_{k+2} - xI)^{-1} \\ & 0 & I_m \end{pmatrix}.$$

Then

$$\hat{T} - xI = F^*(T - xI)F = \begin{pmatrix} T_k - xI_k & 0 & \\ & \hat{a}_{k+1}(x) - x & \\ & 0 & H_{k+2} - xI_m \end{pmatrix},$$

and therefore

$$\pi(T_k - xI) + \pi(H_{k+2} - xI) \leq \pi(T - xI) \leq \pi(T_k - xI) + \pi(H_{k+2} - xI) + 1.$$

However, the inertias of T_k and H_{k+2} do not change inside a nonredundant interval, as in (4), and therefore the number of eigenvalues of T there is bounded by

$$\pi(T - xI) - \pi(T - yI) \leq 1$$

for some $x \leq y$ and $x, y \in (\gamma_{i+j-1}, \gamma_{i+j})$. □

We proceed to reflect on some implications of Theorem 3.1.

(i) Theorem 3.1 is a generalization of Theorem 1 in Hill and Parlett [8], which states that in each interval

$$(\theta_{i-1}, \theta_i), \quad i = 1, \ldots, k+1, \tag{5}$$

there is at least one different eigenvalue of T. In fact their result is a simple consequence of ours. For let us assume that

$$\gamma_{j-2} < \theta_{i-1} = \gamma_{j-1} \leq \gamma_j \leq \cdots \leq \gamma_{j+l-1} \leq \gamma_{j+l} = \theta_i < \gamma_{j+l+1}, \tag{6}$$

where $i \leq j \leq n - (k+1-i)$ and $0 \leq l \leq n - (k+j+1-i)$. (Note that for $j = 1$ we ignore γ_{-1} and for $j + l = n$ we ignore γ_{n+1}.) Then there is at least one index r such that

$$\theta_{i-1} \leq \gamma_{j+r-1} < \gamma_{j+r} \leq \theta_i, \quad 0 \leq r \leq l,$$

because $\theta_{i-1} \neq \theta_i$. Hence, there is at least one eigenvalue of T there. The proof of this statement as given in [8] is long and tedious and involves the inspection of the interlacing properties of four sequences of eigenvalues. Our proof of the much more general result is more simple and straightforward.

(ii) Our bounds are sharper than those in [8]. The number of eigenvalues of T in the ith interval, as in (5), is exactly equal to the number of eigenvalues of H_{k+2} that are contained there, plus one. Moreover, if θ_{i-1} or θ_i or both are equal to an eigenvalue of H_{k+2}, then they are, respectively, also eigenvalues of T.

Proof. Consider the ith interval as in (6). Then we observe that

$$\gamma_{j+r-1} \neq \gamma_{j+r}, \quad r = 1, \ldots, l-1,$$

because they both belong to the set β. Hence,

$$\gamma_{j+r} \in (\theta_{i-1}, \theta_i), \quad r = 1, \ldots, l-2.$$

We conclude that so far, the interval contains $l-1$ eigenvalues of T and $l-2$ eigenvalues of H_{k+2}, which is according to our claim. The remaining two extreme intervals may now increase the size of both of these sets, by the same amount, and the proof now follows. The last assertion is then obvious. □

(iii) COROLLARY 3.5. *In case*

$$\gamma_{i-1} < \gamma_i < \gamma_{i+1},$$

γ_i is not an eigenvalue of T, and there are exactly i eigenvalues of T before it and $n-i$ eigenvalues after it. Otherwise, in case

$$\gamma_{i-2} < \gamma_{i-1} = \gamma_i < \gamma_{i+1},$$

γ_i is an eigenvalue of T, and there are exactly $(i-1)$ eigenvalues before it and $(n-i)$ eigenvalues after it.

(iv) Theorem 3.1 also implies Theorem 2 of Hill and Parlett, which states the following for the special case of $n = k+2$. Let $H_n = a_n = a$, and let a belong to the ith interval of θ, i.e.,

$$a \in [\theta_{i-1}, \theta_i], \quad 1 \le i \le k+1.$$

Then there is exactly one eigenvalue of T in each of the intervals

$$(\theta_{j-1}, \theta_j), \quad 1 \le j \le k+1, \ j \ne i,$$

and in each of the subintervals

$$(\theta_{i-1}, a), \quad (a, \theta_i). \tag{7}$$

This first part of their theorem is the conclusion of ours. Next, consider the extended set of eigenvalues of T_{k+1}, which we denote by χ, i.e.,

$$\chi = \{\chi_0 = -\infty < \chi_1 < \cdots < \chi_{k+1} < \infty = \chi_{k+2}\}.$$

Note that from Cauchy's interlacing theorem the eigenvalues of χ strictly interlace those of θ. Let $a = \chi_i$, i.e.,

$$\chi_{i-1} < \theta_{i-1} < \chi_i = a < \theta_i < \chi_{i+1}.$$

Then, beside the two eigenvalues of T that lie in (7), i.e., in

$$(\theta_{i-1}, \chi_i), \quad (\chi_i, \theta_i),$$

there is exactly one eigenvalue in each subinterval

$$(\theta_{j-1}, \chi_j), \quad j = 1, \ldots, (i-1), \tag{8}$$

and in each subinterval

$$(\chi_j, \theta_j), \quad j = (i+1), \ldots, (k+1). \tag{9}$$

This is the essence of the second part of their theorem, and is again a simple consequence of ours. By Cauchy's interlacing theorem, there is exactly one eigenvalue of T in each of the intervals

$$(\chi_{j-1}, \chi_j), \quad 1 \le j \le k+2,$$

and in particular exactly one in the ith and $(i+1)$th intervals. Hence, there can be no eigenvalue of T in the subintervals

$$(\chi_{i-1}, \theta_{i-1}), \quad (\theta_i, \chi_{i+1}),$$

and the conclusion in (8) and (9) now follows. In case $\chi_i \neq a$, the theorem further bounds the eigenvalues of T in the ith interval of θ above as follows. For $\chi_i > a$ there is exactly one eigenvalue of T in the subintervals

$$(\theta_{i-1}, a), \quad (\chi_i, \theta_i),$$

and for $a > \chi_i$ there is exactly one eigenvalue in the subintervals

$$(\theta_{i-1}, \chi_i), \quad (a, \theta_i).$$

These conclusions are again an obvious result of our theorem. However, we will give a much more general result in the next section for which this theorem of Hill and Parlett is a special case.

(v) Theorem 3.1 slightly modified also holds for general tridiagonal matrices. However, for the sake of simplicity we have decided to consider first this more easily verifiable case. The general case is dealt with in §5.

(vi) There are some generalizations of Theorem 3.1 to specially structured full symmetric matrices. These will also be considered in §5.

(vii) There are some consequences of our theoretical results to the parallel computation of the eigenvalues of very large size matrices. For example, by choosing $k = \lfloor (n-1)/2 \rfloor$, we can compute the eigenvalues of T_k and H_{k+2} in parallel and then use their interlacing properties to get some sharper bounds for the exact eigenvalues of T. For example, consider the famous tridiagonal matrix $T = tridiag(-1, 2, -1)$, whose eigenvalues are given analytically by

$$\lambda_i = 4 \sin^2 \frac{i\pi}{2(n+1)}, \quad i = 1, \ldots, n.$$

Let $n = 1024$, and assume for our demonstration that we are looking for the 307th and 308th eigenvalues which are shown below:

$$\lambda = (\ \lambda_{307} = 0.82195126525, \quad \lambda_{308} = 0.82691048642 \ ).$$

For $k = 511$ and $m = 512$, the related eigenvalues of θ and β are

$$\theta = (\ \theta_{153} = 0.81848059628, \quad \theta_{154} = 0.82840428509 \ )$$

and

$$\beta = (\ \beta_{153} = 0.81552949467, \quad \beta_{154} = 0.82542059370 \ ).$$

We conclude that

$$\lambda_{307} \in (\ \theta_{153} = 0.81848059628, \quad \beta_{154} = 0.82542059370 \ )$$

and that

$$\lambda_{308} \in (\ \beta_{154} = 0.82542059370, \quad \theta_{154} = 0.82840428509 \ ).$$

In fact, we have been able to isolate each of these eigenvalues in a subinterval containing that eigenvalue alone. Thereafter, we may use fast iterative methods such as the QR algorithm to locate that eigenvalue more accurately. We elaborate more on these applications in §6.

4. Refined interlacing properties for the eigenvalues of an unreduced symmetric tridiagonal matrix. We present in this section more refined results relating the spread of the eigenvalues of a matrix to the eigenvalues of its leading and trailing submatrices.

THEOREM 4.1. *Let there be given $r \geq 1$ sequences*

$$\Gamma^{(k)} = \{\gamma_0^{(k)} = -\infty < \gamma_1^{(k)} \leq \cdots \leq \gamma_{n-1}^{(k)} < \infty = \gamma_n^{(k)}\}, \quad k = 1, \ldots, r,$$

that fulfill the conclusions of Theorem 3.1, *that is,*

$$\lambda_i \in (\gamma_{i-1}^{(k)}, \gamma_i^{(k)}), \quad k = 1, \ldots, r, \quad i = 1, \ldots, n.$$

Let us denote their union by $\mathcal{E}$:

$$\mathcal{E} = \{\epsilon_0 = -\infty < \epsilon_1 \leq \cdots \leq \epsilon_{r(n-1)} < \infty = \epsilon_{r(n-1)+1}\}. \tag{10}$$

Then in each interval

$$(\epsilon_{ri}, \epsilon_{ri+1}), \qquad i = 0, \ldots, (n-1),$$

there is exactly one different eigenvalue of T.

We will prove this result using the following lemma.

LEMMA 4.2. *For $1 \leq i \leq j \leq n$, we have*

$$\gamma_{i-1}^{(k_1)} \leq \gamma_j^{(k_2)}, \qquad 1 \leq k_1, k_2 \leq r.$$

Proof. Using Theorem 3.1 we obtain

$$\lambda_i \in (\gamma_{i-1}^{(k_1)}, \gamma_i^{(k_1)}), \quad \lambda_j \in (\gamma_{j-1}^{(k_2)}, \gamma_j^{(k_2)}).$$

Suppose $\gamma_j^{(k_2)} < \gamma_{i-1}^{(k_1)}$. Then $\lambda_i > \lambda_j$, which is a contradiction since $i \leq j$. □

Proof. Using Theorem 3.1 we obtain

$$\lambda_i \in \bigcap_{k=1}^{r} (\gamma_{i-1}^{(k)}, \gamma_i^{(k)}) = (\max_{1 \leq k \leq r} \gamma_{i-1}^{(k)}, \quad \min_{1 \leq k \leq r} \gamma_i^{(k)}).$$

Since by Lemma 4.2

$$\gamma_{i-1}^{(k_1)} \leq \gamma_i^{(k_2)}, \quad 1 \leq k_1, k_2 \leq r, \quad 1 \leq i \leq (n-1),$$

we may assume without loss of generality that

$$\gamma_i^{(k)}, \ k = 1, \ldots, r \quad \text{for } i = 1, \ldots, (n-1)$$

appears consecutively in $\mathcal{E}$. The rest of the proof now follows. □

We proceed to reflect on some of the implications of Theorem 4.1.

(i) Theorem 4.1 is a generalization of Theorem 2 in Hill and Parlett [8], which is reviewed for convenience in the fourth implication after Theorem 3.1. Let $n = k+2$, as is there, choose $k' = n-1$, and apply Theorem 4.1 with $r = 2$. Then the eigenvalues of $T_{k'-1}$ are the eigenvalues in θ, the eigenvalues of $T_{k'}$ are those in χ, $H_{k'+1} = a_{k+2} = a$, and $H_{k'+2}$ is empty. The result now follows word by word from the conclusion of Theorem 4.1, and is much simpler than the previous proof based on Theorem 3.1, which is already a simplification of the proof given in [8].

(ii) Our result is much more general than the one given in [8] since it applies to any number of unrelated sequences.

(iii) Theorem 4.1 slightly modified also holds for general tridiagonal matrices. However, as before, for the sake of simplicity we have decided to consider first this more easily verifiable case. The more general case will be considered in §5.

(iv) Numerical examples. We display in Table 1 some experiments with random matrices of order $n = 128$, created and tested with MATLAB [14]. Here, we consider the sequences corresponding to $k = 64, 96, 112, 120$ and compute the widths of the respective enclosing intervals, taking one to four sequences.

TABLE 1
Interlacing bounds for a random matrix of order $n = 128$.

i	λ_i	$r = 1$	$r = 2$	$r = 3$	$r = 4$
22	−5.1882e−01	4.4719e−02	3.3648e−03	1.3152e−12	1.3152e−12
23	−4.7746e−01	5.7449e−03	1.7208e−15	5.5511e−17	5.5511e−17
24	−4.7172e−01	1.6467e−02	5.6621e−15	5.6621e−15	2.7756e−15
25	−4.5525e−01	1.4844e−02	1.8612e−06	1.8612e−06	1.3323e−15

(v) We note that in the transition from one to two sequences we obtain a tremendous gain, possibly more than from the classical sequential iterative methods. Adding more and more sequences, even in the case they are computed in parallel, may become questionable in terms of the overall computational cost. However, it is possible to compute only partial bounds for specific eigenvalues (see §6), and in that case it may become worthwhile.

5. Generalizations. We extend Theorems 3.1 and 4.1 of the previous sections for general tridiagonal symmetric matrices in §5.1 and give some generalizations to specially structured full symmetric matrices in §5.3.

5.1. Symmetric tridiagonal matrices. We extend Theorem 3.1 for general symmetric tridiagonal matrices in the following.

THEOREM 5.1. *Let $T \in \mathcal{M}(n)$ be a symmetric tridiagonal matrix, and let*

$$\theta = \{\theta_0 = -\infty < \theta_1 \le \cdots \le \theta_k < \infty = \theta_{k+1}\}, \quad 1 \le k \le (n-1), \tag{11}$$

denote the extended set of eigenvalues of T_k. Let

$$\beta = \{\beta_0 = -\infty < \beta_1 \le \cdots \le \beta_m < \infty = \beta_{m+1}\}, \quad m = n - (k+1), \tag{12}$$

denote the extended set of eigenvalues of H_{k+2}. Let γ denote their respective union as in (3). *Then in each interval*

$$[\gamma_{i-1}, \gamma_i], \quad i = 1, \ldots, n, \tag{13}$$

we can choose a different eigenvalue of T in a unique way.

Remarks.

1. We may have $\gamma_{i-1} < \gamma_i$, and yet γ_{i-1} or γ_i or both are eigenvalues of T. This is why we must use the closed parenthesis notation.

2. Let $\hat{\lambda}$ correspond to the sequence of eigenvalues thus chosen. Then we say that $\hat{\lambda}$ is a legal sequence.

3. By uniqueness we mean that if $\hat{\lambda}$ is a legal sequence, then

$$\hat{\lambda}_i = \lambda_i, \quad i = 1, \ldots, n.$$

Proof. We first prove existence and then uniqueness. We show that there is at least one different eigenvalue in each such interval.

(i) Case $b_k, b_{k+1} \neq 0$. For some index $1 \leq k_1 \leq k$ and index $(k+2) \leq k_2 \leq n$,

$$T_k = \begin{pmatrix} T_{1,(k_1-1)} & 0 \\ 0 & T_{k_1,k} \end{pmatrix}, \quad H_{k+2} = \begin{pmatrix} H_{k+2,k_2} & 0 \\ 0 & H_{k_2+1} \end{pmatrix},$$

where $T_{k_1,k}$ and H_{k+2,k_2} are unreduced. Let $\hat{T}$ be the unreduced matrix

$$\hat{T} = \begin{pmatrix} T_{k_1,k} & b_k & & \\ b_k & a_{k+1} & b_{k+1} & \\ & b_{k+1} & H_{k+2,k_2} & \end{pmatrix}.$$

Let $\hat{\theta} \subset \theta$ denote the eigenvalues of $T_{k_1,k}$, and let $\hat{\beta} \subset \beta$ denote the eigenvalues of H_{k+2,k_2}. Let $\hat{\gamma} \subset \gamma$ denote the union of these respective two sets. Applying Theorem 3.1 to $\hat{T}$, with $k' = k - (k_1 - 1)$ and $n' = k_2 - (k_1 - 1)$, we conclude that there is exactly one different eigenvalue of $\hat{T}$ in each interval

$$\hat{\lambda}_i \in (\hat{\gamma}_{i-1}, \hat{\gamma}_i), \quad i = 1, \ldots, k_2 - k_1 + 1. \tag{14}$$

However, the eigenvalues of $\hat{T}$ are also eigenvalues of T, and the remaining eigenvalues of T, namely, those of $T_{1,(k_1-1)}$ and of H_{k_2+1}, are simply the eigenvalues in the set $\gamma - \hat{\gamma}$. We next describe how we choose a different eigenvalue from each interval of (13) that is a subinterval of the same interval of (14). Since these last intervals are disjoint and cover the whole real line, the proof then follows. Given an index $i, 1 \leq i \leq k_2 - k_1 + 1$, there are indices l, r, s such that

$$\begin{aligned} \gamma_{l-2} < \hat{\gamma}_{i-1} &= \gamma_{l-1} \\ &\leq \gamma_l \leq \cdots \leq \gamma_{l+r-1} \leq \hat{\lambda}_i < \gamma_{l+r} \leq \cdots \leq \gamma_{l+s-1} \\ &< \gamma_{l+s} = \hat{\gamma}_i \leq \gamma_{l+s+1}, \end{aligned}$$

where $1 \leq i \leq l \leq n - (k_2 + 1 - k_1 - i)$ and $0 \leq r \leq s \leq n - (k_2 + l + 1 - k_1 - i)$. (Note that for $l = 1$ we omit γ_{l-2}, and for $n = (l+s)$ we omit γ_{l+s+1}.) We now choose

$$\begin{aligned} \gamma_{l+j} &\in (\gamma_{l+j-1}, \gamma_{l+j}], \quad j = 0, \ldots, (r-1), \\ \hat{\lambda}_i &\in [\gamma_{l+r-1}, \gamma_{l+r}), \\ \gamma_{l+j-1} &\in [\gamma_{l+j-1}, \gamma_{l+j}), \quad j = (r+1), \ldots, s, \end{aligned}$$

which is indeed a correct choice.

(ii) Case $b_k = 0, b_{k+1} \neq 0$. Here, $\hat{T} = H_{k+1,k_2}$ and $k' = 0, n' = k_2 - k$ in the notation above. The same proof then holds. The case $b_k \neq 0, b_{k+1} = 0$ is similar.

(iii) Case $b_k = b_{k+1} = 0$. The eigenvalues of T are those of γ together with a_{k+1}. The proof is now trivial.

This ends the proof for the existence of a legal sequence. The uniqueness follows from Corollary 5.2. □

COROLLARY 5.2. *For any legal sequence $\hat{\lambda}$, we must have*

$$\hat{\lambda}_i = \lambda_i, \quad i = 1, \ldots, n,$$

and therefore λ is a legal sequence.

Proof. Let $1 \le i \le n$ be the minimal index such that $\hat{\lambda}_i \neq \lambda_i$. By the minimality of i,

$$\hat{\lambda}_i = \lambda_j, \quad i < j \le n,$$

and therefore $\lambda_i < \hat{\lambda}_i \le \gamma_i$. However, in that case, there is no way to choose λ_i in the subsequent intervals, and this is a contradiction since $\hat{\lambda}$ is a legal sequence. □

COROLLARY 5.3. *Consider the case where, for some index* $1 \le l \le n$,

$$\gamma_{l-2} < \gamma_{l-1} = \cdots = \gamma_{l+r-1} < \gamma_{l+r}.$$

Then λ_l *is an eigenvalue of* T *of multiplicity at least* r *but of no more than* $r+2$.

Proof. We first observe that by Corollary 5.2,

$$\lambda_l = \lambda_j \in [\gamma_{j-1}, \gamma_j], \quad j = l, \ldots, l+r-1,$$

and therefore its multiplicity is at least r. The rest follows from the proof of Theorem 5.1. For example, consider the case where $b_k, b_{k+1} \neq 0$. Then, if there exists an index $i, 1 \le i \le k_2 - k_1 + 1$ such that

$$\hat{\gamma}_{i-1} = \gamma_{l-1} \quad \text{or} \quad \gamma_{l+r-1} = \hat{\gamma}_i,$$

then the multiplicity of λ_l is at most $r+1$, and otherwise, for some index i as above,

$$\hat{\gamma}_{i-1} < \gamma_{l-1} \quad \text{and} \quad \gamma_{l+r-1} < \hat{\gamma}_i,$$

and its multiplicity is at least $r+1$ and at most $r+2$. □

We proceed to generalize Theorem 5.1 in the following.

THEOREM 5.4. *Let there be given an* $r \ge 1$ *sequence as in* (3), *which satisfies the conclusions of Theorem* 5.1. *Let* ϵ *denote their union as in* (10). *Then*

$$\lambda_i \in [\epsilon_{r(i-1)}, \epsilon_{r(i-1)+1}], \quad i = 1, \ldots, n.$$

Proof. It is similar to the proof of Theorem 4.1, but based on Theorem 5.1. □

5.2. Specially structured full symmetric matrices. We will present in this subsection some generalizations of the results of the previous subsections to full symmetric matrices of a special structure.

COROLLARY 5.5. *Let* $A \in \mathcal{M}(n)$ *be a symmetric matrix as follows:*

$$A = \begin{pmatrix} A_k & v_k & \\ v_k^* & a_{k+1} & w_m^* \\ & w_m & B_{k+2} \end{pmatrix}, \qquad \begin{array}{c} A_k \in \mathcal{M}(k), \quad B_{k+2} \in \mathcal{M}(m), \\ v_k \in \mathcal{R}^k, \quad w_m \in \mathcal{R}^m, \\ m = n - (k+1). \end{array}$$

Let us denote the extended set of eigenvalues of A_k *by* θ *as in* (11) *and the extended set of eigenvalues of* B_{k+2} *by* β *as in* (12). *Let* γ *denote their respective union as in* (3). *Then in each interval*

$$[\gamma_{i-1}, \gamma_i], \quad i = 1, \ldots, n,$$

we can choose a different eigenvalue of A *in a unique way.*

Proof. We will show that A is similar to a tridiagonal matrix T, such that

$$\theta(A_k) = \theta(T_k), \quad \beta(B_{k+2}) = \beta(H_{k+2}).$$

The rest then follows from Theorem 5.1. Let $V_k \in \mathcal{M}(k)$ and $U_m \in \mathcal{M}(m)$ be orthogonal matrices such that

$$U_m^* w_m = b_{k+1} e_1^{(m)}, \quad V_k^* v_k = b_k e_k^{(k)}, \quad Q = \begin{pmatrix} V_k & & \\ & 1 & \\ & & U_m \end{pmatrix},$$

where Q is orthogonal. Then

$$\hat{A} = Q^* A Q = \begin{pmatrix} \hat{A}_k & b_k & \\ b_k & a_{k+1} & b_{k+1} \\ & b_{k+1} & \hat{B}_{k+2} \end{pmatrix}, \qquad \begin{aligned} \hat{A}_k &= V_k^* A_k V_k, \\ \hat{B}_{k+2} &= U_m^* B_{k+2} U_m, \end{aligned}$$

and $\hat{A}_k, A_k$ as well as $\hat{B}_{k+2}, B_{k+2}$ are similar. Finally, by a bottom-up tridiagonalization of $\hat{A}_k$ and by a top-down tridiagonalization of $\hat{B}_{k+2}$, we obtain the tridiagonal matrix T similar to $\hat{A}$. □

COROLLARY 5.6. *Let $A \in \mathcal{M}(n)$ be a symmetric matrix as follows:*

$$A = \begin{pmatrix} A_k & C_{mk} \\ C_{mk}^* & B_{k+1} \end{pmatrix}, \qquad \begin{aligned} &B_{k+1} \in \mathcal{M}(m+1), \\ &m = (n - (k+1)), \end{aligned} \tag{15}$$

where C_{mk} is a rank-one matrix. Let θ be as before. Then in each interval

$$[\theta_{i-1}, \theta_i], \quad i = 1, \ldots, (k+1), \tag{16}$$

there is at least one different eigenvalue of A. Similarly, let α denote the extended set of eigenvalues of B_{k+1}. Then also in each interval

$$[\alpha_{i-1}, \alpha_i], \quad i = 1, \ldots, (m+2), \tag{17}$$

there is at least one different eigenvalue of A.

Proof. Since C_{mk} is a rank-one matrix, there exist orthogonal matrices $V_k \in \mathcal{M}(k)$ and $U_{m+1} \in \mathcal{M}(m+1)$, such that

$$U_{m+1}^* C_{mk}^* V_k = \begin{pmatrix} 0 & b_k \\ 0 & 0 \end{pmatrix}, \quad Q = \begin{pmatrix} V_k & 0 \\ 0 & U_{m+1} \end{pmatrix},$$

where Q is orthogonal. Hence,

$$\hat{A} = Q^* A Q = \begin{pmatrix} \hat{A}_k & b_k \\ b_k & \hat{B}_{k+1} \end{pmatrix}, \qquad \begin{aligned} \hat{A}_k &= V_k^* A_k V_k, \\ \hat{B}_{k+1} &= U_{m+1}^* B_{k+1} U_{m+1}, \end{aligned}$$

and $\hat{A}_k, A_k$ as well as $\hat{B}_{k+1}, B_{k+1}$ are similar. Finally $\hat{A}$ is similar to a tridiagonal matrix T as before, and the proof now follows from Theorem 5.1. □

We note in passing that this is the same as Theorem 4 in Hill and Parlett [8]. However, we can say much more in the following.

COROLLARY 5.7. *Let $A \in \mathcal{M}(n)$ be a symmetric matrix as in* (15)*, and let η denote the sequence of mixed eigenvalues of θ and α. Then in each interval*

$$(\eta_{i-1}, \eta_i), \quad \eta_{i-1} < \eta_i, \quad i = 1, \ldots, (n+1), \tag{18}$$

there are either no eigenvalues, one simple eigenvalue, or two simple eigenvalues.

Proof. We consider the intervals in (18) that belong to a given interval of α. Let $1 \le s \le (m+2)$. Then there exist indices i, l, and r, such that

$$\theta_{r-1} < \eta_{i-1} = \alpha_{s-1} \le \eta_i = \theta_r \le \cdots \le \theta_{r+l-1} = \eta_{i+l-1} < \alpha_s = \eta_{i+l} \le \theta_{r+l}.$$

In case there is some index t such that

$$\beta_{s-1} = \eta_{i+t-1}, \quad 0 \le t \le l+1,$$

then by Theorem 5.1 there are either no eigenvalues or one simple eigenvalue in each subinterval

$$(\eta_{i+j-1}, \eta_{i+j}), \quad \eta_{i+j-1} < \eta_{i+j}, \quad 0 \le j \le l.$$

Otherwise, let there be an index t such that

$$\eta_{i+t-1} < \beta_{s-1} < \eta_{i+t}, \quad 0 \le t \le l.$$

Then only here can we have two simple eigenvalues of A. □

We could have generalized these results to obtain a bound on the number of eigenvalues of A in each interval of (16) and (17), in a way similar to Corollary 5.3.

6. Parallel applications. We describe in this section a simple parallel algorithm for computing a few eigenvalues of an unreduced tridiagonal matrix, based on the interlacing theory developed in the previous sections. We assume a parallel architecture with p independent processors, each having its own local memory, and a reasonably fast interconnection network that allows, for example, merging of small sets of numbers to be done efficiently. We then assume for simplicity that $T \in \mathcal{M}(N)$ with $N = np - 1$ and $p = 2^t$. We denote the matrix in block form by

$$T^{(s)} = \begin{pmatrix} T_1^{(s)} & b_1^{(s)} & & & & \\ b_1^{(s)} & a_1^{(s)} & c_1^{(s)} & & & \\ & c_1^{(s)} & \ddots & & & \\ & & & \ddots & b_{l-1}^{(s)} & \\ & & & b_{l-1}^{(s)} & a_{l-1}^{(s)} & c_{l-1}^{(s)} \\ & & & & c_{l-1}^{(s)} & T_l^{(s)} \end{pmatrix}, \quad \begin{array}{c} T_i^{(s)} \in \mathcal{M}(n_s - 1), \\ i = 1, \ldots, l, \\ l = p/2^s, \ n_s = 2^s n, \\ s = 0, \ldots, t, \end{array}$$

where in relation to (1) we have

$$b_i^{(s)} = b_{in_s - 1}, \quad a_i^{(s)} = a_{in_s}, \quad c_i^{(s)} = b_{in_s}.$$

We assume that initially the matrix is subdivided among the processors, each having in its own local memory the corresponding submatrix

$$\begin{pmatrix} b_{i-1}^{(0)} & a_{i-1}^{(0)} & c_{i-1}^{(0)} & & \\ & c_{i-1}^{(0)} & T_i^{(0)} & b_i^{(0)} & \\ & & b_i^{(0)} & a_i^{(0)} & c_i^{(0)} \end{pmatrix}.$$

We then look for an eigenvalue of T closer to some given real number x.

(i) Step $s = 0$. For each submatrix

$$T_i^{(0)}, \quad i = 1, \ldots, p,$$

we find an interval $x \in \mathcal{I}_i^{(0)}$ containing the closest pair of eigenvalues that surrounds x from both sides and that contains no other eigenvalue of $T_i^{(0)}$. Here, each processor computes a different interval in parallel. We note that it is not necessary to compute the pair of eigenvalues exactly, but only to isolate it appropriately. This can be done using a combination of QR with bisection.

(ii) Step $s = 1, \ldots, t$. For each submatrix

$$T_i^s, \quad i = 1, \ldots, p/2^s,$$

we find as above an interval $x \in \mathcal{I}_i^{(s)}$ containing the closest pair of eigenvalues that surrounds x from both sides and that contains no other eigenvalue of $T_i^{(s)}$. Here, however, we make use of the corresponding intervals computed in the previous step. We merge

$$\hat{\mathcal{I}}_i^{(s)} = \mathcal{I}_{2i-1}^{(s-1)} \cup \mathcal{I}_{2i}^{(s-1)}$$

and note that from Theorem 3.1 the new interval contains at least three eigenvalues of $T_i^{(s)}$, surrounding x from both sides. We then use bisection with possibly the QR algorithm to eliminate the unnecessary eigenvalues. We expect in practice to have a small number of eigenvalues in the newly created interval, so this procedure should be very fast. For parallel implementation, we let a corresponding group of 2^s processors compute in parallel their respective intervals using parallel bisection as in Bar-On [1] or parallel QR as in Bar-On and Codenotti [3].

(iii) Step $s = t + 1$. We now have a sharp bound for the eigenvalue near x, and we can use the parallel QR algorithm to locate it accurately.

We note that this algorithm can be extended in a natural way to compute some or all of the eigenvalues of T. More specifically, we may design a new divide and conquer parallel algorithm for computing a partial set, say in a given interval $[a, b]$, or the complete set of eigenvalues of T; see Bar-On [2]. As compared to Cuppen's method, it is useful in both cases.

7. Conclusion. We have presented new theoretical results relating the eigenvalues of a tridiagonal symmetric matrix to those of its leading and trailing submatrices. These theoretical results were also generalized to specially structured full symmetric matrices. We have then applied these results and obtained fast and efficient parallel algorithms for locating the eigenvalues of matrices of very large order. Further research is still required for the investigations of related results for specially structured sparse symmetric matrices.

REFERENCES

[1] I. Bar-On, *A Fast Parallel Bisection Algorithm for Symmetric Band Matrices*, Tech. report 726, Computer Science Department, Technion, Haifa, Israel, 1992.

[2] ———, *A New Divide and Conquer Parallel Algorithm for Computing the Eigenvalues of a Symmetric Tridiagonal Matrix*, Tech. report 832, Computer Science Department, Technion, Haifa, Israel, 1994.

[3] I. Bar-On and B. Codenotti, *A fast and stable parallel QR algorithm for symmetric tridiagonal matrices*, Linear Algebra Appl., 220 (1995), pp. 63–96.

[4] J. CULLUM AND R. WILLOUGHBY, *Lanczos Algorithms for Large Symmetric Eigenvalues Computaions*, Birkhäuser Boston, Cambridge, MA, 1985.
[5] J. CUPPEN, *A divide and conquer method for the symmetric eigenproblem*, Numer. Math., 36 (1981), pp. 177–195.
[6] J. J. DONGARRA AND D. SORENSEN, *A fully parallel algorithm for the symmetric eigenvalue problem*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. s139–s154.
[7] G. H. GOLUB AND C. F. V. LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1989.
[8] R. HILL AND B. PARLETT, *Refined interlacing properties*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 239–247.
[9] R. HORN AND C. JOHNSON, *Matrix Analysis*, Cambridge University Press, London, UK, 1985.
[10] W. KAHAN, *Accurate Eigenvalues of a Symmetric Tridiagonal Matrix*, Tech. report CS41, Computer Science Department, Stanford University, Stanford, CA, July 1966.
[11] A. KRISHNAKUMAR AND M. MORF, *Eigenvalues of a symmetric tridiagonal matrix: A divide and conquer approach*, Numer. Math., 48 (1986), pp. 349–368.
[12] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[13] H. RUTISHAUSER, *Solution of eigenvalue problems with the LR transformation*, Nat Bur. Standards, AMS, 49 (1958), pp. 47–81.
[14] K. SIGMON, *Matlab Primer*, The MATH WORKS Inc., 1994.
[15] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, UK, 1965. Reprinted in Oxford Science Publications, 1988.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 3, pp. 563–569, July 1996

008

ON EIGENVALUES OF QUADRATIC MATRIX POLYNOMIALS AND THEIR PERTURBATIONS*

M. RADJABALIPOUR† AND A. SALEMI‡

Abstract. Following the terminology used by Gohberg, Lancaster, and Rodman, the main results of the paper are as follows. (i) Studying the values of the partial multiplicities of a matrix polynomial $A(\lambda) = \lambda^2 I + \lambda C + K$ with hermitian coefficients at real eigenvalues λ_0 and determining sharp bounds for the highest degree d of the factor $(\lambda - \lambda_0)^d$ in the bivariate polynomial $t(\lambda, \epsilon) = det(A(\lambda) + \lambda\epsilon C)$. (ii) Finding conditions on general matrices C and K implying that the leading exponent in the Puiseux expansion of the zero $\lambda(\epsilon)$ of $t(\lambda, \epsilon) = 0$ near λ_0 is $1/a$, where a is the algebraic multiplicity of λ_0.

Key words. eigenvalues, quadratic matrix polynomials, perturbation

AMS subject classifications. 15A18, 47A56

1. Introduction.

Let C and K be $n \times n$ complex matrices and

$$A(\lambda) = \lambda^2 I + \lambda C + K; \quad T(\lambda, \epsilon) = A(\lambda) + \epsilon\lambda C \tag{1}$$

for $(\lambda, \epsilon) \in \mathbb{C}^2$. A complex number λ_0 is called an *eigenvalue* of the matrix polynomial $A(\lambda)$ of algebraic multiplicity a if

$$detA(\lambda) = (\lambda - \lambda_0)^a f(\lambda) \text{ and } a \geq 1 \tag{2}$$

for some polynomial f with $f(\lambda_0) \neq 0$. For ϵ small enough, the matrix polynomial $T(\lambda, \epsilon)$ regarded as a polynomial in λ has exactly a eigenvalues near λ_0 counting, of course, the multiplicities. If

$$t(\lambda, \epsilon) := detT(\lambda, \epsilon) = (\lambda - \lambda_0)^d h(\lambda, \epsilon) \tag{3}$$

for some bivariate polynomial h satisfying $h(\lambda_0, \epsilon) \not\equiv 0$, then exactly d eigenvalues of $T(\lambda, \epsilon)$ are constantly equal to λ_0. The integer d is called the *triviality degree* of T at λ_0 and is clearly less than or equal to a. The other $a - d$ eigenvalues of $T(\lambda, \epsilon)$ near λ_0 have the form

$$\lambda(\epsilon) = \lambda_0 + \gamma\epsilon^\beta + o(|\epsilon|^\beta), \ \gamma \neq 0, \ \beta > 0. \tag{4}$$

The number β, called the *leading exponent* of $\lambda(\epsilon)$, plays an important role in the study of the perturbation of quadratic matrix polynomials and is our main focus in the present paper. Finding partial multiplicities of $A(\lambda)$ at λ_0 and other multiplicities such as the triviality degree d are helpful in finding the values of β. Aside from this, the concepts of various multiplicities are of independent interest.

In §2, after a brief definition of partial multiplicities and related concepts, we prove that if C and K are hermitian, λ_0 is real, and $C + 2\lambda_0 I$ is semidefinite, then the partial multiplicities are either 1 or 2. The results of §2 are proved with a condition

* Received by the editors January 13, 1995; accepted for publication (in revised form) by P. Lancaster July 20, 1995. The authors acknowledge support from the International Centre for Theoretical Physics (ICTP) (Trieste, Italy) and the International Center for Science, High Technology, and Environmental Sciences (ICST) (Kerman, Iran).

† Department of Mathematics, University of Kerman, Kerman, Iran (radjab@irearn.bitnet).

‡ International Center for Science, High Technology, and Environmental Sciences, Kerman, Iran.

weaker than the semidefiniteness. (See Remark 2.5.) In §3 we find sharp bounds on d. In §4 we assume C and K are general matrices and study conditions implying $\beta = 1/a$. Both sections generalize results due to [5].

Throughout the paper we will fix the notation established in (1)–(4) and may give no further reference. A matrix with no entry is called a *vacuous matrix* and its determinant is defined to be 1. Also, we may have to use expressions such as $P_i(1 \leq i \leq 0)$, which means that no such expression exists.

2. Partial multiplicities. With $A(\lambda)$ as in (1), there exist matrix polynomials $D(\lambda), E(\lambda), F(\lambda)$ such that

$$A(\lambda) = E(\lambda)D(\lambda)F(\lambda), \tag{5}$$

$$D(\lambda) = diag(d_1(\lambda), d_2(\lambda), \ldots, d_n(\lambda)), \tag{6}$$

$$d_i \in \mathbb{C}[x], \quad d_{i+1} | d_i \quad (i = 1, 2, \ldots, n-1), \tag{7}$$

and $E(\lambda), F(\lambda)$ are products of elementary polynomial matrices. (An elementary matrix polynomial can be obtained from the identity matrix by one of the following alterations: (i) interchange of two rows, (ii) multiplication of one row by a nonzero constant, and (iii) replacement of the rth row by row r plus p times row s for any polynomial p and $r \neq s$. It is obvious that the inverse of an elementary matrix polynomial is again an elementary matrix polynomial.) There exist positive integers $g, m_1, m_2, \ldots, m_g$ such that

$$d_i(\lambda) = (\lambda - \lambda_0)^{m_i} f_i(\lambda), \ f_i(\lambda_0) \neq 0 \ (i = 1, 2, \ldots, g), \tag{8}$$

$$d_i(\lambda_0) \neq 0 \ \text{ if } \ g + 1 \leq i \leq n \tag{9}$$

for some polynomials f_i. The integer g is in fact the geometric multiplicity of 0 as an eigenvalue of the (numerical) matrix $A(\lambda_0)$. Also, the integers $m_1, m_2, \ldots, m_g$ are called the *partial multiplicities* of $A(\lambda)$ at $\lambda = \lambda_0$. Note that $m_1 \geq m_2 \geq \cdots \geq m_g \geq 1$. The matrix $D(\lambda)$ is called the Smith form of $A(\lambda)$ and is essentially determined by $m_1, \ldots, m_g$. (See, for example, [2, 4].)

Our first main result determines the values of $m_1, \ldots, m_g$ in case C and K are hermitian, $C + 2\lambda_0 I$ is semidefinite, and $\lambda_0 \in \mathbb{R}$. In [3] a criterion is given for the partial multiplicities to be all equal to 1. From now on, we will further fix the notation established in (5)–(9), which may be used with no reference.

THEOREM 2.1. *Let C and K be $n \times n$ hermitian matrices and let $A(\lambda)$ be as in (1). Let $\lambda_0 \in \mathbb{R}$ be an eigenvalue of $A(\lambda)$ and $C + 2\lambda_0 I$ be semidefinite. Then $m_i = 2$ if $1 \leq i \leq j$ and $m_i = 1$ if $j + 1 \leq i \leq g$, where*

$$j = dim(ker(C_{11} + 2\lambda_0 I)) \tag{10}$$

and C_{11} is the compression of C to $ker A(\lambda_0)$.

Proof. Since $A(\lambda_0)$ is hermitian, we can decompose $\mathbb{C}^n$ as $ker A(\lambda_0) \oplus ker A(\lambda_0)^{\perp}$ with respect to which

$$A(\lambda_0) = \begin{bmatrix} 0 & 0 \\ 0 & R \end{bmatrix}, \ C = \begin{bmatrix} C_{11} & C_{12} \\ C_{12}^* & C_{22} \end{bmatrix}, \tag{11}$$

where R and C_{22} are $(n-g)\times(n-g)$ hermitian matrices and $detR \neq 0$. Thus $kerA(\lambda_0)$ can be further decomposed as $kerA(\lambda_0) = ker(C_{11}+2\lambda_0) \oplus ker(C_{11}+2\lambda_0)^{\perp}$. Hence, up to a unitary similarity, $C_{11} = diag(-2\lambda_0, \ldots, -2\lambda_0, c_{j+1,j+1}, \ldots, c_{gg})$ and

$$B(\mu) = \begin{bmatrix} \mu^2 I & 0 & \mu C_{121} \\ 0 & (\mu^2 + 2\lambda_0\mu)I + \mu\hat{C}_{11} & \mu C_{122} \\ \mu C_{121}^* & \mu C_{122}^* & (\mu^2 + 2\lambda_0\mu)I + \mu C_{22} + R \end{bmatrix}, \tag{12}$$

where $\hat{C}_{11} = diag(c_{j+1,j+1}, \ldots, c_{gg})$, $\mu = \lambda - \lambda_0$, $B(\mu) = A(\mu + \lambda_0)$, and j is as in (10). Since $C + 2\lambda_0 I$ is semidefinite, $C_{121} = 0$.

Fix $t \in \{1, 2, \ldots, n\}$. Let the polynomial $\delta_t \in \mathbb{C}[\mu]$ be the greatest common divisor of all $t \times t$ minors of $B(\mu)$. If $1 \leq t \leq n-g$, then $B(\mu)$ has a $t \times t$ submatrix whose determinant is a polynomial $u_t(\mu)$ such that $u_t(0)$ is the determinant of some invertible $t \times t$ submatrix of R.

If $n-g+1 \leq t \leq n-j$, let $\hat{B}(\mu)$ be an arbitrary $t \times t$ submatrix of $B(\mu)$. Then at least $t-n+g$ rows of $\hat{B}(\mu)$ contain a factor μ and hence $det\hat{B}(\mu) = \mu^{t-n+g}\hat{k}(\mu)$ for some polynomial $\hat{k}$. In particular, if we choose $\hat{B}(\mu)$ by omitting the first $n-t$ rows and columns of $B(\mu)$, then $det\hat{B}(\mu) = \mu^{t-n+g}u_t(\mu)$, where

$$u_t(0) = (detR) \prod_{i=n-t+1}^{g} (2\lambda_0 + c_{ii}) \neq 0. \tag{13}$$

Hence $\delta_t(\mu) = \mu^{t-n+g}k_t(\mu)$ with $k_t(0) \neq 0$.

Finally, if $n-j+1 \leq t \leq n$, then, for every $t \times t$ submatrix $\hat{B}(\mu)$ of $B(\mu)$, $det\hat{B}(\mu) = \mu^{2t-2n+g+j}\hat{k}(\mu)$ for some polynomial $\hat{k}$. In particular, if $\hat{B}(\mu)$ is obtained by omitting the first $n-t$ rows and columns of $B(\mu)$, then $det\hat{B}(\mu) = \mu^{2t-2n+g+j}u_t(\mu)$, where $u_t(0) = u_{n-j}(0) \neq 0$ as in (13). Hence $\delta_t(\mu) = \mu^{2t-2n+g+j}k_t(\mu)$ with $k_t(0) \neq 0$.

Now, it follows from (5)–(9) and the Cauchy–Binet formula that

$$\delta_t(\mu) = \mu^{m_{n-t+1}+m_{n-t+2}+\cdots+m_n}G_t(\mu) \quad (1 \leq t \leq n),$$

where each G_t is a polynomial with $G_t(0) \neq 0$, and $m_{g+1} = \cdots = m_n = 0$. Thus,

$$m_{n-t+1} + m_{n-t+2} + \cdots + m_g = t - n + g, \quad n-g+1 \leq t \leq n-j,$$

$$m_{n-t+1} + m_{n-t+2} + \cdots + m_g = 2t - 2n + g + j, \quad n-j+1 \leq t \leq n.$$

Hence $m_g = m_{g-1} = \cdots = m_{j+1} = 1$ and $m_j = m_{j-1} = \cdots = m_1 = 2$. □

COROLLARY 2.2. *With the hypotheses of the theorem, $g + j = a$. In particular if $g = 1$, then $a = 1$ or 2.*

Proof. In view of (2) and (5)–(9), $a = m_1 + \cdots + m_g = 2j + g - j = g + j$. □

The following example, pointed out by P. Lancaster, shows that Theorem 2.1 is false if $C + 2\lambda_0 I$ is not semidefinite.

Example 2.3. Let $n = 2$ and $C = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$, $K = \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right]$.

For $\lambda_0 = 0$, $D(\lambda) = diag(\lambda^4, 1)$. Thus $g = 1$ and $m_1 = 4$. Also, by altering C and K, one may obtain $m_1 = 1, 2$, or 3.

THEOREM 2.4. *Let C and K be hermitian and $\lambda_0 = 0$ be an eigenvalue of $A(\lambda)$. Then $d \geq g+j$, where $j = dim\{KerC_{11}\}$ and C_{11} is the compression of C to $kerA(0)$. Moreover, if C is semidefinite, then $d = a$.*

Proof. Using (12) and noting that $\mu = \lambda$, it follows that

$$\begin{aligned} t(\lambda,\epsilon) &= detT(\lambda,\epsilon) = det(A(\lambda)+\epsilon\lambda C) \\ &= det(\lambda^2 I + \lambda C + \epsilon\lambda C + A(0)) \\ &= det\begin{bmatrix} \lambda^2 I & 0 & \lambda(1+\epsilon)C_{121} \\ 0 & \lambda^2 I + \lambda(1+\epsilon)\hat{C}_{11} & \lambda(1+\epsilon)C_{122} \\ \lambda(1+\epsilon)C_{121}^* & \lambda(1+\epsilon)C_{122}^* & \lambda^2 I + \lambda(1+\epsilon)C_{22} + R \end{bmatrix} \\ &= \lambda^{g+j}\hat{f}(\lambda,\epsilon), \end{aligned}$$

where $\hat{f}$ is a bivariate polynomial. Thus $d \geq g+j$. By Corollary 2.2, if C is semidefinite, then $d = a$. □

The proof of Theorem 2.1 suggests the following remark.

Remark 2.5. In the proof of Theorem 2.1, the semidefiniteness of $C+2\lambda_0 I$ is used only to establish $C_{121} = 0$. The latter condition is not needed in its full strength. All we must do is show that for each integer $i \in [1,j]$, there exists an i-dimensional subspace W_i of $ker(C_{11}+2\lambda_0 I)$, such that $dim W_i = i$ and

$$det\begin{bmatrix} I & L_i \\ L_i^* & R \end{bmatrix} \neq 0,$$

where L_i is the operator sending each $x \in kerA(\lambda_0)^{\perp}$ to the orthogonal projection of $C_{121}x$ onto W_i. (In the proof of Theorem 2.1, L_i was zero due to the semidefiniteness of $C+2\lambda_0 I$.)

3. Triviality degree. In this section we study the triviality degree of $T(\lambda,\epsilon)$ at $\lambda_0 \in \mathbb{R}\setminus\{0\}$, where C and K are hermitian. We find lower and upper bounds for d and construct examples to show that the bounds are sharp. Since our results will throw some light on the structure of the falling part of the Newton diagram of $t(\lambda,\epsilon)$, we first begin with a definition of this concept. Let

$$t(\lambda,\epsilon) = \sum_{i,j} t_{ij}(\lambda-\lambda_0)^i\epsilon^j \tag{14}$$

be the Taylor expansion of t defined in (3) and let $F = \{(i,j) : t_{ij} \neq 0\}$. Let H be the lower boundary of the convex hull of F in $\mathbb{R}^2$. Note that $(a,0) \in H \cap F$ and $(i,0) \notin H$ for all $i < a$. Let $N(t;\lambda_0) = F \cap H$. The set $N(t;\lambda_0)$, called the falling part of the Newton diagram of t at λ_0, plays an important role in determining the approximate values of the eigenvalues of $T(\lambda,\epsilon)$ near λ_0. (A curious reader can see [1, 7] for the definition of the Newton diagram itself; it is not needed in the present paper.) Reorder the set $N(t;\lambda_0)$ as $\{(x_0,y_0),(x_1,y_1),\ldots,(x_k,y_k)\}$ such that $a = x_0 > x_1 > \cdots > x_k = d$ and $0 = y_0 < y_1 < \cdots < y_k$. (Note that H must intersect the line $x = d$.) It is known that if $k \geq 1$ and $\beta = (y_i - y_{i-1})/(x_{i-1} - x_i)$ for some $i \in \{1,2,\ldots,k\}$, then $T(\lambda,\epsilon)$ has an eigenvalue $\lambda(\epsilon)$ near λ_0 of the form (4). Conversely, if $T(\lambda,\epsilon)$ has an eigenvalue of the form (4), then $\beta = (y_i - y_{i-1})/(x_{i-1} - x_i)$ for some $i \in \{1,2,\ldots,k\}$ [1, 7]. Note that $d = a$ if and only if $N(t;\lambda_0) = \{(a,0)\}$ if and only if all eigenvalues of $T(\lambda,\epsilon)$ near λ_0 are constantly equal to λ_0.

Now we state and prove the main result of this section.

THEOREM 3.1. *Assume C and K are hermitian and λ_0 is a nonzero real eigenvalue of $A(\lambda)$. Then $dim(kerA(\lambda_0) \cap kerC) \leq d \leq nullity C_{11}$, where C_{11} is the compression of C to $kerA(\lambda_0)$.*

Proof. Let

$$r = dim(kerA(\lambda_0) \cap kerC), \tag{15}$$

$$s = nullityC_{11}, \tag{16}$$

$$x = (\lambda - \lambda_0)^2 + 2\lambda_0(\lambda - \lambda_0) = \mu^2 + 2\lambda_0\mu, \tag{17}$$

$$y = (\lambda - \lambda_0)(1 + \epsilon) + \epsilon\lambda_0 = \mu(1 + \epsilon) + \epsilon\lambda_0. \tag{18}$$

Decompose $kerA(\lambda_0)$ as $W_1 \oplus W_2 \oplus W_3$ such that $W_1 = kerA(\lambda_0) \cap kerC$, W_2 is the orthogonal complement of W_1 in $kerC_{11}$, and W_3 is the orthogonal complement of $kerC_{11}$ in $kerA(\lambda_0)$. Then, with respect to this decomposition,

$$\begin{aligned} t(\lambda, \epsilon) &= detT(\lambda, \epsilon) = det(A(\lambda) + \epsilon\lambda C) \\ &= det \begin{bmatrix} xI & 0 & 0 & 0 \\ 0 & xI & 0 & yC_{121} \\ 0 & 0 & xI + yC_{111} & yC_{122} \\ 0 & yC_{121}^* & yC_{122}^* & xI + yC_{22} + R \end{bmatrix} \\ &= x^r \hat{f}(x, y), \end{aligned}$$

where C_{111} is an invertible hermitian matrix, C_{22} and R are as in (11), and $\hat{f}$ is a bivariate polynomial. Thus $(\lambda - \lambda_0)^r | t(\lambda, \epsilon)$ and hence $d \geq r$.

Let $t_{pq}\mu^p\epsilon^q$ be a nonzero term in the MacLaurin expansion of $t(\mu + \lambda_0, \epsilon)$ such that $m = p + q$ is minimal. Let $W_3 = W_{31} \oplus W_{32}$, where $W_{32} \perp W_{31} = ker(C_{111} + 2\lambda_0 I)$. Then, with respect to $\mathbb{C}^n = W_1 \oplus W_2 \oplus W_{31} \oplus W_{32} \oplus kerA(\lambda_0)^\perp$, we have

$$\sum_{p+q=m} t_{pq}\mu^p\epsilon^q = det \begin{bmatrix} 2\lambda_0\mu I & 0 & 0 & 0 & 0 \\ 0 & 2\lambda_0\mu I & 0 & 0 & 0 \\ 0 & 0 & -2\lambda_0^2\epsilon I & 0 & 0 \\ 0 & 0 & 0 & 2\mu\lambda_0 I + \mu\hat{C} + \epsilon\lambda_0\hat{C} & 0 \\ 0 & 0 & 0 & 0 & R \end{bmatrix},$$

where $\hat{C}$ is the part of C_{111} restricted to W_{32}. It follows that $m = g$,

$$t_{s,g-s} = (-1)^{g-s-u}2^{g-u}\lambda_0^{2g-s-u}(det\hat{C})(detR) \neq 0,$$

$$t_{s+u,g-s-u} = (-1)^{g-s-u}2^{g-u}\lambda_0^{2g-s-2u}det(2\lambda_0 I + \hat{C})(detR) \neq 0,$$

and $t_{i,g-i} = 0$ for $i < s$ or $i > s + u$, where $u = dimW_{32}$. Thus $d \leq s$. □

COROLLARY 3.2. *The convex hull of the set* $F = \{(i, j) : t_{ij} \neq 0\}$ *contains the points* $(s, g-s)$ *and* $(s+u, g-s-u)$ *as extremal points, where* $u = g-s-nullity(2\lambda_0 I + C_{11})$ *and* C_{11} *is as in* (11).

COROLLARY 3.3. *If* $s - r = n - g$, *then* $d = r$, *where* r *and* s *are as in* (15) *and* (16).

Proof. If $s - r = n - g$, then C_{121} is a square matrix. With this assumption we claim the coefficient $t_{r,n-r}$ in the MacLaurin expansion of $t(\mu + \lambda_0, \epsilon)$ is nonzero. It can be easily verified that the sum of the terms in which μ has the least power r is

$$\sum_{v \geq 0} t_{r,v}\mu^r\epsilon^v = (2\lambda_0)^r\mu^r det \begin{bmatrix} 0 & 0 & \lambda_0\epsilon C_{121} \\ 0 & \lambda_0\epsilon C_{111} & \lambda_0\epsilon C_{122} \\ \lambda_0\epsilon C_{121}^* & \lambda_0\epsilon C_{122}^* & \lambda_0\epsilon C_{22} + R \end{bmatrix}.$$

In particular, the coefficient $t_{r,n-r}$ of the term $\mu^r \epsilon^{n-r}$ is

$$t_{r,n-r} = 2^r \lambda_0^n (detC_{121}^*)(detC_{121})(detC_{111}).$$

If $C_{121}^* x = 0$, then $Cx = 0$ and $x \in kerA(\lambda_0) \cap kerC$. Thus $x = 0$. This implies that $detC_{121}^*$ and hence $detC_{121}$ are nonzero and thus $t_{r,n-r} \neq 0$. Therefore, $d \leq r$ and hence $d = r$. □

COROLLARY 3.4. *If C is nonnegative, then $d = r$.*

Proof. If C is nonnegative, so is C_{11} and hence $kerC_{11} \subset kerC$, where C_{11} is as in (11). Then $s = r = d$. □

The following examples show that the triviality degree d can be any number between r and s. (See (15)–(16).)

Example 3.5. Let $C = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right]$ and $K = \left[\begin{smallmatrix} -1 & -1 \\ -1 & -1 \end{smallmatrix}\right]$. Then $A(1) = \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right]$. Obviously, $detA(1) = 0$, $r = 0$, and $s = 1$, where r and s are as in (14)–(15). By Corollary 3.3 (or 3.4), $d = 0$. Thus $d = r = 0 < 1 = s$.

Example 3.6. Let

$$C = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad K = \begin{bmatrix} -1 & -1 & 0 \\ -1 & -1 & 1 \\ 0 & 1 & -1 \end{bmatrix}.$$

Then

$$A(1) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad \text{and} \quad T(1,\epsilon) = \begin{bmatrix} 0 & \epsilon & 0 \\ \epsilon & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Here $detA(1) = 0 = detT(1,\epsilon)$ for all $\epsilon \in \mathbb{C}$. Hence $\lambda_0 = 1$ is an eigenvalue of both $A(\lambda)$ and $T(\lambda,\epsilon)$ for all $\epsilon \in \mathbb{C}$. In particular $d \geq 1$. Since $s = 1$, $r = 0 < d = s = 1$.

Example 3.7. Assume $0 \leq p \leq q \leq u$ are given integers. Let C' (resp., K') be the $p \times p$ zero matrix. Let C'' (resp., K'') be the matrix C (resp., K) of Example 3.6. Let C''' (resp., K''') be the matrix C (resp., K) of Example 3.5. Now, let C (resp., K) be the direct sum of one copy of C' (resp., K'), $q - p$ copies of C'' (resp., K''), and $u - q$ copies of C''' (resp., K'''). Then $\lambda_0 = 1$ is an eigenvalue of the corresponding $A(\lambda)$ for which $r = p$, $d = q$, and $s = u$. (See (15)–(16).)

4. Leading exponents. Recall that the positive number β in (4) is called a leading exponent. In this section we assume C and K are general $n \times n$ matrices and study necessary or sufficient conditions to imply $\beta = 1/a$. Note that $\beta = 1/a$ if and only if $N(t;\lambda_0) = \{(a,0),(0,1)\}$. In this case λ_0 is necessarily nonzero. In fact $\lambda_0 = 0$ would imply $0 = detA(0) = detT(0,\epsilon)$ and hence $d \geq 1$. Thus $N(t;0) = \{(a,0),(x_1,y_1),\ldots,(d,p)\}$ for some $p \geq 1$. Therefore, either $d = a$, in which case all eigenvalues of $T(\lambda,\epsilon)$ are constant, or $a > d \geq 1$, in which case the smallest possible β is $1/(a-d-1) \geq 1/(a-1) > 1/a$. Also, we proved in [6, Thm. 3.1] that $N(t;\lambda_0)$ cannot have a point below the line $x + y = g$ and, consequently, if $\beta > 0$ is a leading exponent then $\beta \geq 1/(a-g+1)$. Therefore, if $\beta = 1/a$ then $g = 1$.

Our results in this section are stated in terms of the algebraic multiplicity b of 0 as an eigenvalue of $A(\lambda_0)$. Note that

$$det(zI - A(\lambda_0)) = z^b v(z), \tag{19}$$

where v is a polynomial with $v(0) \neq 0$. Hence $1 \leq g \leq b$. In particular, if $A(\lambda_0)$ is hermitian, then $g = b$. At this end, to justify the study of the eigenvalues of $A(\lambda)$

with non-hermitian coefficients, we remark that if

$$\frac{d^3y}{dt^3} + B\frac{d^2y}{dt^2} + D\frac{dy}{dt} + Ey = 0,$$

then letting $Y = \left[\begin{smallmatrix} y \\ dy/dt \end{smallmatrix}\right]$ we have the second-order differential equation

$$\frac{d^2Y}{dt^2} + \begin{bmatrix} 0 & -I \\ D & B \end{bmatrix}\frac{dY}{dt} + \begin{bmatrix} 0 & 0 \\ E & 0 \end{bmatrix} Y = 0,$$

in which the coefficients need not be hermitian.

Now, we state and prove the main result of this section.

THEOREM 4.1. *Let C and K be arbitrary $n \times n$ matrices and let $\lambda_0 \neq 0$ be an eigenvalue of $A(\lambda)$. Then, with the notation of* (1)–(4) *and* (19), *$\beta = 1/a < 1$ if and only if $b = 1 < a$.*

Proof. Assume $\beta = 1/a < 1$. Then $g = 1$ and $N(t;\lambda_0) = \{(0,1),(a,0)\}$, which implies that $t(\lambda_0,\epsilon) = \epsilon h(\epsilon)$ for some polynomial h with $h(0) \neq 0$. Assume, if possible, that $b > 1$. Using the Jordan canonical form, we can assume without loss of generality that $A(\lambda_0) \simeq J \oplus R$, where J is a $b \times b$ noninvertible Jordan block and R is invertible. Thus $t(\lambda_0,\epsilon) = det(A(\lambda_0) + \epsilon\lambda_0 C) = \epsilon h(\epsilon)$, where $h(0) = c_{b1}\lambda_0 detR$ and $C = (c_{ij})_{i,j=1}^n$. On the other hand, $detA(\lambda) = det[(\lambda^2 - \lambda_0^2)I + (\lambda - \lambda_0)C + A(\lambda_0)] = (\lambda - \lambda_0)\hat{f}(\lambda)$, where $\hat{f}$ is a polynomial such that $\hat{f}(\lambda_0) = c_{b1} detR$. Since $a > 1$, $\hat{f}(\lambda_0) = 0$ and hence $c_{b1} = 0$. Thus $h(0) = 0$, which is a contradiction. Therefore, $b = 1 < a$.

Conversely, assume $b = 1 < a$. In [6, Thm. 1.1], we showed that $N(t;\lambda_0)$ has at least one point on or below the line $x + y = b$. Now, since $(0,0)$ and $(1,0)$ are excluded, $(0,1) \in N(t;\lambda_0)$ and hence $\beta = 1/a$. □

The following examples show that the condition $a > 1$ cannot be waived on either side of Theorem 4.1.

Example 4.2. Let $C = \left[\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right]$ and $K = \left[\begin{smallmatrix} -1 & -1 \\ -1 & -1 \end{smallmatrix}\right]$. Then $\lambda_0 = 1$ is an eigenvalue of the corresponding $A(\lambda)$ and $detA(\lambda) = (\lambda-1)[(\lambda-1)(\lambda^2+1)-1]$. Thus $a = 1$. Also, $A(1) = \left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right]$, which implies that $g = 1$ and $b = 2$. Moreover, $t(1,\epsilon) = det(A(1)+\epsilon C) = -\epsilon$, which implies that $(0,1) \in N(t;1)$. Therefore, $\beta = 1 = 1/a$.

Example 4.3. Let $C = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$ and $K = \left[\begin{smallmatrix} -1 & -1 \\ -1 & 0 \end{smallmatrix}\right]$. Then $detA(\lambda) = (\lambda - 1)(\lambda^3+\lambda^2-\lambda+1)$. Thus $\lambda_0 = 1$ is an eigenvalue of $A(\lambda)$ with $a = 1$. Since $A(1) = \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right]$, $b = g = 1$. Moreover, $t(1,\epsilon) = -\epsilon^2$, which implies that $N(t;1) = \{(1,0),(0,2)\}$. Thus $\beta = 2 \neq 1/a$.

REFERENCES

[1] H. BAUMGÄRTEL, *Analytic Perturbation Theory for Matrices and Operators*, Akademie-Verlag, Berlin, 1984.

[2] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Matrix Polynomials*, Academic Press, New York, 1982.

[3] ———, *Quadratic matrix polynomials*, Adv. Appl. Math., 7 (1986), pp. 253–281.

[4] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices*, Academic Press, New York, 1985.

[5] H. LANGER, B. NAJMAN, AND K. VESELIĆ, *Perturbation of the eigenvalues of quadratic matrix polynomials*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 474–489.

[6] M. RADJABALIPOUR AND A. SALEMI, *On eigenvalues of perturbed quadratic matrix polynomials*, Integral Equations Operator Theory, 22 (1995), pp. 242–247.

[7] M. M. WAINBERG AND W. A. TRENOGIN, *Theorie der Losungsverzweigung bei nichtlinearen Gleichungen*, Academie-Verlag, Berlin, 1973.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 3, pp. 570–593, July 1996

009

COMPUTATIONAL TECHNIQUES FOR REAL LOGARITHMS OF MATRICES*

LUCA DIECI†, BENEDETTA MORINI†, AND ALESSANDRA PAPINI‡

Abstract. In this work, we consider computing the real logarithm of a real matrix. We pay attention to general conditioning issues, provide careful implementation for several techniques including scaling issues, and finally test and compare the techniques on a number of problems. All things considered, our recommendation for a general purpose method goes to the Schur decomposition approach with eigenvalue grouping, followed by square roots and diagonal Padé approximants of the diagonal blocks. Nonetheless, in some cases, a well-implemented series expansion technique outperformed the other methods. We have also analyzed and implemented a novel method to estimate the Frechét derivative of the log, which proved very successful for condition estimation.

Key words. real logarithm of a matrix, conditioning, Padé approximants, series expansions, eigendecomposition approaches, error analysis, implementations

AMS subject classifications. 65F30, 65F35, 65F99

Some notation. $M \in \mathbb{R}^{2n\times 2n}$ is called *Hamiltonian* if $M^TJ + JM = 0$, where

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}.$$

T is called *symplectic* if $T^TJT = J$; equivalently, $T^{-1} = -JT^TJ$. $\Lambda(T) = \{\lambda_i(T),\ i = 1,\dots,n\}$ will indicate the *spectrum* of T and $\rho(T)$ the spectral radius of T. The notation $\mu = \mathcal{O}(x)$ means that $\frac{\mu}{x} \to c \neq 0$ as $x \to 0$, and c is a constant. $A \otimes B = (a_{ij}B)_{i,j=1}^{n} \in \mathbb{R}^{n^2\times n^2}$ is the Kronecker product of A and B. We write $\|A\| = \|A\|_2$ for the 2-norm of a matrix A and $\|A\|_F$ for its Frobenius norm. Analogously, for a linear operator $L(A) : Z \in \mathbb{R}^{n\times n} \to L(A)Z \in \mathbb{R}^{n\times n}$, we write $\|L(A)\| = \max_{\|Z\|=1}\|L(A)Z\|$ for the operator norm induced by the 2-norm of matrices and $\|L(A)\|_f = \max_{\|Z\|_F=1}\|L(A)Z\|_F$ for that induced by the Frobenius norm.

1. Introduction.

In this work, we address the issue of finding a real logarithm of a real matrix. This problem has a precise and complete answer from the theoretical point of view, but from the computational point of view much work is still needed. A main motivation for carrying out the present work is to provide careful implementation for, and assess the performance of, the most promising techniques to compute real logarithms of matrices. We focus on real matrices, but much of what we say in this work can be adapted to the complex arithmetic case.

Undoubtedly, in comparison with other branches of scientific computation, linear algebra software is placed on very solid ground, the LAPACK and LINPACK/EISPACK libraries being the measure of excellence by which to assess quality software. The high-quality Matlab system also has in these computational linear algebra

*Received by the editors August 29, 1994; accepted for publication (in revised form) by N. J. Higham July 27, 1995. This work was supported in part by NSF grant DMS-9306412 and by MURST and CNR grants (Italy).

†School of Mathematics, Georgia Tech, Atlanta, GA 30332 (dieci@math.gatech.edu).

‡Department of Energetica, University of Florence, via C. Lombroso 6-17, 50134 Florence, Italy (ande@ingfi1.ing.unifi.it).

components its work-horse. However, there are some linear algebra problems that have not yet found their way into proper implementation and high-quality software. We believe that finding the logarithm of a matrix is one of these instances. In fact, more generally, computing functions of a matrix requires more work. (Interestingly, this is one of the very rare instances in which the Matlab implementation can give rather inaccurate answers.) The general lack of good software for functions of a matrix is all the more bothersome since computing functions of a matrix is a common engineering requirement (for the logarithm, see [LS1–2], [SS]). We think that a source of trouble is caused by looking at the computational task as a general task rather than addressing it in a case-by-case way, depending on the function at hand. Not surprisingly, the exp-function, which has been singled out for its importance for a long time, enjoys more personalized and robust implementations. We hope that our work will lead toward more robust implementations for the log function.

In the remainder of this section we briefly review some of the theoretical results we need. In §2 we address the sensitivity (or conditioning) issue for the log-function. The key ingredient is naturally the Frechét derivative of the log, and all throughout this work we try to characterize its norm. In §3 we give an algorithmic description of the methods we have chosen to implement and discuss some of the error's issues for them. In §4 we discuss finite precision aspects of the methods and also the general issue of ameliorating convergence and rescaling. We also present a new technique for estimating the condition number of the log problem, which has proven very reliable and somewhat efficient. In §5 we give details of appropriate implementations for the methods, including cost estimates. Finally, §6 contains examples and §7 conclusions.

Given a matrix $T \in \mathbb{R}^{n\times n}$, any $n \times n$ matrix X such that $e^X = T$, with e^X the matrix exponential of X, is a *logarithm* of T, and one writes $X = \log(T)$. As is well known (e.g., see [He] and [Wo]), every invertible matrix has a logarithm (not necessarily real). Among the logarithms of T, in this work we are only interested in those that are *primary matrix functions* of T [HJ], [G], [GvL], [Hi1]. As usual, these can be characterized from the Jordan decomposition of T (e.g., see [GvL, §1.11.1–2]).

Of course, to guarantee that $X = \log(T)$ is real (assuming T is), one needs a further restriction than mere invertibility. The most complete result is the following.

THEOREM 1.1. (see [C], [HJ]). *Let $T \in \mathbb{R}^{n\times n}$ be nonsingular. Then, there exists a real $X = \log(T)$ if and only if T has an even number of Jordan blocks of each size for every negative eigenvalue. If T has any eigenvalue on the negative real axis, then no real logarithm of T can be a primary matrix function of T.* □

We will henceforth assume that we have a real logarithm of T and that it is a primary matrix function of T. Finally, it has to be appreciated that a logarithm can be uniquely characterized once we specify which branch of the log function (acting on complex numbers) we take. For example, there is a unique $X = \log(T)$ such that all of its eigenvalues z satisfy $-\pi < \mathrm{Im}(z) < \pi$; this is known as the *principal logarithm*, and we will restrict ourselves to this case from now on.

In many applications, there is extra structure that one is interested in exploiting. For example, different techniques can be devised for the cases when $\Lambda(I - T)$ is inside the unit circle and/or when $\Re\mathrm{e}(\Lambda(T)) > 0$. Inter alia, the latter case arises for symmetric positive definite T, a situation in which T has a unique symmetric logarithm [HJ]. Also (see [Si] and [YS]), if T is symplectic (orthogonal), then there exists a real Hamiltonian (skew-symmetric) logarithm. Of course, in these cases we would want approximation techniques that guarantee we can recover the desired structure. This question was recently addressed in [D]; in the present work, we will use and extend some of the results in [D].

Not much work has been done on computing logarithms of matrices in comparison to its inverse function, computing matrix exponentials. The references [KL1], [KL2], [LS1], [LS2], [V] are a representative sample of works on the computation of logarithms of matrices. With the exception of [KL1–2], finite precision issues are not considered in these works. To the best of our knowledge, our work is the first attempt to consider finite precision behavior of several techniques and to implement and compare them.

2. Sensitivity of the problem. Naturally, before computing the logarithm of a matrix, it is appropriate to try to understand the intrinsic sensitivity of this function. The works of Kenney and Laub [KL2] and Mathias [M] are important sources of information on the general topic of conditioning of matrix functions. Our presentation is explicitly geared toward the log function and is partly different than these works.

Given a matrix function $F(T)$, where $F : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, the basic issue is to understand how the value of the function changes as the argument T does. This leads to a reliance on the Frechét derivative as a measure of sensitivity. From here on, unless otherwise stated, we use the 2-norm; with minimal changes (if at all), all results hold true for different norms.

DEFINITION 2.1. *Given a matrix function $G : T \in \mathbb{R}^{n\times n} \to G(T) \in \mathbb{R}^{n\times n}$, a linear mapping $G'(T) : Z \in \mathbb{R}^{n\times n} \to G'(T)Z \in \mathbb{R}^{n\times n}$ is the Frechét derivative of G at T if for any $Z \in \mathbb{R}^{n\times n}$ we have*

$$\lim_{\lambda\to 0} \left\| \frac{G(T+\lambda Z) - G(T)}{\lambda} - G'(T)Z \right\| = 0. \tag{2.1}$$

The norm of the Frechét derivative is given by $\|G'(T)\| = \max_{\|Z\|=1} \|G'(T)Z\|$. If G has a Frechét derivative, we say that G is differentiable. □

With this definition, one has a general way to assess sensitivity for matrix functions. This is a general procedure and can be found (essentially identical) in the works [KL2], [Hi1], [MvL], and references there, in special cases.

Let $X \neq 0 : \ G(T) = X$, and consider the perturbed input $T + \Delta T$ with corresponding perturbed output $X + \Delta X : \ X + \Delta X = G(T + \Delta T)$. For $G(T) = \log(T)$, from the relation $\Delta X = G(T + \Delta T) - G(T)$, upon using (2.1) we can obtain

$$\frac{\|\Delta X\|}{\|X\|} \leq \|G'(T)\| \frac{\|T\|}{\|X\|} \frac{\|\Delta T\|}{\|T\|} + \mathcal{O}(\|\Delta T\|^2)\,. \tag{2.2}$$

The quantity

$$\mathrm{cond}(G(T)) := \|G'(T)\| \frac{\|T\|}{\|X\|} \tag{2.3}$$

acts as a relative error magnification factor, and it is therefore natural to call it the *condition number* of the matrix function G at T. (Notice that, strictly speaking, we still have to justify the $\mathcal{O}(\|\Delta T\|^2)$ term in (2.2); we will do this in §3.)

Remarks 2.2. (i) It is clear that $\mathrm{cond}(G(T))$ depends both on G and T and on X. A measure of conditioning that neglects any of these components may be faulty.

(ii) Of course, different functions G might allow for more specialized ways to characterize $\mathrm{cond}(G(T))$, as is clearly evidenced in the work on the matrix exponential (see [vL], [MvL]). One of our tasks in the remainder of this work is to characterize better the Frechét derivative of the log function, hence $\mathrm{cond}(\log(T))$.

(iii) If $X \approx 0$, it is of course more sensible to assess absolute errors and, thus, to replace (2.2) with

$$\|\Delta X\| \leq \|G'(T)\| \, \|\Delta T\| + \mathcal{O}(\|\Delta T\|^2) \, .$$

We begin with the following elementary result, already in [KL2, Lem. B2], which is just the chain rule.

LEMMA 2.3. *Let F and G be matrix functions such that $G(T)$ is in the domain of F. Consider the composite function $H(T) := F(G(T))$. Let $G'(T)$ and $F'(G(T))$ be the Frechét derivatives of the functions G and F, at T and $G(T)$, respectively. Then the Frechét derivative of the composite function is characterized as the linear mapping*

$$H'(T) : \; Z \in \mathbb{R}^{n \times n} \to F'(G(T))G'(T)Z \in \mathbb{R}^{n \times n} . \qquad \square$$

As a consequence of Lemma 2.3, we have (essentially, [KL2, Lem. B1]) the following result.

COROLLARY 2.4. *Let F and G be inverse functions of each other, that is, $F(G(T)) = T \; \forall T$ in the domain of G, and $G(T)$ in the domain of F; and let F and G be differentiable, as in Lemma 2.3. Also, let $F'(G(T))$ be invertible. Then we have*

$$G'(T)Z = (F'(G(T)))^{-1}Z \, , \tag{2.4}$$

and therefore also

$$\|G'(T)\| = \|(F'(G(T)))^{-1}\| \, . \tag{2.5}$$

Proof. Apply the chain rule of Lemma 2.3 to the relation $F(G(T)) = T$. $\square$

LEMMA 2.5. *Let $G(T) = \log(T)$ and $F(Y) = e^Y$. Then we have*

$$\|G'(T)\| \geq \|T^{-1}\| \, , \tag{2.6}$$

and therefore

$$\operatorname{cond}(G(T)) \geq \frac{\operatorname{cond}(T)}{\|\log(T)\|} \, , \quad \textit{where} \quad \operatorname{cond}(T) = \|T\| \, \|T^{-1}\| \, .$$

Proof. From [vL, formula (1.3) and p. 972] we have

$$F'(Y)Z = \int_0^1 e^{Y(1-s)} Z e^{Ys} ds \, , \tag{2.7}$$

and therefore with $Y = \log(T)$ from Corollary 2.4 we have (take $Z = I$ below)

$$\begin{aligned} \|G'(T)\| &= \max_{\|Z\|=1} \left\| \left(\int_0^1 e^{Y(1-s)} Z e^{Ys} ds \right)^{-1} \right\| \\ &\geq \left\| \left(\int_0^1 e^{Y} ds \right)^{-1} \right\| = \|e^{-Y}\| = \|T^{-1}\| \, , \end{aligned}$$

where we have used the identity $-\log(T) = \log(T^{-1})$ (see [HJ]). $\square$

Remarks 2.6. (i) From (2.7) we can get implicit representations for $G'(T)$ in the case $G(T) = \log(T)$, and $F(Y) = e^Y$; for example,

$$Z = \int_0^1 T^{1-s} G'(T) Z T^s \, ds \, .$$

(ii) In §3, we prove that for positive definite matrices, in (2.6) we have equality.

In [KL2], Kenney and Laub consider matrix functions admitting a series representation such as

$$F(X) := \sum_{n=0}^{\infty} a_n X^n \, , \tag{2.8}$$

with associated scalar series absolutely convergent. In this case, they can represent the Frechét derivative as the infinite series

$$F'(X) : Z \to \sum_{n=1}^{\infty} a_n \sum_{k=0}^{n-1} X^k Z X^{n-k-1} \, . \tag{2.9}$$

Next, they unroll the Frechét derivative by column ordering, call $D(X) \in \mathbb{R}^{n^2 \times n^2}$ the resulting matrix acting on the unrolled Z

$$D(X) = \sum_{n=1}^{\infty} a_n \sum_{k=0}^{n-1} (X^T)^{n-1-k} \otimes X^k \, , \tag{2.10}$$

and then focus on the 2-norm of $D(X)$: $\|D(X)\|_2$. To proceed with their analysis, one must realize that $\|D(X)\|_2$ is the same as $\|F'(X)\|_f$ (see the notation at the beginning of this work). They have some general results giving a lower bound for this norm and then show that this lower bound is achieved when X is normal. We highly recommend a careful reading of their work for details. Notice, however, that the assumption on being able to represent $F(X)$ as the series (2.8) rules out a direct application of their theory to the log function. To deal with the Frechét derivative of the function $G(T) = \log(T)$, they rely on (2.5) and are thus able to estimate $\|G'(T)\|_f$ via estimates on the norm of the inverse of the Frechét derivative of the exponential function. Their approach can be profitably used to get some more information on the norm of the Frechét derivative of the log. Although what follows is not explicitly given in [KL2], it can be deduced from their approach.

Consider the case of $G(T)$ and $F(Y)$ inverse functions of each other so that $F(G(T)) = T$. Moreover, let $F(Y)$ be a matrix function for which (2.8)–(2.10) hold. For example, this is true for $G(T) = \log(T)$, and $F(Y) = e^Y$. Let $F'(G(T))$ be invertible, and let $D(G(T))$ be the unrolled Frechét derivative of $F(Y)$ at $G(T)$. Then we have

$$\|G'(T)\|_f = \|(F'(G(T)))^{-1}\|_f = \|(D(G(T)))^{-1}\|_2 \, .$$

Let λ_i be the eigenvalues of $D(G(T))$, and let $|\lambda_1| \geq \cdots \geq |\lambda_{n^2}|$. One always has the inequality

$$\frac{1}{|\lambda_1|} \leq \|(D(G(T)))^{-1}\|_2 \, ;$$

and if we assume that $D(G(T))$ be diagonalizable by the matrix S : $S^{-1}D(G(T))S = \operatorname{diag}(\lambda_i)$, then the inequality

$$\|(D(G(T)))^{-1}\|_2 \leq \operatorname{cond}_2(S)\frac{1}{|\lambda_1|}$$

is also well known. Now, let T be diagonalizable by V, $T = V\Lambda V^{-1}$, and so also $G(T) = VG(\Lambda)V^{-1}$. Hence for $D(G(T))$ one has (use [HJ, Prob. 3, p. 249])

$$D(G(T)) = (V^{-T} \otimes V)\left(\sum_{n=1}^{\infty} a_n \sum_{k=0}^{n-1} (G(\Lambda))^{n-1-k} \otimes G(\Lambda)^k\right)(V^{-T} \otimes V)^{-1}.$$

With $S = V^{-T} \otimes V$, putting it all together, we get that for diagonalizable matrices T the following holds:

$$\frac{1}{|\lambda_1|} \leq \|(D(G(T)))^{-1}\|_2 = \|G'(T)\|_f \leq \operatorname{cond}_2(S)\frac{1}{|\lambda_1|}. \tag{2.11}$$

To complete this discussion, we now recall that normal matrices can be brought to diagonal form (almost diagonal, i.e., diagonal with possibly 2×2 blocks along the diagonal to allow for complex conjugate pairs of eigenvalues if we insist on real arithmetic) with a unitary (orthogonal) matrix. So, let V be unitary above. Moreover, if T is normal, then so is $G(T)$ [HJ, Prob. 2, p. 439]. Finally, if V is unitary, so is V^{-T} and $S = V^{-T} \otimes V$ [HJ, p. 249]. So, for normal matrices, one has the precise characterization

$$\|G'(T)\|_f = \frac{1}{\min\limits_{1 \leq i \leq n^2} |\lambda_i(D(G(T)))|}. \tag{2.12}$$

In fact, to have $\operatorname{cond}_2(S) = 1$ in (2.11) we must have all singular values of S equal 1, and thus (2.12) holds, for the class of diagonalizable matrices, only if T is normal.

Remarks 2.7. (i) In particular, all of the above holds for the function $G(T) = \log(T)$. But the above reasoning also holds for many other matrix functions $G(T)$ not satisfying (2.8) but for which their inverse function satisfies (2.8); among others, $G(T) = T^{1/p}$, $p = 2, \ldots$.

(ii) Characterization of the eigenvalues of $D(G(T))$ in terms of those of $G(T)$ is done in [KL2, Lem. 2.1].

To obtain a relation between $\|G'(T)\|_f$ and the operator norm $\|G'(T)\|$, reason as follows. Let $G'(T)Z = B(Z) \in \mathbb{R}^{n \times n}$, and let $\sigma_i(Z)$, $\sigma_i(B(Z))$ be the (ordered) singular values of Z, $B(Z)$, respectively. Then

$$\|G'(T)\| = \max_{\sigma_1(Z)=1} \sigma_1(B(Z)),$$

$$\|G'(T)\|_f = \max_{\sigma_1^2(Z)+\cdots+\sigma_n^2(Z)=1} (\sigma_1^2(B(Z)) + \cdots + \sigma_n^2(B(Z)))^{1/2},$$

and the following inequalities are then simple to obtain:

$$\|G'(T)\| \leq \|G'(T))\|_f \leq \sqrt{n}\, \|G'(T)\|. \tag{2.13}$$

Notice that (2.13) are the usual inequalities between the Frobenius and spectral norms of matrices.

Of course, in order for a measure of conditioning of the $G(T)$ problem to be an effective computational tool, one should be able to estimate $\|G'(T)\|$ (perhaps in some norm other than the 2-norm) without drastically increasing the expense needed for the computation of $G(T)$. This seems to be a tall task. Nonetheless, some interesting ideas are in [KL2] and [M], and some other possibilities are discussed in the next two sections.

3. Some methods. More on conditioning. Here we present (i) two series expansion techniques [GvL], [LS1–2], [D], (ii) Padé approximation methods [KL1–2], [D], (iii) the Schur decomposition approach [GvL], [Matlab], and (iv) an ODE reformulation approach.

Series expansions. Under appropriate restrictions on the spectrum of T, the principal logarithm of T can be expressed as a series. In particular, two such series have frequently appeared in the literature. Computational procedures arise upon truncating these series.

Series 1. Let $A = I - T$, and assume $\rho(A) < 1$. Then

$$G(T) := \log(T) = \log(I - A) = -\sum_{k=1}^{\infty} \frac{A^k}{k}. \tag{3.1}$$

Subject to obvious restrictions on spectral radii, from (3.1) we get

$$\log(T + Y) = \log(T) + (\log(T))'Y + E(Y),$$

and $\|E(Y)\| \leq \mathcal{O}(\|Y\|^2)$. From this, for the Frechét derivative we obtain the expression:

$$G'(T) : Y \to \sum_{n=1}^{\infty} \frac{1}{n} \sum_{k=0}^{n-1} A^k Y A^{n-1-k}, \qquad A = I - T, \tag{3.2}$$

and if $\|A\| < 1$,

$$\|G'(T)\| \leq \sum_{n=1}^{\infty} \frac{1}{n} \sum_{k=0}^{n-1} \|A\|^{n-1} = \frac{1}{1 - \|A\|} = \frac{1}{1 - \|I - T\|}.$$

From the above, we get that for positive definite matrices $\|G'(T)\| \leq \frac{1}{\min_{\lambda \in \Lambda(T)} |\lambda|}$, that is, $\|G'(T)\| \leq \|T^{-1}\|$, which justifies Remark 2.6(ii) for positive definite matrices for which (3.1) holds.

Series 2. This is obtained from the series expansion (3.1) for $\log(I+X) - \log(I-X) = \log((I+X)(I-X)^{-1})$ via the conformal transformation $T = (X-I)(X+I)^{-1}$, thereby obtaining

$$\log(T) = 2 \sum_{k=0}^{\infty} \frac{1}{2k+1} [(T-I)(T+I)^{-1}]^{2k+1}. \tag{3.3}$$

Notice that the restriction $\rho(A) < 1$ needed for (3.1) is now $\Re\mathrm{e}(\Lambda(T)) > 0$. Reasoning as before, if also $\Re\mathrm{e}(\Lambda(T+Y)) > 0$, for the Frechét derivative of $\log(T)$ we obtain the expression:

(3.4)

$$(\log(T))' : Y \to 2 \sum_{k=0}^{\infty} \frac{1}{2k+1} \sum_{j=0}^{2k} B^j (2CYC) B^{2k-j}, \; C := (T+I)^{-1}, \; B := (T-I)C.$$

Padé approximants. Under the assumption $\rho(I-T)<1$, these consist of approximating the function $\log(I-A)$, $A=I-T$ with the rational matrix polynomial $R_{n,m}(A)=P_n(A)(Q_m(A))^{-1}$, where $P_n(A)$ and $Q_m(A)$ are polynomials in A of degree n and m, respectively, in such a way that $R_{n,m}(A)$ agrees with $n+m$ terms in the series expansion (3.1) of $\log(I-A)$. This is a universal and powerful tool [BG-M] that is well examined in the context of $\log(T)$ in the works [KL1–2]. It is easy, with the help of tools such as *Maple*, to obtain the coefficients of the matrix polynomials $P_n(A)$ and $Q_m(A)$. Based on the error estimates in [KL1], we have only considered diagonal Padé approximants.

To assess the conditioning of the Padé approximants, we can reason as follows. For given n,m, let $R(A)=R_{n,m}(A)=P(A)(Q(A))^{-1}=\sum_{k=0}^{n}a_kA^k\,(\sum_{k=0}^{m}b_kA^k)^{-1}$. Suppose that rather than T we have $T+Y$, that is $A-Y$, instead of A, and $\|Y\|\ll 1$. Then it is easy to obtain

$$R(A-Y)-R(A)=-(E(Y)-R(A)F(Y))(Q(A))^{-1}+H(Y)\,, \tag{3.5}$$

where $\|H(Y)\|\le\mathcal{O}(\|Y\|^2)$, and $E(Y)=\sum_{k=1}^{n}a_k\sum_{j=0}^{k-1}A^jYA^{k-1-j}$ and $F(Y)=\sum_{k=1}^{m}b_k\sum_{j=0}^{k-1}A^jYA^{k-1-j}$ are the first-order perturbation terms for $P(A)$ and $Q(A)$. From (3.5) we obtain

$$\|R(A-Y)-R(A)\|\le(\|E(Y)\|+\|F(Y)\|\,\|R(A)\|)\|(Q(A))^{-1}\|+\mathcal{O}(\|Y\|^2)$$

or, in a relative error sense (if $\|R(A)\|\neq 0$),

$$\frac{\|R(A-Y)-R(A)\|}{\|R(A)\|}\le\left(\frac{\|E(Y)\|}{\|R(A)\|}+\|F(Y)\|\right)\frac{\operatorname{cond}(Q(A))}{\|Q(A)\|}+\mathcal{O}(\|Y\|^2)\,. \tag{3.6}$$

Therefore, we see that for the conditioning of the Padé problem the most important factor is the conditioning of the denominator problem. In [KL1], this issue is investigated in the case $\|A\|<1$; in particular see [KL1, Lem. 3].

To understand better the term $(E(Y)-R(A)F(Y))(Q(A))^{-1}$ in (3.5), we can use first-order perturbation arguments for the matrix function $R(A)$ to obtain

$$(E(Y)-R(A)F(Y))(Q(A))^{-1}\;=\;R'(A)Y\,.$$

We also have the following general result.

LEMMA 3.1. *Let $F(A)=\sum_{k=0}^{\infty}c_kA^k$, and let $R(A)$ be a Padé approximant agreeing with the series of $F(A)$ up to the power A^{n+m} included. Then $R'(A)Y$ agrees with $F'(A)Y$ up to the term $\sum_{j=1}^{n+m}c_j\sum_{l=0}^{j-1}A^lYA^{j-1-l}$.*

Proof. Write $F(A)=R(A)+M(A)$, so that $M(A)$ has a power series with terms beginning with A^{k+m+1}. Now, since $F'(A)Y=R'(A)Y+M'(A)Y$, the result follows. □

Remark 3.2. For the case of the log, since $F(A-Y)=\log(T+Y)$, Lemma 3.1 tells us that the conditioning of the Padé problem (hence also of the truncated series (3.1)) is close to the conditioning of $\log(T)$ (essentially the same if $\|A\|<1$ for high enough $n+m$). No extra pathological behavior is introduced.

Schur decomposition approach. When properly implemented, this is an extremely effective and reliable technique. The basic principles of the technique are general (see [GvL]), but our adaptation to $\log(T)$ seems to be new. Let Q be orthogonal such

that $QTQ^T := R$ is in real Schur form (upper quasi-triangular). Moreover, let R be partitioned as

$$R := \begin{pmatrix} R_{11} & \cdots & R_{1m} \\ & \ddots & \vdots \\ 0 & & R_{mm} \end{pmatrix},$$

where we assume that $\Lambda(R_{ii}) \cap \Lambda(R_{jj}) = \emptyset$, $i \neq j$ (this can be done in standard ways). To obtain $L := \log(R)$, one realizes that L has the same block structure as R and (see [GvL, §11.1]) can get L from the relation $LR = RL$. The following recursion can be used to get L [P].

For $i = 1, 2, \ldots, m$

$$L_{ii} = \log(R_{ii}) \tag{3.7}$$

Endfor i
For $p = 1, 2, \ldots, m-1$
For $i = 1, 2, \ldots, m-p$, with $j = i + p$, solve for the L_{ij}:

$$L_{ij}R_{jj} - R_{ii}L_{ij} = R_{ij}L_{jj} - L_{ii}R_{ij} + \sum_{k=i+1}^{j-1} (R_{ik}L_{kj} - L_{ik}R_{kj}) \tag{3.8}$$

Endfor i
Endfor p.

In general, the R_{ii} can be the 1×1 or 2×2 blocks of eigenvalues or also much larger quasi-triangular blocks. If T is normal, then Q brings T to block diagonal form with either $(1,1)$ or $(2,2)$ diagonal blocks, and only (3.7) is required. Otherwise, solving the Sylvester equation (3.8) is standard (see [GvL, p. 387], and notice that (3.8) is uniquely solvable since $\Lambda(R_{ii}) \cap \Lambda(R_{jj}) = \emptyset$). To obtain L_{ii} from (3.7) is just a function call if R_{ii} is (1×1), and also if $R_{ii} \in \mathbb{R}^{2\times 2}$ with complex conjugate eigenvalues a direct evaluation is possible (see Lemma 3.3), while in all other cases we need some approximation method, e.g., by truncating the previous series or using Padé approximants (if applicable).

LEMMA 3.3. *Let* $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ *with complex conjugate eigenvalues* $\theta \pm i\mu$ $(\mu \neq 0)$. *Then*

$$\log(A) = \alpha I - \beta \frac{2\mu}{4bc + (a-d)^2} \begin{pmatrix} a-d & 2b \\ 2c & -a+d \end{pmatrix},$$

where $\alpha = \log(\rho)$, $\rho^2 = \theta^2 + \mu^2$, *and* $\beta = \cos^{-1}(\frac{\theta}{\rho})$, $0 \leq \beta < \pi$.

Proof. The proof is just a simple calculation. □

COROLLARY 3.4. *Let* $B \in \mathbb{R}^{2\times 2}$ *be normal with complex conjugate eigenvalues, that is,* $B = \left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right)$. *With the notation of Lemma* 3.3, *we have*

$$\log(B) = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}.$$

Moreover, if B *is orthogonal, then* $\alpha = 0$. □

Remarks 3.5. (i) Corollary 3.4, coupled with prior real Schur reduction, guarantees that the computation of a real logarithm of a normal matrix T can be done in such a way that the end result is a real, normal matrix. In particular, this fact

makes such an algorithm interesting for computing the skew-symmetric logarithm of an orthogonal matrix, an approach not considered in [D].

(ii) Of course, $\left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right)$ can be identified with the complex number $z = a + ib$, which makes Corollary 3.4 obvious ($\log z = \log|z| + i \arg z$). This observation also renders more transparent the first part of Lemma 3.8.

ODE approach. This will be a very useful tool to characterize better both $\log(T)$ and its Frechét derivative. The starting point is to embed the problem into a continuous model, similar in spirit to a "homotopy" path.

Let the time-dependent matrix $X(t)$ be implicitly defined as

$$X(t): \ e^{X(t)} = (T - I)t + I, \quad 0 \le t \le 1. \tag{3.9}$$

Notice that $X(1)$ defines $\log(T)$ and that $X(t)$ is well defined and real $\forall t \in [0,1]$, because for $(T-I)t+I$ Theorem 1.1 holds since it holds for T. Since $Te^{X(t)} = e^{X(t)}T$, we also have that $X(t)$ satisfies the ODE

$$\begin{aligned} &\dot{X} = (T - I)e^{-X(t)}, \qquad 0 \le t \le 1, \\ &X(0) = 0. \end{aligned} \tag{3.10}$$

By construction, (3.10) defines the principal log of $(T - I)t + I$. Upon using (3.9), we have the explicit solution of (3.10):

$$X(t) = \int_0^t (T - I)((T - I)s + I)^{-1} ds, \qquad 0 \le t \le 1, \tag{3.11}$$

and therefore we find the following expression for $\log(T)$:

$$\log(T) = X(1) = \int_0^1 (T - I)((T - I)t + I)^{-1} dt. \tag{3.12}$$

Remarks 3.6. (i) Formula (3.12) is also derived in the works by Helton and Wouk [He], [Wo]. Their interest was in showing that every invertible matrix had a logarithm.

(ii) Computational procedures for $\log(T)$ can be obtained by using integration formulas for the ODE (3.10) or quadrature rules on (3.12). We have experimented with explicit Runge–Kutta integrators for the ODE (3.10) and several quadrature rules for (3.12). We found that quadrature rules were consistently less costly. Notice that the midpoint rule on (3.12) gives the $(1,1)$ Padé approximant; see also Theorem 4.3.

Formula (3.12) can also be used to obtain a new formula for the Frechét derivative of $G(T) = \log(T)$. In fact, upon considering (3.12) for $\log(T + Z)$, using first-order perturbation arguments and some algebra yields

$$G'(T)Z = \int_0^1 ((T - I)t + I)^{-1} Z ((T - I)t + I)^{-1} dt. \tag{3.13}$$

We also notice that using (3.12) for $\log(T + \Delta T)$ and expanding the inverse there in powers of ΔT justify the $\mathcal{O}(\|\Delta T\|^2)$ term in (2.2).

Now, from (3.13) with $Z = I$, since

$$\int_0^1 ((T - I)t + I)^{-2} dt = -(T - I)^{-1}[((T - I)t + I)^{-1}]_0^1 = T^{-1},$$

we obtain $\|T^{-1}\| \leq \|G'(T)\|$, and so

$$\|T^{-1}\| \leq \|G'(T)\| \leq \int_0^1 \|((T-I)t+I)^{-1}\|^2 dt\,. \tag{3.14}$$

Moreover, (3.13) and (3.14) can be profitably exploited to gain further insight into $\|G'(T)\|$.

LEMMA 3.7. *If T is positive definite, then*

$$\|G'(T)\| = \|T^{-1}\|\,.$$

Proof. Diagonalize T with orthogonal Q on the right-hand side of (3.14) and perform the integration. □

LEMMA 3.8. *If $T \in \mathbb{R}^{2\times 2}$ is normal with complex conjugate eigenvalues $a \pm ib$, then*

$$\|G'(T)\| = \frac{1}{\rho}\frac{\theta}{\sin(\theta)},$$

where $-\pi < \theta < \pi$ is the argument of the eigenvalues of T and ρ their modulus. If T is normal of dimension n, then

$$\|G'(T)\| \geq \max_k \frac{1}{\rho_k}\frac{\theta_k}{\sin(\theta_k)} \geq \|T^{-1}\|\,, \tag{3.15}$$

where θ_k's are the arguments of the eigenvalues of T (if $\theta_k = 0$, replace $\frac{\theta_k}{\sin(\theta_k)}$ by 1) and ρ_k's their modulus.

Proof. In the (2×2) case T is of the form $T = \left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right)$ with complex conjugate eigenvalues $a \pm ib$ (and let $b \neq 0$; otherwise T is positive definite). Then $\lambda(t) = ((a\pm ib-1)t+1)^{-1}$ are the eigenvalues of $((T-I)t+I)^{-1}$. Now, if we take $Z = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ in (3.13), we get that

$$\|G'(T)\| \geq \int_0^1 |\lambda(t)|^2\, dt.$$

For $a \neq 0$, with some algebra, this integral equals $\frac{1}{b}\tan^{-1}\frac{b}{a} = \frac{1}{\rho}\frac{\theta}{\sin(\theta)}$, where θ belongs to $(0,\pi/2)$, $(-\pi,-\pi/2)$, $(-\pi/2,0)$, $(\pi/2,\pi)$ depending on whether $b/a > 0$, and $b > 0$ or $b < 0$, or $b/a < 0$, and $b < 0$ or $b > 0$. Now, one always has $\|G'(T)\| \leq \int_0^1 \|((T-I)t+I)^{-1}\|^2\, dt$, and, because of normality, the norm of $((T-I)t+I)^{-1}$ equals $|\lambda(t)|$. Therefore, as before, we get the reverse inequality

$$\|G'(T)\| \leq \frac{1}{\rho}\frac{\theta}{\sin(\theta)}$$

subject to the same restriction on the argument. Therefore, the result for T normal and (2×2) follows. If $a = 0$, one simply gets $\|G'(T)\| = \frac{1}{\rho}\frac{\pi}{2}$.

For general $T \in \mathbb{R}^{n\times n}$, normal, let Q bring T to the almost diagonal form QTQ^T. Next, consider all matrices Z given by all zeros, except that on the diagonal they have just one 1 or one $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ block according to the eigenvalue structure of QTQ^T, and then (3.15) follows from the previous (2×2) case. □

Remark 3.9. The bound (3.15) indicates that there are two key factors determining the condition of the log problem. One is, as usual, nearness to singularity, as evidenced by the $\frac{1}{\rho}$ factor; the other is nearness to the negative real axis, as evidenced

by the $\sin(\theta)$ factor in the denominator. This second fact detects ill-conditioning based on the restrictions imposed by the choice of primary matrix functions.

Finally, (3.13) can also be used to estimate $\|G'(T)\|_f$ *directly.* We reason similarly to [KL2] but stress that (3.13) is a representation for $G'(T)Z$ that does not need a power series representation nor to go through the inverse function (the exponential). We have the following result.

THEOREM 3.10. *Let* $A(t) := ((T-I)t+I)^{-1}$, *and let* $D(T) := \int_0^1 (A^T(t) \otimes A(t))\, dt$. *Then we have*

$$\|G'(T)\|_f = \|D(T)\|_2 \,. \tag{3.16}$$

Proof. Let $\mathrm{vec}(Z)$ be the vector obtained by writing the columns of Z one after another, and so $\|G'(T)\|_f = \max_{\|\mathrm{vec}(Z)\|_2=1} \|\mathrm{vec}(G'(T)Z)\|_2$. But by (3.13)

$$\begin{aligned} \mathrm{vec}(G'(T)Z) &= \int_0^1 \mathrm{vec}(A(t)ZA(t))dt = \int_0^1 (A^T(t) \otimes A(t))\mathrm{vec}(Z)dt \\ &= \int_0^1 (A^T(t) \otimes A(t))dt\, \mathrm{vec}Z \,, \end{aligned}$$

and the result follows. □

Remark 3.11. The above result can be used, in the same spirit as in [KL2, p. 192], as a starting point for a procedure to estimate $\|G'(T)\|_f$. In fact, since $\|D(T)\|_2 = (\lambda_{\max}(D^T(T)D(T)))^{1/2}$, a power method approach to get the dominant eigenvalue is suitable. By noticing that $D^T(T) = D(T^T)$, with $A(t)$ given in Theorem 3.10, a cycle of this power method can be compactly written as follows. "Given $Z_0 : \|Z_0\|_F = 1$, let $Z_1 = \int_0^1 A(t)Z_0A(t)dt$, and then $Z_2 = \int_0^1 A^T(t)Z_1A^T(t)dt$, so that $(\|Z_2\|_F)^{1/2}$ is an estimate for $\|G'(T)\|_f$. If more accuracy is required, repeat this cycle with $Z_0 := \frac{Z_2}{\|Z_2\|_F}$." In practice, of course, the integral must be replaced by a quadrature rule, and we experimented with composite trapezoidal and Simpson rules and Gauss–Legendre rules. For the initial Z_0, we used what we would have gotten after one cycle of the procedure had we started with $\frac{1}{\sqrt{n}}I$; that is, one first would get $Z_1 = \frac{1}{\sqrt{n}}T^{-1}$, and then a quadrature rule for the next integral would give some Z_2 (e.g., $Z_2 = \frac{1}{6\sqrt{n}}(T^{-T}T^{-1}T^{-T}+16(T-I)^{-T}T^{-1}(T-I)^{-T}+T^{-1})$ if we use the Simpson rule). Thus, we used $Z_0 := Z_2/\|Z_2\|$. This choice of Z_0 gave consistently better results than starting with a random matrix. We have experimented with this way to estimate $\|G'(T)\|_f$ by using at most 10 equally spaced subdivisions for the quadrature rules. This approach was very inexpensive, of course, but not entirely reliable. Often, it overestimated the true value (interestingly, almost never underestimated it), so it revealed itself as a good indicator of ill-conditioning but did not give a good measure of achieved accuracy. On the other hand, we cannot expect that for arbitrary T, hence $A(t)$ in Theorem 3.10, a quadrature rule with few points will be accurate; naturally, when we raised the number of quadrature points, the estimate got better, but this became too expensive. For these reasons, we turned our attention to a different technique, which we explain in the next section.

4. Finite precision, rescaling, discretizations. For the *two series* (3.1) and (3.3), the asymptotic rates of convergence are determined by $\rho(A)$, $A := I - T$ and $\rho(B)$, $B := (I-T)(I+T)^{-1}$, respectively. However, the finite precision behavior of a truncated expansion is influenced by progressively taking powers, that is, A^k for (3.1) and B^{2k+1} for (3.3). Moreover, for (3.3) there is also the inverse of $I+T$ with which

to contend. A worst case analysis tells us that roundoff might be magnified by powers of $\|A\|$ or $\|B\|$, respectively. If $\|A\| < 1$, then (3.1) leads to a safe computation. Also, when $\|A\| < 1$, for (3.3) we would have $\|B\| < \|(I+T)^{-1}\|$, and this can be easily bounded since $I + T = 2(I - \frac{I-T}{2})$, and so $(I+T)^{-1} = \frac{1}{2}(I - A/2)^{-1}$. Then

$$\|(I+T)^{-1}\| \leq \frac{1}{2}\,\frac{1}{1-\|A\|/2} < 1\,.$$

So, under the assumption $\|A\| = \|I - T\| < 1$, the two series (3.1) and (3.3) lead to a safe computation. Also for the *Padé approximants*, the assumption on $\|A\| < 1$ seems essential to making progress. Under this assumption, the finite precision behavior of Padé approximants is well analyzed in [KL1]. In particular, see Lemma 3 of [KL1].

Because the transformation to the Schur form is a stable process, the finite precision behavior of the *Schur method* is chiefly determined by finding $L_{ii} = \log(R_{ii})$ in (3.7) in case Lemma 3.3 does not apply and by solving (3.8). The former factor is the usual one. The second factor is carefully analyzed in [Hi2]. One must solve the following Sylvester equation for Z:

$$R_{ii}Z - ZR_{jj} = C,$$

where the spectra of R_{ii} and R_{jj} are disjoint. Ideally, we would like to select the block partitioning of the matrix R in such a way that all Sylvester equations to be solved are well conditioned so that no eventual loss of precision in the computation is introduced. But, of course, to assess the conditioning of a Sylvester equation requires the equation and its solution, whereas—for the sake of efficiency—we would like to have a criterion to determine the partitioning of R beforehand. We reasoned as follows. If we call ϕ the Sylvester equation operator, $\phi : Z \to R_{ii}Z - ZR_{jj}$, then $\|\phi^{-1}\|$ is an upper bound for a relative error magnification factor (see [Hi2]). It is also known that $\|\phi^{-1}\| \geq \frac{1}{\min|\lambda-\mu|}$, where $\lambda \in \Lambda(R_{ii})$, $\mu \in \Lambda(R_{jj})$ (see [GvL, p. 389]), and we can easily control this lower bound by making sure that $\Lambda(R_{ii})$ and $\Lambda(R_{jj})$ are sufficiently separated. Of course, this does not suffice to make the Sylvester equation well conditioned. Still, after extensive computational experiments, we decided to cluster the eigenvalues so that $|\Lambda(R_{ii}) - \Lambda(R_{jj})| \geq 1/10$, and we have *never* encountered a problem where a system (3.8) was ill conditioned, but the log was well conditioned. For this reason, we think the method should be regarded as stable.

For the *ODE approach*, a quadrature rule must replace the integral in (3.12). That is,

$$\log(T) = \int_0^1 (T-I)((T-I)t + I)^{-1}\,dt := \int_0^1 F(t)\,dt \tag{4.1}$$

must be approximated by a rule of the type

$$Q := \sum_{k=1}^{N} c_k F(t_k). \tag{4.2}$$

For example, consider a composite Simpson rule (identical reasoning applies to different quadratures) to approximate (4.1). Let $F(t) = (T-I)((T-I)t + I)^{-1} =: (T-I)A(t)$, with $A(t) = ((T-I)t + I)^{-1}$. The composite Simpson rule with equal spacing $h = 1/N$ (N even) is

$$CS := \frac{h}{3}\,(F(0) + 4(h) + 2F(2h) + 4F(3h) + \cdots + 2F((N-2)h) + 4F((N-1)h) + F(1)).$$

It is easy to bound the error as

$$\|\log(T) - CS\| \leq \frac{nh^4}{180} \max_{0\leq t\leq 1} \|F^{iv}(t)\|. \tag{4.3}$$

We can verify that $F^{(k)}(t) = (-1)^k k!((T-I)A(t))^{k+1}$, from which

$$F^{iv}(t) = 24[(T-I)A(t)]^5\,, \tag{4.4}$$

which can be used in (4.3) to get error estimates. In case $\|I-T\| = \omega < 1$, the error bound can be sharpened. In fact, we easily get $\|A(t)\| \leq \frac{1}{1-\omega t}$, so that $\|F^{iv}(t)\| \leq 24(\frac{\omega}{1-\omega})^5$, and therefore

$$\|\log(T) - CS\| \leq \frac{4n}{45} h^4 \left(\frac{\omega}{1-\omega}\right)^5. \tag{4.5}$$

Remark 4.1. A direct computational procedure based on a composite quadrature rule discretization of (3.12) can eventually be very accurate, but in general it will be expensive unless T is not far from the identity. Still, for low accuracy, a formula like (4.2) can be profitably used. For example, a modification of the above proved very useful in estimating the norm of the Frechét derivative of $\log(T)$, as we will see later.

To complete the discussion on quadrature rules, we now give a new equivalence result about Gauss–Legendre quadratures on (4.1) and diagonal Padé approximants. Aside from its theoretical interest, this fact allows for a new representation of the error for diagonal Padé approximants.

LEMMA 4.2. *Any quadrature rule of the type* (4.2) *is equivalent to a rational approximation of* $\log(T)$.

Proof. We have $Q := \sum_{k=1}^{N} c_k F(t_k)$, and $F(t) = (T-I)((T-I)t+I)^{-1}$. Since $F(t_i)F(t_j) = F(t_j)F(t_i)\ \forall i,j$, we can rewrite Q as

$$Q = (T-I)\left[\sum_{k=1}^{N} c_k \prod_{i=1, i\neq k}^{N} ((T-I)t_i + I)\right]\left[\prod_{i=1}^{N}((T-I)t_i+I)\right]^{-1},$$

from which the claim follows. □

THEOREM 4.3. *Let* $\rho(I-T) < 1$, *and let* Q *in* (4.2) *be the* N*-point Gauss–Legendre quadrature rule for* $\log(T)$. *Then* Q *is the* (N,N) *diagonal Padé approximant to* $\log(T)$.

Proof. With previous notation, and under the stated assumptions, we have

$$F(t) = (T-I)\sum_{k=0}^{\infty}(-1)^k(T-I)^k t^k\,,$$

where the series converges. Therefore,

$$\log(T) = \sum_{k=1}^{\infty}(T-I)\int_0^1 (-1)^k (T-I)^k t^k\, dt\,.$$

Since N-point Gauss–Legendre rules are exact for polynomials of degree up to t^{2N-1}, we immediately realize that Q agrees with $\log(T)$ up to the term $(T-I)^{2N+1}$ excluded.

From Lemma 4.2, Q is a rational approximation to $\log(T)$, and thus it must be the (N,N) diagonal Padé approximant. □

COROLLARY 4.4. *Under the assumptions of Theorem* 4.3, *we have the following error estimate for the* (N,N) *diagonal Padé approximants* Q *to* $\log(T)$*:*

$$\log(T) - Q = \frac{(N!)^4}{(2N+1)((2N)!)^3} \sum_{k=0}^{\infty} (2N+k)\dots(k+1)A^{2N+k+1}\eta^k,$$

where $0 \le \eta \le 1$ *and* $A = I - T$.

Proof. From standard quadrature errors for Gauss–Legendre rules (e.g., see [AS]) and differentiating under the series of Theorem 4.3, the result follows at once. □

Remark 4.5. The previous results hint that a possible way to use quadrature rules is first to pass to their rational form equivalent. On the other hand, for diagonal Padé approximants, it might instead be more desirable to pass to their quadrature formula equivalent (4.2) to avoid ill-conditioning in the denominator of the rational function. Moreover, from Theorem 4.3 we see that Gauss formulas are an excellent candidate for a parallel implementation of Padé approximants.

From the preceding discussion, it has become clear that it would be generally desirable to have T close to I. This would make the finite precision behavior of the above techniques much better.

Scaling. An ideal scaling strategy, in the context of computing $\log(T)$, is to precondition the problem so that (for a modified matrix T) $T \approx I$. In any case, a reasonable scaling ought to give a T for which $\|I - T\| < 1$.

One approach is to find, inexpensively, some X_1 approximating $\log(T)$ such that $X_1 T = T X_1$ and then to consider $e^{-X_1}T$, find its logarithm, and finally recover $\log(T) = X_1 + \log(e^{-X_1}T)$. Some ideas on this are in [D]. Also (3.12) can be used in this light, since any quadrature rule of the type (4.2) gives $X_1 : X_1 T = T X_1$.

A more systematic approach results from the *inverse scaling and squaring* procedure of Kenney and Laub [KL2]. The basic idea of this approach is to "flatten out" the matrix T. It is based upon the identity $\log(T) = \log((T^{1/2^k})^{2^k}) = 2^k \log(T^{1/2^k})$ and the realization that, eventually, $T^{1/2^k} \to I$. With respect to this scaling procedure, we must consider (i) how to take square roots and which square roots should we take, (ii) when should we take square roots, (iii) what is the conditioning of the overall procedure, and (iv) if there are risks involved with this scaling strategy.

With respect to the first issue, we have adopted the choice made by Higham (see [Hi1] and also [BH]), thereby relying on a real Schur approach. Under the assumptions of Theorem 1.1, there are many square roots of T; see [Hi1, Thms. 5 and 7]. However, in our context, to find the principal branch of $\log(T)$ eventually, there is only one choice. We *must* select the square root(s) according to the lemma below (see also [KL2, Lem. A1]).

LEMMA 4.6. *Let* $B \in \mathbb{R}^{m\times m}$ *be invertible with no eigenvalues on the negative real axis. Then* B *has a unique* 2^k*th root* S, *i.e.,* $S^{2^k} = B$, *which is a primary matrix function of* B, *and such that if* $\nu \in \Lambda(S)$, *then*

(a) $-\pi/2^k < \arg(\nu) < \pi/2^k$ *and*

(b) $\Re\mathrm{e}(\nu) > 0$ *for* $k = 1, 2, \dots$.

Proof. A constructive proof can be based upon the method of Higham (see [Hi1, p. 417] for details). □

When to take square roots? Ultimately, it all depends on what algorithm we use to approximate $\log(T)$. For algorithms fully based on truncated series or Padé approximants, square roots of the full matrix T have to be taken to ensure numerical

stability and rapid convergence. When using a Schur decomposition approach, the procedure is only needed to obtain L_{ii} in (3.7) in those cases for which approximation techniques are required for the L_{ii}. One thing to keep in mind is that, asymptotically, taking square roots gives a decrease in norm by a factor of 2. Therefore, how many square roots to take depends on which algorithm we eventually use for computing the log of the scaled matrix.

To examine the conditioning of the inverse scaling and squaring procedure, we must look at the Frechét derivative of $M(T) := 2^k \log(T^{1/2^k})$. Let $T_j = T^{1/2^j}$, $j = 0, 1, \ldots, k$ (so $T_0 = T$); and let G and F be the log and square root functions, respectively. Then, upon repeated use of Lemma 2.3, we have

$$M'(T)Z = 2^k G'(T_k)F'(T_{k-1}) \ldots F'(T_0)Z\,.$$

In other words, unavoidably, the better value for the norm of the Frechét derivative of the log (because $T_k \approx I$) is being paid by the Frechét derivatives of the square roots. The problem of estimating the Frechét derivative of the square root function can be based on Corollary 2.4 by considering $S(X) := X^2$ and the identity $S(F(T)) = T$. Therefore, we have the equalities

$$F'(T_0)Z = (S'(F(T_0)))^{-1}Z, \quad F'(T_1)(F'(T_0)Z) = (S'(F(T_1)))^{-1}F'(T_0)Z\,, \ldots,$$

$$F'(T_{k-1})(F'(T_{k-2}) \ldots F'(T_0)Z) = (S'(F(T_{k-1})))^{-1}(S'(F(T_{k-2})))^{-1} \ldots (S'(F(T_0)))^{-1}Z\,,$$

and thus we have

$$G'(T)Z = 2^k G'(T_k)\left\{(S'(F(T_{k-1})))^{-1} \ldots (S'(F(T_0)))^{-1}Z\right\}\,. \tag{4.6}$$

Formula (4.6) forms the basis of the following algorithm to estimate $\|G'(T)Z_0\|$ for a given Z_0, and hence to estimate $\mathrm{cond}(G(T))$. This procedure gave us much better results (both in terms of accuracy and expense) than one directly based on Theorem 3.10.

Let $T_0 = T$, $T_j = T^{1/2^j}$, $j = 1, \ldots, k$, where the index k must be chosen so that $\|I - T_k\| = \omega < 1$; and let Z_0 be given.

(a) Solve

$$F(T_j)Z_{j+1} + Z_{j+1}F(T_j) = Z_j\,, \quad j = 0, 1, \ldots, k-1 \tag{4.7}$$

(notice that the $F(T_j)$ stay quasi-triangular if T_0 is such; also, one might already have the T_j from scaling via taking square roots, but only if square roots of all of T had been taken).

(b) Since $G'(T)Z_0 = 2^k G'(T_k)Z_k$, we approximate $G'(T_k)Z_k$ by using a quadrature rule on (3.13).

It is obvious that the algorithm is well defined, since the Sylvester equations (4.7) are uniquely solvable. In terms of computational cost, by using a composite quadrature rule with N points, at leading order one needs $\frac{1}{6}(k+N)n^3$ flops, plus the cost of computing the T_j's if they are not available, which might amount to another $\frac{1}{6}kn^3$ flops, plus the initial cost of the Schur reduction of T.

Next, we show that the above eventually provides a good estimate of $\|G'(T)Z_0\|$. We show this for the composite Simpson rule, but the reasoning applies to any other quadrature rule.

THEOREM 4.7. *Let $T \in \mathbb{R}^{n\times n}$ be given such that $\|I - T\| = \omega < 1$, and let $G(T) = \log(T)$. Let Z be given, and let $G'(T)Z$ be given by* (3.13). *Let CS be the composite Simpson rule with N points (N even) approximating* (3.13) *so that $h = 1/N$ below. Then we have*

$$\|G'(T)Z - CS\| \leq \frac{2nh^4}{3}\frac{\omega^4}{(1-\omega)^6}\|Z\|\,. \tag{4.8}$$

Proof. We have

$$G'(T)Z = \int_0^1 A(t)ZA(t)\,dt = \int_0^1 F(t,Z)\,dt,$$

where we have set $A(t) = ((T-I)t + I)^{-1}$ and $F(t,Z) = A(t)ZA(t)$. From standard quadrature errors, we have

$$\|G'(T)Z - CS\| \leq \frac{nh^4}{180}\max_{0\leq t\leq 1}\|F^{(iv)}(t,Z)\|.$$

Now, we can verify that $A^{(j)}(t) = (-1)^j j!A(t)((T-I)A(t))^j$ and that

$$\begin{aligned} F^{(k)}(t,Z) &= \sum_{j=0}^{k}\binom{k}{j}A^{(k-j)}(t)ZA^{(j)}(t) \\ &= (-1)^k k!\sum_{j=0}^{k}A(t)((T-I)A(t))^{k-j}ZA(t)((T-I)A(t))^j\,, \end{aligned}$$

from which it is easy to get

$$\|F^{iv}(t,Z)\| \leq 120\frac{\omega^4}{(1-\omega)^6}\|Z\|\,,$$

and the result follows. □

THEOREM 4.8. *Let $T \in \mathbb{R}^{n\times n}$, $G(T) = \log(T)$, and $E(T) = e^T$. Let Z_0 of norm 1 be given; and let k be such that $\|I - T_k\| = \omega < 1$, with $T_k := T^{1/2^k}$. Let Z_k be obtained from* (4.7) *so that $G'(T)Z_0 = 2^kG'(T_k)Z_k$. Let CS be the composite Simpson rule with N points (N even) approximating $G'(T_k)Z_k$ from* (3.13) *so that $h = 1/N$ below, and let $G'(T_k)$ be invertible. Then we have*

$$\frac{\|G'(T)Z - 2^kCS\|}{\|G'(T)Z_0\|} \leq \frac{nh^4}{9}\frac{\omega^4(1+\omega)}{(1-\omega)^6}\,. \tag{4.9}$$

Proof. From Theorem 4.7, we have

$$\|G'(T)Z_0 - 2^kCS\| \leq \frac{2nh^4}{3}\frac{\omega^4}{(1-\omega)^6}\|2^kZ_k\|\,.$$

On the other hand, from $G'(T)Z_0 = 2^kG'(T_k)Z_k$, we also have $\|2^kZ_k\| \leq \|G'(T)Z_0\|\,\|(G'(T_k))^{-1}\|$, and from Corollary 2.4 we get $\|(G'(T_k))^{-1}\| = \|E'(G(T_k))\|$. Therefore, we have

$$\frac{\|G'(T)Z_0 - 2^kCS\|}{\|G'(T)Z_0\|} \leq \frac{2nh^4}{3}\frac{\omega^4}{(1-\omega)^6}\|E'(G(T_k))\|\,. \tag{4.10}$$

Now, using (2.7) we have

$$\|E'(G(t_k))\| = \max_{Y:\,\|Y\|=1} \left\| \int_0^1 T_k(1-s)YT_k\, s\, ds \right\| \leq \frac{1}{6}\|T_k\| \leq \frac{1+\omega}{6}.$$

Using this in (4.10) gives the result. □

Example 4.9. If $h^{-1} \approx (n)^{1/4}$, then $\omega = .25$ gives three digits accuracy and $\omega = .35$ gives two digits. This is more than acceptable for condition estimation. □

Remark 4.10. Using (4.9) to achieve a good estimate of $\|G'(T)\|$ requires an appropriate choice of Z_0. We have found that selecting Z_0 according to Remark 3.11 always gave excellent results, and no need arose to iterate the process further. For our experiments in §6, we always used this choice of Z_0 along with (4.9) to estimate $\text{cond}(G(T))$. This strategy seems to be both very reliable and efficient in comparison with existing alternatives [KL2].

To complete this section, we ought to warn against some possible risks involved with the "inverse scaling and squaring" procedure. Its main limitation is exactly its power. That is, one progressively flattens out the spectrum of the matrices $T_j = T^{1/2^j}$. This may lead to an unwanted loss of numerical significance in those cases in which the original T has close eigenvalues (but not identical) and several square roots are required to obtain a $T_j : \|I - T_j\| < 1$. The risk is that, after many square roots, all eigenvalues have numerically converged to 1 and are no longer distinct. Our experience has shown that this might occasionally happen, but only for ill-conditioned problems, for which $\|T^{1/2^j}\|$ increases with j, before decreasing.

5. Implementation and expense. In our implementations to approximate $\log(T)$, we have always first reduced the matrix T to ordered quasi-triangular form via a real Schur reduction. The ordered Schur reduction is standard, and we used routines from EISPACK and from [St], thereby ordering eigenvalues according to their modulus. Unless more information is available on T, we always recommend a Schur reduction prior to an approximation technique; inter alia, it allows for an immediate solution of the problem if T is normal (see Corollary 3.4), and it renders transparent whether or not some methods are suitable for the given problem. In what follows, we will therefore assume that T is quasi-triangular and not normal. In tune with our discussion on scaling, we will also assume hereafter that T has been scaled so that $\|A\| < 1$, where $A = I - T$. Typically, this has been achieved by progressively taking square roots of T. To assess the computational expense, we give the leading order flops' count of the algorithms; a flop is the combined expense of one floating point multiplication and one floating point addition.

Both for truncated expansions of the two series and for diagonal Padé approximants, one needs to evaluate matrix polynomials. Ignoring finite precision considerations, let us first discuss what degree is needed to obtain a desired accuracy for a given $\|A\|$. We fixed the accuracy to 10^{-18}.

Figure 1 is a graph showing which degrees q are needed as functions of $\|A\|$, to be guaranteed an absolute error less than 10^{-18} for approximation resulting by

(i) truncating the series (3.1)

$$S_1 := \sum_{k=1}^{q} \frac{A^k}{k}; \tag{5.1}$$

(ii) truncating the series (3.3)

$$S_2 := 2\sum_{k=0}^{m} \frac{B^{2k+1}}{2k+1}, \quad B = (T-I)(T+I)^{-1}, \quad q = 2m+1; \tag{5.2}$$

(iii) considering the diagonal Padé approximant $R_{q,q}(A)$.

To obtain the degrees q, we have made sure that the remainders contributed less than the desired accuracy. This is easy enough to do for (5.1) and (5.2), and for the Padé approximants we used the explicit form of the remainder from [KL1, Thm. 5].

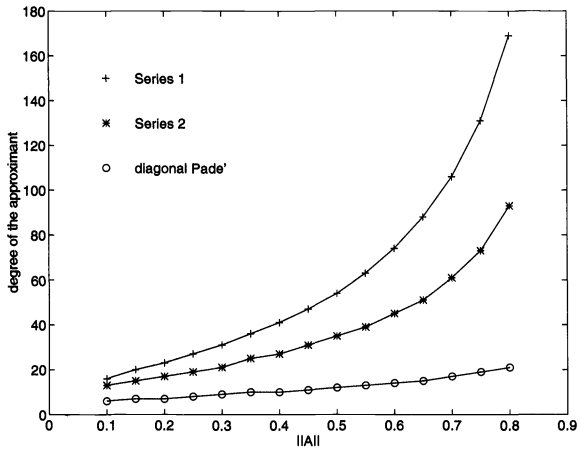


FIG. 1.

As an example, for $\|A\| \leq 0.35, 0.3$, we need $q = 36, 31$ for S_1, $q = 25, 21$ for S_2, and $q = 10, 9$ for $R_{q,q}$. (If $\|A\| = 0.35$, the $(9,9)$ Padé guarantees an error of 1.152×10^{-18}.)

Naturally, for Padé one also needs to be aware of the condition number of the denominator $Q(A)$, since this matrix needs to be inverted. Borrowing from [KL1, Lem. 3], an upper bound on $\text{cond}(Q(A))$ is given by $Q(-\|A\|)/Q(\|A\|)$. Figure 2 shows this upper bound on $\text{cond}(Q(A))$ for the case of $q = 9$ for $\|A\| \in (0,1)$. For example, for $\|A\| = 0.35, 0.3$, one has that $\text{cond}(Q(A)) \leq 25.34, 15.66$.

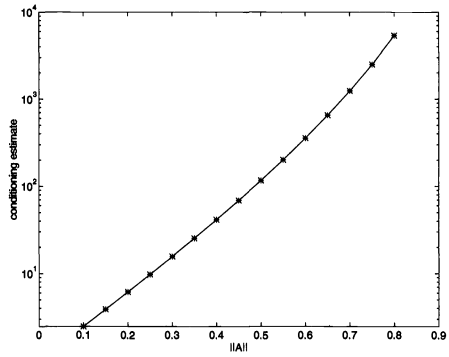


FIG. 2.

Next, we need to consider the expense associated with evaluating polynomials of degree q and the $q \times q$ diagonal Padé. As usual, let T be quasi-triangular of dimension n. The algorithm we used to evaluate the polynomials is taken from [GvL, §11.2], and it requires the explicit computation of $A^2, A^3, \ldots, A^s$, where s is a given integer satisfying $1 \leq s \leq \sqrt{q}$. Let $r = \lfloor q/s \rfloor$; then, following [GvL], it is easy to show that,

at leading order, the evaluation of S_1 requires $(r+s-2)\frac{1}{6}n^3$ flops if $sr = q$ and $(r+s-1)\frac{1}{6}n^3$ flops otherwise. The choice $s = \lfloor\sqrt{q}\rfloor$ ensures the minimal flop count.

The cost associated with S_2 can be obtained in a similar way, taking into account the cost of the evaluation of $B = (T-I)(T+I)^{-1}$ (about $\frac{1}{6}n^3$ flops) and observing that only odd powers of B are required. With $q = 2m+1$ now we have $s = \lfloor\sqrt{m}\rfloor$, $r = \lfloor m/s\rfloor$, and a leading cost of $(r+s+1)\frac{1}{6}n^3$ flops if $sr = m$ and $(r+s+2)\frac{1}{6}n^3$ flops otherwise.

Finally, the cost associated with $R_{q,q}(A)$ can be obtained observing that $A^2, A^3, \ldots, A^s$ must be computed only once for the two polynomials $P(A)$ and $Q(A)$ and adding the cost of the evaluation of $P(A)(Q(A))^{-1}$. With the above notation, we have a leading cost of $(2r+s-2)\frac{1}{6}n^3$ flops if $sr = q$ and $(2r+s)\frac{1}{6}n^3$ flops otherwise. In this case, a better compromise for s is $s = \lceil\sqrt{q}\rceil$, which permits us to gain something in the flop count with respect to taking $s = \lfloor\sqrt{q}\rfloor$.

Figure 3 shows the asymptotic cost associated with S_1, S_2, and $R_{q,q}(A)$ having an error less than 10^{-18} in function of $\|A\|$. For example, if $\|A\| \leq 0.35$, 0.3, S_1 requires about $10\frac{1}{6}n^3$ flops, S_2 needs $8\frac{1}{6}n^3$ flops, and $R_{q,q}(A)$ needs $q = 10$ and $8\frac{1}{6}n^3$ flops for $\|A\| = 0.35$, whereas $q = 9$ and $7\frac{1}{6}n^3$ flops suffice when $\|A\| = 0.3$. It is interesting to observe that also using a $(12,12)$ Padé gives a leading flop count of about $8\frac{1}{6}n^3$ flops.

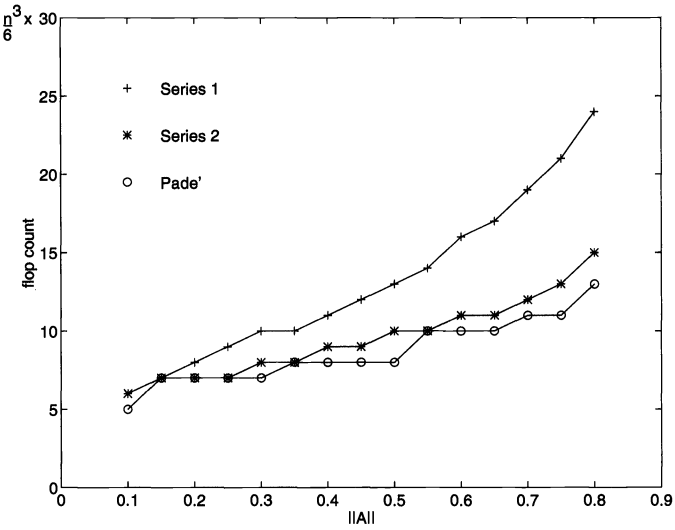


FIG. 3.

Finally, we must consider the cost of the real Schur decomposition and of taking square roots. The cost of solving (3.8) is a complicated function of the block sizes; for distinct eigenvalues, i.e., the triangular case, it amounts to $\frac{1}{3}n^3$ flops. In any case, the bulk of the expense is the ordered real Schur decomposition, which costs about $15n^3$ flops. Then one square root costs about $\frac{1}{6}n^3$ flops (see [Hi1]). Since taking square roots, asymptotically, decreases the norm by 1/2, then we see that it makes better sense, from the point of view of the cost, to take square roots rather than to use a high-degree approximant. We found that a good compromise is to take square roots having up to $\|A\| \leq 0.35$, followed by the $(9,9)$ Padé or S_2.

6. Examples. In this section we report on some of the problems we have solved numerically. All computations have been done on a Sparc10 in double precision ($EPS \approx 2.2 * 10^{-16}$).

We mainly report on results obtained by the methods that have proven robust enough to handle the largest portion of all problems considered; for example, we do not report on results obtained by using (5.1) or by using the ODE approach in either formulation (3.10) or (3.12) (but see Theorem 4.3). Thus, unless otherwise noted, all problems below have been solved by the following general strategy.

(i) Schur reduction with eigenvalues' clustering according to increasing modulus. We used the software in [St] (with minimal modifications) to do this step. The tolerance for the QR algorithm was set to $2 * EPS$.

(ii-a) Scaling of diagonal blocks by taking square roots, obtaining up to $\|A\| \leq 0.35$, followed by the 9×9 diagonal Padé approximant for these blocks, inverse scaling, and the use of (3.8). Diagonal blocks in the real Schur form have been considered distinct if the minimum distance between their eigenvalues was greater than $1/10$. Needless to say, if—after grouping—all diagonal blocks were either 1×1 or 2×2 of complex conjugate eigenvalues, then we used Lemma 3.3 instead of scaling and Padé approximants.

(ii-b) Truncated expansion (5.2) on the whole matrix in lieu of scaling by square roots and Padé approximants if convergence criteria for such series were met.

(iii) Back transformation.

As a measure of accuracy for the computed logarithms, we considered err $:= \|e^{\log_c(T)} - T\|/\|T\|$, where $\log_c(T)$ is our computed approximation to the log. This essentially boils down to assessing the absolute error in the log itself. To approximate the exponential function, we used both Matlab functions *expm* and *expm2*, which are a Schur-based technique and a series expansion technique. Typically, *expm* performed better, but on occasions *expm2* was needed. We also used our own implementation of the method of scaling and squaring followed by a diagonal Padé approximant to the exponential, following [GvL, Alg. 11.3.1, p. 558]. In the examples below, we also report the estimates "cond" of the condition number (2.3). This is done according to Theorem 4.8.

Many tests were done on random matrices. These were generated by exponentiating the matrices obtained with the *randn* function of Matlab, which returns entries in $[-1, 1]$ according to the normal distribution. If a particular structure was desired (e.g., orthogonal) these random matrices were further manipulated (e.g., taking their QR factorization).

In the tables below, for the computed logarithm $\log_c T$, we report L= $\|\log_c T\|$, cond, nbl/nrad (the number of diagonal blocks, and the most square roots taken on any of these blocks), err, q (the number of terms taken for (5.2) directly on T, if applicable), err_2 (the error for (5.2)), and err_m (the error obtained by using the Matlab function *logm* to approximate $\log T$). Exponential notation is used throughout; e.g., 2.3×10^7 is written as 2.3E7. All results are given for the Frobenius norm to conform to previously published results.

Example 6.1. *"Easy" problems.* A set of randomly generated positive definite and orthogonal matrices was considered just to test the technique based on Corollary 3.4. In all cases, accuracy to machine precision was obtained. We also generated more than 60 general random matrices of dimension between 5 and 100. Also in these cases we obtained accuracy to full machine precision.

Example 6.2. *Symplectic* T. We generated a dozen random symplectic matrices by exponentiating (via diagonal Padé approximants) randomly generated Hamiltonian matrices. For some of these matrices we got a very large condition number (3.2). Nonetheless, we obtained very accurate answers for the computed logarithms. However, the end result was often far from being a Hamiltonian matrix; that is, the

TABLE 1.

Test	L	cond	nrad	err	q	err_2	err_m
1	6.98E7	4.75E10	28	1.2E-8	239	1.1E-8	82.54
2	5.32	5.0865	4	0	19	0	9E-3
3	6.56	0.9511	5	2.7E-15	129	3.7E-16	1.7E-2
4	5E9	5.67E14	34	5.9E-4	5	5E-13	1.4E15

TABLE 2.

Test	L	cond	nbl/nrad	err	q	err_2	err_m
1	7.48	5.08	2/4	3.5E-15	7949	1.7E-13	1.1E-15
2	53.85	9E6	3/0	9.E-15	–	–	7.1E-15
3	575.95	6.44E9	3/0	6.2E-14	–	–	5.2E-13
4	2.9997	3.76	1/4	2.5E-15	19	1.5E-16	6.7E-9
5	1E6	3.33E11	1/22	0	1	0	6.2E-6
6	172.68	5.94E6	1/9	3.7E-13	229	2.3E-10	6.4E-10

relevant structure got lost. For these problems, when applicable, using (5.2) directly was also an effective way to proceed; even though some of the linear algebra (such as matrix inversion) was done by nonsymplectic methods, the end result was much more nearly a Hamiltonian matrix than with the Schur method (see [D]).

Example 6.3. *"Harder" problems.* These problems have been chosen to illustrate some of the dangers in using the *logm* function of Matlab. In Table 1, Tests 1–3 refer to a triangular matrix of dimension 20, with all 1's above the diagonal, and $1/4, 1$, and 4 on the diagonal, respectively. Of course, for these matrices, no Schur reduction or grouping occurred. Test 4, instead, was chosen to illustrate the potential danger of taking too many square roots. It is the matrix

$$\begin{pmatrix} 1+10^{-7} & 10^5 & 10^4 \\ 0 & 1 & 10^5 \\ 0 & 0 & 1 \end{pmatrix}.$$

In this case, (5.2) is clearly preferred.

Example 6.4. *Examples from the literature.* These problems have previously appeared in the literature; see [Wa] and [KL2]. We tested our method to confirm independently the results of [KL2] about conditioning. In Table 2, Tests 1–6 refer to Examples 1–6 of [KL2]. We notice that our estimates for cond are in perfect agreement with the results in [KL2]. For Tests 1, 2, and 3, we also used scaling by square roots and the 9×9 diagonal Padé approximant on the whole matrix; this required 5, 8, and 11 square roots, respectively, for the same accuracy.

7. Conclusions. In this work, we provided analysis and implementation of techniques for computing the principal branch of a real logarithm of a matrix T, $\log(T)$. Some of the techniques considered have been around for a while, like Padé approximants and series expansion. Some other techniques have not been previously analyzed or even introduced, in particular, the Schur method with eigenvalue grouping followed by a back recursion and integral-based representations for both the logarithm and its Frechét derivative. This latter aspect is related to the conditioning of the problem, an issue we have addressed in detail, and on which we have given many new results that

better characterize it. In fact, from the theoretical point of view, our main contributions are the results about conditioning and those related to the integral representation of $\log(T)$.

From the computational point of view, all things considered, we think that the most reliable and efficient general-purpose method is one based on the real Schur decomposition with eigenvalues' grouping, scaling of the diagonal blocks via square roots, and diagonal Padé approximants. Also using S_2 (see (5.2)), instead of the Padé approximant, is a sound choice. Moreover, using S_2 was definitely the most appealing choice for poorly conditioned problems. Although all of the programs we have written are of an experimental nature, we believe they are robust enough to be indicative of the typical behavior. We hope that our work will prove valuable to people interested in mathematical software, the moreso since the only existing software tool that computes the logarithm of a matrix ([Matlab]) does not use a foolproof algorithm to do so. Moreover, the implementation of Matlab nearly always produces complex matrices for answers because it uses unitary reduction to complex Schur form.

The problem of reliably estimating the Frechét derivative of $\log(T)$ at a fraction of the cost of computing $\log(T)$, or at least without a drastic increase in cost, is truly an outstanding difficulty. None of the methods of which we are aware succeeds in this. One technique we have considered, based on Theorem 3.10 and Remark 3.11, is usually very inexpensive but not always reliable. The other technique we introduced, based on Theorem 4.8, has at least proven very reliable but, in general, it is at least as expensive as computing the log itself.

Finally, in this work we focused on the problem of computing *one* logarithm of *one* matrix. Different conclusions are reached if one is interested in computing a branch of logarithms of slowly varying matrices. In such cases, of course, one should favor an approach that uses the previously computed logarithms and, thus, more carefully consider iterative techniques and different scaling strategies. We anticipate some work in this direction.

Acknowledgments. We thank N. Higham, A. Iserles, and C. Kenney for insightful comments on this work. Also, thanks are due to N. Higham for pointing out to us [C], [He], [SS], and [Wo].

REFERENCES

[AS] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 10th ed., J. Wiley & Sons, New York, 1972.

[BG-M] G. A. Baker and P. Graves-Morris, *Padé Approximants, Parts I and II*, Encyclopedia of Mathematics, vols. 13–14, Addison-Wesley, Reading, MA, 1981.

[BH] A. Bjorck and S. Hammarling, *A Schur method for the square root of a matrix*, Linear Algebra Appl., 52/53 (1983), pp. 127–140.

[C] W. J. Culver, *On the existence and uniqueness of the real logarithm of matrix*, Proc. Amer. Math. Soc., 17 (1966), pp. 1146–1151.

[D] L. Dieci, *Considerations on computing real logarithms of matrices, Hamiltonian logarithms, and skew-symmetric logarithms*, Linear Algebra Appl., to appear.

[G] F. R. Gantmacher, *Théorie des Matrices*, vols. 1 and 2, Dunod, Paris, 1966.

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

[He] B. W. Helton, *Logarithms of matrices*, Proc. Amer. Math. Soc., 19 (1968), pp. 733–738.

[Hi1] N. J. Higham, *Computing real square roots of a real matrix*, Linear Algebra Appl., 88/89 (1987), pp. 405–430.

[Hi2] ———, *Perturbation theory and backward error for $AX - XB = C$*, BIT, 33 (1993), pp. 124–136.

[HJ] R. A. HORN AND C. R. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, London and New York, 1991.

[KL1] C. KENNEY AND A. J. LAUB, *Padé error estimates for the logarithm of a matrix*, International J. Control, 50-3 (1989), pp. 707–730.

[KL2] ———, *Condition estimates for matrix functions*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 191–209.

[LS1] G. J. LASTMAN AND N. K. SINHA, *Transformation algorithm for identification of continuous-time multivariable systems from discrete data*, Electron. Lett., 17 (1981), pp. 779–780.

[LS2] ———, *Infinite series for logarithm of matrix, applied to identification of linear continuous-time multivariable systems from discrete-time models*, Electron. Lett., 27-16 (1991), pp. 1468–1470.

[M] R. MATHIAS, *Condition estimation for matrix functions via the Schur decomposition*, SIAM J. Matrix Anal. Appl., 16-2 (1995), pp. 565–578.

[Matlab] *Matlab Reference Guide*, The MathWorks, Inc., 1992.

[MvL] C. B. MOLER AND C. VAN LOAN, *Nineteen dubious ways to compute the exponential of a matrix*, SIAM Rev., 20 (1978), pp. 801–836.

[P] B. N. PARLETT, *A recurrence among the elements of functions of triangular matrices*, Linear Algebra Appl., 14 (1976), pp. 117–121.

[Si] Y. SIBUYA, *Note on real matrices and linear dynamical systems with periodic coefficients*, J. Math. Anal. Appl., 1 (1960), pp. 363–372.

[SS] B. SINGER AND S. SPILERMAN, *The representation of social processes by Markov models*, Amer. J. Sociology, 82-1 (1976), pp. 1–54.

[St] G. W. STEWART, *HQR3 and EXCHNG: Fortran subroutines for calculating and ordering the eigenvalues of a real upper Hessenberg matrix*, ACM Trans. Math. Software, 2 (1970), pp. 275–280.

[vL] C. VAN LOAN, *The sensitivity of the matrix exponential*, SIAM J. Numer. Anal., 14 (1977), pp. 971–981.

[V] E. I. VERRIEST, *The matrix logarithm and the continuation of a discrete process*, in Proc. 1991 American Control Conference, 1991, pp. 184–189.

[YS] V. A. YAKUBOVICH AND V. M. STARZHINSKII, *Linear Differential Equations with Periodic Coefficients*, vol. 1, John Wiley, New York, 1975.

[Wa] R. C. WARD, *Numerical computation of the matrix exponential with accuracy estimates*, SIAM J. Numer. Anal., 14 (1977), pp. 600–610.

[Wo] A. WOUK, *Integral representation of the logarithm of matrices and operators*, J. Math. Anal. Appl., 11 (1965), pp. 131–138.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 3, pp. 594–609, July 1996

010

AN ANALYSIS OF ZERO SET AND GLOBAL ERROR BOUND PROPERTIES OF A PIECEWISE AFFINE FUNCTION VIA ITS RECESSION FUNCTION*

M. SEETHARAMA GOWDA†

This paper is dedicated to Professor Richard W. Cottle on the occasion of his 60th birthday.

Abstract. For a piecewise affine function $f : R^n \to R^m$, the recession function is defined by

$$f^\infty(x) := \lim_{\lambda \to \infty} \frac{f(\lambda x)}{\lambda}.$$

In this paper, we study the zero set and error bound properties of f via f^∞. We show, for example, that f has a zero when f^∞ has a unique zero (at the origin) with a nonvanishing index. We also characterize the global error bound property of a piecewise affine function in terms of the recession cones of the zero sets of the function and its recession function.

Key words. piecewise affine function, recession function, error bounds, affine variational inequality, linear complementarity problem

AMS subject classifications. 90C30, 90C33, 49J40, 54C60

1. Introduction.

Consider a piecewise affine function $f : R^n \to R^m$. This means that f is *continuous* and R^n admits a polyhedral subdivision $\{\Omega_1, \Omega_2, \ldots, \Omega_L\}$ such that

$$f(x) = A_j x + a_j \quad \text{on} \quad \Omega_j \quad (j = 1, 2, \ldots, L), \tag{1}$$

where $A_j \in R^{m \times n}$ and $a_j \in R^m$. As in the case of a real valued convex function [26], f admits a recession function defined by [27]

$$f^\infty(x) = \lim_{\lambda \to \infty} \frac{f(\lambda x)}{\lambda}. \tag{2}$$

Similar to the recession cone of a polyhedral set, the recession function of a piecewise affine function deals with the behavior of the function at ∞; for f given by (1), f^∞ is described by (some or all) A_js which correspond to unbounded Ω_js. The recession function appears naturally in the investigations of one-to-one and onto properties. For example, Kojima and Saigal [9], [10] and Schramm [29] investigate the homeomorphism property of f by imposing conditions on the matrices corresponding to unbounded Ω_js. Scholtes, in [27], formally introduces the notion of recession function and proves that when f is coherently oriented (meaning that $m = n$ and all the A_js have the same nonzero determinantal sign), the injectivity (one-to-oneness) of f is equivalent to that of f^∞.

The main objective of this article is to describe zero set and global error bound properties of f via f^∞.

Motivated by existence and stability results in the study of (affine) complementarity problems (such as linear, horizontal, mixed, and vertical complementarity problems), affine variational inequalities [2], [6]–[8], [31], [32], and the surjectivity (onto)

* Received by the editors December 21, 1994; accepted for publication (in revised form) by R. Cottle August 1, 1995. This research was supported by National Science Foundation grant CCR-9307685.

† Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD 21228 (gowda@math.umbc.edu).

results for piecewise affine functions [1], [18], [29], we prove (in Theorem 4.2) that if f^∞ has a unique zero (namely, the origin) and if the index of f^∞ at the origin is nonzero, then f and all piecewise affine functions equivalent to f will have nonempty zero sets. Based on the index of f^∞, in Theorem 4.5, we give a necessary and sufficient condition for f to be a homeomorphism.

The error bound results are useful, particularly in optimization, for sensitivity analysis, exact penalization, convergence analysis of iterative schemes, etc. The literature on this subject is vast and we confine ourselves to quoting a few that are relevant to our study. Many local error bound results in the areas of (affine) complementarity problems and affine variational inequalities follow from Robinson's result on the upper Lipschitzian property of polyhedral multifunctions [24]. When stated in terms of a piecewise affine function f, it says that locally a constant multiple of $||f(x)||$ bounds the distance between a test vector x and the zero set of f. Motivated by various applications, Luo and Tseng [13], [15] considered, for functions f arising from the linear complementarity problem (LCP) and the affine variational inequality problem (AVI), the question of deciding when a constant multiple of $||f(x)||$ acts as a *global* error bound for the distance between a test vector x and the zero set of f. In [13], Luo and Tseng answer this question for the LCP with $f(x) := x \wedge (Mx + q)$ in terms of the recession cones of the zero sets of f and $g(x) := x \wedge Mx$. In §5 of this paper we extend the analysis of Luo and Tseng to piecewise affine functions. By specializing our main result (Theorem 5.4), we derive new necessary and sufficient conditions for global error bounds in the contexts of AVI, LCP, and linear programming (LP). As a consequence of Theorem 5.4 we prove an error bound characterization of **P**-matrices.

2. Preliminaries. For a comprehensive treatment of piecewise affine functions, we refer to Scholtes [27]. Note that the term "piecewise linear" is also widely used. We shall say that a finite set $\{\Omega_1, \Omega_2, \ldots, \Omega_L\}$ is a *polyhedral subdivision of* R^n if each Ω_j is a polyhedral set in R^n with nonempty interior, the union of these Ω_js is all of R^n, and the intersection of any two Ω_js is either empty or a proper common face of both. A piecewise affine function can also be introduced without referring to any polyhedral subdivision [27]: A continuous mapping f from R^n to R^m is piecewise affine if there exist affine functions $f_j : R^n \to R^m$ $(j = 1, 2, \ldots, L)$ such that for each $x \in R^n$,

$$f(x) \in \{f_1(x), f_2(x), \ldots, f_L(x)\}.$$

(This equivalent formulation is particularly useful while describing the recession function.) We write $\mathcal{PA}(R^n, R^m)$ for the set of all piecewise affine functions from R^n into R^m and write $\mathcal{PA}$ for the union of all $\mathcal{PA}(R^n, R^m)$ as m and n vary over all natural numbers.

For a piecewise affine function $f : R^n \to R^m$ and $q \in R^m$, $f - q$ denotes the function $f(x) - q$. The zero set of f is denoted by $\mathcal{Z}(f)$. Clearly, $\mathcal{Z}(f)$ is a finite union of polyhedral sets and hence closed. If in addition f is positively homogeneous, then $\mathcal{Z}(f)$ is also a cone.

We shall say that f, as given by (1), is *coherently oriented* if $m = n$ and the determinants of matrices $A_1, A_2, \ldots, A_L$ have the same nonzero sign.

The following result is well known in the literature; see, e.g., Proposition 2.2.7 and Theorem 2.3.1 in [27].

THEOREM 2.1. *Let $f : R^n \to R^m$ be piecewise affine. Then*

(a) *f is Lipschitz continuous.*

(b) *When $m = n$, f is coherently oriented if it is one-to-one.*

(c) *When $m = n$, f is a homeomorphism if and only if it is one-to-one.*

Every polyhedral set Ω (in a finite-dimensional space) admits a decomposition [26]

$$\Omega = \mathcal{C}(\Omega) + 0^+\Omega,$$

where $\mathcal{C}(\Omega)$ is compact and polyhedral (actually, the convex hull of the extreme points of Ω) and $0^+\Omega$ is the recession cone of Ω. (Recall that for a nonempty set E, a vector r belongs to 0^+E if for some $u \in E$, $u + \lambda r \in E$ for all $\lambda \geq 0$.)

For vectors x and y, $\langle x, y\rangle$ denotes the usual inner product and $||x||$ denotes the Euclidean norm of the vector x in the space (e.g., R^n) under consideration. When the inner product between vectors x and y is zero, we write $x \perp y$. Throughout this paper, B denotes the closed unit ball. For a nonempty set E, the dual cone is defined by

$$E^* := \{y : \langle y, x\rangle \geq 0 \quad \text{for all} \quad x \in E\}.$$

For two nonempty sets X and Y and a vector z, we define

$$d(z, Y) := \inf_{y \in Y} ||z - y|| \quad \text{and} \quad e(X, Y) := \sup_{x \in X} d(x, Y).$$

The Hausdorff distance between sets X and Y is defined as

$$\mathcal{H}(X, Y) := \max\{e(X, Y), e(Y, X)\}.$$

Note that some of these quantities may take the value ∞.

For any two vectors x and y,

$$x \wedge y := \min\{x, y\},$$

that is, $x \wedge y$ denotes the componentwise minimum of x and y. For a polyhedral set $\mathcal{K} \subseteq R^n$, $\Pi_{\mathcal{K}}(x)$ denotes the orthogonal projection of x onto $\mathcal{K}$. Note that when $\mathcal{K}$ is the nonnegative orthant, $\Pi_{R^n_+}(x) = \max\{x, 0\} = x^+$ and $\Pi_{R^n_+}(x) - x = \max\{-x, 0\} = x^-$.

3. The recession function. For a piecewise affine function $f : R^n \to R^m$, the *recession function* of f is defined by (2). The recession function is well defined, since for each x and large λ (depending on x) λx belongs to one polyhedral set on which f is affine. Note that $f^\infty = f$ when f is positively homogeneous.

Before we begin our formal study of properties, we present some examples from complementarity theory and affine variational inequalities. In each example, we specify both the function whose zeros solve the problem and its recession function. In each case, the specified functions are piecewise affine. (In Examples 5 and 6, the projection mapping onto a polyhedral set is piecewise affine; see [25].) While the computation of the recession function in Examples 1–4 and 7–8 is straightforward, for Examples 5 and 6 we use the formula $\lim_{\lambda \to \infty} \frac{1}{\lambda}\Pi_{\mathcal{K}}(\lambda x) = \Pi_{0^+\mathcal{K}}(x)$ (which follows from the definition of projection and the decomposition of the polyhedral set $\mathcal{K}$ into its compact part and the recession cone; see [27], p. 64).

Example 1. *The LCP* [2]. For $M \in R^{n \times n}$ and $q \in R^n$, the problem LCP(M, q) is to find an x such that

$$x \geq 0, \quad Mx + q \geq 0, \quad \text{and} \quad \langle x, Mx + q\rangle = 0.$$

For this problem,

$$f(x) = x \wedge (Mx + q) \quad \text{and} \quad f^{\infty}(x) = x \wedge Mx.$$

Equivalently, we may consider

$$g(x) = Mx^{+} - x^{-} + q \quad \text{and} \quad g^{\infty}(x) = Mx^{+} - x^{-}.$$

If a vector u solves $g(x) = 0$ then u^{+} solves LCP(M, q); conversely, if x^{*} is a solution of LCP(M, q), then $z = x^{*} - Mx^{*} - q$ solves $g(x) = 0$.

Example 2. *The horizontal linear complementarity problem (HLCP)* [31], [32]. For matrices A and B in $R^{m \times n}$, and $q \in R^{m}$, the problem is to find vectors x and y such that

$$x \geq 0,\ y \geq 0, \quad \langle x, y \rangle = 0,$$
$$Ax - By = q.$$

For this problem we have, with $z = (x, y)$,

$$f(z) = \begin{bmatrix} x \wedge y \\ Ax - By - q \end{bmatrix} \quad \text{and} \quad f^{\infty}(z) = \begin{bmatrix} x \wedge y \\ Ax - By \end{bmatrix}.$$

Example 3. *The vertical linear complementarity problem (VLCP)* [8]. For matrices $M_1, M_2, \ldots, M_k$ in $R^{m \times n}$ and vectors $q_1, q_2, \ldots, q_k$ in R^{m}, the problem is to solve the equation $f(x) = 0$, where

$$f(x) = (M_1 x + q_1) \wedge (M_2 x + q_2) \wedge \cdots \wedge (M_k x + q_k),$$
$$f^{\infty}(x) = M_1 x \wedge M_2 x \wedge \cdots \wedge M_k x.$$

Example 4. *The mixed linear complementarity problem (MLCP)* [7]. For matrices $A \in R^{m \times n}$, $B \in R^{m \times k}$, $C \in R^{k \times n}$, $D \in R^{k \times k}$ and vectors $a \in R^{m}$ and $b \in R^{k}$ the problem is to find $z = (x, y)$ such that

$$Ax + By + a = 0,$$
$$y \geq 0, \quad Cx + Dy + b \geq 0, \quad \langle y, Cx + Dy + b \rangle = 0.$$

For this problem, we have

$$f(z) = \begin{bmatrix} Ax + By + a \\ y \wedge (Cx + Dy + b) \end{bmatrix} \quad \text{and} \quad f^{\infty}(z) = \begin{bmatrix} Ax + By \\ y \wedge (Cx + Dy) \end{bmatrix}.$$

Example 5. *The extended linear complementarity problem (XLCP)* [5], [16]. Given matrices M and N in $R^{m \times n}$, a vector $q \in R^{m}$, and a polyhedral set $\mathcal{K}$ find $z = (x, y)$ such that

$$x \geq 0,\ y \geq 0, \quad \langle x, y \rangle = 0,$$
$$Mx - Ny \in \mathcal{K} + q.$$

Note that this problem already includes the LCP, HLCP, and the MLCP. The functions for this problem are given by

$$f(z) = \begin{bmatrix} x \wedge y \\ Mx - Ny - \Pi_{\mathcal{K}+q}(Mx - Ny) \end{bmatrix} \quad \text{and}$$
$$f^{\infty}(z) = \begin{bmatrix} x \wedge y \\ Mx - Ny - \Pi_{0^{+}\mathcal{K}}(Mx - Ny) \end{bmatrix}.$$

Example 6. *The AVI* [3], [7], [15], [25], *etc.* For a matrix $M \in R^{n\times n}$, a vector $q \in R^n$, and a polyhedral set $\mathcal{K} \subseteq R^n$, AVI$(M, \mathcal{K}, q)$ is to find an $x^* \in \mathcal{K}$ such that

$$\langle Mx^* + q, x - x^* \rangle \geq 0 \quad \text{for all} \quad x \in \mathcal{K}.$$

For this problem,

$$f(x) = x - \Pi_{\mathcal{K}}(x - Mx - q) \quad \text{and} \quad f^\infty(x) = x - \Pi_{0^+\mathcal{K}}(x - Mx).$$

Alternatively, we can consider the equation $g(x) = 0$, where Robinson's normal map g and its recession function are given by

$$g(x) = M(\Pi_{\mathcal{K}}(x)) + x - \Pi_{\mathcal{K}}(x) + q \quad \text{and} \quad g^\infty(x) = M(\Pi_{0^+\mathcal{K}}(x)) + x - \Pi_{0^+\mathcal{K}}(x).$$

Note that if u solves $g(x) = 0$, then $\Pi_{\mathcal{K}}(u)$ solves AVI$(M, \mathcal{K}, q)$; conversely, if x^* solves AVI$(M, \mathcal{K}, q)$, then $z = x^* - Mx^* - q$ solves $g(x) = 0$.

Example 7. *The zero-one integer feasibility problem.* For a polyhedral set $\mathcal{F} = \{x : Ax \leq a, Bx = b\} \subseteq R^n$ with $\mathbf{1}$ denoting the vector of ones,

$$f(x) = \begin{bmatrix} x \wedge (\mathbf{1} - x) \\ (Ax - a)_+ \\ Bx - b \end{bmatrix} \quad \text{and} \quad f^\infty(x) = \begin{bmatrix} x \wedge (-x) \\ (Ax)_+ \\ Bx \end{bmatrix}.$$

Example 8. *The affine complementarity system.* This is a system defined by a finite number of linear equalities, inequalities, and complementarity conditions. Here "complementarity condition" refers to a condition where the minimum of a finite number of variables is set to zero. Examples of such systems include all of the previous ones and more. Clearly, finding a solution of this system is equivalent to finding a zero of a piecewise affine function. As an illustration, consider the problem of finding vectors x, y, z, and u such that

$$\begin{gathered} x \wedge y \wedge z = 0, \\ (d - x) \wedge u = 0, \\ Ax + By + Cu + p \leq 0, \\ Dx + My + Nz + q = 0. \end{gathered}$$

Here (d, p, q) is a triplet of vectors and the capital letters denote matrices. For this system,

$$f(w) = \begin{bmatrix} x \wedge y \wedge z \\ (d - x) \wedge u \\ (Ax + By + Cu + p)_+ \\ Dx + My + Nz + q \end{bmatrix} \quad \text{and} \quad f^\infty(w) = \begin{bmatrix} x \wedge y \wedge z \\ (-x) \wedge u \\ (Ax + By + Cu)_+ \\ Dx + My + Nz \end{bmatrix}.$$

In the next result we record some known properties of the recession function (cf. Propositions A.2.1, A.2.2 and Theorem A.2.1 in [27] and Proposition 3.1 in [28]).

THEOREM 3.1. *Let f be a piecewise affine function from R^n into R^m. Then*

(i) *f^∞ is piecewise affine and positively homogeneous.*

(ii) *For any piecewise affine function $g : R^k \to R^n$,*

$$(f \circ g)^\infty(x) = (f^\infty \circ g^\infty)(x).$$

(iii) *For any* $x \in R^n$ *and* $x^0 \in R^n$,

$$f^\infty(x) = \lim_{\lambda\to\infty} \frac{f(\lambda x + x^0)}{\lambda}. \tag{3}$$

(iv) *If* $m = n$ *and* f *is coherently oriented, then* f *is a homeomorphism if and only if* f^∞ *is a homeomorphism.*

Remarks. (a) Item (iii) in the previous theorem can be easily deduced from (ii) by taking $g(x) = x + x^0$. Applying (3) we see that if $r \in 0^+\Omega$ and $f(x) = Ax + a$ on Ω, then $f^\infty(r) = Ar$. Since $R^n = \cup_{j=1}^L \Omega_j$, we have $R^n = \cup_{j=1}^L 0^+\Omega_j$; hence the matrices defining f^∞ come from the set of matrices of f corresponding to the unbounded Ω_js. It should be noted that not every matrix in this set appears in the description of f^∞. (An example, due to a referee, is $f(x,y) := (\min\{4x+2, 2x, x+1\}, y)$ on R^2. This function has three matrices corresponding to unbounded polyhedral sets, and only two of these are needed to describe the recession function $f^\infty(x,y) = (\min\{4x, x\}, y)$.)

(b) In the process of establishing more properties of the recession function, we shall provide a different proof of (iv); see Theorems 3.3 and 4.5.

Our next result shows that $f - f^\infty$ is a bounded (piecewise affine) function and that $f^\infty(x)$ can be computed by sequences not necessarily lying on a ray.

PROPOSITION 3.2. *For a piecewise affine function* f *we have*

(i) $\sup_{x\in R^n} ||f(x) - f^\infty(x)|| < \infty$ *and*

(ii) $[\mu_k \downarrow 0,\ \mu_k x^k \to x^*] \Longrightarrow [\mu_k f(x^k) \to f^\infty(x^*)]$.

Proof. Consider the polyhedral sets Ω_j (as in §2), with f defined by (1). Let $\mathcal{C} := \cup \mathcal{C}(\Omega_j)$ and fix $x \neq 0$. Then x is a recession direction for some Ω_i; thus $e_i + \lambda x \in \Omega_i$ for all $\lambda \geq 0$ and $e_i \in \mathcal{C}(\Omega_i)$. Fixing e_i, we have $f^\infty(x) = \lim_{\lambda\to\infty} (f(e_i + \lambda x))/\lambda = A_i x$ and

$$||f(x) - f^\infty(x)|| = ||f(x) - f(e_i + x) + f(e_i + x) - f^\infty(x)|| \tag{4}$$

$$\leq ||f(x) - f(e_i + x)|| + ||A_i x + A_i e_i + a_i - A_i x|| \tag{5}$$

$$\leq \theta||e_i|| + ||f(e_i)||, \tag{6}$$

where θ is the Lipschitzian constant of f. With $\Delta := \sup_{e\in\mathcal{C}} \theta||e|| + ||f(e)||$, which is clearly finite, we have $\sup_{x\in R^n} ||f(x) - f^\infty(x)|| \leq \Delta$. We thus have (i). The inequality

$$||\mu_k f(x^k) - f^\infty(x^*)|| \leq \mu_k ||f(x^k) - f^\infty(x^k)|| + ||f^\infty(\mu_k x^k) - f^\infty(x^*)||$$

(which holds because of the positive homogeneity of f^∞) and (i) prove (ii). □

In many problems involving piecewise affine functions, the question of knowing whether the given function is one-to-one/onto arises. The following result gives necessary conditions in terms of the recession function.

THEOREM 3.3. *The following statements hold.*

(a) *Suppose* $f \in \mathcal{PA}(R^n, R^n)$. *If* f *is one-to-one, then so is* f^∞ *and*

$$(f^\infty)^{-1} = (f^{-1})^\infty. \tag{7}$$

(b) *Suppose* $g \in \mathcal{PA}(R^n, R^m)$. *If* g *is onto, then so is* g^∞.

Proof. (a) If f is one-to-one, then (by Theorem 2.1) it is a homeomorphism. Since f^{-1} (the inverse of f) is piecewise affine, the chain rule $(f \circ f^{-1})^\infty = f^\infty \circ (f^{-1})^\infty$ proves the one-to-oneness of f^∞ and (7). To see (b), suppose that g is onto. Let $q \in R^m$ be arbitrary. Then for each natural number k, there exists $x^k \in R^n$ such that $g(x^k) = kq$. Without loss of generality, we may assume that $\{x^k\}$ belongs to a

polyhedral set Ω on which g is affine, say, $g(x) = Ax + a$. Writing $x^k = e^k + kr^k$, where $e^k \in \mathcal{C}(\Omega)$ and $r^k \in 0^+\Omega$, we have $Ae^k + kAr^k + a = kq$ for all k. This results in $q \in A(0^+\Omega)$ (since the latter set is polyhedral and hence closed). Thus $q = Ar$ for some $r \in 0^+\Omega$. Since $g^\infty(r) = Ar$ (see the previous remarks), we see that g^∞ is onto. □

Remarks. We note that (a) reproves a part of Theorem 3.1 (iv); the other part will be covered in Theorem 4.5. Easy examples on the real line can be constructed to show that f need not be one-to-one when f^∞ is one-to-one. The converse implication in part (b) is false. An example due to Sznajder [30] shows that in the LCP setting, one can have a matrix M for which $g^\infty(x) = x \wedge Mx$ is onto but $g(x) := x \wedge (Mx+q)$ does not take the value zero for some q. In Theorem 4.2, we specify sufficient conditions for the converse to hold.

To continue our analysis, we need the following definition.

DEFINITION. *Let f and g be two piecewise affine functions. We say that f and g are* equivalent *and write $f \sim g$ if $f^\infty = g^\infty$. For example, when f is piecewise affine and p and q are vectors, $f - p \sim f - q$.*

PROPOSITION 3.4. *Let f and g belong to $\mathcal{PA}(R^n, R^m)$. Then $f \sim g$ if and only if*

$$\sup_{x \in R^n} ||f(x) - g(x)|| < \infty.$$

Proof. Suppose $f \sim g$. Using triangle inequality we see that

$$\sup ||f(x) - g(x)|| \leq \sup ||f(x) - f^\infty(x)|| + \sup ||g(x) - g^\infty(x)||,$$

where the supremum is taken over all of R^n. Since the two quantities on the right side of the above inequality are finite (cf. Proposition 3.2), the left side is also finite. Conversely, suppose that $||f(x) - g(x)|| \leq \Delta < \infty$ for all $x \in R^n$. If we replace x by λx, divide the inequality by λ, and let λ go to ∞, we get $f^\infty(x) = g^\infty(x)$. □

Notation. For two functions f and g in $\mathcal{PA}(R^n, R^m)$ that are equivalent, we write

$$||f - g|| := \sup_{x \in R^n} ||f(x) - g(x)||.$$

As an illustration, let $f(x) = x \wedge (Mx + q)$ and $g(x) = x \wedge (Mx + p)$. Then the inequality $|\lambda \wedge \mu - \lambda \wedge \nu| \leq |\mu - \nu|$ (for real numbers λ, μ, ν) implies $||f - g|| \leq ||q - p||$. To give another example, let f be piecewise affine, so that for $h = f - q$ and $g = f - p$ we have $||g - h|| = ||p - q||$.

4. Zero sets. We have already noted that the zero set of a piecewise affine function is a finite union of polyhedral sets. The following result specifies a sufficient condition for the zero set of a piecewise affine function to be bounded.

PROPOSITION 4.1. *Suppose that $f \in \mathcal{PA}(R^n, R^m)$ is piecewise affine. Then the following are equivalent.*

(a) $\mathcal{Z}(f^\infty) = \{0\}$.

(b) *For all $g \sim f$, $\mathcal{Z}(g)$ is bounded.*

(c) *For all $q \in R^m$, the set $\mathcal{Z}(f - q)$ $(= \{x : f(x) = q\})$ is bounded.*

(Note that in (b) and (c) the sets may be empty.)

Proof. Suppose that $g \sim f$ and $\mathcal{Z}(g)$ is unbounded. Then there exists a sequence $\{x^k\}$ such that $g(x^k) = 0$ and $||x^k|| \to \infty$. Without loss of generality, we may assume that the sequence $\mu_k x^k$ (of normalized vectors) converges to, say, x^*, where $\mu_k := ||x^k||^{-1}$. By Proposition 3.2, $g^\infty(x^*) = 0$. Since $f^\infty = g^\infty$ and $||x^*|| = 1$, (a)

cannot hold. Thus we have (a) $\Longrightarrow$ (b). Since for any q, $f-q \sim f$ we have (b) $\Longrightarrow$ (c). Suppose that (a) is not true, so that $f^\infty(u) = 0$ for some nonzero u. Then u is a recession direction of some polyhedral set on which f is given by the affine function $Ax + a$. It follows that $0 = f^\infty(u) = Au$. For any u^0 in this polyhedral set and for all k, $f(u^0 + ku) = f(u^0)$. Thus with $q = f(u^0)$ condition (c) fails. This proves the implication (c) $\Longrightarrow$ (a). $\square$

Coming to the existence of zeros, we present the following result, whose hypothesis and proof technique have become standard in the study of existence and stability aspects of nonsmooth equations; see [5]–[8], [20], [21]. For notions such as degree and index we refer the reader to [12] or [19].

THEOREM 4.2. *Let $f \in \mathcal{PA}(R^n, R^n)$. Suppose that*

(i) $\mathcal{Z}(f^\infty) = \{0\}$, *and*

(ii) *index* $(f^\infty, 0) \neq 0$.

Then for all $g \sim f$, $\mathcal{Z}(g)$ is nonempty and bounded. Moreover, every such g is onto.

Proof. Consider any $g \sim f$. The previous proposition proves the boundedness of $\mathcal{Z}(g)$. For $x \in R^n$ and $t \in [0,1]$, define $h(x,t) := tf^\infty(x) + (1-t)g(x)$. Then for each t, h is piecewise affine in x. A standard argument using normalized vectors (like the one in the previous proposition) along with condition (i) shows that the zeros of $h(\cdot, t)$ lie in some bounded (open) set $\mathcal{D}$. On this set, h is a homotopy joining f^∞ and g. Since no zero of h can lie on the boundary of $\mathcal{D}$ by the homotopy invariance property of degree, the degree of g at 0 relative to this open set is the same as the index of f^∞ at zero. Therefore by condition (ii), this degree of g is nonzero, which means that g will have a zero. Thus the nonemptiness of $\mathcal{Z}(g)$ is established. Applying this to $g-q$ with $q \in R^n$ arbitrary, we prove the ontoness of g. $\square$

Remarks. Condition (i) in the previous theorem reduces to the so-called $\mathbf{R}_0$-condition in the linear, horizontal, vertical, mixed, and extended linear complementarity problems. In the presence of (i), condition (ii) can be replaced by the following equivalent condition.

(ii)$'$ For some $g^* \sim f$ and a bounded open set $\mathcal{D} \supseteq \mathcal{Z}(g^*)$, deg $(g^*, \mathcal{D}, 0) \neq 0$.

(The proof of the previous theorem shows that (ii) implies (ii)$'$. The reverse implication can be seen by considering the homotopy $h(x,t) := (1-t)g^*(x) + tf^\infty(x)$ on a suitable bounded open set and using the excision property of the degree; see Theorem 2.2.1 in [12].) We also remark that the nearness property of the degree [12, Thm. 2.1.2] allows us to state the following stability principle: Let the conditions of the theorem hold. Let $g \sim f$, $\mathcal{Z}(g) \subseteq \mathcal{O}$, where $\mathcal{O}$ is a bounded open set. Then for all continuous functions h (from $\overline{\mathcal{O}} \to R^n$) that are close to g on $\overline{\mathcal{O}}$ we have $\mathcal{Z}(h) \cap \mathcal{O} \neq \emptyset$. For some recent results on the (local) stability of a nonsmooth function at a zero see [21]; Theorem 1 in this reference, specialized to f^∞, gives the ontoness of f^∞.

In the following corollary, the hypothesis is akin to the **R**-condition of the LCP.

COROLLARY 4.3. *Let $f \in \mathcal{PA}(R^n, R^n)$. Suppose that*

(i) $\mathcal{Z}(f^\infty) = \{0\}$ *and*

(ii) *there is a $g^* \sim f$ such that $\mathcal{Z}(g^*) = \{x^*\}$ and in a neighborhood of x^*, g^* is affine.*

Then the conclusions of the previous theorem hold.

Proof. We verify condition (ii)$'$ described above. Let $\mathcal{D}$ be the (open) neighborhood of x^* on which g^* is given by the affine function $Ax+a$. Since x^* is the only zero of g^*, A is nonsingular. By definition, deg $(g^*, \mathcal{D}, 0) =$ sgn det $A \neq 0$. This completes the proof. $\square$

As an illustration, suppose f, described by (1), has all A_js nonsingular. If for

some $e \in \operatorname{int} \Omega_j$ (for some j), $f(x) = f(e) \Longrightarrow x = e$, then taking $g^* := f - f(e)$ and $x^* := e$, we see that f is onto. In the previous corollary, the condition that g^* *is affine in a neighborhood of* x^* *defines the notion of "nondegeneracy,"* studied in [33]. The conclusion of the previous corollary remains the same if $\mathcal{Z}(g^*)$, instead of being a singleton, consists of an odd number of nondegenerate points. Many existence results in complementarity theory and affine variational inequalities, such as Theorem 1 in [8], Theorem 1 in [7], Proposition 3 in [6], and Theorem 5.2.4 in [31], follow from the previous two results.

It is well known [23], [27] that a coherently oriented piecewise affine function is onto. The same conclusion is obtained if the coherency condition is imposed on the matrices corresponding to the unbounded polyhedral sets Ω_js; see Chien and Kuh [1] and Ohtsuki, Fujisawa, and Kumagai [18]. The following result obtains the same conclusion with further weakening of the hypothesis.

COROLLARY 4.4. *Suppose that for* $f \in \mathcal{PA}(R^n, R^n)$, f^∞ *is coherently oriented. Then conditions* (i) *and* (ii) *of Theorem* 4.2, *and hence its conclusions hold. In particular,* f *is onto.*

Proof. Since the matrices involved in the function f^∞ are nonsingular, $\mathcal{Z}(f^\infty) = \{0\}$. Pick a vector q in the range of f^∞ that does not come from any of the boundaries of the polyhedral sets defining f^∞. (This can be done because the matrices of f^∞ are nonsingular, and the images of the polyhedral sets defining f^∞ are n-dimensional while the images of the boundaries are at most $(n-1)$-dimensional.) Let $g^* := f^\infty - q$. Then $\mathcal{Z}(g^*)$ is finite and every zero of g^* is nondegenerate. As in the proof of the previous theorem, we see that for any bounded open set $\mathcal{D}$ containing $\mathcal{Z}(g^*)$, index $(f^\infty, 0)$ is the same as deg $(g^*, \mathcal{D}, 0)$; the latter number is the sum of the indexes of g^* at the zeros of g^*. Following the proof of Corollary 4.3, we see that these indexes have the same nonzero sign. Thus the hypothesis and conclusions of Theorem 4.2 hold. □

Remarks. Apart from the onto property, coherently oriented piecewise affine functions have other interesting properties. It is well known (see, for example, Theorem 2.3.1 in [27]) that a piecewise affine function is coherently oriented if and only if it is an open map. Schramm [29] has shown that when f is coherently oriented, the cardinality of $\mathcal{Z}(f - q)$ is the same as q varies over the complement of an exceptional set (of measure zero). Another property (that we shall use in our next result) is the following. Let f be coherently oriented and let q^* be arbitrary; let $x^* \in \mathcal{Z}(f - q^*)$. Then index $(f - q^*, x^*)$ is nonzero. This is known [9, Thm. 3.3] and can be seen as follows. x^* is an isolated zero of f as the matrices corresponding to f are nonsingular. If U is an open neighborhood of x^* not containing other zeros of $f - q^*$, then $f(U)$ contains q^* and is open. We can pick a vector p in $f(U)$ sufficiently close to q^* so that each element in $\mathcal{Z}(f - p)$ is nondegenerate; see the argument in the proof of Corollary 4.4. By the nearness property of the degree [12, Thm. 2.1.2], index $(f - q^*, x^*)$ is equal to deg $(f - p, U, 0)$, which in turn is equal to the sum of the indexes of $f - p$ at each of its (nondegenerate) zeros in U. Since f is coherently oriented, all these indexes have the same nonzero sign. Thus index $(f - q^*, x^*)$ is nonzero and its sign is the same as the sign of the determinant of any matrix defining f.

The following result (whose proof is based on degree theory) recovers Theorem 3.1 (iv).

THEOREM 4.5. *Suppose that* $f \in \mathcal{PA}(R^n, R^n)$. *Then the following are equivalent.*

(a) f *is one-to-one.*

(b) f *is coherently oriented and* f^∞ *is one-to-one.*

(c) *f is coherently oriented and index $(f^\infty, 0) = \pm 1$.*

Proof. Suppose (a) holds. It is well known that f is coherently oriented [27, Thm. 2.3.1]. That f^∞ is one-to-one follows from Theorem 3.3. Thus (a) $\Longrightarrow$ (b). Since the index of a one-to-one function about any point is ± 1, we have (b) $\Longrightarrow$ (c). Now suppose (c). (We remark that under the assumption that f is coherently oriented, f^∞ is coherently oriented and hence $\mathcal{Z}(f^\infty) = \{0\}$. So the index of f^∞ at zero is defined.) By the previous remarks (applied to f^∞), the index of f^∞ at the origin is nonzero and its sign is the same as the sign of the determinant of any matrix describing f^∞ (also of f). The previous corollary shows that f is onto. For any $q \in R^n$, let $\mathcal{D}$ denote an open set containing the finite set $\mathcal{Z}(f - q)$. As in the proof of Theorem 4.2 we consider a homotopy joining $f - q$ and f^∞ and conclude that deg $(f - q, \mathcal{D}, 0) =$ index $(f^\infty, 0) = \pm 1$. Now the degree of $f - q$ at zero over $\mathcal{D}$ is the sum of the indexes of $f - q$ at each of its zeros. Since these indexes (which are nonzero by the previous remarks) and index of f^∞ at zero have the same sign, we conclude that there can be only one element in $\mathcal{Z}(f - q)$. This completes the proof of the theorem. □

We remark that the implication (c) $\Longrightarrow$ (a) in the previous theorem improves a result of Kojima and Saigal [9] which says that f is one-to-one when f is coherently oriented and for some matrix B, the matrices $tA_j + (1 - t)B$ (with $t \in [0, 1]$ and A_js corresponding to unbounded polyhedral sets defining f) are all nonsingular. This is because, under this nonsingularity condition, $h(x, t) := (1 - t)f^\infty(x) + tBx$ defines a homotopy between f^∞ and the mapping $x \mapsto Bx$ on some bounded open set containing the origin so that the index condition in (c) holds.

Although the previous theorem specifies a necessary and sufficient condition one-to-oneness, in practice it is not easy to verify these conditions. Note that these conditions are imposed on the matrices defining f. For characterizations based on conditions on the polyhedral sets Ω_j, the interested reader may consult [10], [11], [28], and [29]. As far as the normal map g of Example 6 is concerned, Robinson [25] has shown that g is one-to-one whenever it is coherently oriented. It would be interesting to see if this result could be proved by verifying the index condition in part (c) of the previous theorem using, say, homotopy arguments.

5. Global error bounds. Our starting point is the following result due to Robinson [24].

THEOREM 5.1. *Suppose that f is piecewise affine. Then there exist positive numbers α and ρ such that*

$$d(x, \mathcal{Z}(f)) \leq \alpha ||f(x)|| \quad \text{whenever} \quad ||f(x)|| \leq \rho. \tag{8}$$

An application of this result to the recession function of f along with the observation that f^∞ is positively homogeneous proves Corollary 5.2.

COROLLARY 5.2. *Suppose that f is piecewise affine. Then there exists a positive number α^∞ such that*

$$d(x, \mathcal{Z}(f^\infty)) \leq \alpha^\infty ||f^\infty(x)|| \quad \text{for all} \quad x. \tag{9}$$

In preparation for our main result in this section, we prove the following elementary proposition.

PROPOSITION 5.3. *Let X and Y be two nonempty sets in R^n that are unions of a finite number of polyhedral sets. Then $e(X, Y) < \infty$ if and only if $0^+X \subseteq 0^+Y$. Hence $\mathcal{H}(X, Y) < \infty$ if and only if $0^+X = 0^+Y$.*

Proof. Suppose that $e(X, Y) < \infty$ and let w be a recession direction for X. Then for some $x^0 \in X$ and all natural numbers k, $x^k := x^0 + kw \in X$. Corresponding to x^k,

there exists a $y^k \in Y$ such that $d(x^k, y^k) \leq e(X,Y)$. Without loss of generality, we can assume that all y^ks belong to one polyhedral set, say, Y_1 contained in Y. Now writing $y^k = e^k + kr^k$, where $e^k \in \mathcal{C}(Y_1)$ and $r^k \in 0^+Y_1$, we see that $||x^0 + kw - (e^k + kr^k)|| \leq e(X,Y)$. Dividing this inequality by k and letting $k \to \infty$, we get $w = \lim r^k$, proving $w \in 0^+Y_1 \subseteq 0^+Y$. This argument shows that $0^+X \subseteq 0^+Y$. Conversely, suppose that $0^+X \subseteq 0^+Y$; let $x^* \in X_1 \subseteq X$ with X_1 polyhedral. We write $x^* = e^* + r^*$ with $e^* \in \mathcal{C}(X_1)$ and $r^* \in 0^+X_1$. Then $d(x^*, Y) \leq d(x^*, y^*)$, where $y^* = d^* + r^*$ belongs to Y and $d^* \in \mathcal{C}(Y)$. Since $d(x^*, y^*) = d(e^*, d^*) \leq \gamma := \sup\{d(u,v) : u \in \mathcal{C}(X), v \in \mathcal{C}(Y)\} < \infty$, we see that $e(X,Y) \leq \gamma < \infty$. The statement involving the Hausdorff distance is immediate. □

THEOREM 5.4. *Suppose that f is piecewise affine from R^n into R^m with $\mathcal{Z}(f) \neq \emptyset$. Then the following are equivalent.*

(a) *There exists $\beta > 0$ such that*

$$d(x, \mathcal{Z}(f)) \leq \beta ||f(x)|| \quad \text{for all} \quad x \in R^n. \tag{10}$$

(b) *There exists $\beta > 0$ such that*

$$\mathcal{Z}(g) \subseteq \mathcal{Z}(f) + \beta ||f - g|| B \quad \text{for all} \quad g \sim f. \tag{11}$$

(c) *There exists $\beta > 0$ such that*

$$\mathcal{Z}(f^\infty) \subseteq \mathcal{Z}(f) + \beta ||f - f^\infty|| B. \tag{12}$$

(d) $\mathcal{H}(\mathcal{Z}(f^\infty), \mathcal{Z}(f)) < \infty$.

(e) $0^+\mathcal{Z}(f) = \mathcal{Z}(f^\infty)$.

Proof. (a) $\Longrightarrow$ (b): Assume (a) and take any $g \sim f$. If $x \in \mathcal{Z}(g)$, then $d(x, \mathcal{Z}(f)) \leq \beta ||f(x) - g(x)|| \leq \beta ||f - g||$. The inclusion in (b) is immediate. Since $f^\infty \sim f$, (b) $\Longrightarrow$ (c). Assuming (c), we get $e(\mathcal{Z}(f^\infty), \mathcal{Z}(f)) \leq \beta ||f - f^\infty|| < \infty$. From Corollary 5.2 and the implication (a) $\Longrightarrow$ (b) above, we deduce that $e(\mathcal{Z}(f), \mathcal{Z}(f^\infty)) < \infty$. Thus (d) follows. The previous proposition shows that (d) and (e) are equivalent. We now show that (d) $\Longrightarrow$ (a). Assume that (d) holds, and let α and ρ be as in Theorem 5.1. We show that (a) holds with

$$\beta = \max\left\{\alpha, \frac{e(\mathcal{Z}(f^\infty), \mathcal{Z}(f)) + \alpha^\infty(\rho + ||f - f^\infty||)}{\rho}\right\}.$$

Fix an $x \in R^n$. In view of (8), we can assume that $||f(x)|| \geq \rho$. The triangle inequality gives

$$\frac{d(x, \mathcal{Z}(f))}{||f(x)||} \leq \frac{d(x, \mathcal{Z}(f^\infty))}{||f(x)||} + \frac{e(\mathcal{Z}(f^\infty), \mathcal{Z}(f))}{||f(x)||}.$$

The second term on the right side of the above inequality is less than or equal to $\frac{e(\mathcal{Z}(f^\infty), \mathcal{Z}(f))}{\rho}$. If $f^\infty(x) = 0$, i.e., $d(x, \mathcal{Z}(f^\infty)) = 0$, then the left-hand side of the above inequality is bounded by β. When $f^\infty(x) \neq 0$, we have

$$\frac{d(x, \mathcal{Z}(f^\infty))}{||f(x)||} = \frac{d(x, \mathcal{Z}(f^\infty))}{||f^\infty(x)||} \frac{||f^\infty(x)||}{||f(x)||} \tag{13}$$

$$\leq \alpha^\infty \left(1 + \frac{||f - f^\infty||}{\rho}\right). \tag{14}$$

It follows that in all cases, $d(x, \mathcal{Z}(f)) \leq \beta||f(x)||$. This completes the proof of the theorem. □

DEFINITION. *A piecewise affine function f is said to have the* global error bound property *(GEBP, for short) if condition* (a) *in the previous theorem holds.*

To obtain an easy consequence of Theorem 5.4, assume that $\mathcal{Z}(f^\infty) = \{0\}$. Then by Proposition 4.1, for any $g \sim f$, $\mathcal{Z}(g)$ is bounded. When $\mathcal{Z}(g) \neq \emptyset$, the implication (e) $\Longrightarrow$ (a) of the previous theorem proves the following.

COROLLARY 5.5. *Suppose that f is piecewise affine and $\mathcal{Z}(f^\infty) = \{0\}$. For any $g \sim f$ with $\mathcal{Z}(g) \neq \emptyset$, there exists a positive constant $\beta(g)$ such that*

$$d(x, \mathcal{Z}(g)) \leq \beta(g)||g(x)|| \quad \text{for all} \quad x.$$

6. Global error bounds for the AVI, LCP, and LP. In this section we apply our previous error bound analysis to the AVIs, LCPs, and LP. Recall (Example 6) that AVI$(M, \mathcal{K}, q)$ is to find a solution of the equation $f(x) = 0$, where

$$f(x) = x - \Pi_{\mathcal{K}}(x - Mx - q). \tag{15}$$

Fixing M and $\mathcal{K}$, we wish to find all q such that the global error bound property (10) holds for $f(x)$. Let $\mathcal{S}(q) := \mathcal{Z}(f) \neq \emptyset$ and $\hat{\mathcal{S}} := \mathcal{Z}(f^\infty)$. By the description of f^∞ in Example 6, we see that $\hat{\mathcal{S}} = \{r : \langle Mr, s - r\rangle \geq 0 \quad \forall\ s \in 0^+\mathcal{K}\}$. Since $0^+\mathcal{K}$ is a cone, we have

$$\hat{\mathcal{S}} = \{r : r \in 0^+\mathcal{K}, Mr \in (0^+\mathcal{K})^*, \langle Mr, r\rangle = 0\}.$$

If $\hat{\mathcal{S}} = \{0\}$, then by Corollary 5.5, for every q, f will have the global error bound property.

So assume that $\hat{\mathcal{S}} \neq \{0\}$. We proceed to find all q satisfying the condition $\hat{\mathcal{S}} = 0^+\mathcal{S}(q)$. Since $0^+\mathcal{S}(q)$ is always a subset of $\hat{\mathcal{S}}$ (as a consequence of (3)), we assume that $\hat{\mathcal{S}} \subseteq 0^+\mathcal{S}(q)$. Fix $r \in \hat{\mathcal{S}}$. Then for some u in $\mathcal{S}(q)$ (depending on r) we have $u + \lambda r \in \mathcal{S}(q)$ for all $\lambda \geq 0$. This means that

$$\langle M(u + \lambda r) + q, x - (u + \lambda r)\rangle \geq 0 \quad \text{for all} \quad \lambda \geq 0,\ x \in \mathcal{K}.$$

With $\langle Mr, r\rangle = 0$ (recall $r \in \hat{\mathcal{S}}$) this becomes

$$\langle Mr, x - u\rangle - \langle Mu + q, r\rangle \geq 0 \quad \text{for all} \quad x \in \mathcal{K}.$$

Since $\langle Mu + q, x - u\rangle \geq 0$ for all $x \in \mathcal{K}$, the previous statement is equivalent to

$$\langle Mu + q, r\rangle = 0 \quad \text{and} \quad \langle Mr, x - u\rangle \geq 0 \quad \text{for all} \quad x \in \mathcal{K}. \tag{16}$$

(For full derivation, see Proposition 4 in [6].)

To summarize, f has the global error bound property for a q if and only if for each $r \in \hat{\mathcal{S}}$, there exists a $u \in \mathcal{S}(q)$ such that (16) holds. Note that when $\mathcal{S}(q)$ is convex, (16) should be satisfied for all $r \in \hat{\mathcal{S}}$ and for all $u \in \mathcal{S}(q)$.

For further analysis, we need some definitions. We shall say that a matrix M is *copositive* on a set if the quadratic form $\langle Mx, x\rangle$ is nonnegative on that set. Note that when M is copositive on the cone $0^+\mathcal{K}$, we have

$$r \in 0^+\mathcal{K},\ \langle Mr, r\rangle = 0 \Longrightarrow (M + M^T)r \in (0^+\mathcal{K})^*.$$

We shall say that M is *copositive-star* on $0^+\mathcal{K}$ [4] if M is copositive on $0^+\mathcal{K}$ and

$$r \in \hat{\mathcal{S}} \Longrightarrow -M^T r \in (0^+\mathcal{K})^*.$$

For example, M is copositive-star when it is either positive semidefinite or *copositive-plus* on $0^+\mathcal{K}$ (defined, in addition to the copositivity, by the condition $r \in \hat{\mathcal{S}} \Longrightarrow Mr + M^T r = 0$.) Following Iusem and Pang, we shall say that M is a *positive-semidefinite-plus* matrix if it is positive semidefinite and

$$\langle Mr, r\rangle = 0 \Longrightarrow Mr = 0$$

or, equivalently, if $M = E^T Q E$, where Q is a positive-definite matrix and E is arbitrary; see [15]. Note that for a positive-semidefinite matrix M, $\langle Mr, r\rangle = 0 \Longrightarrow (M + M^T)r = 0$.

THEOREM 6.1. *Suppose that one of the following holds.*

(a) *M is positive semidefinite-plus.*

(b) *$\mathcal{K}$ has an extreme point c such that M is copositive on $\mathcal{K} - c$ and*

$$r \in \hat{\mathcal{S}} \Longrightarrow Mr = M^T r = 0. \tag{17}$$

Then for a $q \in R^n$, $\mathcal{S}(q) \neq \emptyset$ and f (given by (15)) has the GEBP if and only if $q \perp \hat{\mathcal{S}}$.

Proof. Suppose that $\mathcal{S}(q) \neq \emptyset$ and the GEBP holds for f. For any $r \in \hat{\mathcal{S}}$, there exists an $u \in \mathcal{S}(q)$ such that $\langle Mu + q, r\rangle = 0$. Since $M^T r = 0$ under both conditions (a) and (b), we have $\langle q, r\rangle = 0$. Thus $q \perp \hat{\mathcal{S}}$. For the converse, suppose that $q \perp \hat{\mathcal{S}}$. Note that M is copositive-plus on $0^+\mathcal{K}$. (In the case of (b), the copositivity of M on $\mathcal{K} - c$ implies that of M on $0^+\mathcal{K}$.) From $q \perp \hat{\mathcal{S}}$ and (17), we have $Mu^* + q \in \hat{\mathcal{S}}^*$ for any $u^* \in \mathcal{K}$. This shows that $q \in \hat{\mathcal{S}}^* - M(\mathcal{K})$, i.e., AVI$(M, \mathcal{K}, q)$ is feasible; see [6]. Using Theorem 7 and Corollary 7 in [6], we see that $\mathcal{S}(q) \neq \emptyset$. For any $u \in \mathcal{S}(q)$ and any $r \in \hat{\mathcal{S}}$, we easily verify (16). Thus f has the GEBP. □

Our next result deals with the (generalized) LCP on a polyhedral cone. Note that in the case of the nonnegative orthant, the function f given by (15) reduces to

$$f(x) = x \wedge (Mx + q). \tag{18}$$

THEOREM 6.2. *Let $\mathcal{K}$ be a polyhedral cone.*

(a) *Suppose M is copositive-star on $\mathcal{K}$, $\mathcal{S}(q) \neq \emptyset$, and f has the GEBP. Then $q \perp \hat{\mathcal{S}}$.*

(b) *Suppose M is positive semidefinite. Then $\mathcal{S}(q) \neq \emptyset$ and f has the GEBP if and only if $q \perp \hat{\mathcal{S}}$ and $\mathcal{S}(q) \perp M(\hat{\mathcal{S}})$.*

Proof. (a) Fix $r \in \hat{\mathcal{S}}$. Then for some $u \in \mathcal{S}(q)$, we have (16). Since $\mathcal{K}$ is a cone, upon putting $x = 0$ and $x = 2u$ successively in (16), we get $\langle Mr, u\rangle = 0$. From the first equation in (16) and the fact that $-M^T r \in \mathcal{K}^*$, we have $\langle q, r\rangle \geq 0$. Adding the first equation in (16) and $\langle Mr, u\rangle = 0$ and noting that $(M + M^T)r \in \mathcal{K}^*$, we see that $\langle q, r\rangle \leq 0$. We thus have $q \perp \hat{\mathcal{S}}$.

(b) Assume $\mathcal{S}(q) \neq \emptyset$ and that f has the GEBP. The proof of the previous theorem shows that $q \perp \hat{\mathcal{S}}$. To show $\mathcal{S}(q) \perp M(\hat{\mathcal{S}})$, let $r \in \hat{\mathcal{S}}$. Then for some $u \in \mathcal{S}(q)$, (16) holds. Since $\mathcal{S}(q)$ is convex (because M is positive semidefinite) we can replace u by any v in this set. This leads to, as in the first part, $\langle Mr, v\rangle = 0$, proving $\mathcal{S}(q) \perp Mr$. Since r is arbitrary, we have $\mathcal{S}(q) \perp M(\hat{\mathcal{S}})$.

To see the converse, suppose that $q \perp \hat{\mathcal{S}}$. Since M is positive semidefinite and $q \in \hat{\mathcal{S}}^*$, AVI$(M, \mathcal{K}, q)$ (which is the same as the generalized LCP$(M, \mathcal{K}, q)$) has a

solution (cf. [4, Prop. 8]). The condition $\mathcal{S}(q) \perp M(\hat{\mathcal{S}})$ along with $(M + M^T)r = 0$ for all $r \in \hat{\mathcal{S}}$ proves (16) for any $u \in \mathcal{S}(q)$ and $r \in \hat{\mathcal{S}}$. Thus f has the GEBP. □

Our next result deals with the LCP formulation of the primal-dual LP [2]. For a matrix A and vectors b and c, we consider

$$M = \begin{bmatrix} 0 & -A^T \\ A & 0 \end{bmatrix} \quad \text{and} \quad q = \begin{bmatrix} c \\ -b \end{bmatrix}. \tag{19}$$

In this setting,

$$\hat{\mathcal{S}} = \left\{ \begin{pmatrix} r \\ s \end{pmatrix} : r, s \geq 0,\ Ar \geq 0, A^T s \leq 0 \right\}$$

and

$$\mathcal{S}(q) = \left\{ \begin{pmatrix} u \\ v \end{pmatrix} : u, v \geq 0,\ Au \geq b, A^T v \leq c,\ \langle c, u \rangle = \langle b, v \rangle \right\}.$$

THEOREM 6.3. *Consider f given by* (18) *with M and q previously described. Then $\mathcal{S}(q) \neq \emptyset$ and f has the GEBP if and only if*

$$\begin{aligned} r \geq 0,\ Ar \geq 0 &\Longrightarrow \langle c, r \rangle = 0, \\ s \geq 0,\ A^T s \leq 0 &\Longrightarrow \langle b, s \rangle = 0. \end{aligned} \tag{20}$$

Proof. In view of part (b) of the previous theorem, it is enough to show that (20) is equivalent to $q \perp \hat{\mathcal{S}}$ and $\mathcal{S}(q) \perp M(\hat{\mathcal{S}})$. Since, in $\hat{\mathcal{S}}$, r and s can vary independently of each other, the condition $q \perp \hat{\mathcal{S}}$ is easily seen to be equivalent to (20). To complete the proof we show that $q \perp \hat{\mathcal{S}}$ implies $\mathcal{S}(q) \perp M(\hat{\mathcal{S}})$. To this end, let $q \perp \hat{\mathcal{S}}$ and consider $(r, s) \in \hat{\mathcal{S}}$ and $(u, v) \in \mathcal{S}(q)$. Then $\langle c, r \rangle = \langle b, s \rangle$ and

$$M \begin{pmatrix} r \\ s \end{pmatrix} = \begin{pmatrix} -A^T s \\ Ar \end{pmatrix}.$$

We have

$$0 \leq \langle Ar, v \rangle = \langle r, A^T v \rangle \leq \langle r, c \rangle = \langle s, b \rangle \leq \langle s, Au \rangle = \langle u, A^T s \rangle \leq 0.$$

This proves, in particular, $\langle Ar, v \rangle = \langle A^T s, u \rangle$, i.e., $(u, v) \perp M(\hat{\mathcal{S}})$. This completes the proof. □

The previous theorem shows that f given by (18) is not a good choice as far as the global error bound analysis is concerned in the LP setting. Fortunately, for the LP, the set $\mathcal{S}(q)$ can be described as the solution set of a finite number of linear inequalities; the well-known Hoffman's error bound analysis is applicable.

7. An error bound characterization of P-matrices. Matrices with all principal minors positive are called **P**-matrices. It is known in the context of LCP [2, Prop. 5.10.5] that for a **P**-matrix M, $\mathcal{S}(q)$ (the solution set of LCP(M, q)) is nonempty for every q and there is a positive number β independent of q such that

$$d(x, \mathcal{S}(q)) \leq \beta ||x \wedge (Mx + q)|| \quad \text{for all} \quad x. \tag{21}$$

The next result proves the converse of this statement.

THEOREM 7.1. *Suppose that for a matrix M, $\mathcal{S}(q) \neq \emptyset$ for all q, and for some $\beta > 0$,* (21) *holds for all q. Then M is a* **P***-matrix.*

Proof. Let $f(x) = x \wedge (Mx + q)$ and $g(x) = x \wedge (Mx + p)$, where q and p denote vectors. Then $f \sim g$ and $||f - g|| \leq ||p - q||$. The proof of the implication (a) $\Longrightarrow$ (b) in Theorem 5.4 shows that

$$\mathcal{S}(p) \subseteq \mathcal{S}(q) + \beta||p - q||B \quad \text{for all} \quad p, q.$$

This Lipschitzian property along with the assumption that $\mathcal{S}(q) \neq \emptyset$ for all q implies, thanks to a recent result due to Murthy, Parthasarathy, and Sabatini [17], that M is a **P**-matrix. □

To derive a consequence of the previous theorem, suppose that for some positive vector e, LCP(M, e) has a unique solution, namely, zero. (We note that positive-semidefinite matrices, copositive matrices, and semimonotone matrices share this property.) If (21) holds for this M and for all q, then Theorem 5.4 shows that LCP$(M, 0)$ has zero as the only solution. It follows that M is a regular matrix [2] and hence $\mathcal{S}(q) \neq \emptyset$ for all q. The previous theorem is applicable and we conclude that M is a **P**-matrix.

Acknowledgments. I would like to thank R. Sznajder, P. Tseng, and Y. Zhang for discussions and helpful comments, and the referees for their suggestions.

REFERENCES

[1] M. J. Chien and E. S. Kuh, *Solving piecewise linear equations for resistive networks,* Circuit Theory Appl., 3 (1976), pp. 3–24.

[2] R. W. Cottle, J.-S. Pang, and R. E. Stone, *The Linear Complementarity Problem*, Academic Press, Boston, MA, 1992.

[3] M. Ferris and J.-S. Pang, *Nondegenerate solutions and related concepts in affine variational inequalities,* SIAM J. Control Optim., 34 (1996), pp. 244–263.

[4] M.S. Gowda, *Pseudomonotone and copositive star matrices,* Linear Algebra Appl., 113 (1989), pp. 107–118.

[5] ———, *On the extended linear complementarity problem*, Math. Programming, forthcoming.

[6] M. S. Gowda and J.-S. Pang, *On the boundedness and stability of solutions to the affine variational inequality problem,* SIAM J. Control Optim., 32 (1994), pp. 421–441.

[7] ———, *Stability analysis of variational inequalities and nonlinear complementarity problems via the mixed linear complementarity problem and degree theory*, Math. Oper. Res., 19 (1994), pp. 831–879.

[8] M. S. Gowda and R. Sznajder, *The generalized order linear complementarity problem,* SIAM J. Matrix Anal. Appl., 15 (1994), pp. 779–795.

[9] M. Kojima and R. Saigal, *A study of PC^1 homeomorphisms on subdivided polyhedrons*, SIAM J. Math. Anal., 10 (1979), pp. 1299–1312.

[10] ———, *On the relationship between conditions that insure a PL mapping is a homeomorphism*, Math. Oper. Res., 5 (1980), pp. 101–109.

[11] D. Kuhn and R. Löwen, *Piecewise affine bijections of R^n and the equation $Sx^+ - Tx^- = y$,* Linear Algebra Appl., 96 (1987), pp. 109–129.

[12] N. G. Lloyd, *Degree Theory*, Cambridge University Press, Cambridge, UK, 1978.

[13] X.-D. Luo and P. Tseng, *On a global projection-type error bound for the linear complementarity problem,* Linear Algebra Appl., forthcoming.

[14] Z.-Q. Luo and J.-S. Pang, *Error bounds for analytic systems and their applications*, Math. Programming, 67 (1994), pp. 1–28.

[15] Z.-Q. Luo and P. Tseng, *On the global error bound for a class of monotone affine variational inequality problems*, Oper. Res. Lett., 11 (1992), pp. 159–165.

[16] O. L. Mangasarian and J.-S. Pang, *The extended linear complementarity problem,* SIAM J. Matrix Anal. Appl., 16 (1995), pp. 359–368.

[17] G. S. R. Murthy, T. Parthasarathy, and M. Sabatini, *On Lipschitzian **Q**-Matrices*, Tech. report, Statistical Quality Control and Operations Research Unit, Indian Statistical Institute, Madras 600 034, India, July 1994.

[18] T. Ohtsuki, T. Fujisawa, and S. Kumagai, *Existence theorems and a solution algorithm for piecewise-linear resistor networks*, SIAM J. Math. Anal., 8 (1977), pp. 69–99.

[19] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[20] J.-S. PANG, *A degree-theoretic approach to parametric nonsmooth equations with multivalued perturbed solution sets*, Math. Programming, 62 (1993), pp. 359–383.

[21] ———, *Necessary and sufficient conditions for solution stability of parametric nonsmooth equations*, in Recent Advances in Nonsmooth Optimization, D.Z. Du, L. Qi, and R.S. Wommersley, eds., World Scientific Publishers, Singapore, 1995, pp. 261–288.

[22] J. REN, *Computable Error Bounds in Mathematical Programming*, Ph.D. thesis, Computer Sciences Department, University of Wisconsin, Madison, WI, August 1993.

[23] W. C. RHEINBOLDT AND J. S. VANDERGRAFT, *On piecewise affine mappings in R^n*, SIAM J. Appl. Math., 29 (1975), pp. 680–689.

[24] S. M. ROBINSON, *Some continuity properties of polyhedral multifunctions*, Math. Programming Study, 14 (1981), pp. 206–214.

[25] S. ROBINSON, *Normal maps induced by linear transformations*, Math. Oper. Res., 17 (1992), pp. 691–714.

[26] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[27] S. SCHOLTES, *Introduction to Piecewise Differentiable Equations*, Institute für Statistik und Mathematische Wirtschaftstheorie, Universität Karlsruhe, 7500 Karlsruhe, Germany, May 1994, Preprint 53/1994.

[28] ———, *Homeomorphism Conditions for Coherently Oriented Piecewise Affine Mappings*, Research report, Institute für Statistik und Mathematische Wirtschaftstheorie, Universität Karlsruhe, 7500 Karlsruhe, Germany, June 1994.

[29] R. SCHRAMM, *On piecewise linear functions and piecewise linear equations*, Math. Oper. Res., 5 (1980), pp. 510–522.

[30] R. SZNAJDER, Private communication, September 1994.

[31] ———, *Degree Theoretic Analysis of the Vertical and Horizontal Linear Complementarity Problem*, Ph.D. thesis, Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD, May 1994.

[32] R. SZNAJDER AND M. S. GOWDA, *Generalizations of $\mathbf{P}_0$ and $\mathbf{P}$-properties; extended vertical and horizontal LCPs*, Linear Algebra Appl., 223/224 (1995), pp. 695–715.

[33] ———, *Nondegeneracy Concepts for Zeros of Piecewise Affine Functions*, Research report, Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD, October 1994.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 3, pp. 610–620, July 1996

011

A CHAIN RULE FOR MATRIX FUNCTIONS AND APPLICATIONS*

ROY MATHIAS†

Abstract. Let f be a not necessarily analytic function and let $A(t)$ be a family of $n \times n$ matrices depending on the parameter t. Conditions for the existence of the first and higher derivatives of $f(A(t))$ are presented together with formulae that represent these derivatives as a submatrix of $f(B)$, where B is a larger block Toeplitz matrix. This block matrix representation of the first derivative is shown to be useful in the context of condition estimation for matrix functions. The results presented here are slightly stronger than those in the literature and are proved in a considerably simpler way.

Key words. derivative, matrix function, condition estimation, Jordan structure

AMS subject classifications. 15A99, 47A55, 47A56

1. Introduction.

Let f be an analytic function. It is well known that

$$f\begin{pmatrix} \lambda & w \\ 0 & \lambda \end{pmatrix} = \begin{pmatrix} f(\lambda) & wf'(\lambda) \\ 0 & f(\lambda) \end{pmatrix}.$$

We generalize this by showing that

$$f\begin{pmatrix} A & W \\ 0 & A \end{pmatrix} = \begin{pmatrix} f(A) & \frac{d}{dt}f(A+tW)\big|_{t=0} \\ 0 & f(A) \end{pmatrix} \tag{1.1}$$

when A and W are square matrices. One can generalize this idea to obtain formulae for higher derivatives.

Our results improve on the results in the literature[1] in several ways. First, [3, 2, 7] all require that $A(t)$ be continuously differentiable in order to conclude that $f(A(t))$ is merely differentiable. Second, our method of proof and our expression for the derivative are considerably simpler than those in [3, 2, 7]. Finally, our formula for the derivative is more useful (easier to evaluate and probably more accurate) for numerical computations—see §5 and the discussion following Theorem 2.1.

Throughout we let D denote an open subset of $\mathbb{C}$ or $\mathbb{R}$. We let M_n denote the set of $n \times n$ complex matrices and $M_n(D, m)$ denote the set of $n \times n$ matrices that have spectrum contained in D and largest Jordan block of size at most m. Let f be $m-1$ times continuously differentiable on D. Given $A \in M_n(D, m)$ we define $f(A)$ by

$$f(A) = r_{A,f}(A), \tag{1.2}$$

where $r_{A,f}$ is any polynomial that interpolates f and its derivatives at the roots of the minimal polynomial of A. That is, if λ is an eigenvalue of A of index p then

$$f^{(i)}(\lambda) = r^{(i)}_{A,f}(\lambda), \quad i = 0, 1, \ldots, p-1.$$

We discuss the many ways to define $f(A)$ later in the section. By considering the Jordan canonical form of A one can check that the right-hand side of (1.2) is indeed

* Received by the editors March 20, 1995; accepted for publication (in revised form) by N.J. Higham August 30, 1995. This research was supported in part by National Science Foundation grant DMS-9201586 and by a Summer Research Grant from the College of William and Mary.

† Department of Mathematics, College of William and Mary, Williamsburg, VA 23187 (na.mathias@na-net.ornl.gov).

[1] The book *Matrix Differential Calculus with Applications in Statistics and Econometrics* [11] does not address the subject of this paper.

independent of the interpolating polynomial chosen—see, for example, [7, Thm. 6.1.9 (b)] for the details.

We now give three useful properties of functions defined on matrices by (1.2). A simple generalization of [7, Thm. 6.1.28], using the continuity of the divided differences that arise in the interpolation problem, yields Lemma 1.1.

LEMMA 1.1. *Let f be $m-1$ times continuously differentiable on D. Then f is continuous on $M_n(D,m)$.*

This fact is crucial in obtaining (1.1). The definition (1.2) implies the desirable property

$$f(SAS^{-1}) = S\,f(A)\,S^{-1}. \tag{1.3}$$

This will also be used in the proof of (1.1). An immediate corollary of the definition (1.2) is that $f(A)$ depends on f only through its first few derivatives on the spectrum of A, as follows in Lemma 1.2.

LEMMA 1.2. *Let $A \in M_n(D,m)$. If for each $k = 1,2,\ldots,m$*

$$f^{(i)}(\lambda) = g^{(i)}(\lambda), \quad i = 0,1,\ldots,k-1$$

for all eigenvalues λ of A of index k then

$$f(A) = g(A).$$

There are a number of ways to extend a scalar-valued function to matrices. Rinehart discusses eight different definitions and shows that many are identical and that all but one are essentially the same in the sense that if for some function f and matrix A two definitions are applicable then the resulting value of $f(A)$ is the same in either case [15]. Some of these definitions are also mentioned in [7, Probs. 6.1.14–15, 6.2.1 and Thm. 6.2.28].

In [7, Def. 6.2.4] the notion of a *primary matrix function* derived from a scalar stem function was defined—it is essentially the same as our definition of $f(A)$. However, the starting point in [7] was (1.3) and the requirement that f be a continuous function on $M_n(D,n)$. The relation (1.2) was proved as a consequence of these two requirements. The reason for our approach is that if $m < n$ then we can consider functions that are defined on $M_n(D,m)$ but not on $M_n(D,n)$.

One could define $f(A)$ via a contour integral or via a power series, assuming in either case that the scalar function could be expressed in the same way. If one were to use these definitions then one could derive formulae for the derivative of $f(A(t))$ quite easily. However, if $A(t)$ is a Hermitian family of matrices then its spectrum would be real for all t and so it would be reasonable to consider $f(A(t))$, where f is defined only on a subset of the real line rather than an open subset of $\mathbb{C}$, and so f could be differentiable without being infinitely differentiable. The question of differentiating such functions of a matrix arises in the study of monotone matrix functions (see, e.g., [7, §6.6]). For such functions we would not be able to define $f(A)$ by a contour integral or a power series.

Section 2 contains our main result. Theorem 2.1 is a formal statement of the formula (1.1). This is perhaps the most important result in the paper. The question of differentiating $f(A(t))$ has also been considered by Horn and Johnson [7, §6.6], Daleckiĭ and Kreĭn [3] (Hermitian case only), and Daleckiĭ [2]. We compare Theorem 2.1 with their results.

In §3 we give an upper bound on the size of the Jordan blocks of certain block upper triangular matrices. This bound is used in §§2 and 4 and is only necessary

because we want to consider functions f that are not infinitely differentiable and want to require only the weakest possible differentiability conditions on f. We generalize Theorem 2.1 to higher derivatives in Theorem 4.1.

In §5 we present an application of Theorem 2.1.

2. The first derivative. In this section we give a basic formula for the first derivative of $f(A(t))$. In §4 we generalize it to the kth derivative of $f(A(t))$. The following theorem is our basic result.

THEOREM 2.1. *Let f be $2m-1$ times continuously differentiable on D. Let $A(t)$ be differentiable at t_0 and assume that $A(t) \in M_n(D,m)$ for all t in some neighborhood of t_0. Then*

$$\frac{d}{dt} f(A(t))\Big|_{t=t_0} = \left[f\begin{pmatrix} A(t_0) & A'(t_0) \\ 0 & A(t_0) \end{pmatrix} \right]_{12}. \tag{2.1}$$

The $_{12}$ on the right-hand side means "take the 1,2 block of the matrix" on the right-hand side.

Proof. Take $\epsilon \neq 0$ and let

$$S = \begin{pmatrix} I & \epsilon^{-1} I \\ 0 & I \end{pmatrix}.$$

Then

$$\begin{aligned} f\begin{pmatrix} A(t_0) & \frac{A(t_0+\epsilon)-A(t_0)}{\epsilon} \\ 0 & A(t_0+\epsilon) \end{pmatrix} &= Sf\left(S^{-1}\begin{pmatrix} A(t_0) & \frac{A(t_0+\epsilon)-A(t_0)}{\epsilon} \\ 0 & A(t_0+\epsilon) \end{pmatrix} S \right) S^{-1} \\ &= Sf\left(\begin{pmatrix} A(t_0) & 0 \\ 0 & A(t_0+\epsilon) \end{pmatrix} \right) S^{-1} \\ &= S\begin{pmatrix} f(A(t_0)) & 0 \\ 0 & f(A(t_0+\epsilon)) \end{pmatrix} S^{-1} \\ &= \begin{pmatrix} f(A(t_0)) & \frac{f(A(t_0+\epsilon))-f(A(t_0))}{\epsilon} \\ 0 & f(A(t_0+\epsilon)) \end{pmatrix}. \end{aligned}$$

Now let $\epsilon \to 0$. Because f is $2m-1$ times continuously differentiable and the largest Jordan block of the matrix on the left-hand side is at most $2m$ (Lemma 3.1), the continuity of f (Lemma 1.1) implies that the limit of the left-hand side exists and is

$$f\begin{pmatrix} A(t_0) & A'(t_0) \\ 0 & A(t_0) \end{pmatrix}.$$

Since the limit on the left-hand side exists so does the limit on the right-hand side. The 1,2 block of this limit is $\frac{d}{dt} f(A(t))|_{t=t_0}$. This gives the desired result. □

We have used Lemma 3.1, a bound on Jordan block size, in proving this result. If we had made the stronger assumption that f is $2n-1$ times continuously differentiable (rather than merely $2m-1$ times) then it would not have been necessary to use Lemma 3.1.

Typically, one will know only that the size of the largest Jordan block of $A(t)$ is bounded by n, so one would usually apply this result with $m = n$. That is, in general f must be $2n-1$ times continuously differentiable in order that $f(A(t))$ be differentiable. If $A(t)$ is Hermitian for all t, then it is also diagonalizable, and hence we may apply the result with $m = 1$ and can conclude that f need only be continuously

differentiable in order that $f(A(t))$ be differentiable. In this case, or more generally when $A(t)$ is diagonalizable, the derivative can be expressed in a form involving a Hadamard product [7, Thm. 6.6.30].

Let us compare our result with those in the literature—Daleckiĭ and Kreĭn [3, Thm. 1] (Hermitian case only), Daleckiĭ [2], and Horn and Johnson [7, Thm. 6.6.14]. For comparison we state part of [7, Thm. 6.6.14], which is representative of the other two results also.

THEOREM 2.2. *Let f be $2n-1$ times continuously differentiable on D. Let $A(t)$ be continuously differentiable on D. Then*

1. *$f(A(t))$ is continuously differentiable on D.*
2. *Let $t_0 \in D$ be given and let $p_{A(t_0)\oplus A(t_0)}(\cdot)$ be the Newton interpolating polynomial that interpolates f and its derivatives at the zeros of the characteristic polynomial of $A(t_0) \oplus A(t_0)$. Then*

$$\frac{d}{dt} f(A(t))\bigg|_{t=t_0} = \frac{d}{dt} p_{A(t_0)\oplus A(t_0)}(A(t))\bigg|_{t=t_0}.$$

3. *For each $t \in D$ let $\lambda_1(t), \ldots, \lambda_{\mu(t)}(t)$ denote the distinct eigenvalues of $A(t)$ and let $r_1(t), \ldots, r_{\mu(t)}(t)$ denote their respective multiplicities as zeros of the minimal polynomial of $A(t)$. Let $A_1(t), \ldots, A_{\mu(t)}(t)$ denote the Frobenius covariants of $A(t)$ (defined in [7, Eq. (6.1.40)]) and let $\Delta f(u,v)$ denote the divided difference $(f(u) - f(v))/(u-v)$. Then*

$$\frac{d}{dt} f(A(t)) = \sum_{j,k=1}^{\mu(t)} \sum_{l=0}^{r_j(t)-1} \sum_{m=0}^{r_k(t)-1} \frac{1}{l!m!} \frac{\partial^{l+m}}{\partial u^l \partial v^m} \Delta f(u,v)\bigg|_{u=\lambda_j(t), v=\lambda_k(t)}$$
$$\times A_j(t)[A(t) - \lambda_j(t)I]^l \frac{d}{dt} A(t)\; A_k(t)[A(t) - \lambda_k(t)I]^m.$$

All the results in [3, 2, 7] require that $A(t)$ be continuously differentiable at t_0 in order to conclude that $f(A(t))$ is differentiable at $t = t_0$. Our result is stronger than theirs in this respect since we require only that $A(t)$ be differentiable at t_0. Horn and Johnson go on to show that under the stronger assumption of continuous differentiability $f(A(t))$ is also continuously differentiable.[2] In Corollary 2.3 we show that the formula (2.1) easily yields the continuous differentiability of $f(A(t))$ when $A(t)$ is continuously differentiable. In fact, the continuous differentiability of $f(A(t))$ seems quite natural given the formula (2.1), while it seems rather surprising if one looks at a formula for the derivative like those in [7, 3, 2] which involve Frobenius covariants or eigenprojections–quantities that may not even be continuous.

Theorem 2.1 shows that if one can evaluate f at a matrix then one can also compute the derivative of $f(A(t))$ using the same method—we exploit this in the last section. From a computational point of view our formula (2.1) is superior to those in [2, 3, 7]. In particular, there is no need to know the eigenvalues of $A(t_0)$, as is required by the formula in part 2 of Theorem 2.2. Part 3 of Theorem 2.2 requires that one also know the Frobenius covariants/eigenprojections of $A(t_0)$. Having the formula depend on the eigenvalues and possibly eigenprojections could be a source of serious error in numerical computation since the eigenvalues and eigenprojections may be very ill conditioned.

[2] One can check that the continuous differentiability of $A(t)$ is used in an essential way in proving the differentiability of $f(A(t))$. See [7, top of p. 525], for example.

Our proof of Theorem 2.1 is much simpler than the proofs of the corresponding results in [7, 3, 2] because most of the work is in proving that f is continuous on $M_{2n}(D, 2m)$. Another nice feature of the formula (2.1) is that it allows one to obtain a similar formula for higher derivatives by a simple inductive argument. We indicate how to this at the beginning of §4.

Theorem 2.1 covers the Hermitian and non-Hermitian cases together. The arguments in [2, 3, 7] do not. So it may appear that our approach is superior in this respect. It is not. If one were to develop the arguments in [2] or [7, proof of Thm. 6.6.14] more carefully then one would see that the differentiability of $f(A(t))$ is guaranteed by f having $2m_i - 1$ continuous derivatives at each eigenvalue λ_i of $A(t_0)$, where m_i is such that for all t in some neighborhood of t_0 every Jordan block corresponding to an eigenvalue in a neighborhood of λ_i of $A(t)$ has size at most m_i.[3] In particular, the more careful argument would cover the Hermitian case. (This more careful approach still requires the continuous differentiability of $A(t)$.)

A possible weakness of all these results, Theorem 2.1 included, is that they require f to be *continuously* differentiable in order to conclude that $f(A(t))$ is differentiable. Whereas if $A(t)$ were a scalar function then it would be sufficient that f be merely differentiable.

Now we show that the continuous differentiability of $A(t)$ guarantees that of $f(A(t))$.

COROLLARY 2.3. *Let f be $2n-1$ times continuously differentiable on D and $A(t) \in M_n(D, n)$ be a continuously differentiable function of t. Then $f(A(t))$ is continuously differentiable.*

Proof. From Theorem 2.1 we know that the derivative of $f(A(t))$ is the 1, 2 block of $f(\hat{A}(t))$, where

$$\hat{A}(t) = \begin{pmatrix} A(t) & A'(t) \\ 0 & A(t) \end{pmatrix}.$$

The matrix $\hat{A}(t)$ is a continuous function of t since $A(t)$ is continuously differentiable. Since f is $2n-1$ times continuously differentiable we know that $f(\hat{A}(t))$ is continuous, and thus

$$\frac{d}{dt} f(A(t)) = [f(\hat{A}(t))]_{12}$$

is also continuous. □

We shall say no more about continuous differentiability.

3. Bounds on Jordan block size. It is useful to have a bound on the size of the Jordan blocks of block upper triangular matrices. The bound can be derived from results due to Friedland and Hershkowitz [4, §3] and Hershkowitz, Rothblum, and Schneider [5, Thm. 5.9]. Meyer and Rose also prove this result [13, Thm. 2.1]. For completeness we include a simple proof, which is different from those in the previously mentioned papers.

LEMMA 3.1. *Let A be a block upper triangular matrix with square main diagonal blocks $A_{ii}, i = 1, 2, \ldots, m$ that are not necessarily of the same size. Fix $\lambda \in \mathbb{C}$ and let k_i be the index of λ in A_{ii}. Then the index of λ in A is at most $k_1 + k_2 + \cdots + k_m$.*

Proof. It is sufficient to consider the case $\lambda = 0$ and $m = 2$. The general case can be derived from this by considering $A - \lambda I$ and by using induction on m.

[3] This point has been noted [2, between lines 18 and 19].

To show that the index of 0 in A is at most $k_1 + k_2$ it is sufficient to show that

$$\text{rank } (A^{k_1+k_2}) = \text{rank } (A^{k_1+k_2+1}).$$

This is implied by

$$\text{rank } (A^{k_1+k_2}) \leq \text{rank } (A^{k_1+k_2+1}) \tag{3.1}$$

since rank $(XY) \leq$ rank (X) for any matrices X and Y for which XY is defined. We shall prove (3.1).

Let r_i be the number of nonzero eigenvalues of A_{ii}. Then using the block upper triangularity of A we have

$$\text{rank } (A^{k_1+k_2+1}) \geq \text{rank } (A_{11}^{k_1+k_2+1}) + \text{rank } (A_{22}^{k_1+k_2+1}) \geq r_1 + r_2. \tag{3.2}$$

Let $k = k_1 + k_2$. Then

$$\begin{aligned}
A^{k_1+k_2} &= \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}^k \\
&= \begin{pmatrix} A_{11}^k & \sum_{j=0}^k A_{11}^j A_{12} A_{22}^{k-j} \\ 0 & A_{22}^k \end{pmatrix} \\
&= \begin{pmatrix} A_{11}^k & \sum_{j=k_1}^k A_{11}^j A_{12} A_{22}^{k-j} \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & \sum_{j=0}^{k_1} A_{11}^j A_{12} A_{22}^{k-j} \\ 0 & A_{22}^k \end{pmatrix} \\
&= \begin{pmatrix} A_{11}^{k_1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} A_{11}^{k_2} & \sum_{j=0}^{k_2} A_{11}^j A_{12} A_{22}^{k-j} \\ 0 & 0 \end{pmatrix} \\
&\quad + \begin{pmatrix} 0 & \sum_{j=0}^{k_1} A_{11}^j A_{12} A_{22}^{k_1-j} \\ 0 & A_{22}^{k_1-1} \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & A_{22}^{k_2} \end{pmatrix}.
\end{aligned}$$

Since the index of 0 in A_{11} is k_1 it follows that the rank of $A_{11}^{k_1}$ is r_1 and hence the rank of the first term in the sum is at most r_1. In the same way the rank of the second term is at most r_2. Since rank is subadditive we have

$$\text{rank } (A^{k_1+k_2}) \leq r_1 + r_2 \leq \text{rank } (A^{k_1+k_2+1}),$$

as required. The second inequality is from (3.2). □

4. Higher derivatives. Now let us consider higher derivatives. One approach is to use induction and Theorem 2.1. This would give us

$$\frac{d^2}{dt^2} f(A(t))\Big|_{t=t_0} = \left[f \begin{pmatrix} A(t_0) & A'(t_0) & A'(t_0) & A''(t_0) \\ 0 & A(t_0) & 0 & A'(t_0) \\ 0 & 0 & A(t_0) & A'(t_0) \\ 0 & 0 & 0 & A(t_0) \end{pmatrix} \right]_{14}$$

for the second derivative. Since we have a $4n \times 4n$ matrix on the right-hand side one might expect that $4n - 1$ continuous derivatives are required of f, but a careful analysis of the Jordan structure of the $4n \times 4n$ matrix shows that $3n - 1$ derivatives are sufficient. This approach can be generalized to higher derivatives and one can derive Theorem 4.1 from it, but this is a rather roundabout and unnatural development.

Given $n \times n$ matrices $A_0, A_1, \ldots, A_k$ let $T[A_0, A_1, \ldots, A_k]$ denote the $n(k+1) \times n(k+1)$ block upper triangular block Toeplitz matrix with i,j block equal to A_{j-i} for $j \geq i$. So, for example,

$$T[A_0, A_1, A_2] = \begin{pmatrix} A_0 & A_1 & A_2 \\ 0 & A_0 & A_1 \\ 0 & 0 & A_0 \end{pmatrix}.$$

THEOREM 4.1. *Let $A(t)$ be k times differentiable at t_0 and assume that $A(t) \in M_n(D,m)$ for all t in some neighborhood of t_0. Assume that f is $(k+1)m-1$ times continuously differentiable on D. Then $f(A(t))$ is k times differentiable at $t = t_0$ and*

$$f\left(T\left[A(t_0), \frac{A^{(1)}(t_0)}{1!}, \ldots, \frac{A^{(k)}(t_0)}{k!}\right]\right) = T\left[f(A(t_0)), \frac{d}{dt}f(A(t_0)), \ldots, \frac{1}{k!}\frac{d^k}{dt^k}f(A(t_0))\right]. \tag{4.1}$$

Proof. Take $0 = \epsilon_0 < \epsilon_1 < \cdots < \epsilon_k$. Let $\Delta_t^j A(\epsilon_i, \epsilon_{i+1}, \ldots, \epsilon_{i+j})$ denote the jth divided difference of A at the $j+1$ points $t+\epsilon_i, \ldots, t+\epsilon_{i+j}$. That is, $\Delta_t^0 A = A(t)$ and for $j > 0$

$$\Delta_t^j A(\epsilon_i, \epsilon_{i+1}, \ldots, \epsilon_{i+j}) = \frac{\Delta_t^{j-1} A(\epsilon_{i+1}, \ldots, \epsilon_{i+j}) - \Delta_t^{j-1} A(\epsilon_i, \ldots, \epsilon_{i+j-1})}{(t+\epsilon_{i+j}) - (t-\epsilon_i)}.$$

Let $T_A(\epsilon)$, $T_f(\epsilon)$, and $S(\epsilon)$ denote the $(k+1)n \times (k+1)n$ block upper triangular matrices with i,j block equal to $\Delta_t^{j-i} A(\epsilon_i, \ldots, \epsilon_j)$, $\Delta_t^{j-i}(f(A(\epsilon_i, \ldots, \epsilon_j))$, and

$$\left[\prod_{l=i}^{j-1}(\epsilon_{j-1} - \epsilon_{l-1})\right]^{-1} I,$$

respectively, for $i \leq j$. If $i > j$ the ij block is 0 because the matrix is block upper triangular. Let $D(\epsilon)$ be the block diagonal matrix with i,i block equal to $A(t_0 + (i-1)\epsilon)$. All these matrices depend on ϵ, but we suppress this dependence in the case of S for simplicity of notation. Note also that in the limit as ϵ goes to 0

$$T_A(\epsilon) \to T\left[A(t_0), \frac{A^{(1)}(t_0)}{1!}, \ldots, \frac{A^{(k)}(t_0)}{k!}\right].$$

We now demonstrate that

$$T_A(\epsilon)S = SD(\epsilon) \tag{4.2}$$

by induction on k. The result is immediate when $k = 0$ since then $S = I$ and $T_A(\epsilon) = D(\epsilon)$.

Let us assume that (4.2) is true for $k-1$ and prove it for k. Since by assumption the result is true for $k-1$, every block on the right-hand side must be the same as that n the left-hand side except perhaps for the $1, k+1$ block. We shall show that this block is also the same by explicitly computing it. The $1, k+1$ block on the left-hand

side is

$$
\begin{aligned}
(T_A(\epsilon)S)_{1,k+1} &= \sum_{j=1}^{k+1} \Delta_{t_0}^{j-1} A(\epsilon_0, \ldots, \epsilon_{j-1}) \left[\prod_{l=j}^{k} (\epsilon_k - \epsilon_{l-1}) \right]^{-1} \\
&= \left[\prod_{l=1}^{k} (\epsilon_k - \epsilon_{l-1}) \right]^{-1} \sum_{j=1}^{k+1} \Delta_{t_0}^{j-1} A(\epsilon_0, \ldots, \epsilon_{j-1}) \prod_{l=1}^{j-1} (\epsilon_k - \epsilon_{l-1}) \\
&= \left[\prod_{l=1}^{k} (\epsilon_k - \epsilon_{l-1}) \right]^{-1} \sum_{j=1}^{k+1} \Delta_{t_0}^{j-1} A(\epsilon_0, \ldots, \epsilon_{j-1}) \prod_{l=1}^{j-1} (t_0 + \epsilon_k - (t_0 + \epsilon_{l-1})) \\
&= \left[\prod_{l=1}^{k} (\epsilon_k - \epsilon_{l-1}) \right]^{-1} A(t_0 + \epsilon_k),
\end{aligned}
$$

which is the $1, k+1$ block of the right-hand side, as desired. The last equality follows from the fact that the penultimate quantity is a multiple of the Newton form of the polynomial that interpolates $A(t)$ at the points $t_0 + \epsilon_0, t_0 + \epsilon_1, \ldots, t_0 + \epsilon_k$ evaluated at the point $t_0 + \epsilon_k$. This can be found in most numerical analysis texts; see, for example, [1, Eq. (3.11)].

Since S is nonsingular it follows from (4.2) that $S^{-1}T_A(\epsilon)S = D(\epsilon)$. Thus we have

$$
f(T_A(\epsilon)) = Sf(S^{-1}T_A(\epsilon)S)S^{-1} = Sf(D(\epsilon))S^{-1} = T_f(\epsilon). \tag{4.3}
$$

One can check that

$$
\lim_{\epsilon \to 0} T_A(\epsilon) = T\left[A(t_0), \frac{A^{(1)}(t_0)}{1!}, \ldots, \frac{A^{(k)}(t_0)}{k!} \right].
$$

Lemma 3.1 ensures that the largest Jordan blocks of $T_A(\epsilon)$ are of size at most $(k+1)m$, and so Lemma 1.1 ensures that $f(T_A(\epsilon))$ is continuous at $\epsilon = 0$. As $\epsilon \to 0$ that is the limit of the term on the extreme left in (4.3), and so the limit of the extreme right term must also exist and be the same. If

$$
\lim_{\epsilon \to 0} \Delta_{t_0}^{j} f(A(\epsilon_i, \ldots, \epsilon_{i+j})
$$

exists then $f(A(t))$ is necessarily j times differentiable at t_0 and the limit is the derivative. This gives the result. □

Notice that the right-hand side of (4.1) depends on f only through $f^{(i)}(\lambda)$ for $i = 0, 1, \ldots, (k+1)m - 1$ and λ in the spectrum of $A(t_0)$. Consequently, if

$$
f^{(i)}(\lambda) = g^{(i)}(\lambda), \quad i = 0, 1, \ldots, (k+1)m - 1
$$

for all λ in the spectrum of $A(t_0)$ then

$$
\frac{d^j}{dt^j} f(A(t))\bigg|_{t=t_0} = \frac{d^j}{dt^j} g(A(t))\bigg|_{t=t_0}, \quad j = 0, 1, \ldots, k.
$$

In the case $k = 2$ this observation is [7, Thm. 6.6.14, part 4]. If we further specialize to the case where g is the polynomial that interpolates f and its derivatives at the eigenvalues (counting multiplicities) of $A(t_0) \oplus A(t_0)$ then we obtain part 3 of the same theorem in [7].

5. Applications to condition estimation. Often one wishes to compute the condition number for the problem of computing $f(A)$. That is, one wishes to find

$$\inf_{\epsilon \downarrow 0} \max_{\|E\| \le \epsilon} \frac{\|f(A+E) - f(A)\|}{\epsilon} \tag{5.1}$$

for some norm $\|\cdot\|$. (Actually, the relative condition number, i.e., the quantity in (5.1) multiplied by the factor $\|A\|/\|f(A)\|$, is more commonly used. It is easily obtained given (5.1) and so we will consider only (5.1).) One can show that (5.1) is equal to

$$\max_{\|E\| \le \epsilon} \|L_f(A;E)\|, \tag{5.2}$$

where $L_f(A;\cdot)$ is the Fréchet derivative of f at A and can be evaluated by

$$L_f(A;E) = \frac{d}{dt} f(A+tE)\Big|_{t=0}. \tag{5.3}$$

If one can evaluate $L_f(A;E)$ for various values of E and if one takes $\|\cdot\|$ to be the Frobenius norm then one can use a power method [8, 12] or a Lanczos-type method [12] to estimate the quantity in (5.2). However, we know that

$$L_f(A;E) = \left[f\begin{pmatrix} A & E \\ 0 & A \end{pmatrix} \right]_{12}. \tag{5.4}$$

The utility of this observation is that there are special methods to compute $f(X)$ when f is a function with special properties—for example, the sine or cosine [16], the exponential [14], the logarithm [8], the square root [6], and the matrix sign ([9] and the references therein). These special methods immediately yield methods for computing the directional derivative. Furthermore, one can use error analysis and perturbation theory for the function f to obtain error analysis and perturbation theory for its derivative. We illustrate this with the matrix sign function.

The matrix sign function is the matrix function obtained by taking D to be the complex plane, excluding the imaginary axis, and f to be defined by $f(z) = \text{sign}\,(\text{Re}(z))$. It is defined for any matrix with no eigenvalues on the imaginary axis. Note that f is infinitely differentiable on D.

One way to compute sign (A) is by the Newton iteration

$$A_0 = A, \quad A_{i+1} = \frac{1}{2}(A_i + A_i^{-1}), \quad i = 0, 1, \ldots, \tag{5.5}$$

which is globally convergent to sign (A), assuming of course that A has no eigenvalues on the imaginary axis. This iteration can be accelerated by scaling; see [10] for the details. For simplicity we omit scaling here. The iteration is quadratically convergent to $S = \text{sign}\,(A)$. One can show that

$$\|A_{i+1} - S\| \le \frac{1}{8}\|S\|\,\|A_i - A_i^{-1}\|^2 + O(\|A_i - A_i^{-1}\|^3), \tag{5.6}$$

where $\|\cdot\|$ is the spectral norm (or any other submultiplicative norm). The main ideas in the proof of (5.6) are the use of the Neumann series for the inverse and the fact that all the quantities that arise (A_i, A_i^{-1}, and S) are polynomials in A and therefore commute. When $\|A_i - A_i^{-1}\|$ is small we have $\|S\| \approx \|A_{i+1}\|$, and so (5.6) gives an approximate upper bound on the error in A_{i+1} as an approximation to S.

One can compute $L_{\text{sign}}(A, E)$ by applying the Newton iteration (5.5) to

$$B_0 \equiv B \equiv \begin{pmatrix} A & E \\ 0 & A \end{pmatrix}. \tag{5.7}$$

By induction we have

$$(B_i)_{11} = (B_i)_{22} = A_i.$$

Let $E_i = (B_i)_{12}$. Explicitly computing $B_{i+1} = (B_i + B_i^{-1})/2$ gives

$$E_{i+1} = (B_{i+1})_{12} = \frac{1}{2}(E_i - A_i^{-1}E_iA_i^{-1}). \tag{5.8}$$

This iteration for E_i is precisely what was derived in [8, Thm. 3.3]. One can obtain a stopping criterion by applying the error bound (5.6) to the matrices B_i. In particular,

$$\begin{aligned}
\|E_{i+1} - L_{\text{sign}}(A; E)\| &= \|(B_{i+1})_{12} - (\text{sign } (B))_{12}\| \\
&\le \|B_{i+1} - \text{sign } (B)\| \\
&\le \frac{1}{8}\|\text{sign } (B)\| \, \|B_i - B_i^{-1}\|^2 + O(\|B_i - B_i^{-1}\|^3) \\
&\le \frac{1}{8}(\|S\| + \|L_{\text{sign}}(A, E)\|) \, (\|A_i - A_i^{-1}\| + \|E_i + A_i^{-1}E_iA_i^{-1}\|)^2 \\
&\quad + O((\|A_i - A_i^{-1}\| + \|E_i + A_i^{-1}E_iA_i^{-1}\|)^3).
\end{aligned}$$

Notice that

$$\|S\| + \|L_{\text{sign}}(A, E)\| \approx \|A_{i+1}\| + \|E_{i+1}\|$$

so we have an approximate upper bound on $\|E_{i+1} - L_{\text{sign}}(A; E)\|$ in terms of the known quantities A_i, A_i^{-1}, A_{i+1}, and E_{i+1}. This is useful because we generally do not need to compute $L_{\text{sign}}(A; E)$ as accurately as sign(A), and so can stop the iteration (5.8) before the iteration (5.5). Although the iteration (5.8) is not new, the bound on $\|E_{i+1} - L_{\text{sign}}(A; E)\|$ is new.

Acknowledgment. The present proof of Theorem 4.1 is based on an idea provided by an anonymous referee. The original proof was very roundabout and unnatural.

REFERENCES

[1] R. Burden and J. Faires, *Numerical Analysis*, PWS-Kent, Boston, MA, 1993.

[2] Ju. L. Daleckiĭ, *Differentiation of non-hermitian matrix functions depending on a parameter*, Amer. Math. Soc. Transl. Ser. 2, 47 (1965), pp. 73–87.

[3] Ju. L. Daleckiĭ and S. G. Kreĭn, *Integration and differentiation of functions of Hermitian matrices and applications to the theory of perturbations*, Amer. Math. Soc. Transl. Ser. 2, 47 (1965), pp. 1–30. (Russian version published in 1958.)

[4] S. Friedland and D. Hershkowitz, *The rank of powers of matrices in a block triangular form*, Linear Algebra Appl., 107 (1988), pp. 17–22.

[5] D. Hershkowitz, U. Rothblum, and H. Schneider, *The combinatorial structure of the generalized nullspace of a block triangular matrix*, Linear Algebra Appl., 116 (1989), pp. 9–26.

[6] N. J. Higham, *Newton's method for the matrix square root*, Math. Comp., 46 (1986), pp. 537–549.

[7] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.

[8] C. KENNEY AND A. J. LAUB, *Polar decomposition and matrix sign function condition estimates*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 488–504.

[9] ———, *Rational iteration methods for the matrix sign function*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 273–291.

[10] ———, *On scaling Newton's method for polar decomposition and the matrix sign function*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 688–706.

[11] J. R. MAGNUS AND H. NEUDECKER, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley, New York, 1988.

[12] R. MATHIAS, *Evaluating the Fréchet derivative of the matrix exponential*, Numer. Math., 63 (1992), pp. 213–226.

[13] C. MEYER AND N. ROSE, *The index and Drazin inverse of block triangular matrices*, SIAM J. Appl. Math., 33 (1976), pp. 1–7.

[14] C. MOLER AND C. VAN LOAN, *Nineteen dubious ways to compute the exponential of a matrix*, SIAM Rev., 20 (1978), pp. 801–836.

[15] R. F. RINEHART, *The equivalence of definitions of a metric function*, Amer. Math. Monthly, 62 (1955), pp. 395–413.

[16] S. SERBIN AND S. BLALOCK, *An algorithm for computing the matrix cosine*, SIAM J. Sci. Statist. Comput., 1 (1980), pp. 198–204.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 3, pp. 621–631, July 1996

012

FURTHER STUDY AND GENERALIZATION OF KAHAN'S MATRIX EXTENSION THEOREM*

DAO-SHENG ZHENG†

Abstract. In 1967, Kahan obtained a matrix extension theorem: Suppose $H \in \mathbb{C}^{l\times l}$ is Hermitian and $B \in \mathbb{C}^{s\times l}$. Denote the spectral norm of

$$R = \begin{bmatrix} H \\ B \end{bmatrix}$$

by $\|R\|_2$. Then there exists a $W \in \mathbb{C}^{s\times s}$ such that

$$A = \begin{bmatrix} H & B^* \\ B & W \end{bmatrix}$$

is Hermitian and $\|A\|_2 = \|R\|_2$. Kahan did not give an explicit expression for W. We show that one may take

$$W = -BH(\varrho^2 I - H^2)^\dagger B^*, \tag{1}$$

where $A^\dagger$ denotes the Moore–Penrose generalized inverse of A. Furthermore, the inequality

$$B(\varrho I + H)^\dagger B^* - \varrho I \le W \le \varrho I - B(\varrho I - H)^\dagger B^* \tag{2}$$

gives the "general solution formula" for W in Kahan's theorem, where $A \ge B$ means A and B are Hermitian and $A - B$ is positive semidefinite. A by-product of (2) is the inequality

$$2\varrho I \ge B[(\varrho I + H)^\dagger + (\varrho I - H)^\dagger]B^*. \tag{3}$$

In this paper we also consider the following problem: Suppose $H \in \mathbb{C}^{l\times l}$ is normal, $B \in \mathbb{C}^{s\times l}$, and

$$R = \begin{bmatrix} H \\ B \end{bmatrix}.$$

How can we find a Hermitian W and a matrix B_1 such that $\|B_1\|_2 = \|B\|_2$ and $\|A\|_2 = \|R\|_2$, where

$$A = \begin{bmatrix} H & B_1^* \\ B & W \end{bmatrix}?$$

Key words. Hermitian matrix, matrix extension theorem, general solution of extension theorem, pseudoinverse form of solution

AMS subject classifications. 47A20, 15A09, 65F30

1. Introduction.

In this paper, $A \in \mathbb{C}^{m\times n}$ means A is an m-by-n complex matrix. $A > B$ ($A \ge B$) means A, B are both Hermitian matrices and $A - B$ is positive definite (positive semidefinite). The Moore–Penrose generalized inverse of matrix A is denoted by $A^\dagger$ [2, Chap. 1]. $\|A\|_2$ is the 2-norm (spectral norm or largest singular value) of A [3, pp. 56–57].

In [5, pp. 231–233], Parlett quoted an important matrix extension theorem, which is given by Kahan in [4] and has not been published in any journal. The theorem is called "Kahan's theorem" in this paper.

* Received by the editors January 13, 1994; accepted for publication (in revised form) by R. Horn August 31, 1995.

† Department of Mathematics, East China Normal University, Shanghai 200062, China.

THEOREM (Kahan [4], 1967). *Suppose* $H \in \mathbb{C}^{l\times l}$ *is Hermitian,* $B \in \mathbb{C}^{s\times l}$,

$$R = \begin{bmatrix} H \\ B \end{bmatrix},$$

and $\varrho = \|R\|_2$. *Then there exists a Hermitian matrix* $W \in \mathbb{C}^{s\times s}$ *such that*

$$A = \begin{bmatrix} H & B^* \\ B & W \end{bmatrix} \tag{1.1}$$

satisfies

$$\|A\|_2 = \|R\|_2. \tag{1.2}$$

In his proof, Kahan pointed out that for any $\sigma > \varrho$, if one sets

$$W_\sigma = -BH(\sigma^2 I - H^2)^{-1}B^*,$$

$$A_\sigma = \begin{bmatrix} H & B^* \\ B & W_\sigma \end{bmatrix},$$

then $\sigma \geq \|A_\sigma\|_2 \geq \|W_\sigma\|_2$. Using meromorphic function theory, Kahan proved that

$$\lim_{\sigma\to\varrho^+} W_\sigma \equiv W \tag{$*$}$$

exists, and W satisfies (1.1) and (1.2).

Kahan's theorem has important applications in matrix perturbation theory [5, Chap. 11]. But Kahan did not give an explicit form for W; his result was purely existential.

It is natural to seek an explicit representation for W. Indeed, an explicit expression for W in $(*)$ can be given by the formula

$$W = -BH(\varrho^2 I - H^2)^\dagger B^*. \tag{1.3}$$

In general (contrary to a statement given without proof in Parlett's book [5, p. 232]), solutions to Kahan's extension problem are not unique. A "general solution formula" for Kahan's theorem is given by the inequality

$$B(\varrho I + H)^\dagger B^* - \varrho I \leq W \leq \varrho I - B(\varrho I - H)^\dagger B^*. \tag{1.4}$$

Both (1.3) and (1.4) can be used to compute W numerically, but from the perturbation theory of generalized inverses [9, pp. 136–140], [8], we can show that (1.4) is better than (1.3) for numerical computation.

From (1.4) we can obtain a by-product. If $H = H^*$,

$$R = \begin{bmatrix} H \\ B \end{bmatrix},$$

and $\varrho = \|R\|_2$, then

$$2\varrho I \geq B[(\varrho I + H)^\dagger + (\varrho I - H)^\dagger]B^*. \tag{1.5}$$

If in Kahan's theorem the matrix H is positive definite, then there is a positive-definite extension A if and only if

$$BH^\dagger B^* < W < \varrho I - B(\varrho I - H)^\dagger B^*. \tag{1.6}$$

The inequality (1.6) can be used to compute W, e.g., $W = \frac{1}{2}(BH^{-1}B^* + \varrho I - B(\varrho I - H)^\dagger B^*)$ satisfies (1.6).

Kahan's theorem can also be modified to consider the following problem: Suppose $H \in \mathbb{C}^{l\times l}$ is non-Hermitian and $B \in \mathbb{C}^{s\times l}$. How can we find a Hermitian W and a matrix B_1 such that $\|B_1\|_2 = \|B\|_2$ and $\|A\|_2 = \|R\|_2$ if

$$A = \begin{bmatrix} H & B_1^* \\ B & W \end{bmatrix}?$$

Section 2 is preliminary. Formula (1.3) is obtained in §3. Inequality (1.4) is obtained in §4. Equation (1.6) is obtained in §5. Extension to a normal matrix is discussed in §6.

2. Preliminary.

DEFINITION 2.1. *$A > 0$ ($A \geq 0$) means that the Hermitian matrix A is positive definite (positive semidefinite). $A \geq B$ means A, B are both Hermitian and $A - B \geq 0$.*

LEMMA 2.1 [7, pp. 315–316]. *If $A \geq B$ and $B \geq C$, then $A \geq C$.*

DEFINITION 2.2 [2, Chap. 1]. *Suppose $A \in \mathbb{C}^{m\times n}$. If $X \in \mathbb{C}^{n\times m}$ satisfies the following equations:*

$$AXA = A,\ XAX = X,\ (AX)^* = AX,\ \text{and}\ (XA)^* = XA, \tag{2.1}$$

then X is called the Moore–Penrose generalized inverse (or pseudoinverse) of A and X is denoted by $A^\dagger$.

LEMMA 2.2 [2, Exercise 22 of Chap. 1]. *Suppose $U \in \mathbb{C}^{m\times m}$, $V \in \mathbb{C}^{n\times n}$ are two unitary matrices, and $A \in \mathbb{C}^{m\times n}$. Then*

$$(UAV)^\dagger = V^* A^\dagger U^*. \tag{2.2}$$

LEMMA 2.3 [10], [7, p. 288]. *Suppose $H \in \mathbb{C}^{n\times n}$; then $HH^* = H^*H$ if and only if there exists an [io], matrix Q such that*

$$Q^*HQ = \Lambda = diag(h_1, \ldots, h_n). \tag{2.3}$$

LEMMA 2.4. *Suppose*

$$R = \begin{bmatrix} H \\ B \end{bmatrix} \in \mathbb{C}^{(n+s)\times n},$$

$H = diag(h_1, \ldots, h_r, \ldots, h_n)$, $|h_1| = |h|_2 = \cdots = |h|_r > |h_{r+1}| \geq \cdots \geq |h_n|$. $B = (B_1, B_2)$ and $B_1 = (b_1, \ldots, b_r)$. If $\|R\|_2 = \|H\|_2$, then $B_1 = 0$.

Proof. Denote $\varrho = \|R\|_2$. From [7, Chap. 6], [3, p. 60], $\|R\|_2^2 \geq \|H\|_2^2$. So $\varrho^2 = \|R\|_2^2 = \|R^*R\|_2 = \|H^*H + B^*B\|_2 \geq \|H^*H\|_2 = |h_1|^2 = \varrho^2$. The diagonal elements of $H^*H + B^*B$ are $\{|h_i|^2 + b_i^*b_i\}$ for $i = 1, \ldots, n$. Hence

$$\varrho^2 \geq |h_i^2| + b_i^*b_i \geq |h_i|^2 = \varrho^2\ (i = 1, \ldots, r).$$

Thus we have $b_i = 0$, $i = 1, \ldots, r$. □

LEMMA 2.5 [1]. *Suppose*

$$A = \begin{bmatrix} E & F^* \\ F & G \end{bmatrix}$$

is Hermitian, $E \in \mathbb{C}^{n\times n}$, and $G \in \mathbb{C}^{k\times k}$. Then $A \geq 0$ if and only if

$$E \geq 0,\ G - FE^\dagger F^* \geq 0,\ \text{and}\ rank(E, F^*) = rank(E). \tag{2.4}$$

LEMMA 2.6. *Suppose A is the same as in Lemma 2.5. Then $A > 0$ if and only if*

$$E > 0\ \text{and}\ G - FE^\dagger F^* > 0. \tag{2.5}$$

LEMMA 2.7 [6], [9, p. 136]. *Suppose* $A, \{A_i\} \in \mathbb{C}^{m\times n}$ *and* $\lim_{i\to\infty} A_i = A$. *Then*

$$\lim_{i\to\infty} A_i^\dagger = A^\dagger \tag{2.6}$$

if and only if

$$\lim_{i\to\infty} rank(A_i) = rank(A). \tag{2.7}$$

LEMMA 2.8 [9, p. 140], [8]. *Suppose* $A \in \mathbb{C}^{m\times n}$, $\tilde{A} = A + E$, *and* $rank(A) \neq rank(\tilde{A})$; *then*

$$\|\tilde{A}^\dagger - A^\dagger\|_2 \geq \frac{1}{\|E\|_2}. \tag{2.8}$$

Moreover, if $rank(\tilde{A}) > rank(A)$, *then*

$$\|\tilde{A}^\dagger\|_2 \geq \frac{1}{\|E\|_2}. \tag{2.9}$$

3. Pseudo-inverse form of W.

THEOREM 3.1. *Suppose* $H \in \mathbb{C}^{l\times l}$ *is Hermitian,* $Q_l^* Q_l = I$ *and* $Q_l^* H Q_l = diag(h_1, \ldots, h_l) = \Lambda$, $B \in \mathbb{C}^{s\times l}$,

$$R = \begin{bmatrix} H \\ B \end{bmatrix},$$

and $\varrho = \|R\|_2$. *Let*

$$W = -BH(\varrho^2 I - H^2)^\dagger B^*, \tag{3.1}$$

$$A = \begin{bmatrix} H & B^* \\ B & W \end{bmatrix}. \tag{3.2}$$

Then

$$W = -(BQ_l)\Lambda(\varrho^2 I - \Lambda^2)^\dagger (BQ_l)^* = W^* \text{ and } \|A\|_2 = \|R\|_2. \tag{3.3}$$

Proof. It is well known [3, p. 60] that if X is a submatrix of Y, then $\|Y\|_2 \geq \|X\|_2$. So $\|A\|_2 \geq \|R\|_2 = \varrho$. To prove $\|A\|_2 = \|R\|_2 = \varrho$, we need only show that $\|A\|_2 \leq \varrho$. Taking

$$Q = \begin{bmatrix} Q_l & 0 \\ 0 & I_s \end{bmatrix},$$

we have

$$\tilde{A} = Q^* A Q = \begin{bmatrix} \Lambda & (BQ_l)^* \\ BQ_l & W \end{bmatrix}. \tag{3.4}$$

From Lemma 2.2, we have

$$W = -BQ_l Q_l^* H Q_l Q_l^* (\varrho^2 I - Q_l \Lambda^2 Q_l^*)^\dagger Q_l Q_l^* B^* = -BQ_l \Lambda(\varrho^2 I - \Lambda^2)^\dagger (BQ_l)^* = W^*. \tag{3.5}$$

Equations (3.4) and (3.5) mean that in the proof of Theorem 3.1 we can assume $H = \Lambda$ and $Q_l = I$.

In order to prove $\|A\|_2 \leq \varrho$, we need a lemma.

LEMMA 3.1. *Suppose*

$$R = \begin{bmatrix} \Lambda \\ B \end{bmatrix},$$

$\Lambda \in \mathbb{C}^{l\times l}$, $\varrho = \|R\|_2 > \|\Lambda\|_2$, $\sigma \geq \varrho$, *and*

$$A_\sigma = \begin{bmatrix} \Lambda & B^* \\ B & W_\sigma \end{bmatrix},$$

with

$$W_\sigma = -B\Lambda(\sigma^2 I - \Lambda^2)^{-1}B^*.$$

Then

$$\|A_\sigma\|_2 \leq \sigma. \tag{3.6}$$

Proof. We first set

$$D = B\Lambda(\sigma^2 I - \Lambda^2)^{-1}.$$

Then

$$W_\sigma = -DB^* = -BD^*.$$

Consider the matrix

$$M = \begin{bmatrix} I & 0 \\ D & I \end{bmatrix} [\sigma^2 I - A_\sigma^2] \begin{bmatrix} I & D^* \\ 0 & I \end{bmatrix}.$$

Because of

$$(\sigma^2 I - \Lambda^2 - B^*B)D^* - \Lambda B^* - B^*W_\sigma = (\sigma^2 I - \Lambda^2)(\sigma^2 I - \Lambda^2)^{-1}\Lambda B^* - \Lambda B^* = 0$$

and

$$\begin{aligned} D(-\Lambda B^* - B^*W_\sigma) + \sigma^2 I - BB^* - W_\sigma^2 \\ &= D(-\Lambda B^*) + \sigma^2 I - BB^* - DB^*(-DB^*) - DB^*DB^* \\ &= -B\Lambda(\sigma^2 I - \Lambda^2)^{-1}\Lambda B^* - BB^* + \sigma^2 I \\ &= \sigma^2[I - B(\sigma^2 I - \Lambda^2)^{-1}B^*], \end{aligned}$$

we have

$$M = \begin{bmatrix} \sigma^2 I - \Lambda^2 - B^*B & 0 \\ 0 & \sigma^2 X \end{bmatrix},$$

with

$$X = I - B(\sigma^2 I - \Lambda^2)^{-1}B^*. \tag{3.7}$$

Since $\|R\|_2 = \varrho \leq \sigma$, we have $\sigma^2 I - R^*R = \sigma^2 I - \Lambda^2 - B^*B \geq 0$. If we can show that $X \geq 0$, then $M \geq 0$, $\sigma^2 I - A_\sigma^2 \geq 0$, and $\|A_\sigma\|_2 \leq \sigma$. To this end, let

$$N = \begin{bmatrix} I & 0 \\ D & I \end{bmatrix} [\sigma^2 I - RR^*] \begin{bmatrix} I & D^* \\ 0 & I \end{bmatrix}.$$

A computation shows that

$$0 \leq N = \begin{bmatrix} \sigma^2 I - \Lambda^2 & 0 \\ 0 & \sigma^2 X \end{bmatrix}, \tag{3.8}$$

where X is given by (3.7). It follows from (3.8) that $X \geq 0$. □

Now we continue to prove $\|A\|_2 \leq \varrho$. There are two cases.

Case I. $\varrho = \|R\|_2 > \|H\|_2 = |h_1|$. In this case, taking $\sigma = \varrho$ in Lemma 3.1, we obtain

$$\|A\|_2 \leq \varrho. \tag{3.9}$$

Thus, Theorem 3.1 is proved for Case I.

Case II. $\|H\|_2 = \|R\|_2 = \varrho$. In this case, suppose $h_1, \ldots, h_l$ are the eigenvalues of Λ and

$$|h_1| = \cdots = |h_r| > |h_{r+1}| \geq \cdots \geq |h_l|. \tag{3.10}$$

From Lemma 2.4, we have

$$b_1 = \cdots = b_r = 0. \tag{3.11}$$

Assume

$$b_1 = \cdots = b_t = 0,\ b_{t+1} \neq 0,\ r \leq t \leq l. \tag{3.12}$$

If $t = l$, then $B = 0$. We can take $W = 0$ and $A = \left[\begin{smallmatrix} \Lambda & 0 \\ 0 & 0 \end{smallmatrix}\right]$. And then $\|A\|_2 = \|R\|_2 = \|\Lambda\|_2$. Consequently, only the case $t < l$ needs to be further considered.

Write

$$B_2 = (b_{t+1}, \ldots, b_l),\ \Lambda = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix},$$

with

$$\Lambda_1 = diag(h_1, \ldots, h_t). \tag{3.13}$$

Then we have

$$A = \begin{bmatrix} \Lambda_1 & 0 & 0 \\ 0 & \Lambda_2 & B_2^* \\ 0 & B_2 & W \end{bmatrix} = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & A_2 \end{bmatrix}, \tag{3.14}$$

with

$$A_2 = \begin{bmatrix} \Lambda_2 & B_2^* \\ B_2 & W \end{bmatrix}.$$

From Lemma 2.4, we have

$$|h_{t+1}| = \|\Lambda_2\|_2 < \left\| \begin{matrix} \Lambda_2 \\ B_2 \end{matrix} \right\|_2 \leq \|R\|_2 = \varrho.$$

Let $\sigma = \varrho$, and

$$W_2 = -B_2\Lambda_2(\varrho^2 I - \Lambda_2^2)^{-1}B_2^*,\ \tilde{A}_2 = \begin{bmatrix} \Lambda_2 & B_2^* \\ B_2 & W_2 \end{bmatrix}. \tag{3.15}$$

From Lemma 3.1, we have

$$\|\tilde{A}_2\|_2 \leq \varrho.$$

Because of

$$\Lambda(\varrho^2 I - \Lambda^2)^\dagger = \begin{bmatrix} \Lambda_1(\varrho^2 I - \Lambda_1^2)^\dagger & 0 \\ 0 & \Lambda_2(\varrho^2 I - \Lambda_2^2)^\dagger \end{bmatrix}$$

$$= diag\left(0, \ldots, 0, \frac{h_{r+1}}{\varrho^2 - |h_{r+1}|^2}, \ldots, \frac{h_l}{\varrho^2 - |h_l|^2}\right), \tag{3.16}$$

we have

$$W = -(0, B_2)\begin{bmatrix}\Lambda_1 & 0\\ 0 & \Lambda_2\end{bmatrix}(\varrho^2 I - \Lambda^2)^{\dagger}(0, B_2)^* \tag{3.17}$$
$$= -B_2\Lambda_2(\varrho^2 I - \Lambda_2^2)^{-1}B_2^* = W_2.$$

So we have

$$A_2 = \tilde{A}_2.$$

Hence we have

$$\|A\|_2 = \max\{\|\Lambda_1\|_2, \|A_2\|_2\} = \varrho,$$

and Theorem 3.1 is proved for Case II. □

It is not difficult to prove the following corollary.

COROLLARY 3.1. *Suppose* $H \in \mathbb{C}^{l\times l}$ *is Hermitian,*

$$B \in \mathbb{C}^{s\times l},\ R = \begin{bmatrix}H\\ B\end{bmatrix},\ \varrho = \|R\|_2 = \|H\|_2,$$

$$H = Q_l\Lambda Q_l^*,\ \Lambda = diag(h_1, \ldots, h_l),\ |h_1| \geq \cdots \geq |h_l|,$$

$$\tilde{B} = BQ_l = (0, \ldots, 0, b_{t+1}, \ldots, b_l) = (0, B_2),\ b_{t+1} \neq 0, and\ B_2 = (b_{t+1}, \ldots, b_l).$$

For any σ, $\varrho \geq \sigma > |h_{t+1}|$, *take*

$$W_\sigma = -\tilde{B}\Lambda(\sigma^2 I - \Lambda^2)^{\dagger}\tilde{B}^* = -B_2\Lambda_2(\sigma^2 I - \Lambda_2^2)^{-1}B_2^*, \tag{3.18}$$

$$A_\sigma = \begin{bmatrix}H & B^*\\ B & W_\sigma\end{bmatrix}. \tag{3.19}$$

Then $\|A_\sigma\|_2 = \|R\|_2$.

Corollary 3.1 means that when $|h_{t+1}| < \sigma \leq \varrho$, any W_σ given by (3.18) satisfies Kahan's theorem.

Notice that when $\sigma > \varrho$, (3.18) is still valid. From Theorem 3.1 we obtain the following proposition.

PROPOSITION 3.1. *In Theorem* 3.1 *we have*

$$\lim_{\sigma\to\varrho^{+0}} W_\sigma = \lim_{\sigma\to\varrho^{+0}}[-BH(\sigma^2 I - H^2)^{-1}B^*] = -BH(\varrho^2 I - H^2)^{\dagger}B^* = W. \tag{3.20}$$

Proof. If $\|R\|_2 > \|H\|_2$, then $\sigma^2 I - \Lambda^2 > \varrho^2 I - \Lambda^2 > 0$ and

$$W = -B\Lambda(\varrho^2 I - \Lambda^2)^{-1}B^*$$

so we can easily prove (3.20).

If $\|R\|_2 = \|H\|_2$, from Lemma 2.4 and (3.18), we have

$$W = -B_2\Lambda_2(\varrho^2 I - \Lambda_2^2)^{-1}B_2^*$$

and

$$W_\sigma = -B_2\Lambda_2(\sigma^2 I - \Lambda_2^2)^{-1}B_2^*. \qquad \square$$

Remark 3.1. Theoretically, the matrix W in Proposition 3.1 coincides with the matrix W in Kahan's theorem. But if we want to compute W by (3.20) numerically, a difficulty might appear when $\varrho = \|H\|_2$. In this case, we have $rank(\varrho^2 I - H^2) < l$ for $H \in \mathbb{C}^{l\times l}$. From Lemmas 2.7 and 2.8, if the computed value of $rank(\varrho^2 I - H^2)$ is greater than its theoretical value, the computed value of $\|(\varrho^2 I - H^2)^{\dagger}\|_2$ must be very large. In fact, when $\|R\|_2 = \|H\|_2$, by Lemma 2.4, we have $b_1 = \cdots = b_r = 0$ in (3.12).

However, because of rounding error, we will not obtain equation (3.12) exactly. And when $\sigma > \varrho$ and $\sigma \doteq \varrho$, from (3.14)–(3.17), the computed W might be very large.

Fortunately, this difficult can always be removed by using the results of §4 in this paper, except in only one extreme case.

4. General solution formula for Kahan's extension theorem.

THEOREM 4.1. *Suppose* $H \in \mathbb{C}^{l\times l}$ *is Hermitian,*

$$R = \begin{bmatrix} H \\ B \end{bmatrix},$$

$\varrho = \|R\|_2$, $W \in \mathbb{C}^{s\times s}$ *is Hermitian, and*

$$A = \begin{bmatrix} H & B^* \\ B & W \end{bmatrix}.$$

Then $\|A\|_2 = \|R\|_2$ *if and only if*

$$B(\varrho I + H)^\dagger B^* - \varrho I \le W \le \varrho I - B(\varrho I - H)^\dagger B^* \tag{4.1}$$

or, equivalently,

$$BQ_l(\varrho I + \Lambda)^\dagger (BQ_l)^* - \varrho I \le W \le \varrho I - BQ_l(\varrho I - \Lambda)^\dagger (BQ_l)^*, \tag{4.1$'$}$$

where $Q_l^* Q_l = I$ *and* $\Lambda = Q_l^* H Q_l$.

Proof. From [3, p. 60] we know that $\|A\|_2 \ge \|R\|_2$; hence $\|A\|_2 = \|R\|_2 \Leftrightarrow \|A\|_2 \le \|R\|_2 = \varrho \Leftrightarrow$

$$\varrho I - A \ge 0 \text{ and } \varrho I + A \ge 0. \tag{4.2}$$

Now

$$\varrho I - A = \begin{bmatrix} \varrho I - H & -B^* \\ -B & \varrho I - W \end{bmatrix}.$$

From Lemma 2.5, $\varrho I - A \ge 0$ is equivalent to the three conditions: (1) $\varrho I - H \ge 0$, (2) $\varrho I - W - B(\varrho I - H)^\dagger B^* \ge 0$, (3) $rank(\varrho I - H, -B^*) = rank(\varrho I - H)$.

In the theorem, (1) is always true. We can show (3) is also true. In fact, if $\varrho > \|H\|_2$, then $\varrho I - H$ is nonsingular, and (3) holds. If $\varrho = \|H\|_2$, from Lemma 2.4, (3) is also true. Hence we obtain

$$\varrho I - A \ge 0 \Leftrightarrow W \le \varrho I - B(\varrho I - H)^\dagger B^*.$$

Similarly we have

$$\varrho I + A \ge 0 \Leftrightarrow W \ge B(\varrho I + H)^\dagger B^* - \varrho I. \qquad \square$$

Remark 4.1. We may use Theorem 4.1 to compute numerically a W that satisfies Kahan's theorem. From (4.1), we know that

$$W_u = \varrho I - B(\varrho I - H)^\dagger B^*$$

and

$$W_l = B(\varrho I + H)^\dagger B^* - \varrho I$$

are two solutions of Kahan's theorem.

When we compute W_u or W_l, there may be three different cases.

(1) $\|R\|_2 - \|H\|_2$ is not too small. In this case both W_u and W_l can be computed with high accuracy.

(2) $\|R\|_2 - \|H\|_2$ is very small,

$$R = \begin{bmatrix} H \\ B \end{bmatrix},$$

and the eigenvalues of H are

$$\lambda_1 \geq \cdots \geq \lambda_l.$$

Assume $\max\{\|R\|_2 - \lambda_1,\ \|R\|_2 + \lambda_l\}$ is not too small. Then one of the W_u and W_l can be computed with high accuracy.

(3) The conditions of R and H are the same as in case (2), but $\max\{\|R\|_2 - \lambda_1,\ \|R\|_2 + \lambda_l\}$ is very small. Then it is very difficult to obtain a good approximation matrix of W_u or W_l.

We regard the case $\|R\|_2 = \lambda_1 = -\lambda_l$ in case (3) as the extreme case.

From Remarks 3.1 and 4.1, we can say that (1.4) is better than (1.3) for computing W.

Combining Theorem 3.1 with Theorem 4.1 we obtain the following theorem.

THEOREM 4.2. *Suppose H, R, B are the same as in Theorem* 3.1*; then*

$$B(\varrho I + H)^\dagger B^* - \varrho I \leq -BH(\varrho^2 I - H^2)^\dagger B^* \leq \varrho I - B(\varrho I - H)^\dagger B^* \tag{4.3}$$

and

$$2\varrho I \geq B[(\varrho I + H)^\dagger + (\varrho I - H)^\dagger]B^*. \tag{4.4}$$

5. Extension to a positive-definite (semidefinite) matrix.

THEOREM 5.1. *Suppose $0 < H \in \mathbb{C}^{l\times l}$, $B \in \mathbb{C}^{s\times l}$,*

$$R = \begin{bmatrix} H \\ B \end{bmatrix},$$

$\varrho = \|R\|_2$, $W \in \mathbb{C}^{s\times s}$ is Hermitian, and

$$A = \begin{bmatrix} H & B^* \\ B & W \end{bmatrix}.$$

Then $A > 0$ and $\|A\|_2 = \|R\|_2$ if and only if

$$BH^{-1}B^* < W \leq \varrho I - B(\varrho I - H)^\dagger B^*. \tag{5.1}$$

Proof. From Lemma 2.6 and Theorem 4.1, $A > 0$, and $\|A\|_2 = \|R\|_2 \Leftrightarrow A > 0$ and (4.1) holds $\Leftrightarrow$ (1) $H > 0$, (2) $W - BH^\dagger B^* > 0$, and (3) $B(\varrho I + H)^\dagger B^* - \varrho I \leq W \leq \varrho I - B(\varrho I - H)^\dagger B^*$.

Since $H > 0$, we obtain [7, p. 143] $H^{-1} > 0$ and $\varrho I + H > H$. So $(\varrho I + H)^{-1} < H^{-1}$, $B(\varrho I + H)^{-1}B^* - \varrho I \leq BH^{-1}B^*$. □

Similarly we have Theorem 5.2.

THEOREM 5.2. *Suppose $0 \leq H \in \mathbb{C}^{l\times l}$, $B \in \mathbb{C}^{s\times l}$,*

$$R = \begin{bmatrix} H \\ B \end{bmatrix},$$

$\varrho = \|R\|_2$, $W \in \mathbb{C}^{s\times s}$ is Hermitian, and

$$A = \begin{bmatrix} H & B^* \\ B & W \end{bmatrix}.$$

Then $A \geq 0$ *and* $\|A\|_2 = \|R\|_2$ *if and only if*

$$\begin{aligned} &\text{(a) } rank(H, B^*) = rank(H), \\ &\text{(b) } BH^\dagger B^* \leq W, \\ &\text{(c) } B(\varrho I + H)^\dagger B^* - \varrho I \leq W \leq \varrho I - B(\varrho I - H)^\dagger B^*. \end{aligned} \tag{5.2}$$

Proof. By Lemma 2.5 and Theorem 4.1, $A \geq 0$ and $\|A\|_2 = \|R\|_2$ if and only if (1) $H \geq 0$, (2) $W - BH^\dagger B^* \geq 0$, (3) $rank(H, B^*) = rank(H)$, and (4) $B(\varrho I + H)^\dagger B^* - \varrho I \leq W \leq \varrho I - B(\varrho I - H)^\dagger B^*$. □

Remark 5.1. For a given matrix

$$R = \begin{bmatrix} H \\ B \end{bmatrix},$$

there may be no W that satisfies the conditions in Theorems 5.1 and 5.2. For example, in Theorem 5.1, assume $R = \left[\begin{smallmatrix} 1 \\ 5 \end{smallmatrix}\right]$. Then $\varrho = \sqrt{26}$, $BH^{-1}B^* = 25$, $\varrho I - B(\varrho I - H)^\dagger B^* = -1$.

Remark 5.2. In (5.2), condition (a) cannot be removed. For example, take $R = \left[\begin{smallmatrix} 0 \\ 5 \end{smallmatrix}\right]$. Then, if (a) is removed, we obtain $0 \leq W \leq 0$ and $A = \left[\begin{smallmatrix} 0 & 5 \\ 5 & 0 \end{smallmatrix}\right]$, which is not positive semidefinite.

6. Extension to a normal matrix. In this section, we consider the following problem: Suppose $H \in \mathbb{C}^{l \times l}$ is normal,

$$R = \begin{bmatrix} H \\ B \end{bmatrix},$$

and $B \in \mathbb{C}^{s \times l}$. How can we find a matrix

$$A = \begin{bmatrix} H & B_1^* \\ B & W \end{bmatrix},$$

with $W = W^*$ such that $\|B_1\|_2 = \|B\|_2$ and $\|A\|_2 = \|R\|_2$? Here we don't require that A be normal.

When $H \in \mathbb{C}^{l \times l}$ is normal, Lemma 2.3 ensures that there exists a unitary matrix Q_l such that

$$Q_l^* H Q_l = \Lambda = diag(h_1, \ldots, h_p, 0, \ldots, 0), \quad |h_1| \geq \cdots \geq |h_p| > 0. \tag{6.1}$$

It is easy to show that there exists a unitary diagonal matrix Ω such that

$$D = \Omega\Lambda = diag(\pm|h_1|, \ldots, \pm|h_p|, 0, \ldots, 0). \tag{6.2}$$

The two simplest cases of (6.2) are

$$D = \pm diag(|h_1|, \ldots, |h_p|, 0, \ldots, 0).$$

THEOREM 6.1. *Suppose* (6.1) *and* (6.2) *hold,* $H \in \mathbb{C}^{l \times l}$ *is normal,* $B \in \mathbb{C}^{s \times l}$,

$$R = \begin{bmatrix} H \\ B \end{bmatrix},$$

and $\varrho = \|R\|_2$. *Take*

$$B_1 = BQ_l \Omega^{-*} Q_l^*, \; W = -BQ_l D(\varrho^2 I - \Lambda^2)^\dagger (BQ_l)^* = W^* \tag{6.3}$$

or

$$BQ_l(\varrho I + D)^\dagger (BQ_l)^* - \varrho I \le W \le \varrho I - BQ_l(\varrho I - D)^\dagger (BQ_l)^*. \tag{6.4}$$

Let

$$A = \begin{bmatrix} H & B_1^* \\ B & W \end{bmatrix}. \tag{6.5}$$

Then $\|B_1\|_2 = \|B\|_2$ *and* $\|A\|_2 = \|R\|_2$.

Proof. It is obvious that $\|B_1\|_2 = \|B\|_2$.

To prove $\|A\|_2 = \|R\|_2$, as in Theorem 3.1, take

$$Q = \begin{bmatrix} Q_l & 0 \\ 0 & I \end{bmatrix},$$

so that

$$\tilde{A} = Q^* A Q = \begin{bmatrix} \Lambda & Q_l^* B_1^* \\ BQ_l & W \end{bmatrix}.$$

From (6.2), we have

$$\begin{bmatrix} \Omega & 0 \\ 0 & I \end{bmatrix} \tilde{A} = M = \begin{bmatrix} D & \Omega Q_l^* B_1^* \\ BQ_l & W \end{bmatrix} = \begin{bmatrix} D & (BQ_l)^* \\ BQ_l & W \end{bmatrix}. \tag{6.6}$$

Here M and W can be regarded as A and W in Theorems 3.1 or 4.1, respectively. Hence we obtain

$$\|M\|_2 = \left\| \begin{matrix} D \\ BQ_l \end{matrix} \right\|_2 = \|R\|_2 \text{ and } \|A\|_2 = \|M\|_2 = \|R\|_2. \qquad \square$$

Acknowledgment. I am grateful to Professor R. A. Horn and the referees for their help and suggestions.

REFERENCES

[1] A. ALBERT, *Conditions for positive and nonnegative definiteness in terms of pseudoinverses*, SIAM J. Appl. Math., 17 (1969), pp. 434–440.

[2] A. BEN-ISRAEL AND T. N. E. GREVILLE, *Generalized Inverses: Theory and Applications*, John Wiley, New York, 1974.

[3] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1989.

[4] W. M. KAHAN, *Inclusion Theorems for Clusters Eigenvalues of Hermitian Matrices*, Computer Science report, Univ. of Toronto, Toronto, Canada, 1967.

[5] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[6] G. W. STEWART, *On the continuity of the generalized inverse*, SIAM J. Appl. Math., 17 (1969), pp. 33–45.

[7] ———, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[8] ———, *On the perturbation of pseudo-inverses, projections and linear least squares problems*, SIAM Rev., 19 (1977), pp. 634–662.

[9] G. W. STEWART AND J. G. SUN, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[10] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, UK, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 3, pp. 632–639, July 1996

013

ON THE JACOBI MATRIX INVERSE EIGENVALUE PROBLEM WITH MIXED GIVEN DATA*

SHU-FANG XU†

Abstract. In this paper, we discuss the problem of constructing a $2n \times 2n$ Jacobi matrix J_{2n} such that its eigenvalues are given distinct values $\lambda_1, \lambda_2, \ldots, \lambda_{2n}$ and its leading $n \times n$ principal submatrix is a given $n \times n$ Jacobi matrix J_n. We give some sufficient and necessary conditions for the solubility of the problem and propose a new fast algorithm for solving this problem. We also present some numerical results.

Key words. Jacobi matrix, eigenvalue, inverse eigenvalue problem

AMS subject classifications. 15A18, 15A57, 65F15

1. Introduction.

An $m \times m$ matrix J_m is called a Jacobi matrix if it is of the form

$$J_m = \begin{bmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \ddots & \mathbf{0} & \\ & \ddots & \ddots & \ddots & \\ & \mathbf{0} & \ddots & \alpha_{m-1} & \beta_m \\ & & & \beta_m & \alpha_m \end{bmatrix}, \tag{1}$$

where α_i and β_i are real and all β_i are positive.

In this paper, we will discuss a class of inverse eigenvalue problems for Jacobi matrices as follows.

PROBLEM DD (double dimension). *Given an $n \times n$ Jacobi matrix J_n and a set of distinct eigenvalue $\{\lambda_i\}_1^{2n}$, construct a $2n \times 2n$ Jacobi matrix J_{2n} whose eigenvalues are the given values $\{\lambda_i\}_1^{2n}$ and whose leading $n \times n$ principal submatrix is exactly J_n.*

The continuum version of this problem is concerned with the following practical question: given a violin string of variable density in $0 \le x \le L$, can the string be extended to $0 \le x \le 2L$ and vibrate with given tones $\lambda_1, \lambda_2, \ldots$?

In 1979, Hochstadt [6] proved that the solution of Problem DD is unique if it exists. In 1984, Dieft and Nanda [2] showed that the space of matrices J_n which can be completed a J_{2n} with fixed spectrum $\{\lambda_i\}_1^{2n}$ is the section of a cone by a hyperplane. In 1987, Boley and Golub [1] proposed a numerical method for solving Problem DD, but this method needs to compute all the eigenvalues and eigenvectors of J_n, which is very time consuming. In 1989, Dai [3] gave sufficient and necessary conditions for the solubility of Problem DD, but the result is somewhat complicated and its proof is very lengthy and tedious. In view of this case, in this paper we first give some new sufficient and necessary conditions for the solubility of Problem DD that seem to be more concise with a simpler proof. Then we propose a new numerical method for finding its solution which does not need to compute the eigenvalues and eigenvectors of J_n. Hence, it is faster than the method proposed by Boley and Golub.

Throughout this paper we will use e_i to denote the ith column of the identity matrix of size implied by context and x^T to denote the transpose of a vector x.

* Received by the editors October 7, 1991; accepted for publication (in revised form) by G. H. Golub September 1, 1995. This paper is a China state major key project for basic researches.

† Department of Mathematics, Peking University, Beijing 100871, People's Republic of China (xsf@sxx0.math.pku.edu.cn).

2. Preliminary lemmas. In this section we first give some preliminary results which play a fundamental role in this paper. Much of this material can be found elsewhere (see, e.g., [2]–[4]) and is included here for the reader's convenience.

LEMMA 2.1. *Let J_m be an $m \times m$ Jacobi matrix defined by* (1) *($m \geq 2$). Then for any integer k with $1 \leq k < m$ the vector*

$$(\xi_1^{(k)}, \ldots, \xi_m^{(k)})^T \equiv J_m^k e_1$$

satisfies that $\xi_{k+2}^{(k)} = \cdots = \xi_m^{(k)} = 0$, $\xi_{k+1}^{(k)} = \beta_{k+1} \cdots \beta_2$, and $\xi_1^{(k)}, \ldots, \xi_k^{(k)}$ are determined completely by $\alpha_1, \ldots, \alpha_k$ and $\beta_2, \ldots, \beta_k$.

Proof. The proof is by induction of k. The lemma is trivially true for $k = 1$. Noting that $J_m^l e_1 = J_m J_m^{l-1} e_1$, we have

$$\begin{aligned}
\xi_1^{(l)} &= \alpha_1 \xi_1^{(l-1)} + \beta_2 \xi_2^{(l-1)}, \\
\xi_i^{(l)} &= \beta_i \xi_{i-1}^{(l-1)} + \alpha_i \xi_i^{(l-1)} + \beta_{i+1} \xi_{i+1}^{(l-1)}, \quad i = 2, 3, \ldots, m-1, \\
\xi_m^{(l)} &= \beta_m \xi_{m-1}^{(l-1)} + \alpha_m \xi_m^{(l-1)},
\end{aligned}$$

where $2 \leq l < m$. Thus it follows from these relations that the lemma is also true for $k = l$, assuming that it is true for $k = l - 1$. By mathematical induction, the lemma is established. □

LEMMA 2.2. *Let J_{2n} be a $2n \times 2n$ Jacobi matrix, and let J_n be the leading principal submatrix of J_{2n} of order n. Then*

$$e_1^T J_{2n}^k e_1 = e_1^T J_n^k e_1 \tag{2}$$

for each $k = 1, 2, \ldots, 2n-1$.

Proof. Partition J_{2n} as follows:

$$J_{2n} = \begin{bmatrix} J_n & \beta_{n+1} e_n e_1^T \\ \beta_{n+1} e_1 e_n^T & J \end{bmatrix}.$$

By Lemma 2.1 it is easy to know that

$$J_{2n}^k e_1 = \begin{bmatrix} J_n^k e_1 \\ 0 \end{bmatrix} \tag{3}$$

for each $k = 1, 2, \ldots, n-1$ and

$$J_{2n}^n e_1 = \begin{bmatrix} J_n^n e_1 \\ \beta e_1 \end{bmatrix}, \quad \beta = \beta_{n+1} \beta_n \cdots \beta_2. \tag{4}$$

Notice that

$$e_1^T J_{2n}^{n+k} e_1 = (J_{2n}^k e_1)^T (J_{2n}^n e_1)$$

for each $k = 1, 2, \ldots, n-1$. The lemma immediately follows from (3) and (4). □

LEMMA 2.3. *Let J_n and $\widetilde{J}_n$ be two $n \times n$ Jacobi matrices. If*

$$e_1^T J_n^k e_1 = e_1^T \widetilde{J}_n^k e_1 \tag{5}$$

for $k = 1, 2, \ldots, 2n-1$, then $J_n = \widetilde{J}_n$.

Proof. Using (5) for $k = 1, 2$, we have

$$\alpha_1 = e_1^T J_n e_1 = e_1^T \widetilde{J}_n e_1 = \widetilde{\alpha}_1$$

and

$$\alpha_1^2 + \beta_2^2 = e_1^T J_n^2 e_1 = e_1^T \widetilde{J}_n^2 e_1 = \widetilde{\alpha}_1^2 + \widetilde{\beta}_2^2,$$

and so $\beta_2 = \widetilde{\beta}_2$ since β_2, $\widetilde{\beta}_2$ are positive. Assume that we have proved that

$$\begin{aligned} &\alpha_i = \widetilde{\alpha}_i, \quad i = 1, 2, \ldots, m-1, \\ &\beta_i = \widetilde{\beta}_i, \quad i = 2, 3, \ldots, m, \end{aligned} \tag{6}$$

where $2 \le m \le n$. Then, by Lemma 2.1, we have that

$$J_n^{m-1} e_1 = \widetilde{J}_n^{m-1} e_1 = (\xi_1, \ldots, \xi_m, 0, \ldots, 0)^T \tag{7}$$

with $\xi_m = \beta_m \cdots \beta_2 > 0$, $\xi_1, \ldots, \xi_{m-1}$ determined completely by $\alpha_1, \ldots, \alpha_{m-1}$ and $\beta_2, \ldots, \beta_{m-1}$.

Now we first prove that $\alpha_m = \widetilde{\alpha}_m$. Using (5) for $k = 2m-1$ we have

$$\begin{aligned} (J_n^{m-1} e_1)^T J_n (J_n^{m-1} e_1) &= e_1^T J_n^{2m-1} e_1 = e_1^T \widetilde{J}_n^{2m-1} e_1 \\ &= (\widetilde{J}_n^{m-1} e_1)^T \widetilde{J}_n (\widetilde{J}_n^{m-1} e_1). \end{aligned} \tag{8}$$

Substituting (7) into (8) and using (6) we obtain

$$\begin{aligned} &x^T J_{m-1} x + 2\xi_m \beta_m \xi_{m-1} + \xi_m^2 \alpha_m \\ &= x^T J_{m-1} x + 2\xi_m \beta_m \xi_{m-1} + \xi_m^2 \widetilde{\alpha}_m, \end{aligned}$$

where

$$x = (\xi_1, \ldots, \xi_{m-1})^T,$$

$$J_{m-1} = \begin{bmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \ddots & \mathbf{0} & \\ & \ddots & \ddots & \ddots & \\ & \mathbf{0} & \ddots & \alpha_{m-2} & \beta_{m-1} \\ & & & \beta_{m-1} & \alpha_{m-1} \end{bmatrix}.$$

Thus $\xi_m^2 \alpha_m = \xi_m^2 \widetilde{\alpha}_m$, and so it must have $\alpha_m = \widetilde{\alpha}_m$ since $\xi_m > 0$.

Next we prove that $\beta_{m+1} = \widetilde{\beta}_{m+1}$. It follows from (7) that

$$\begin{aligned} J_n^m e_1 &= J_n J_n^{m-1} e_1 = (\eta_1, \ldots, \eta_m, \eta_{m+1}, 0, \ldots, 0)^T, \\ \widetilde{J}_n^m e_1 &= \widetilde{J}_n \widetilde{J}_n^{m-1} e_1 = (\eta_1, \ldots, \eta_m, \widetilde{\eta}_{m+1}, 0, \ldots, 0)^T, \end{aligned}$$

where $\eta_{m+1} = \xi_m \beta_{m+1}$, $\widetilde{\eta}_{m+1} = \xi_m \widetilde{\beta}_{m+1}$, and $\eta_1, \ldots, \eta_m$ are determined completely by $\alpha_1, \ldots, \alpha_m$ and $\beta_2, \ldots, \beta_m$. Hence, using (5) for $k = 2m$ we get

$$\begin{aligned} \eta_1^2 + \cdots + \eta_m^2 + \eta_{m+1}^2 &= e_1^T J_n^{2m} e_1 = e_1^T \widetilde{J}_n^{2m} e_1 \\ &= \eta_1^2 + \cdots + \eta_m^2 + \widetilde{\eta}_{m+1}^2, \end{aligned}$$

and so

$$\xi_m^2\beta_{m+1}^2 = \eta_{m+1}^2 = \widetilde{\eta}_{m+1}^2 = \xi_m^2\widetilde{\beta}_{m+1}^2,$$

which implies that $\beta_{m+1} = \widetilde{\beta}_{m+1}$ since β_{m+1}, $\widetilde{\beta}_{m+1}$, and ξ_m are positive. By mathematical induction, the lemma is established. □

Lemma 2.4 is a basic and important result from the inverse theory for the Jacobi matrix.

LEMMA 2.4. *Let $\lambda_1, \ldots, \lambda_{2n}$ be $2n$ distinct real numbers, with $\lambda_1 < \lambda_2 < \cdots < \lambda_{2n}$, and let $\omega_1, \omega_2, \ldots, \omega_{2n}$ be $2n$ positive numbers with $\sum_{i=1}^{2n} \omega_i^2 = 1$. Then there exists a unique $2n \times 2n$ Jacobi matrix J_{2n} such that its eigenvalues are the given values λ_i and the first components of its normalized eigenvectors are exactly ω_i.*

Proof. See [4] for the proof. □

3. Sufficient and necessary conditions. In this section we will devote ourselves to establishing some necessary and sufficient conditions for the solubility of Problem DD. The following theorem is the main result in this section.

THEOREM 3.1. *Suppose an $n \times n$ Jacobi matrix J_n and a set of distinct real numbers $\lambda_1 < \lambda_2 < \cdots < \lambda_{2n}$ are given. Then there exists a unique $2n \times 2n$ Jacobi matrix J_{2n} such that J_{2n} has the given eigenvalues λ_i and its leading $n \times n$ principal submatrix is exactly J_n if and only if*

$$\Delta_i \equiv \det \begin{bmatrix} 1 & \cdots & 1 & 1 & 1 & \cdots & 1 \\ \lambda_1 & \cdots & \lambda_{i-1} & c_1 & \lambda_{i+1} & \cdots & \lambda_{2n} \\ \lambda_1^2 & \cdots & \lambda_{i-1}^2 & c_2 & \lambda_{i+1}^2 & \cdots & \lambda_{2n}^2 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \lambda_1^{2n-1} & \cdots & \lambda_{i-1}^{2n-1} & c_{2n-1} & \lambda_{i+1}^{2n-1} & \cdots & \lambda_{2n}^{2n-1} \end{bmatrix} > 0 \tag{9}$$

for $i = 1, 2, \ldots, 2n$, where

$$c_k = e_1^T J_n^k e_1, \quad k = 1, 2, \ldots, 2n-1. \tag{10}$$

Remark 1. Dai [3] has proved that

$$\Delta_i \Delta_{i+1} > 0, \quad i = 1, 2, \ldots, 2n-1, \tag{11}$$

are the sufficient and necessary conditions for the solubility of Problem DD. Obviously, conditions (9) imply conditions (11). And we will see that the proof of Theorem 3.1 is very simple.

Proof of Theorem 3.1. Let us first prove that conditions (9) are necessary. Assume that J_{2n} is the unique solution of Problem DD. Then by Lemma 2.2 we have

$$e_1^T J_{2n}^k e_1 = e_1^T J_n^k e_1 \tag{12}$$

for $k = 1, 2, \ldots, 2n-1$ and

$$J_{2n} = Q\Lambda Q^T, \tag{13}$$

where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_{2n})$ and Q is the orthogonal matrix of the normalized eigenvectors.

Substituting (13) into (12), we have

$$\omega_1^2\lambda_1^k + \omega_2^2\lambda_2^k + \cdots + \omega_{2n}^2\lambda_{2n}^k = c_k, \quad k = 1, 2, \ldots, 2n-1, \tag{14}$$

where c_k are defined by (10) and

$$(\omega_1, \ldots, \omega_{2n})^T = Q^T e_1, \tag{15}$$

i.e., ω_i is the first component of the ith normalized eigenvector. Hence, we have

$$\omega_1^2 + \omega_2^2 + \cdots + \omega_{2n}^2 = 1. \tag{16}$$

Writing (14) and (16) in matrix-vector form, we have

$$V(\lambda_1, \ldots, \lambda_{2n})x = c, \tag{17}$$

where $V(\lambda_1, \ldots, \lambda_{2n})$ denotes the Vandermonde matrix, i.e.,

$$V(\lambda_1, \ldots, \lambda_{2n}) = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_{2n} \\ \vdots & \vdots & & \vdots \\ \lambda_1^{2n-1} & \lambda_2^{2n-1} & \cdots & \lambda_{2n}^{2n-1} \end{bmatrix}$$

and

$$x = (\omega_1^2, \ldots, \omega_{2n}^2)^T, \quad c = (1, c_1, \ldots, c_{2n-1})^T.$$

Let Δ be the determinant of $V(\lambda_1, \ldots, \lambda_{2n})$. By Gramer's rule we have

$$\omega_i^2 = \Delta_i / \Delta, \quad i = 1, 2, \ldots, 2n, \tag{18}$$

where Δ_i are defined by (9). Since

$$\Delta = \det(V(\lambda_1, \ldots, \lambda_{2n})) = \prod_{1 \le i < j \le 2n} (\lambda_j - \lambda_i) > 0$$

and the first component of any eigenvector of a Jacobi matrix is nonzero, we know that

$$\Delta_i = \Delta \omega_i^2 > 0, \quad i = 1, \ldots, 2n.$$

This shows that conditions (9) are necessary for the solubility of Problem DD.

We now prove that conditions (9) are also sufficient. Assume that conditions (9) are true. Then the linear system (17) has a positive solution x, i.e., if $x = (x_1, \ldots, x_{2n})^T$ satisfies (17), then $x_i > 0$ for $i = 1, \ldots, 2n$. Thus we can define the first components of the normalized eigenvectors of J_{2n}, which we want to find as

$$\omega_i = \sqrt{x_i}. \tag{19}$$

By Lemma 2.4 there exists a unique $2n \times 2n$ Jacobi matrix J_{2n} such that its eigenvalues are the given values $\lambda_1, \lambda_2, \ldots, \lambda_{2n}$ and the first components of its eigenvectors are exactly ω_i. Let $\widetilde{J}_n$ be the leading $n \times n$ principal submatrix of J_{2n}. Then, by Lemma 2.2 and the definition of ω_i, we have

$$e_1^T \widetilde{J}_n^k e_1 = e_1^T J_{2n}^k e_1 = \sum_{i=1}^{2n} \lambda_i^k \omega_i^2 = \sum_{i=1}^{2n} \lambda_i^k x_i = c_k = e_1^T J_n^k e_1$$

for $k = 1, \ldots, 2n-1$. Thus, by Lemma 2.3, we know that $J_n = \widetilde{J}_n$. Hence, J_{2n} is the solution of Problem DD. This completes the proof of Theorem 3.1. □

4. Numerical methods. The process of the proof of Theorem 3.1 provides us with a recipe for finding the solution of Problem DD if it exists. In summary, the method is as follows.

METHOD I.

Step 1. Compute $c_k = e_1^T J_n^k e_1$ for $k = 1, \ldots, 2n-1$.

Step 2. Solve the Vandermonde system (17) for $x = (\omega_1^2, \ldots, \omega_{2n}^2)^T$.

Step 3. Construct the Jacobi matrix J_{2n} from the spectrum data $\{\lambda_i\}_1^{2n}$ and $\{\omega_i\}_1^{2n}$, i.e., compute a $2n \times 2n$ Jacobi matrix J_{2n} such that J_{2n} has eigenvalues λ_i and eigenvectors with the first components ω_i.

In practice, Step 2 of Method I can be completed by using Algorithm 5.6-2 of [5]. As has been pointed out in [5], the algorithm is a fast Vandermonde solver and frequently produces surprisingly accurate solutions, even when the Vandermonde matrix is very ill conditioned.

There are many effective numerical methods to complete Step 3 of Method I, such as the algorithm of [8] and the Rutishauser algorithm of [1]. For details see [1], [8], and references contained therein.

Now we show how to carry out Step 1 of Method I in practice. Let $u^{(0)} = e_1$. We then iteratively define $u^{(k)}$ by

$$u^{(k)} = J_n u^{(k-1)} \tag{20}$$

for $k = 1, \ldots, n$. It is easy to see that

$$c_k = \begin{cases} e_1^T u^{(k)}, & 1 \le k \le n, \\ (u^{(k-n)})^T u^{(n)}, & n < k \le 2n-1. \end{cases} \tag{21}$$

Thus we can use (21) to calculate c_k without computing the power J_n^k. But this process often suffers from overflow. Note that for any real number $r \neq 0$, the Vandermonde system

$$V(r\lambda_1, r\lambda_2, \ldots, r\lambda_{2n})x = (1, rc_1, r^2 c_2, \ldots, r^{2n-1} c_{2n-1})^T \tag{22}$$

has the same solution as the system (17), so an easy way to avoid overflow in the calculation of $u^{(k)}$ is to determine $\widetilde{u}^{(k)}$ from $\widetilde{u}^{(k)} = \widetilde{J}_n \widetilde{u}^{(k-1)}$, where $\widetilde{u}^{(0)} = e_1$ and $\widetilde{J}_n = \|J_n\|_\infty^{-1} J_n$. As a result, instead of solving system (17) in Step 2 of Method I, we need to solve the Vandermonde system

$$V(r\lambda_1, r\lambda_2, \ldots, r\lambda_{2n})x = d, \tag{23}$$

where $d = (1, \widetilde{c}_1, \ldots, \widetilde{c}_{2n-1})^T$, $r = \|J_n\|_\infty^{-1}$, and

$$\widetilde{c}_k = \begin{cases} e_1^T \widetilde{u}^{(k)}, & 1 \le k \le n, \\ (\widetilde{u}^{(k-n)})^T \widetilde{u}^{(n)}, & n < k \le 2n-1. \end{cases}$$

In order to compare this method with the method in [1], we briefly describe the method of [1] as follows.

METHOD II.

Step 1. Compute the eigenvalues $\widetilde{\lambda}_i$ and the normalized eigenvectors $v^{(i)}$ of J_n, and set $\widetilde{\omega}_i = e_1^T v^{(i)}$, where $\widetilde{\lambda}_1 < \widetilde{\lambda}_2 < \cdots < \widetilde{\lambda}_n$.

TABLE 1
The operation count for Methods I *and* II.

	Step 1 (flops)	Step 2 (flops)	Step 3 (flops)	Total (flops)
Method I	$\frac{7}{2}n^2$	$10n^2$	$48n^2$	$\frac{123}{2}n^2$
Method II	$28n^2$	$16n^2$	$48n^2$	$92n^2$

Step 2. Compute

$$\theta_{si} = \prod_{j \neq i} (\lambda_j - \tilde{\lambda}_s) \Big/ \prod_{j \neq i} (\lambda_j - \lambda_i)$$

for $i = 1, \ldots, 2n$, $s = 1, \ldots, n$, and compute

$$\omega_i^2 = \sum_{s=1}^{n} \tilde{\omega}_i^2 \theta_{si}$$

for $i = 1, \ldots, 2n$.

Step 3. This is the same as Step 3 of Method I.

Table 1 shows the approximate amount of arithmetic for each step of Methods I and II, in which a flop means a floating point operation. The amount of arithmetic for Step 2 of Method II is based on the results of [7]—$18n^2$ to compute the eigenvalues plus $10n^2$ to find the eigenvectors by inverse iteration. Observe that the cost of Steps 1 and 2 of Method II is about three times as high as Method I.

5. Numerical example. We have applied Methods I and II to a number of examples collected from the literature. To illustrate their behavior, we give two examples here, both of which were carried out on a PC 486/33 microcomputer in single precision using NDP Fortran 386.

In the following examples $\tilde{J}_{2n}$ denotes the computed solution of Problem DD found by either Method I or Method II, and the matrix norm $\|\cdot\|$ is defined as $\|A\| = \min_{1 \leq i,j \leq m} |\alpha_{ij}|$ for an $m \times m$ matrix $A = [\alpha_{ij}]$.

Example 5.1. It is well known that the $2n \times 2n$ Jacobi matrix J_{2n} with all diagonal elements -2 and all codiagonal elements 1 has eigenvalues

$$\lambda_j = 2\left(\cos\frac{j\pi}{2n+1} - 1\right), \quad j = 1, \ldots, 2n. \tag{24}$$

We have used Methods I and II to reconstruct this Jacobi matrix from its $n \times n$ leading principal submatrix and the λ_j defined by (24) for $n = 5$, 10, 15, 20, 25, 30, 35, 40, and 45. The numerical results are shown in Table 2.

Example 5.2. Let the $2n \times 2n$ Jacobi matrix J_{2n} be as follows:

$$J_{2n} = \begin{bmatrix} -1 & 1 - \frac{2n-1}{2n} & & & 0 \\ 1 - \frac{2n-1}{2n} & \frac{2n-1}{2n} - 2 & 1 - \frac{2n-2}{2n} & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 - \frac{1}{2n} \\ 0 & & & 1 - \frac{1}{2n} & \frac{1}{2n} - 2 \end{bmatrix}.$$

TABLE 2
The computational results of Example 5.1.

	Method I		Method II	
n	$\|\widetilde{J}_{2n} - J_{2n}\|$	CPU time (seconds)	$\|\widetilde{J}_{2n} - J_{2n}\|$	CPU time (seconds)
5	0.8750×10^{-6}	0.06	0.1114×10^{-5}	0.06
10	0.1187×10^{-5}	0.10	0.3609×10^{-5}	0.11
15	0.1350×10^{-5}	0.18	0.6394×10^{-5}	0.22
20	0.1269×10^{-4}	0.21	0.4250×10^{-5}	0.27
25	0.9730×10^{-4}	0.35	0.7700×10^{-5}	0.55
30	0.3670×10^{-4}	0.61	0.1467×10^{-5}	0.88
35	0.8750×10^{-4}	0.99	0.1269×10^{-4}	1.32
40	0.1387×10^{-4}	1.53	0.8457×10^{-5}	1.92
45	0.9870×10^{-4}	1.98	0.9244×10^{-5}	2.64

TABLE 3
The computational results of Example 5.2.

	Method I		Method II	
n	$\|\widetilde{J}_{2n} - J_{2n}\|$	CPU time (seconds)	$\|\widetilde{J}_{2n} - J_{2n}\|$	CPU time (seconds)
10	0.3669×10^{-4}	0.28	0.3250×10^{-5}	0.37
20	0.2330×10^{-4}	0.40	0.6782×10^{-5}	0.59
30	0.5670×10^{-4}	0.69	0.2587×10^{-5}	0.97
40	0.1727×10^{-4}	1.58	0.3657×10^{-5}	2.12

To the best of our knowledge, there is no simple formula for its eigenvalues. Therefore, we have first applied the EISPACK subroutine TQL1 to compute its eigenvalues, then applied Methods I and II to reconstruct it from its $n \times n$ leading principal submatrix and the computed eigenvalues for $n = 10, 20, 30, 40$. The numerical results are given in Table 3.

The previous examples show that the computational time requirement for Method I is really less than Method II, but its computational precision is slightly lower than Method II.

Acknowledgments. The author is very grateful to the referee for valuable comments. He also thanks Professor G. H. Golub for helpful suggestions.

REFERENCES

[1] B. BOLEY AND G. H. GOLUB, *A survey of matrix inverse eigenvalue problems*, Inverse Problems, 3 (1987), pp 595–622.

[2] P. DEIFT AND T. NANDA, *On the determination of a tridiagonal matrix from its spectrum and a submatrix*, Linear Algebra Appl., 60 (1984), pp. 43–55.

[3] H. DAI, *On the Construction of a Jacobi Matrix From Its Spectrum and a Submatrix*, Tech. report NHJB-89-5760, Nanjing Aeronautical Institute, P. R. China, 1989.

[4] G. M. L. GLADWELL, *Inverse Problems in Vibration*, Martinus Nijhoff, Dordrecht, The Netherlands, Boston, MA, 1986.

[5] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[6] H. HOCHSTADT, *On the construction of a Jacobi matrix from mixed given data*, Linear Algebra Appl., 28 (1979), pp 113–115.

[7] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[8] Q. J. SHEN, *A new fast algorithm of reducing a symmetric bordered diagonal matrix to tridiagonal form*, in Proceedings of the 1991 Tianjin Conference on Computational Mathematics, Tianjin, China, pp. 521–523.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 3, pp. 640–648, July 1996

014

ON THE DERIVATIVES OF MATRIX POWERS*

PAOLA SEBASTIANI†

Abstract. We compute the derivatives of rational powers of a positive-definite matrix when this is either a function of a matrix or a scalar variable. In the latter case the result is generalised to real powers. The usefulness of the results found is shown in two statistical applications.

Key words. powers of a positive-definite matrix, matrix functions, matrix derivatives, generalised eigenvalues, optimal experimental design

AMS subject classifications. 15A54, 15A18

1. Introduction.

Matrix derivatives are often used in statistical applications, and explicit formulae for the derivatives of functions commonly used are particularly welcome. In one sense (see, for instance, [3]), the theory of matrix derivatives is not necessary, because each matrix derivative is simply a collection of scalar derivatives. However, as it is generally recognised that matrices ease algebraic manipulations and greatly simplify results, matrix derivatives can be a powerful tool in many statistical applications such as multivariate analysis, likelihood theory, or the theory of optimal design of experiments; see [5].

There are a number of currently used definitions of a derivative of a matrix function of a matrix variable, which differ in the arrangement of the scalar derivatives in a matrix form. We follow the approach suggested by Magnus and Neudecker [5] and define the derivative via the vec-operator of the differential. This is essentially an extension to matrix functions of the concept of differentiability of a vector function of two or more variables. Using this approach, we find the derivatives of rational powers of a positive-definite matrix that is a function of a matrix variable. We further compute the derivatives of real powers of a matrix that is a function of a real variable. The expression found for the derivative generalises the well-known result that $df(x)^p/dx = pf(x)^{p-1}df(x)/dx$, where $f(x)$ is a real function of the real variable x. We also characterise the set of positive-definite matrices such that the derivative of the square root assumes a simple form.

In the next section we recall the rules of differentiation of matrices. In §3 we apply these rules to compute the derivatives of powers of matrices. In §3.1 we consider rational powers of a matrix that is a function of a matrix variable. In §3.2 we consider real powers of a matrix that is a function of a real variable. Special attention is then directed to the square root. The usefulness of the results presented is shown in two statistical applications given in §§4 and 5.

2. Some preliminaries.

Let $A = [a_{ij}]$ $(1 \leq i, j \leq n)$ be an $n \times n$ symmetric matrix. We denote by $A \geq O$ a nonnegative-definite matrix and by $A > O$ a positive-definite matrix.

Suppose $A > O$ and let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be the eigenvalues of A, with corresponding eigenvectors γ_i. We denote by Γ the matrix whose columns are the eigenvectors of A, so that $\Gamma\Gamma^T = \Gamma^T\Gamma = I_n$, and by Λ the diagonal matrix of the

* Received by the editors March 8, 1995; accepted for publication (in revised form) by R. Bhatia September 22, 1995.

† Department of Actuarial Science and Statistics, The City University, Northampton Square, London EC1V 0HB, UK (p.sebastiani@city.ac.uk).

eigenvalues of A, say $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Then the spectral decomposition of A is

$$A = \Gamma \Lambda \Gamma^T. \tag{1}$$

If $A > O$ and (1) holds, the rational power of A is defined as follows; see [6].

DEFINITION 2.1. *If $A > O$, the matrix $A^{p/s}$, $p = \pm 1, \pm 2, \ldots$ and $s \in \mathcal{N}$, exists and is given by*

$$A^{p/q} = \Gamma \Lambda^{p/q} \Gamma^T, \ \Lambda^{p/q} = \ diag\ (\lambda_1^{p/q}, \ldots, \lambda_n^{p/q}). \tag{2}$$

For $p = 1$ and $q = 2$, (2) is the *square root* of A. Note that $A^{p/q}$ and A have the same eigenvectors. If $A > O$ then $A^{p/q} > O$. If $A \geq O$ then $A^{p/q} \geq O$ and is defined only for $p \in \mathcal{N}$. Clearly Definition 2.1 can be generalised to any real power.

Let vec denote the vec-operator and let $\otimes$ denote the Kronecker product, defined as $A \otimes B = [a_{ij}B]$. We shall be considering matrix functions of the real variable $x \in \mathcal{X} \subset \mathcal{R}$, say $A(x)$. If $a_{ij}(x)$ are differentiable for all i, j, then the derivative of A with respect to x exists and is the $n \times n$ matrix

$$\frac{dA}{dx} = \left[\frac{da_{ij}}{dx}\right]; \tag{3}$$

see [5, p. 174]. The matrix application $A(x)$ is a special case of a matrix function of the matrix variable X, say $A(X)$, $X \in \mathcal{S} \subset \mathcal{R}^{n \times q}$. If $a_{ij}(X)$ are differentiable functions of X for all i, j, we can compute the derivative of $A(X)$ with respect to X. This has been defined in various manners (see, for instance, [7]). We will follow the approach suggested by Magnus and Neudecker in Chapter 5 in [5] and define the derivative via the vec-operator of the differential as follows.

Suppose that $A(X)$ is an $m \times p$ matrix, and let $\tilde{X} \in \mathcal{S}$. If A is differentiable at X, the differential of A at X with increment $\tilde{X}$ is the $m \times p$ matrix $dA(X, \tilde{X})$, defined by

$$\text{vec}dA(X, \tilde{X}) = J(X)\text{vec}\tilde{X}. \tag{4}$$

The $mp \times nq$ matrix $J(X)$ in (4) is called the derivative of A at X and we use the notation

$$\frac{dA}{dX} = J(X). \tag{5}$$

Note that this approach generalises the concept of differentiability of a vector function of two or more variables; in fact $J(X)$ is the Jacobian of vec$A$ with respect to vec$X$. When $X = x$, $x \in \mathcal{R}$, (5) is vecdA/dx, from which dA/dx can be recovered.

Thus if we can compute $dA(X, \tilde{X})$, the derivative of A at X is then easily obtained by vectorising the differential to obtain the product form (4), so that the derivative is $J(X)$. Magnus and Neudecker in Chapter 8 in [5] give the differentials of some important matrix functions. In some cases the differential $dA(X, \tilde{X})$ can be easily computed via the directional derivative of A at X in the direction $\tilde{X}$, defined as

$$DA(X)(\tilde{X}) := \left.\frac{d}{dt}\right|_{t=0} A(X + t\tilde{X}). \tag{6}$$

An application of the chain rule shows that

$$\text{vec}DA(X)(\tilde{X}) = \text{vec}dA(X, \tilde{X}) = J(X)\text{vec}\tilde{X}. \tag{7}$$

This approach seems to be particularly useful when $A(X)$ is a matrix power; see [1].

Suppose now that X is itself a matrix function of the matrix variable $Y \in \mathcal{P} \subset \mathcal{R}^{r\times s}$, and consider $A\{X(Y)\}$. It is well known (see [5, p. 91]) that, assuming differentiability, the derivative of $A\{X(Y)\}$ can be computed via the chain rule and is the $mp \times rs$ matrix given by

$$\frac{dA}{dY} = \frac{dA}{dX}\frac{dX}{dY}. \tag{8}$$

When $Y = y$, $y \in \mathcal{R}$, (8) becomes $\operatorname{vec} dA/dy = (dA/dX)\operatorname{vec} dX/dy$, which yields the derivative of A with respect to y in vector form and then the matrix form (3) can be recovered.

3. The derivative of the rational power of a matrix. In this section A will be an $n \times n$ positive-definite matrix. We first give some results on the derivative of powers of matrices.

3.1. Matrix functions of a matrix variable. The proof of the next two lemmas can be found in [5], or easily derived from Examples 2.1 and 2.2 in [1], by using (7).

LEMMA 3.1. *If s is a positive integer, $s \in \mathcal{N}$ say,*

$$\frac{dA^s}{dA} = \sum_{j=1}^{s}(A^{s-j} \otimes A^{j-1}) \text{ and } \frac{dA^{-1}}{dA} = -A^{-1} \otimes A^{-1}.$$

By applying Lemma 3.1 and the chain rule (8), it is easy to prove the next theorem.

THEOREM 3.2. *Let $s, p \in \mathcal{N}$, $r := -s$, and $q := -p$. Then*

$$\frac{dA^r}{dA} = -\sum_{j=1}^{-r}(A^{-j} \otimes A^{r+j-1}), \tag{9}$$

$$\frac{dA^{1/s}}{dA} = \left(\sum_{j=1}^{s}(A^{(s-j)/s} \otimes A^{(j-1)/s})\right)^{-1}, \tag{10}$$

$$\frac{dA^{1/r}}{dA} = -(A^{1/r} \otimes A^{1/r})\left(\sum_{j=1}^{-r}(A^{(r+j)/r} \otimes A^{-(j-1)/r})\right)^{-1}, \tag{11}$$

$$\frac{dA^{p/s}}{dA} = \left(\sum_{j=1}^{p}(A^{(p-j)/s} \otimes A^{(j-1)/s})\right)\left(\sum_{j=1}^{s}(A^{(s-j)/s} \otimes A^{(j-1)/s})\right)^{-1}, \tag{12}$$

$$\frac{dA^{q/s}}{dA} = -\left(\sum_{j=1}^{-q}(A^{-j/s} \otimes A^{(q+j-1)/s})\right)\left(\sum_{j=1}^{s}(A^{(s-j)/s} \otimes A^{(j-1)/s})\right)^{-1}. \tag{13}$$

For example, let $s = 2$ in (10) so that we have

$$\frac{dA^{1/2}}{dA} = \left(I_n \otimes A^{1/2} + A^{1/2} \otimes I_n\right)^{-1}.$$

The results given can then be used to compute the derivatives of more complex matrix functions by using the chain rule (8). Of particular interest are real valued functions of rational powers of A. See Chapter 15 in [8] for general results.

Suppose now that $X = x$. Then (8) leads to $\mathrm{vec} dA^{p/s}/dx = dA^{p/s}/dA\ \mathrm{vec} dA/dx$, which collects the scalar derivatives in a vector form. Unfortunately there does not seem to be a simple way to arrange its elements in a matrix. This is the subject of the next section.

3.2. Matrix functions of a real variable. In this section A will denote an $n \times n$ matrix whose elements are regular functions of the real variable x, $x \in \mathcal{X}$. We suppose that $A > O$ for all $x \in \mathcal{X}$, so that $A^{p/s}$ exists for all s, p. The next result follows by the definition of a derivative given in (3).

LEMMA 3.3. *Let* $A(x) = \mathrm{diag}\,(a_{11}(x), \ldots, a_{nn}(x)) > O$. *Then*

$$\frac{dA^{p/s}}{dx} = \frac{p}{s} A^{\alpha} \frac{dA}{dx} A^{(p-s)/s-\alpha}, \quad \alpha \in \mathcal{R}. \tag{14}$$

Remark. Special cases now follow:

$$\alpha = \frac{p-s}{2s}; \text{ then } \frac{dA^{p/s}}{dx} = \frac{p}{s} A^{(p-s)/(2s)} \frac{dA}{dx} A^{(p-s)/(2s)},$$

$$\alpha = 0; \text{ then } \frac{dA^{p/s}}{dx} = \frac{p}{s} \frac{dA}{dx} A^{(p-s)/s},$$

$$\alpha = \frac{p-s}{s}; \text{ then } \frac{dA^{p/s}}{dx} = \frac{p}{s} A^{(p-s)/s} \frac{dA}{dx}.$$

We want to extend (14) to matrices that are not diagonal. We will need the following lemma, whose proof is straightforward.

LEMMA 3.4. *Let* Γ *be an orthogonal* $n \times n$ *matrix with elements that are regular functions of* x. *Then*

1. $d\Gamma/dx = -\Gamma(d\Gamma^T/dx)\Gamma$ *and* $d\Gamma^T/dx = -\Gamma^T(d\Gamma/dx)\Gamma^T$.
2. $(d\Gamma/dx\Gamma^T)^T = -d\Gamma/dx\Gamma^T$, *that is,* $d\Gamma/dx\Gamma^T$ *is skew symmetric.*

In the next theorem we extend the result given in Lemma 3.3. If $\Gamma\Lambda\Gamma^T$ is the spectral decomposition of $A(x)$, in general both Γ and Λ will be functions of x. If $d\Gamma/dx \neq O$, we define

$$F := \frac{d\Gamma}{dx}\Gamma^T \tag{15}$$

and note that, from Lemma 3.4, $F^T = -F$.

THEOREM 3.5. *Let* $A(x) > O$ *for all* $x \in \mathcal{X}$. *Then*

$$\frac{dA^{p/s}}{dx} = \frac{p}{s} A^{\alpha} \left(\frac{dA}{dx} + H_{\alpha,p,s}\right) A^{(p-s)/s-\alpha}, \quad \alpha \in \mathcal{R}, \tag{16}$$

where

$$H_{\alpha,p,s} := \frac{s}{p} A^{-\alpha} F A^{\alpha+1} - \frac{s}{p} A^{p/s-\alpha} F A^{\alpha-(p-s)/s} - FA + AF. \tag{17}$$

Proof. Using the general rules of differentiation we have

$$\frac{dA}{dx} = \frac{d\Gamma\Lambda\Gamma^T}{dx} = FA + \Gamma\frac{d\Lambda}{dx}\Gamma^T - AF, \tag{18}$$

and similarly

$$\frac{dA^{p/s}}{dx} = FA^{p/s} + \Gamma\frac{d\Lambda^{p/s}}{dx}\Gamma^T - A^{p/s}F. \tag{19}$$

By applying Lemma 3.3 we have that (19) simplifies to

$$\frac{dA^{p/s}}{dx} = FA^{p/s} + \frac{p}{s}\Gamma\Lambda^{\alpha}\frac{d\Lambda}{dx}\Lambda^{(p-s)/s-\alpha}\Gamma^T - A^{p/s}F,\ \alpha \in \mathcal{R}, \tag{20}$$

so that

$$\frac{dA^{p/s}}{dx} - \frac{p}{s}A^{\alpha}\frac{dA}{dx}A^{(p-s)/s-\alpha}$$
$$= \frac{p}{s}A^{\alpha}\left(\frac{s}{p}A^{-\alpha}FA^{1+\alpha} - \frac{s}{p}A^{p/s-\alpha}FA^{\alpha-(p-s)/s} + AF - FA\right\}A^{(p-s)/s-\alpha}. \qquad \square$$

Note that when $\alpha = (p-s)/(2s)$, the matrix $H_{\alpha,p,s}$ is symmetric. If $s = 1$, $p = -1$, and $\alpha = -1$, then (16) gives the well-known result: $dA^{-1}/dx = -A^{-1}(dA/dx)A^{-1}$. If $p = 1$, $s = 2$, and $\alpha = -1/2$

$$\frac{dA^{1/2}}{dx} = \frac{1}{2}A^{-1/2}\left(\frac{dA}{dx} + H_{-.5,1,2}\right) \tag{21}$$

with

$$H_{-.5,1,2} = 2A^{1/2}FA^{1/2} - AF - FA, \tag{22}$$

which is a skew symmetric matrix. A simple consequence of Theorem 3.5 is Corollary 3.6.

COROLLARY 3.6. *If the eigenvectors of A do not depend on x, then* (14) *holds.*

The converse of Corollary 3.6 does not hold in general; the case $s = 1, p = -1$, and $\alpha = -1$ is a counterexample. We can show that the converse of Corollary 3.6 does not hold if the eigenvalues of A are not simple. Suppose in fact that $A > O$, so that (14) holds if and only if $H_{\alpha,p,s} = O$ for all α, that is, if and only if

$$\frac{s}{p}A^{-\alpha}FA^{1+\alpha} - \frac{s}{p}A^{p/s-\alpha}FA^{\alpha-(p-s)/s} + AF - FA = O\ \forall\ \alpha \in \mathcal{R}. \tag{23}$$

If we take the vec-operator, the matrix equation (23) is equivalent to the system of linear equations:

$$\left\{\frac{s}{p}(A^{\alpha+1} \otimes A^{-\alpha} - A^{\alpha-(p-s)/s} \otimes A^{p/s-\alpha}) + (I_n \otimes A) - (A \otimes I_n)\right\}\mathrm{vec}F = 0. \tag{24}$$

Because the matrices in (23) commute, they can be diagonalised by the common orthogonal transformation $\tilde{\Gamma} = \Gamma \otimes \Gamma$, so that (24) can be transformed into

$$\tilde{\Gamma}(\Lambda_1 - \Lambda_2 + \Lambda_3 - \Lambda_4)\tilde{\Gamma}^T \tag{25}$$

with Λ_i's the diagonal matrices with the eigenvalues of the corresponding matrices. Note that the elements of Λ_1 are a permutation of those of Λ_2, as are those of Λ_3 and Λ_4. Thus the rank of the matrix (25) is at most $2n(n-1)$ if the eigenvalues are all simple. Suppose now $\lambda_1 = \lambda_2$. Then the rank of (25) is at most $2n(n-2)$, which is less than n^2 if, for instance, $n = 3$. This is sufficient to conclude that the null matrix is not in general the unique solution of (23). A sufficient condition to have $H_{\alpha,p,s} = O$ for all α is given in the next corollary.

COROLLARY 3.7. *If the matrices A and F commute, then $H_{\alpha,p,s} = O$ for all α, so that*

$$\frac{dA^{p/s}}{dx} = \frac{p}{s} A^{\alpha} \frac{dA}{dx} A^{\frac{p-s}{s}-\alpha}, \quad \alpha \in \mathcal{R}.$$

Finally, a trivial consequence of Theorem 3.5 is Corollary 3.8.

COROLLARY 3.8. *If $H_{\alpha,p,s} = O$ for all α, then the matrices dA/dx and $A^{(p-s)/s}$ commute.*

Clearly, the result given in Theorem 3.5 and the corollaries following it can be generalised to any real power r of the matrix A:

$$\frac{dA^r}{dx} = rA^{\alpha}\left(\frac{dA}{dx} + H_{\alpha,r}\right) A^{r-1-\alpha}, \quad \alpha \in \mathcal{R}, \tag{26}$$

where

$$H_{\alpha,r} := \frac{1}{r} A^{-\alpha} F A^{\alpha+1} - \frac{1}{r} A^{r-\alpha} F A^{\alpha-r+1} - FA + AF. \tag{27}$$

Example 1. Let $x > 0$ and

$$A(x) = \frac{x}{2}\begin{bmatrix} 1+x & 1-x \\ 1-x & 1+x \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x & 0 \\ 0 & x^2 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$

Clearly $A(x) > O$ for all $x > 0$, and for any real r

$$A^r = \frac{x^r}{2}\begin{bmatrix} 1+x^r & 1-x^r \\ 1-x^r & 1+x^r \end{bmatrix}.$$

For any real α we can compute

$$A^{\alpha} = \frac{x^{\alpha}}{2}\begin{bmatrix} 1+x^{\alpha} & 1-x^{\alpha} \\ 1-x^{\alpha} & 1+x^{\alpha} \end{bmatrix}$$

and

$$A^{r-1-\alpha} = \frac{x^{r-1-\alpha}}{2}\begin{bmatrix} 1+x^{r-1-\alpha} & 1-x^{r-1-\alpha} \\ 1-x^{r-1-\alpha} & 1+x^{r-1-\alpha} \end{bmatrix}.$$

It follows that

$$\frac{dA^r}{dx} = \frac{rx^{r-1}}{2}\begin{bmatrix} 1+2x^r & 1-2x^r \\ 1-2x^r & 1+2x^r \end{bmatrix} \text{ and } \frac{dA}{dx} = \frac{1}{2}\begin{bmatrix} 1+2x & 1-2x \\ 1-2x & 1+2x \end{bmatrix}.$$

It is then straightforward to verify that

$$\frac{dA^r}{dx} = rA^{\alpha}\frac{dA}{dx}A^{r-\alpha-1} \;\forall\; \alpha.$$

To end this section we consider a special property of the derivative of the square root of A.

COROLLARY 3.9. *Expression* (21) *simplifies to*

$$\frac{dA^{1/2}}{dx} = \frac{1}{2}A^{-1/2}\frac{dA}{dx} \tag{28}$$

if and only if the matrices $dA^{1/2}/dx$ *and* $A^{1/2}$ *commute.*

Proof. Since $A(x) > O$, so is $A(x)^{-1/2}$ and hence (28) holds if and only if $H_{-.5,1,2} = O$. Thus, from (22), (28) holds if and only if

$$2A^{1/2}FA^{1/2} = AF + FA. \tag{29}$$

The identity (29) holds if and only if

$$A^{1/2}FA^{1/2} - AF = FA - A^{1/2}FA^{1/2}.$$

If we now add the matrix $\Gamma\Lambda^{1/2}(d\Lambda^{1/2}/dx)\Gamma^T$ to both sides, we obtain

$$\begin{aligned} &A^{1/2}\left(FA^{1/2} + \Gamma\frac{d\Lambda^{1/2}}{dx}\Gamma^T - A^{1/2}F\right) \\ &= \left(FA^{1/2} + \Gamma\Lambda^{1/2}\frac{d\Lambda^{1/2}}{dx}\Lambda^{-1/2}\Gamma^T - A^{1/2}F\right)A^{1/2}. \end{aligned}$$

By noting that $\Lambda^{1/2}(d\Lambda^{1/2}/dx)\Lambda^{-1/2} = d\Lambda^{1/2}/dx$, we have from (19) that the last identity holds if and only if

$$A^{1/2}\frac{dA^{1/2}}{dx} = \frac{dA^{1/2}}{dx}A^{1/2}. \qquad \square$$

4. The derivatives of the generalised eigenvalues. The perturbation theory of generalised eigenproblems is the mathematical background underlying the derivation of influence functions in multivariate analysis; see [9] and [10]. In this section we start by recalling the definition of a generalised eigenproblem, and we then find an explicit expression for the derivative of the generalised eigenvalues and related eigenvectors when they are functions of a real variable.

DEFINITION 4.1. *Let* A *and* M *be* $n \times n$ *symmetric matrices and* $M > O$. *The generalised eigenvalues of* A *are the solutions of the equation*

$$\det(A - \lambda M) = 0. \tag{30}$$

All vectors $a \neq 0$, *such that* $Aa = \lambda Ma$, *with* λ *a solution of* (30), *are called generalised eigenvectors of* A *relative to* λ.

Note that if $M > O$, then the solutions of equation (30) are the eigenvalues of $M^{-1/2}AM^{-1/2}$ with the same multiplicities. If v is an eigenvector of $M^{-1/2}AM^{-1/2}$, with eigenvalue λ, then $a = M^{-1/2}\mathbf{v}$ is a generalised eigenvector of A relative to λ. Moreover, if $v^Tv = 1$ then $a^TMa = 1$.

Suppose now that A and M are regular functions of x. Hence the generalised eigenvalues of A will be functions of x. We shall compute the derivative of the generalised eigenvalues of A in closed form using (21).

THEOREM 4.2. *Let A and M be symmetric matrices, $n \times n$ and $M > O$. Suppose that the eigenvalues and eigenvectors of M are regular functions of $x \in U$, U an open subset of $\mathcal{R}$. If, for $x = x_0$, $\lambda_0 := \lambda(x_0)$ is a simple solution of (30), with corresponding generalised eigenvector $a_0 := a_0(x_0)$, then there exists a set $N(x_0)$ containing x_0 and two functions $\lambda : N(x_0) \to \mathcal{R}^+$ and $a : N(x_0) \to \mathcal{R}^n$, $\lambda, a \in C^\infty$ with $\lambda(x_0) = \lambda_0$ and $a(x_0) = \mathbf{a}_0$, such that*

$$Aa = \lambda Ma, \quad a^T Ma = 1 \ \forall x \in N(x_0).$$

Moreover, for the eigenvalues with multiplicity 1

$$\frac{d\lambda}{dx} = a^T \left(\frac{dA}{dx} - \lambda \frac{dM}{dx} \right) a \ \forall x \in N(x_0).$$

Proof. Let B be $M^{-1/2} A M^{-1/2}$. If, for $x = x_0$, λ_0 is a simple solution of (30), then it is a simple eigenvalue of B. Let v_0 be a corresponding normalised eigenvector. B is a regular function of x and hence (see [5, p. 158]) there exists a set $N(x_0)$ containing x_0 and two functions $\lambda : N(x_0) \to \mathcal{R}^+$ and $v : N(x_0) \to \mathcal{R}^n$, $\lambda, v \in C^\infty$ such that $Bv = \lambda v$ and $v^T v = 1$ for all $x \in N(x_0)$, with $\lambda(x_0) = \lambda_0$ and $v(x_0) = v_0$. Also,

$$\frac{d\lambda}{dx} = v^T \frac{dB}{dx} v \ \forall \ x \in N(x_0). \tag{31}$$

By recalling that $a = M^{-1/2} v$, with M differentiable, the first part follows. Now put $v = M^{-1/2} a$ and $B = M^{-1/2} A M^{-1/2}$ in equation (31). Then

$$\begin{aligned} \frac{d\lambda}{dx} &= a^T M^{1/2} \frac{d(M^{-1/2} A M^{-1/2})}{dx} M^{1/2} a \\ &= a^T M^{1/2} \left(\frac{dM^{-1/2}}{dx} A M^{-1/2} + M^{-1/2} \frac{dA}{dx} M^{-1/2} + M^{-1/2} A \frac{dM^{-1/2}}{dx} \right) M^{1/2} a \\ &= a^T M^{1/2} \frac{dM^{-1/2}}{dx} Aa + a \frac{dA}{dx} a + a^T A \frac{dM^{-1/2}}{dx} M^{1/2} a. \end{aligned} \tag{32}$$

From Theorem 3.5,

$$\frac{dM^{-1/2}}{dx} = -\frac{1}{2} M^{-1/2} \left(\frac{dM}{dx} + H_{-.5,-1,2} \right) M^{-1} = -\frac{1}{2} M^{-1} \left(\frac{dM}{dx} + H_{-1,-1,2} \right) M^{-1/2}$$

with

$$H_{-.5,-1,2} = -2A^{1/2} F A^{1/2} + FA + AF = -H_{-1,-1,2},$$

which are both skew symmetric matrices. If we substitute $Aa = \lambda Ma$ in (32) we obtain

$$a^T \frac{dA}{dx} a - \frac{1}{2} a^T \left(\frac{dM}{dx} + H_{-.5,-1,2} + H_{-1,-1,2} \right) a,$$

which equals

$$a^T \frac{dA}{dx} a - \lambda a^T \frac{dM}{dx} a. \qquad \square$$

5. The directional derivative of $(\text{trace}A^r)^{1/r}$. In the classical theory of optimal design of experiments it is common to consider optimality criteria that are nondecreasing and concave real functions defined in a subset of the positive-definite matrices, say $\mathcal{V} \subset \mathcal{R}^{n\times n}$; see [11]. A family of coherent optimality criteria (see [2]) is

$$\Phi(V) := \begin{cases} \left(\frac{1}{n}\text{trace}V^r\right)^{1/r}, & r \le 1,\ r \neq 0, \\ \det V^{1/n}, & r = 0. \end{cases}$$

A powerful tool for verifying that a design is Φ-optimal is the directional derivative of Φ at V_1 in the direction of V_2, that is,

$$F_\Phi(V_1, V_2) = \left.\frac{d}{dt}\right|_{t=0} \text{trace}\Phi(V(t)),$$

where $V(t) := V_1 + tV_2$. We now compute the directional derivative of $\Phi(\cdot)$. The proof of the following result is straightforward.

LEMMA 5.1. *If $A(x) > O$ for all x, then*

$$\frac{d\,(\text{trace}A^r)^{1/r}}{dx} = (\text{trace}A^r)^{1/r-1}\text{trace}\left(A^{r-1}\frac{dA}{dx}\right). \tag{33}$$

If we now let $r \to 0$ in (33) we have the well-known result; see [4, p. 356]:

$$\frac{d\det A^{1/n}}{dx} = \frac{1}{n}\det A^{1/n}\text{trace}\left(A^{-1}\frac{dA}{dx}\right).$$

An application of Lemma 5.1 yields

$$F_\Phi(V_1, V_2) = \begin{cases} \frac{1}{n^r}(\text{trace}V_1^r)^{1/r-1}\text{trace}(V_1^{r-1}V_2), & r \le 1,\ r \neq 0, \\ \frac{1}{n}\det V_1^{1/n}\text{trace}(V_1^{-1}V_2), & r = 0. \end{cases}$$

Acknowledgments. The author is grateful to the editor and the referee for their valuable suggestions which helped in improving a first version of this article.

REFERENCES

[1] R. BHATIA, *First and second order perturbation bounds for the operator absolute value*, Linear Algebra Appl., 208 (1994), pp. 367–376.
[2] A. P. DAWID AND P. SEBASTIANI, *Proper Criteria for Optimal Experimental Design*, Statistical research paper #2, Department of Actuarial Science and Statistics, City University, London, UK, March 1996.
[3] P. S. DWYER, *Some applications of matrix derivatives in multivariate analysis*, J. Amer. Statist. Assoc., 62 (1967), pp. 607–625.
[4] F. A. GRAYBILL, *Matrices with Applications in Statistics*, 2nd ed., Wadsworth, Belmont, CA, 1983.
[5] J. R. MAGNUS AND M. NEUDECKER, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley, New York, 1988.
[6] K. V. MARDIA, J. T. KENT, AND J. M. BIBBY, *Multivariate Analysis*, Academic Press, New York, 1979.
[7] C. R. RAO, *Matrix derivatives*, in Encyclopedia of Statistical Science, Vol. 5, S. Kotz, N. I. Johnson, and C. B. Read, eds., John Wiley, New York, 1985, pp. 320–325.
[8] G. R. ROGERS, *Matrix Derivatives*, Marcel Dekker, New York, 1980.
[9] M. ROMANAZZI, *Influence in canonical variates analysis*, Comput. Statist. Data Anal., 11 (1991), pp. 143–164.
[10] ———, *Influence in canonical correlation analysis*, Psychometrika, 57 (1993), pp. 237–259.
[11] S. D. SILVEY, *Optimal Designs*, Chapman and Hall, London, UK, 1980.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 3, pp. 649–657, July 1996

015

SECOND-ORDER SYSTEMS WITH SINGULAR MASS MATRIX AND AN EXTENSION OF GUYAN REDUCTION*

SANJAY P. BHAT† AND DENNIS S. BERNSTEIN†

Abstract. The set of consistent initial conditions for a second-order system with singular mass matrix is obtained. In general, such a system can be decomposed (i.e., partitioned) into three coupled subsystems of which the first is algebraic, the second is a regular system of first-order differential equations, and the third is a regular system of second-order differential equations. Under specialized conditions, these subsystems are decoupled. This result provides an extension of Guyan reduction to include viscous damping.

Key words. second-order differential equation, singular mass matrix

AMS subject classifications. 34A30, 70J05, 93A99

Notation.

$\mathcal{R}$ ($\mathcal{C}$)	real (complex) numbers,
$\mathcal{R}^n(\mathcal{R}^{n\times n})$	real vectors (matrices) of dimension n ($n \times n$),
$(A)_{ij}$	ijth element of the matrix A,
rank A (def A, ind A)	rank (defect, index) of the matrix A,
$\mathcal{N}(A)$ ($\mathcal{R}(A)$)	nullspace (range) of the matrix A,
A^{T}	transpose of the matrix A,
$A > (\geq)0$	symmetric positive- (nonnegative-) definite matrix,
$\mathcal{S}_1 \perp \mathcal{S}_2$	subspace $\mathcal{S}_1$ orthogonal to the subspace $\mathcal{S}_2$,
$\mathcal{S}_1 \oplus \mathcal{S}_2$	direct sum of the subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$,
$\mathcal{S}_1 \cap \mathcal{S}_2$	intersection of the subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$,
$\triangleq$	equal by definition.

1. Introduction.

Singular linear systems, that is, linear systems of the form $E\dot{x} = Ax$, where the matrix E is singular, have been studied extensively. Such systems arise in singular perturbation problems [1], optimal control [2], and large scale interconnected systems and economics [3].

An interesting property of singular systems is the existence of impulsive behavior for certain initial conditions. Although for *consistent* initial conditions the system behaves like a regular linear system, initial conditions that are not consistent lead to impulsive behavior by which the state is instantaneously transferred to the set of consistent initial conditions. A familiar example is the sparking that often occurs when two electrical subsystems are suddenly connected together.

In the present paper we study the matrix second-order equation $M\ddot{q}+C\dot{q}+Kq = 0$, where M, C, and K denote nonnegative-definite mass, damping, and stiffness matrices, respectively. This equation represents a special case of a singular system when the mass matrix M is singular. A second-order system with singular mass matrix may arise from a singular perturbation problem [4] or may represent a large scale system with algebraic constraints placed on the state variables of the component subsystems. Our goal is to investigate the properties of this second-order equation in the case in which M is singular.

* Received by the editors May 20, 1994; accepted for publication (in revised form) by C. Meyer October 3, 1995. This research was supported in part by Air Force Office of Scientific Research grant F49620-95-1-0019.

† Department of Aerospace Engineering, University of Michigan, Ann Arbor, MI 48109-2118 (dsbaero@engin.umich.edu).

In applications it is often the case that M is not singular but rather contains terms that are numerically small. It is then standard engineering practice to assume that these terms are zero, in which case M is singular. However, if the initial conditions are restricted to lie in the set of consistent initial conditions, then the behavior of the system is governed by a regular system of linear differential equations of reduced dimension. This is the idea behind Guyan reduction [5], which is a model-reduction technique widely used for analyzing structural vibrations of large undamped mechanical systems. Certain finite element modeling techniques involving massless coordinates may also give rise to second-order models with singular mass matrices [6, pp. 107–109]. Although numerical simulations of such systems can be problematic because of the impulsive behavior of the model, such problems can be avoided by restricting the initial conditions appropriately. Singular mass matrices also arise in nonlinear multi-degree-of-freedom mechanical systems [7].

The purpose of this paper is to determine the set of consistent initial conditions for matrix second-order systems with nonnegative-definite mass, damping, and stiffness matrices and to construct a reduced model for such systems when the states are restricted to lie in this set. These results are obtained by specializing known results relating to singular systems to the second-order case. It is shown that a second-order system can be decomposed (i.e., partitioned) into three coupled subsystems of equations—the first is algebraic, the second is a regular system of first-order differential equations, and the third is a regular system of second-order differential equations. This result is used to obtain an extension of Guyan reduction to include viscous damping.

2. Preliminaries. We begin by introducing some definitions concerning the linear system

$$E\dot{x}(t) = Ax(t), \quad x(0) = c, \tag{1}$$

where $t \geq 0$, $x(t) \in \mathcal{R}^n$, E, $A \in \mathcal{R}^{n\times n}$, and where E may be singular. In the definitions to follow, a solution is assumed to be analytic. In general, the singular system (1) admits nonanalytic solutions in the form of distributions [8].

A vector $c \in \mathcal{R}^n$ is a *consistent initial condition* if the initial value problem (1) possesses at least one solution. It is easy to see that the set of consistent initial conditions of (1) is a linear subspace. The system (1) is *tractable* if the initial value problem (1) possesses exactly one solution for every consistent initial condition c. The following proposition, which is stated and proved as Theorem 9.2.1 in [9], gives a necessary and sufficient condition for (1) to be tractable.

PROPOSITION 1. *The system* (1) *is tractable if and only if there exists* $\lambda \in \mathcal{C}$ *such that* $\operatorname{rank}(\lambda E - A) = n$.

If $\lambda \in \mathcal{C}$ and $\operatorname{rank}(\lambda E - A) = n$, then we define $\hat{E}_\lambda \triangleq (\lambda E - A)^{-1}E$. Recall that the *index* of a matrix A, denoted by $\operatorname{ind} A$, is the smallest nonnegative integer k such that $\operatorname{rank} A^k = \operatorname{rank} A^{k+1}$. The following lemma gives some properties of $\hat{E}_\lambda$ that are independent of λ whenever $\hat{E}_\lambda$ is defined. This lemma is stated and proved as Theorem 9.2.2 in [9].

LEMMA 1. *Suppose* $\lambda_1, \lambda_2 \in \mathcal{C}$ *satisfy* $\operatorname{rank}(\lambda_1 E - A) = \operatorname{rank}(\lambda_2 E - A) = n$. *Then* $\operatorname{ind} \hat{E}_{\lambda_1} = \operatorname{ind} \hat{E}_{\lambda_2}$ *and* $\mathcal{R}(\hat{E}_{\lambda_1}^k) = \mathcal{R}(\hat{E}_{\lambda_2}^k)$, *where* $k = \operatorname{ind} \hat{E}_{\lambda_1}$.

The following proposition, which follows from Theorem 9.2.3 of [9], characterizes the set of consistent initial conditions of (1).

PROPOSITION 2. *Suppose that* (1) *is tractable, let* $\lambda \in \mathcal{C}$ *be such that* $\operatorname{rank}(\lambda E -$

$A) = n$, *and let* $k = \operatorname{ind} \hat{E}_\lambda$. *Then the set of consistent initial conditions of* (1) *is given by* $\mathcal{R}(\hat{E}_\lambda^k)$.

3. Second-order systems with singular mass matrix. In this section, the results stated in the previous section are specialized to the matrix second-order system

$$M\ddot{q} + C\dot{q} + Kq = 0, \tag{2}$$

where $q \in \mathcal{R}^r$ and M, C, $K \in \mathcal{R}^{r\times r}$ denote symmetric nonnegative-definite mass, damping, and stiffness matrices, respectively. This system can be rewritten in the first-order form (1) by defining

$$x \triangleq \begin{bmatrix} q \\ \dot{q} \end{bmatrix}, \quad E \triangleq \begin{bmatrix} I & 0 \\ 0 & M \end{bmatrix}, \quad A \triangleq \begin{bmatrix} 0 & I \\ -K & -C \end{bmatrix}.$$

Note that if the mass matrix M is singular then E is also singular.

Before proceeding further, we state the following useful lemma.

LEMMA 2. *Suppose that* P, $Q \in \mathcal{R}^{r\times r}$ *and* $P \geq 0$ *and* $Q \geq 0$. *Then* $\mathcal{N}(P+Q) = \mathcal{N}(P) \cap \mathcal{N}(Q)$ *and*

$$\operatorname{rank}(P+Q) = \operatorname{rank} \begin{bmatrix} P \\ Q \end{bmatrix}.$$

The following theorem is an application of Proposition 1 to (2).

THEOREM 1. *The system* (2) *is tractable if and only if* $M + C + K > 0$. *In this case,* $E - A$ *is invertible.*

Proof. Since

$$\lambda E - A = \begin{bmatrix} I & 0 \\ -(\lambda M + C) & I \end{bmatrix} \begin{bmatrix} 0 & -I \\ \lambda^2 M + \lambda C + K & 0 \end{bmatrix} \begin{bmatrix} I & 0 \\ -\lambda I & I \end{bmatrix},$$

it follows that

$$\operatorname{rank}(\lambda E - A) = r + \operatorname{rank}(\lambda^2 M + \lambda C + K). \tag{3}$$

If $M + C + K > 0$, then $\operatorname{rank}(\lambda^2 M + \lambda C + K) = r$ for $\lambda = 1$. The result now follows from (3) and Proposition 1.

Conversely, suppose there exists nonzero $x \in \mathcal{R}^r$ such that $(M + C + K)x = 0$. Then, since M, C, and K are nonnegative definite, it follows from Lemma 1 that $Mx = Cx = Kx = 0$. Thus $(\lambda^2 M + \lambda C + K)x = 0$ for every $\lambda \in \mathcal{C}$. Consequently, (3) implies that $\operatorname{rank}(\lambda E - A) < 2r$ for every $\lambda \in \mathcal{C}$. It now follows from Proposition 1 that (2) is not tractable, as required.

If $M + C + K > 0$, then it follows from (3) that $\operatorname{rank}(E - A) = r + \operatorname{rank}(M + C + K) = 2r$ so that $E - A$ is invertible. □

Since we are interested only in systems possessing unique solutions, we shall assume that $M + C + K > 0$ throughout the rest of this paper. In this case, it follows from Theorem 1 that the matrix $E - A$ is invertible. We define $\hat{M} \triangleq M + C + K$ and $\hat{E} \triangleq (E - A)^{-1}E$. Note that

$$\hat{E} = \begin{bmatrix} \hat{M}^{-1}(M + C) & \hat{M}^{-1}M \\ -\hat{M}^{-1}K & \hat{M}^{-1}M \end{bmatrix}.$$

The following lemma gives a few properties of $\hat{E}$.

LEMMA 3. *The matrix* $\hat{E}$ *satisfies* $\operatorname{rank} \hat{E}^2 = \operatorname{rank} M + \operatorname{rank}(M + C)$. *Furthermore, the following statements are valid.*

i) *ind* $\hat{E} \leq 2$.

ii) *ind* $\hat{E} \leq 1$ *if and only if* $M + C > 0$.

iii) *ind* $\hat{E} = 0$ *if and only if* $M > 0$.

Proof. Let

$$y_k = \begin{bmatrix} y_{k1} \\ y_{k2} \end{bmatrix}$$

and $y_{k+1} \triangleq \hat{E}^{k+1} y_k$ for $k = 0, 1, 2$, where $y_{k1}, y_{k2} \in \mathcal{R}^r$ for $k = 0, 1, 2, 3$. Now, suppose $\hat{E}^3 y_0 = \hat{E}^2 y_1 = \hat{E} y_2 = y_3 = 0$. Then $E y_2 = (E - A) y_3 = 0$, that is,

$$y_{21} = \hat{M}^{-1}[(M + C) y_{11} + M y_{12}] = 0 \tag{4}$$

and

$$M y_{22} = M \hat{M}^{-1}(-K y_{11} + M y_{12}) = 0. \tag{5}$$

Therefore $0 = M y_{21} - M y_{22} = M y_{11}$. Premultiplying (4) by $y_{11}^{\mathrm{T}} \hat{M}$ yields $y_{11}^{\mathrm{T}} C y_{11} = 0$. Since C is nonnegative definite, it follows that $C y_{11} = 0$. Using $(M + C) y_{11} = 0$ in (4) gives $M y_{12} = 0$. Thus

$$M y_{11} = C y_{11} = M y_{12} = 0. \tag{6}$$

Note that in deriving (6), no use was made of the fact that $y_1 = \hat{E} y_0$. Thus it is true in general that $\hat{E}^2 y_1 = 0$ implies

$$\begin{bmatrix} C & M \\ M & 0 \end{bmatrix} y_1 = 0.$$

The converse can be easily verified. Thus

$$\mathcal{N}(\hat{E}^2) = \mathcal{N} \begin{bmatrix} C & M \\ M & 0 \end{bmatrix}.$$

Hence

$$\operatorname{rank} \hat{E}^2 = \operatorname{rank} M + \operatorname{rank} \begin{bmatrix} C \\ M \end{bmatrix}.$$

Since M and C are nonnegative definite, it follows from Lemma 1 that $\operatorname{rank} \hat{E}^2 = \operatorname{rank} M + \operatorname{rank}(M + C)$.

To prove i) it suffices to show that $\mathcal{N}(\hat{E}^3) \subseteq \mathcal{N}(\hat{E}^2)$. Using (6) we compute $y_{11}^{\mathrm{T}} \hat{M} y_{11} = y_{01}^{\mathrm{T}}(M + C) y_{11} + y_{02}^{\mathrm{T}} M y_{11} = 0$. Since $\hat{M}$ is positive definite, it follows that $y_{11} = 0$. This together with (6) implies that $y_{22} = \hat{M}^{-1}(-K y_{11} + M y_{12}) = 0$. It now follows from (4) that $y_2 = 0$. Thus $\hat{E}^3 y = 0$ implies that $\hat{E}^2 y = 0$. This proves i).

To prove ii) note that the index of $\hat{E}$ is less than 2 if and only if $\operatorname{rank} \hat{E}^2 = \operatorname{rank} \hat{E}$. Since $\operatorname{rank} \hat{E}^2 = \operatorname{rank} M + \operatorname{rank}(M + C)$ and $\operatorname{rank} \hat{E} = \operatorname{rank} E = \operatorname{rank} M + r$, it follows that $\operatorname{ind} \hat{E} < 2$ if and only if $M + C$ is positive definite.

Finally, $\operatorname{ind} \hat{E} = 0$ if and only if $\operatorname{rank} \hat{E} = 2r$. Since $\operatorname{rank} \hat{E} = \operatorname{rank} E = r + \operatorname{rank} M$, it follows that $\operatorname{rank} \hat{E} = 2r$ if and only if M is positive definite. □

The following theorem uses Lemma 3 to determine the set of consistent initial conditions of (2).

THEOREM 2. *The set of consistent initial conditions of* (2) *is given by* $\mathcal{R}(\hat{E}^2)$. *Furthermore, if* $M+C>0$, *then the set of consistent initial conditions of* (2) *is given by* $\mathcal{R}(\hat{E})$.

Proof. The results follow from Proposition 2 and Lemma 3. □

The second part of Theorem 2 is a special case of Proposition 8.2.1 in [10].

COROLLARY 1. *The dimension of the subspace of consistent initial conditions of* (2) *is* rank M + rank$(M+C)$.

Proof. The result follows from Theorem 2 and Lemma 3. □

4. Model reduction. In this section it is shown that the second-order system (2) can be decomposed into a system of algebraic equations, a regular first-order system of differential equations, and a regular second-order system of differential equations. It is also shown that under special assumptions the algebraic subsystem can be eliminated to obtain a regular second-order system having fewer degrees of freedom.

For convenience, define $r_1 \triangleq \operatorname{def}(M+C)$, $r_2 \triangleq \operatorname{def} M - r_1$, and $r_3 \triangleq \operatorname{rank} M$. Note that $r_1+r_2+r_3=r$. It can be seen from Corollary 1 that the subspace of consistent initial conditions has dimension $2r_3+r_2$. In this section we assume that M is singular but nonzero, in which case $r_1+r_2>0$ and $r_3>0$.

THEOREM 3. i) *Suppose* $\operatorname{def} M > \operatorname{def}(M+C) > 0$. *Then there exists an orthogonal matrix* $U \in \mathcal{R}^{r\times r}$ *such that*

$$U^{\mathrm{T}}MU = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & M_3 \end{bmatrix}, \quad U^{\mathrm{T}}CU = \begin{bmatrix} 0 & 0 & 0 \\ 0 & C_2 & C_{23} \\ 0 & C_{23}^{\mathrm{T}} & C_3 \end{bmatrix},$$

$$U^{\mathrm{T}}KU = \begin{bmatrix} K_1 & K_{12} & K_{13} \\ K_{12}^{\mathrm{T}} & K_2 & K_{23} \\ K_{13}^{\mathrm{T}} & K_{23}^{\mathrm{T}} & K_3 \end{bmatrix}, \tag{7}$$

where $M_3 \in \mathcal{R}^{r_3\times r_3}$, $C_2 \in \mathcal{R}^{r_2\times r_2}$, *and* $K_1 \in \mathcal{R}^{r_1\times r_1}$ *are positive definite. Furthermore,* $K_{12}=0$, $K_2=0$, *and* $K_{23}=0$ *if and only if* $\mathcal{N}(M+C) \perp \mathcal{N}(M+K)$ *and* $\mathcal{N}(M)=\mathcal{N}(M+C)\oplus\mathcal{N}(M+K)$.

ii) *Suppose* $\operatorname{def} M = \operatorname{def}(M+C) > 0$. *Then there exists an orthogonal matrix* $U \in \mathcal{R}^{r\times r}$ *such that*

$$U^{\mathrm{T}}MU = \begin{bmatrix} 0 & 0 \\ 0 & M_2 \end{bmatrix}, \quad U^{\mathrm{T}}CU = \begin{bmatrix} 0 & 0 \\ 0 & C_2 \end{bmatrix}, \quad U^{\mathrm{T}}KU = \begin{bmatrix} K_1 & K_{12} \\ K_{12}^{\mathrm{T}} & K_2 \end{bmatrix}, \tag{8}$$

where $M_2 \in \mathcal{R}^{r_3\times r_3}$ *and* $K_1 \in \mathcal{R}^{r_1\times r_1}$ *are positive definite.*

iii) *Suppose* $\operatorname{def} M > \operatorname{def}(M+C) = 0$. *Then there exists an orthogonal matrix* $U \in \mathcal{R}^{r\times r}$ *such that*

$$U^{\mathrm{T}}MU = \begin{bmatrix} 0 & 0 \\ 0 & M_2 \end{bmatrix}, \quad U^{\mathrm{T}}CU = \begin{bmatrix} C_1 & C_{12} \\ C_{12}^{\mathrm{T}} & C_2 \end{bmatrix}, \tag{9}$$

where $M_2 \in \mathcal{R}^{r_3\times r_3}$ *and* $C_1 \in \mathcal{R}^{r_2\times r_2}$ *are positive definite.*

Proof. i) In this case, $r_1>0$ and $r_2>0$. Let $x_1, x_2, \ldots, x_r$ be an orthonormal basis for $\mathcal{R}^r$ such that $x_1, x_2, \ldots, x_{r_1+r_2}$ is an orthonormal basis for $\mathcal{N}(M)$ and $x_1, x_2, \ldots, x_{r_1}$ is an orthonormal basis for $\mathcal{N}(M+C)$. Let $U =$

$[\ x_1\ \ x_2\ \ \dots\ \ x_r\ ]$. Then it can easily be verified that $U^{\mathrm{T}}U = I$. Note that for every matrix P, $(U^{\mathrm{T}}PU)_{ij} = x_i^{\mathrm{T}}Px_j$. The sizes and placement of the zero subblocks in $U^{\mathrm{T}}MU$ and $U^{\mathrm{T}}CU$ now follow from the choice of the vectors $x_1,\ x_2, \dots,\ x_r$. Since M_3, C_2, and K_1 are principal submatrices of the nonnegative-definite matrices $U^{\mathrm{T}}MU$, $U^{\mathrm{T}}CU$, and $U^{\mathrm{T}}KU$, respectively, it follows that M_3, C_2, and K_1 are nonnegative definite. Now rank M_3 = rank $M = r_3$, which is also the dimension of M_3. Hence $M_3 > 0$. To show that $C_2 > 0$, suppose that $C_2y_2 = 0$ for some $y_2 \in \mathcal{R}^{r_2}$. Then $y_2^{\mathrm{T}}C_2y_2 = z^{\mathrm{T}}Cz = 0$, where $z = U[\ 0\ \ y_2^{\mathrm{T}}\ \ 0\ ]^{\mathrm{T}}$. The nonnegative definiteness of C leads to $Cz = 0$. Also $U^{\mathrm{T}}Mz = 0$. Thus $(M+C)z = 0$. By construction, every vector in $\mathcal{N}(M + C)$ is of the form $U[\ y_1^{\mathrm{T}}\ \ 0\ \ 0\ ]^{\mathrm{T}}$, where $y_1 \in \mathcal{R}^{r_1}$. Therefore, $y_2 = 0$ and hence $C_2 > 0$. Finally, K_1 is a principal submatrix of the positive-definite matrix $U^{\mathrm{T}}(M + C + K)U$ and hence positive definite. This proves the first part of i).

If $\mathcal{N}(M + C) \perp \mathcal{N}(M + K)$ and $\mathcal{N}(M) = \mathcal{N}(M + C) \oplus \mathcal{N}(M + K)$, then the vectors $x_{r_1+1},\ x_{r_1+2}, \dots,\ x_{r_1+r_2}$ form a basis for $\mathcal{N}(M + K)$. By Lemma 1, these vectors also lie in $\mathcal{N}(K)$. Since every element of K_{12}, K_2, and K_{23} is of the form $x_i^{\mathrm{T}}Kx_j$, where either $r_1 + 1 \le i \le r_1 + r_2$ or $r_1 + 1 \le j \le r_1 + r_2$, it follows that $K_{12} = 0$, $K_2 = 0$, and $K_{23} = 0$. If $K_{12} = 0$, $K_2 = 0$, and $K_{23} = 0$, then since M_3 and K_1 are positive definite, it follows that $\mathcal{N}(M + K)$ consists of vectors of the form $z = U[\ 0\ \ y_2^{\mathrm{T}}\ \ 0\ ]^{\mathrm{T}}$, where $y_2 \in \mathcal{R}^{r_2}$. Thus the vectors $x_{r_1+1},\ x_{r_1+2}, \dots,\ x_{r_1+r_2}$ form a basis for $\mathcal{N}(M + K)$ and the result follows.

The proofs of ii) and iii) are similar. □

Theorem 3 gives conditions under which M, C, and K may be assumed without loss of generality to be of the form given by (7). Note that the first r_1 equations are algebraic while the remaining equations represent a regular first-order system of dimension r_2 coupled with a regular r_3-degree-of-freedom second-order system. The following corollary shows that under special assumptions the algebraic equations can be eliminated to obtain a regular second-order system with a reduced number of degrees of freedom.

COROLLARY 2. *Suppose* def $M >$ def$(M+C) > 0$ *and assume that* $\mathcal{N}(M+C) \perp \mathcal{N}(M + K)$ *and* $\mathcal{N}(M) = \mathcal{N}(M + C) \oplus \mathcal{N}(M + K)$. *Then there exists a matrix* $S \in \mathcal{R}^{r\times r_3}$ *such that* $S^{\mathrm{T}}MS > 0$, $S^{\mathrm{T}}CS \ge 0$, *and* $S^{\mathrm{T}}KS \ge 0$.

Proof. Under the stated assumptions, there exists a matrix $U \in \mathcal{R}^{r\times r}$ such that $U^{\mathrm{T}}MU$, $U^{\mathrm{T}}CU$, and $U^{\mathrm{T}}KU$ are given by (7) with $K_{12} = 0$, $K_2 = 0$, and $K_{23} = 0$. Define $S \in \mathcal{R}^{r\times r_3}$ by

$$S = U \begin{bmatrix} -K_1^{-1}K_{13} \\ -C_2^{-1}C_{23} \\ I \end{bmatrix}.$$

Then $S^{\mathrm{T}}MS = M_3$ is positive definite by Theorem 3, and $S^{\mathrm{T}}CS$ and $S^{\mathrm{T}}KS$ are nonnegative definite since C and K are nonnegative definite. □

The following corollary gives another case in which a reduction in the number of degrees of freedom can be achieved.

COROLLARY 3. *Suppose* def $M =$ def$(M + C) > 0$. *Then there exists a matrix* $S \in \mathcal{R}^{r\times r_3}$ *such that* $S^{\mathrm{T}}MS > 0$, $S^{\mathrm{T}}CS \ge 0$, *and* $S^{\mathrm{T}}KS \ge 0$.

Proof. Since def $M =$ def$(M+C) > 0$, there exists a matrix $U \in \mathcal{R}^{r\times r}$ such that $U^{\mathrm{T}}MU$, $U^{\mathrm{T}}CU$ and $U^{\mathrm{T}}KU$ are as given by (8). Define $S \in \mathcal{R}^{r\times r_3}$ by

$$S = U \begin{bmatrix} -K_1^{-1}K_{12} \\ I \end{bmatrix}.$$

Then $S^{\mathrm{T}}MS = M_2$ is positive definite by ii) of Theorem 3. Finally, $S^{\mathrm{T}}CS$ and $S^{\mathrm{T}}KS$ are nonnegative definite since C and K are nonnegative definite. □

Remark. The matrix S in Corollaries 2 and 3 gives the transformation that reduces the r-degree-of-freedom system (2) to a regular second-order system having fewer (r_3) degrees of freedom. It is worth pointing out that if the conditions of Corollary 2 are satisfied, then the dimension of the state–space of the reduced system ($2r_3$) is less than the dimension of the subspace of consistent initial conditions ($2r_3 + r_2$). In this case, the reduced system does not give solutions to all possible consistent initial conditions of the full system (2). However, it is often the case in applications that only the response of the reduced system is of interest. This response is completely determined by the initial conditions in the reduced state–space. This point will be illustrated in the examples. Finally, it should be noted that if $C = 0$, then Corollary 3 reduces to the well-known Guyan reduction.

5. Examples. In this section, we present two examples to illustrate Theorem 3 and Corollary 2.

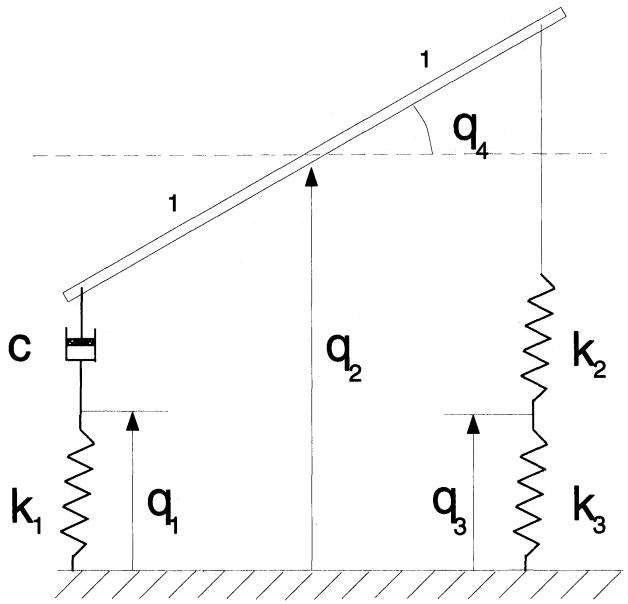


FIG. 1.

Example 1. To illustrate Theorem 3, consider Figure 1, which shows a uniform rod of length 2 units having mass m and moment of inertia J about its center of mass. The motion of the rod takes place under the action of linear springs with positive spring constants k_1, k_2, and k_3 and a linear viscous damper with positive damping coefficient c as shown. Assuming small motions, the unforced motion of this system is governed by (2) with

$$M = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & m & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & J \end{bmatrix}, \quad C = \begin{bmatrix} c & -c & 0 & c \\ -c & c & 0 & -c \\ 0 & 0 & 0 & 0 \\ c & -c & 0 & c \end{bmatrix},$$

$$K = \begin{bmatrix} k_1 & 0 & 0 & 0 \\ 0 & k_2 & -k_2 & k_2 \\ 0 & -k_2 & k_2 + k_3 & -k_2 \\ 0 & k_2 & -k_2 & k_2 \end{bmatrix},$$

and $q = [\, q_1 \; q_2 \; q_3 \; q_4 \,]^{\mathrm{T}}$. For this system $M + C + K > 0$, $r_1 = r_2 = 1$, and $r_3 = 2$. Letting U be given by

$$U = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

so that $U^{\mathrm{T}}U = I$, it follows that

$$U^{\mathrm{T}}MU = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & m & 0 \\ 0 & 0 & 0 & J \end{bmatrix}, \quad U^{\mathrm{T}}CU = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & c & -c & c \\ 0 & -c & c & -c \\ 0 & c & -c & c \end{bmatrix},$$

$$U^{\mathrm{T}}KU = \begin{bmatrix} k_2 + k_3 & 0 & -k_2 & -k_2 \\ 0 & k_1 & 0 & 0 \\ -k_2 & 0 & k_2 & k_2 \\ -k_2 & 0 & k_2 & k_2 \end{bmatrix}.$$

This decomposition illustrates i) in Theorem 3.

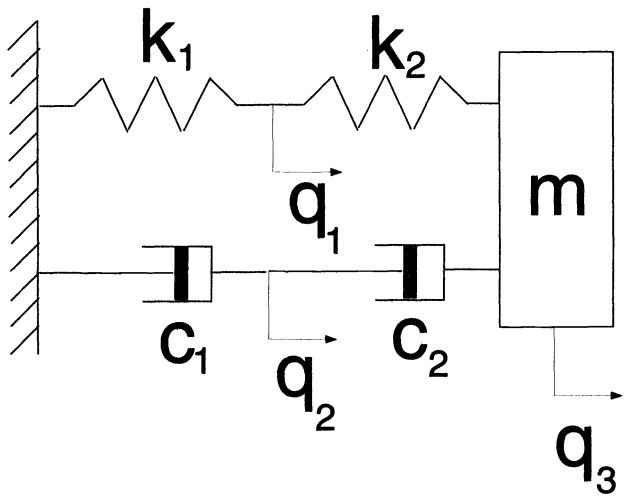


FIG. 2.

Example 2. Consider the lumped-parameter system shown in Figure 2 consisting of a mass m with displacement q_3, linear springs with positive spring constants k_1 and k_2, and linear viscous dampers with positive damping coefficients c_1 and c_2. The massless joint between the dampers c_1 and c_2 has a displacement q_1, while the massless joint between the springs k_1 and k_2 has a displacement q_2. The equations of motion

for this system can be written in the form (2) with

$$M = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & m \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 0 & 0 \\ 0 & c_1 + c_2 & -c_2 \\ 0 & -c_2 & c_2 \end{bmatrix},$$

$$K = \begin{bmatrix} k_1 + k_2 & 0 & -k_2 \\ 0 & 0 & 0 \\ -k_2 & 0 & k_2 \end{bmatrix}, \quad q = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix}.$$

It can easily be verified that $\mathcal{N}(M) = \text{span}\{[1\ 0\ 0]^{\mathrm{T}}, [0\ 1\ 0]^{\mathrm{T}}\}$, $\mathcal{N}(M + C) = \text{span}\{[1\ 0\ 0]^{\mathrm{T}}\}$, and $\mathcal{N}(M + K) = \text{span}\{[0\ 1\ 0]^{\mathrm{T}}\}$. Thus the hypotheses of Corollary 2 are satisfied. The matrix S in Corollary 2 is given by $S = [\ \frac{k_2}{k_1+k_2}\ \ \frac{c_2}{c_1+c_2}\ \ 1\ ]^{\mathrm{T}}$. The transformation $q = S\tilde{q}$ reduces (2) to

$$m\ddot{\tilde{q}} + \frac{c_1 c_2}{c_1 + c_2}\dot{\tilde{q}} + \frac{k_1 k_2}{k_1 + k_2}\tilde{q} = 0, \tag{10}$$

whose coefficients are consistent with the well-known formulas for series combinations of springs and dashpots. For this example, the subspace of consistent initial conditions has dimension 3. This follows from Corollary 1 by noting that rank $M = 1$ and $\text{rank}(M + C) = 2$. Thus only three independent quantities need to be specified at the initial instant, specifically, either q_2, q_3, and $\dot{q}_3$ or q_2, $\dot{q}_2$, and q_3. However, $q_3(t) = \tilde{q}(t)$ satisfies the reduced order equation (10) and is completely determined by the initial values of q_3 and $\dot{q}_3$. Consequently, $q_3(t)$ is independent of the initial value of q_2. In physical applications, the displacement of the mass is of primary interest. In such cases, the reduction procedure automatically eliminates the unwanted variable q_2. This illustrates the extension of Guyan reduction to systems with damping.

Acknowledgment. We wish to thank William Anderson and an anonymous reviewer for several helpful comments.

REFERENCES

[1] S. L. Campbell and N. J. Rose, *Singular perturbation of autonomous linear systems*, SIAM J. Math. Anal., 10 (1979), pp. 542–551.

[2] S. L. Campbell, *Optimal control of autonomous linear processes with singular matrices in the quadratic cost functional*, SIAM J. Control Optim., 14 (1976), pp. 1092–1106.

[3] P. Bernhard, *On singular implicit linear dynamical systems*, SIAM J. Control Optim., 20 (1982), pp. 612–633.

[4] S. L. Campbell and N. J. Rose, *A second order singular linear system arising in electric power systems analysis*, Internat. J. Systems Sci., 13 (1982), pp. 101–108.

[5] R. J. Guyan, *Reduction of stiffness and mass matrices*, AIAA J., 3 (1965), p. 380.

[6] J. Argyris and H. Mlejnek, *Dynamics of Structures*, North–Holland, Amsterdam, 1991.

[7] S. K. Agrawal, *Inertia matrix singularity of series-chain spatial manipulators with point masses*, J. Dynamic Systems, Measurement, Control, 115 (1993), pp. 723–725.

[8] D. Cobb, *On the solutions of linear differential equations with singular coefficients*, J. Differential Equations, 46 (1982), pp. 310–323.

[9] S. L. Campbell and C. D. Meyer Jr., *Generalized Inverses of Linear Transformations*, Pitman, Boston, MA, 1979.

[10] S. L. Campbell, *Singular Systems of Differential Equations*, Pitman, Boston, MA, 1980.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 3, pp. 658–679, July 1996

016

MULTIFRONTAL COMPUTATION WITH THE ORTHOGONAL FACTORS OF SPARSE MATRICES*

SZU-MIN LU† AND JESSE L. BARLOW†

Abstract. This paper studies the solution of the linear least squares problem for a large and sparse m by n matrix A with $m \geq n$ by QR factorization of A and transformation of the right-hand side vector b to $Q^T b$. A multifrontal-based method for computing $Q^T b$ using Householder factorization is presented. A theoretical operation count for the K by K unbordered grid model problem and problems defined on graphs with $\sqrt{n}$-separators shows that the proposed method requires $O(N_R)$ storage and multiplications to compute $Q^T b$, where $N_R = O(n \log n)$ is the number of nonzeros of the upper triangular factor R of A. In order to introduce BLAS-2 operations, Schreiber and Van Loan's storage-efficient WY representation [*SIAM J. Sci. Stat. Comput.*, 10 (1989), pp. 53–57] is applied for the orthogonal factor Q_i of each frontal matrix F_i. If this technique is used, the bound on storage increases to $O(n(\log n)^2)$. Some numerical results for the grid model problems as well as Harwell–Boeing problems are provided.

Key words. multifrontal QR factorization, $\sqrt{n}$-separable graphs, Householder matrices

AMS subject classifications. 15A23, 65F50

1. Introduction.

We study the linear least squares problem

$$\min_x ||Ax - b||_2, \tag{1}$$

where x is a real n-vector and b is a real m-vector. Orthogonal factorization is often used in methods for solving the linear least squares and eigenvalue problems. Let A be an m by n large sparse matrix of full column rank with $m \geq n$. The QR factorization of A is $A = Q\binom{R}{0}$, where R is an n by n upper triangular matrix and Q is an m by m orthogonal matrix. We apply QR factorization of A and transform the right-hand side vector b to $Q^T b$, as follows.

$$\begin{aligned}\min_x ||Ax - b||_2^2 &= \left\| \begin{pmatrix} R \\ 0 \end{pmatrix} x - Q^T b \right\|_2^2 \\ &= \left\| \begin{pmatrix} R \\ 0 \end{pmatrix} x - \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \right\|_2^2 \\ &= ||Rx - c_1||_2^2 + ||c_2||_2^2.\end{aligned}$$

The least squares solution is given by solving

$$Rx = c_1.$$

The problem is that if b is not known in advance or if we have more than one b, we need to save the orthogonal matrix Q. Unfortunately, Q is often larger and much denser than the factor R. Instead of storing the first n columns of Q, one often stores the orthogonal factor Q implicitly, as follows. Let $A = H_1 H_2 \cdots H_n R$. The orthogonal matrix Q is then expressed as

$$Q = H_1 H_2 \cdots H_n,$$

* Received by the editors December 9, 1993; accepted for publication (in revised form) by J. W. H. Liu October 3, 1995. The authors' research was supported by National Science Foundation grant CCR-9201612.

† Computer Science and Engineering Department, Pennsylvania State University, University Park, PA 16802 (lu@cse.psu.edu, barlow@cse.psu.edu).

where $H_i = I - h_i h_i^T$ is a Householder reflection that zeros out column i of A below the main diagonal. The vector h_i is often referred to as a Householder vector and is zero in positions 1 through $i-1$. The orthogonal factor Q can therefore be represented implicitly by the m by n lower trapezoidal matrix H:

$$H = (h_1 \; h_2 \; \cdots \; h_n),$$

which is referred to as the Householder matrix. A matrix-vector product $Q^T b$ can be computed efficiently from H and b. The LINPACK routines SQRDC and SQRSL employ H [6]. In [13], Gilbert, Ng, and Peyton analyzed the nonzero counts of the factors Q, R, and H in terms of the sizes of separators in the column intersection graph $G_\cap(A)$ of A, where $G_\cap(A)$ is an undirected graph in which an edge joins two vertices whose columns share a nonzero row in A. This graph corresponds to the matrix of the normal equations $A^T A$. If A is such that $G_\cap(A)$ has $\sqrt{n}$-separators for all its subgraphs and if $m - n$ is of the same order as n, then H is smaller than Q only by a constant factor [13]. That is, both $|Q|$, the number of nonzeros in the first n columns of Q, and $|H|$, the number of nonzeros in H, are of $O(n\sqrt{n})$. Moreover, the difference between $|Q|$ and $|H|$ is likely to be relatively small if m is much larger than n. Other results on the nonzero structures of the Householder matrix H and the orthogonal factor Q for a sparse matrix are given in [11, 22]. In this paper, we study the computation of orthogonal factors using the multifrontal QR factorization [16, 20]. Associated with each row of the upper triangular factor R is a frontal matrix F_i. Likewise for each F_i, there is a frontal Householder matrix Y_i. Note that Y_i is the H matrix for F_i. Figure 1 is a small sample matrix A and its column intersection graph. Figure 2 is the Householder matrix H of A and the elimination tree of $A^T A$. The frontal Householder matrices Y_i's are given in Figure 3. The size of H is $|H| = 106$, where the sum of sizes of Y_i's is $\sum_{i=1}^{n} |Y_i| = 65$. The results of this paper provide an explanation for this dramatic difference. We are going to present an efficient method for computing $Q^T b$ by using the frontal Householder matrices Y_i's. In addition, this method is suitable for parallel computation because of the special structure of multifrontal matrices [21].

For the theoretical part, we study the K by K grid model problem and problems that are defined on $\sqrt{n}$-separable graphs under one assumption for the initial step. We are going to describe these problems in §2. An $O(n \log n)$ bound is proven on the number of nonzeros of all frontal Householder matrices Y_i's. We also count the number of nonzeros used in the WY representations of Bischof and Van Loan [4] and Schreiber and Van Loan [24]. We prove that the bounds for the K by K grid model problem and problems that are defined on $\sqrt{n}$-separable graphs are $O(n \log n)$ and $O(n(\log n)^2)$, respectively. Note that these bounds are valid even if $m - n$ is of the same order of n.

The rest of the paper is organized as follows. Section 2 briefly reviews the model problem and the $\sqrt{n}$-separator problems, the multifrontal Householder QR factorization, and the application of supernodes. Section 3 proposes a multifrontal-based method for computing $Q^T b$. Section 4 proves an upper bound on the nonzero counts of all Y_i's. This section builds on the work of Lewis, Pierce, and Wah [16] for the K by K grid problem. We extend their result to the $\sqrt{n}$-separator problem. Section 5 introduces BLAS-2 operations in computing $Q^T b$ by using the YTY representation of Schreiber and Van Loan [24] for the orthogonal factor Q_i of each frontal matrix F_i. The upper bound on operation counts of that representation is also included. In §6, we provide some numerical test results.

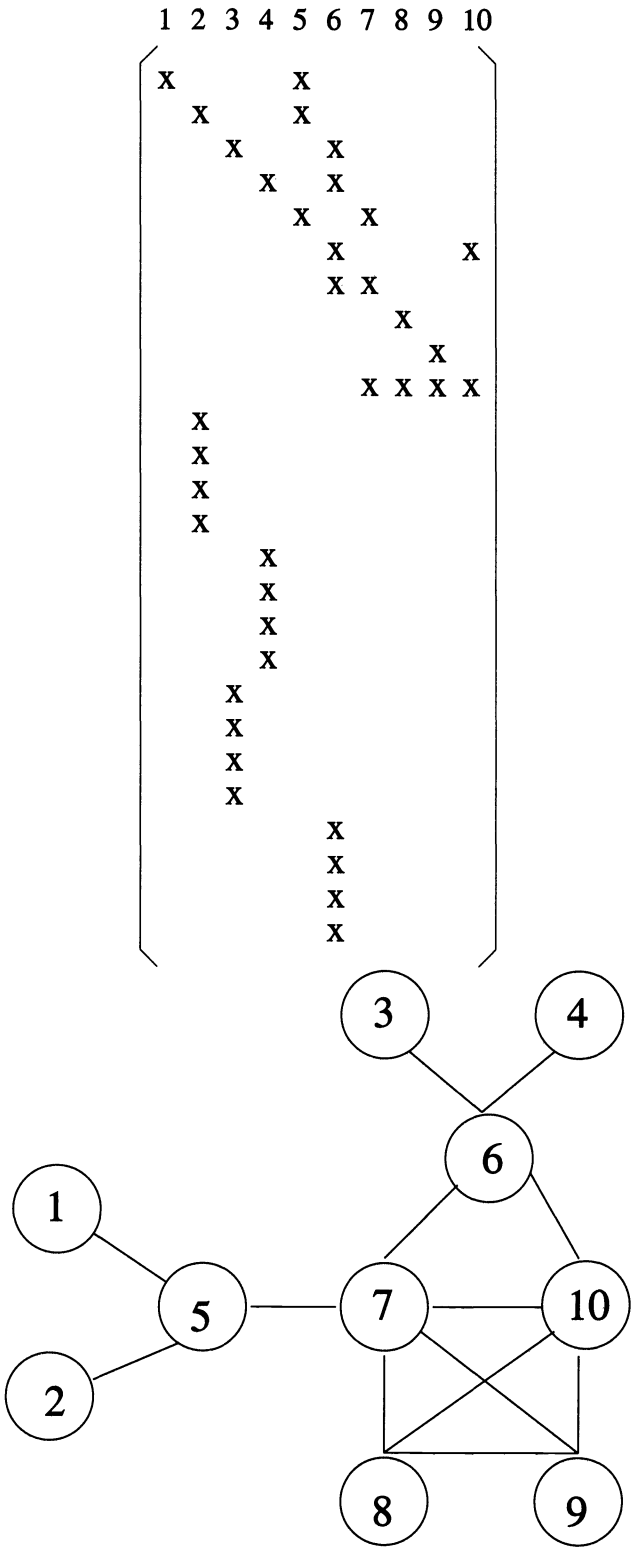


FIG. 1. *A sample matrix A and its column intersection graph.*

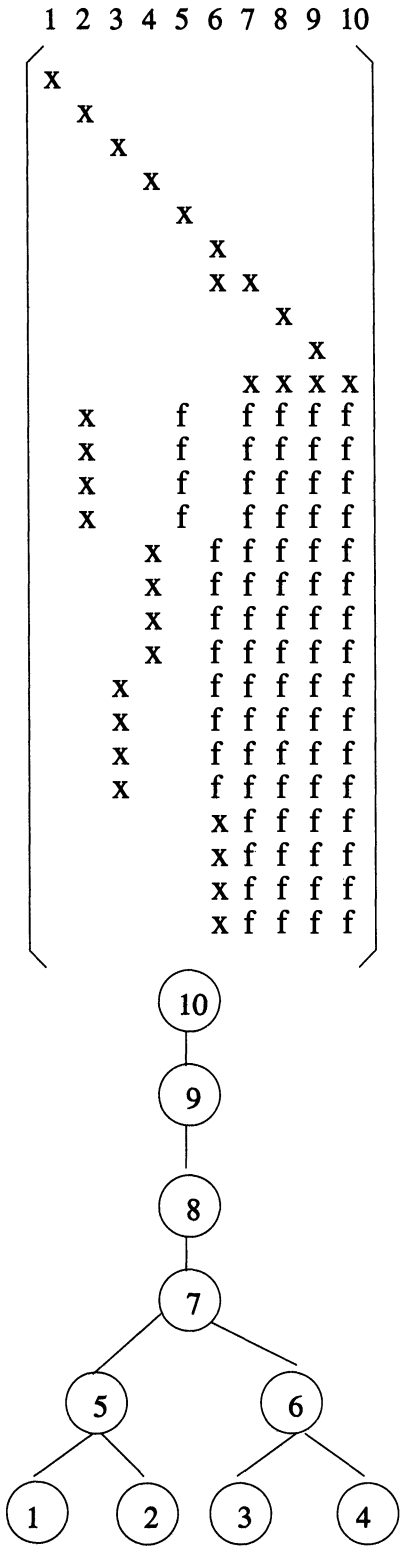


FIG. 2. *The Householder matrix H and the elimination tree for the matrix of Figure* 1.

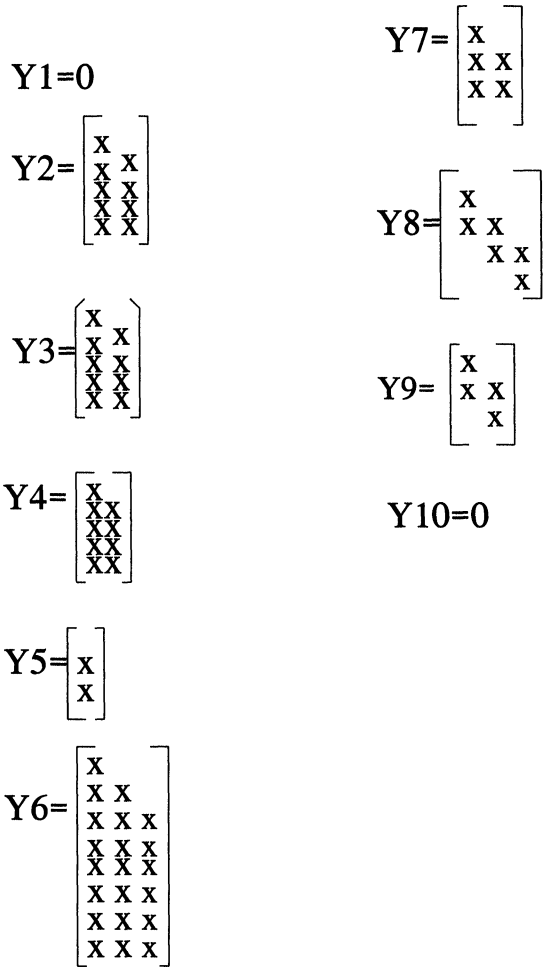


FIG. 3. *The frontal Householder matrices of A.*

2. Background.

2.1. The model problem. Since the sparsity patterns of general sparse matrices are difficult to predict, our theoretical operation counts for sparse matrices are based on the model problem which is described in this section. The model problem is motivated by the finite element method. Consider a K by K regular grid with $(K-1)^2$ small squares. A variable is assigned to each grid point. Associated with each square is a set of s equations involving the four variables at the corners of the square. The assembly of these equations results in a large overdetermined system of equations:

$$Ax = b, \tag{2}$$

where A is m by n with $m = s(K-1)^2$ and $n = K^2$.

In our examples, we let $s = 4$ as in [16]. Figure 4 is an example for a 3 by 3 grid ordered by nested dissection ordering. The corresponding matrix A and the upper triangular matrix R are given in Figure 5.

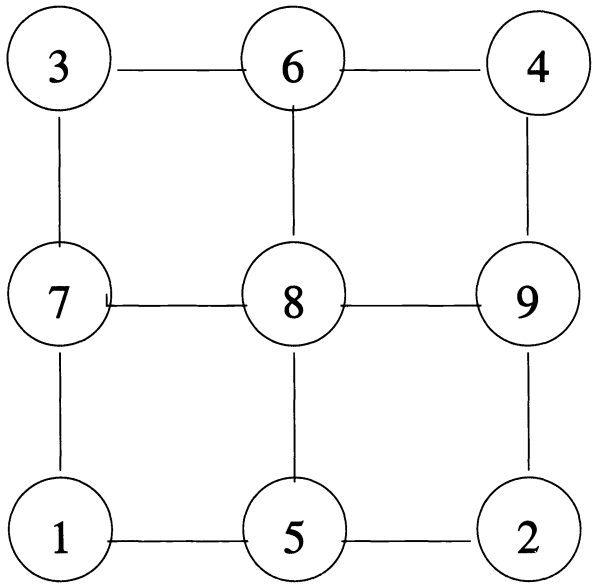


FIG. 4. *A 3 by 3 nested dissection ordered grid.*

2.2. The extended problem. In addition to the model problem, we would like to study more general problems. Consider problems that are defined by graphs. Let S be a class of graphs closed under the subgraph relation. That is, if $G_1 \in S$ and G_2 is a subgraph of G_1, then $G_2 \in S$.

DEFINITION 2.1 ($\sqrt{n}$-separable graph). *A $\sqrt{n}$-separable graph is an n-vertex connected graph $G \in S$ with the following properties: There exist constants $\alpha < 1, \beta > 0$ such that G can be partitioned into three sets A, B, C such that no edge joins a vertex in A with a vertex in B, neither A nor B contains more than αn vertices, and C contains no more than $\beta\sqrt{n}$ vertices.*

Throughout the paper, we refer to the $\sqrt{n}$-separator matrices as the set of matrices whose column intersection graphs are members of the set of $\sqrt{n}$-separable graphs with the constants α and β defined above.

2.3. Multifrontal Householder *QR* factorization method. We describe a multifrontal-based method by [16] for computing the upper triangular factor R of A. The factorization uses the frontal structure inherent in multifrontal Cholesky algorithms and Householder transformations. A theoretical operation count for the model problem which indicates that the factorization algorithm requires half the multiplications as Liu's algorithm [7, 18] is given by Lewis, Pierce, and Wah [16]. We begin this subsection with the definition of the elimination tree.

DEFINITION 2.2 (elimination tree). *Given an m by n matrix A, such that A^TA is irreducible, the elimination tree of A^TA is a tree consisting of n vertices each uniquely labeled by an integer from $1, 2, 3, \ldots, n$. Let R denote the upper triangular factor of the QR factorization of A. Then j is the parent of vertex i in the elimination tree if j is the leading off-diagonal nonzero in the ith row of R.*

Consider the matrix A and the factor R of A given in Figure 5. The corresponding elimination tree is given in Figure 6. The elimination tree is a tool for ordering and organizing the computation in the multifrontal method. In order to compute the ith

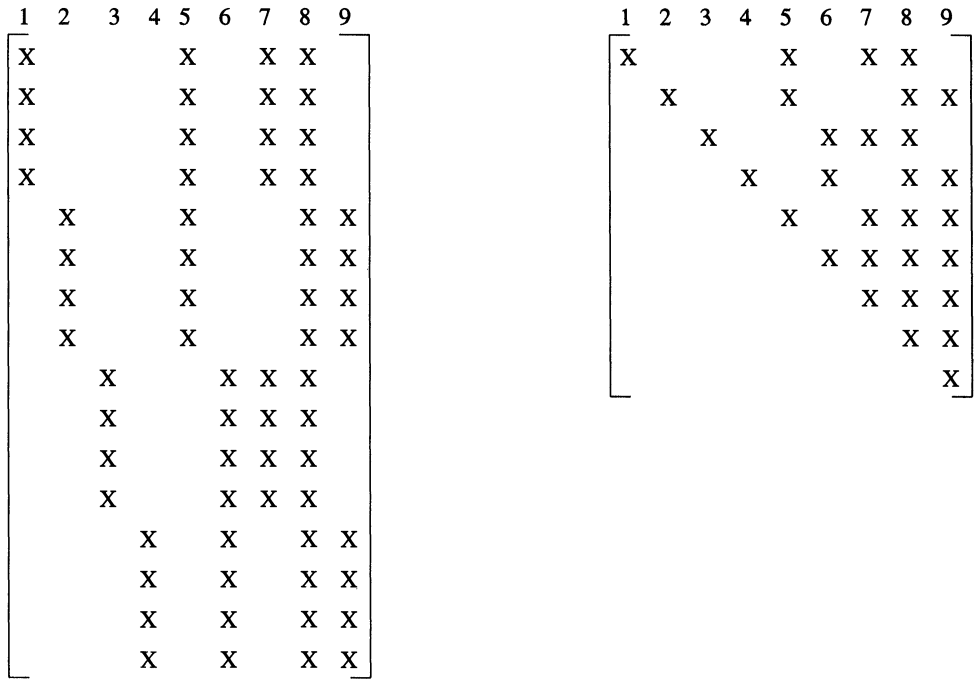


FIG. 5. *A sample matrix A and its upper triangular factor R.*

row of R, all the rows corresponding to node i's descendants in the elimination tree must be computed. That is, row i cannot be computed until its children's rows are computed. The multifrontal QR factorization method uses the elimination tree to determine the required information for forming each frontal matrix. We explain this in detail next.

We begin by defining the following notation.

1. Let i denote node i in the elimination tree as well as the ith column of A.
2. Let $A[i]$ be the matrix whose rows are those rows of A that have their leading nonzeros in column i.

Let j be a leaf in the elimination tree. During each frontal stage, only $A[j]$ contributes to building the frontal matrix F_j. That is, the nonzero structure of the jth row of the upper triangular factor R is completely dependent on $A[j]$. One then computes the QR factorization of F_j, resulting in $Q_j^T F_j = R_j$, where R_j is an upper triangular or usually trapezoidal factor. The first row of R_j corresponds to the jth row of the factor R; the remaining part of R_j is saved as update matrix U_j, which is used by j's parent. Now consider an internal node i. We assemble the frontal matrix F_i by collecting all rows of $A[i]$ and all the update matrices from the children of i. We then compute the QR factorization of F_i, use the first row of the upper triangular factor R_i to fill the ith row of R, and save the update matrix U_i for i's parent. The update matrices can be stored and retrieved in a last-in/first-out (i.e., stack) manner if the nodes of the elimination tree are ordered by a postordering. The use of a stack for update matrices is due to Duff and Reid [8]. Now we outline the multifrontal QR factorization in an algorithm.

ALGORITHM 2.3. **(Multifrontal *QR* Factorization):**
For $j = 1$ **To** *number of tree nodes* **Do**

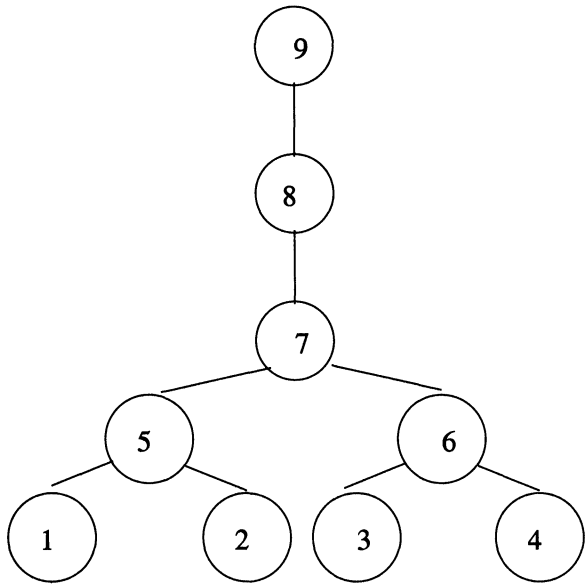


FIG. 6. *The elimination tree of R in Figure* 5.

1. *Assemble the frontal matrix for vertex j, which consists of all rows of A with first nonzero in column j and the update matrices from the children of vertex j.*

$$\begin{pmatrix} \leftarrow A[j] \rightarrow \\ \leftarrow U_{c_1} \rightarrow \\ \vdots \\ \leftarrow U_{c_s} \rightarrow \end{pmatrix},$$

where the children of vertex j are vertices $c_1, \ldots, c_s$.

2. *Compute the QR factorization of the frontal matrix such that*

$$Q_j^T F_j = R_j = \begin{pmatrix} r_{jj} & r_{jj_1} & \cdots & r_{jj_t} \\ 0 & U_j & & \\ \vdots & & & \\ 0 & & & \end{pmatrix}.$$

3. *Save the first row of* R_j, $(r_{jj}, r_{jj_1}, \ldots, r_{jj_t})$ *for the jth row of R; save the remaining part as update matrix* U_j *for j's parent.*

End For

The data flow of multifrontal QR factorization is given in Figure 7.

Matstoms [20] implemented the multifrontal method and solved (1) by the corrected seminormal equation (CSNE).

2.4. Supernodes. In order to use dense operations and reduce data movement, we can apply the supernode concept to the frontal method in §2.3. We begin by defining the fundamental supernodes as follows.

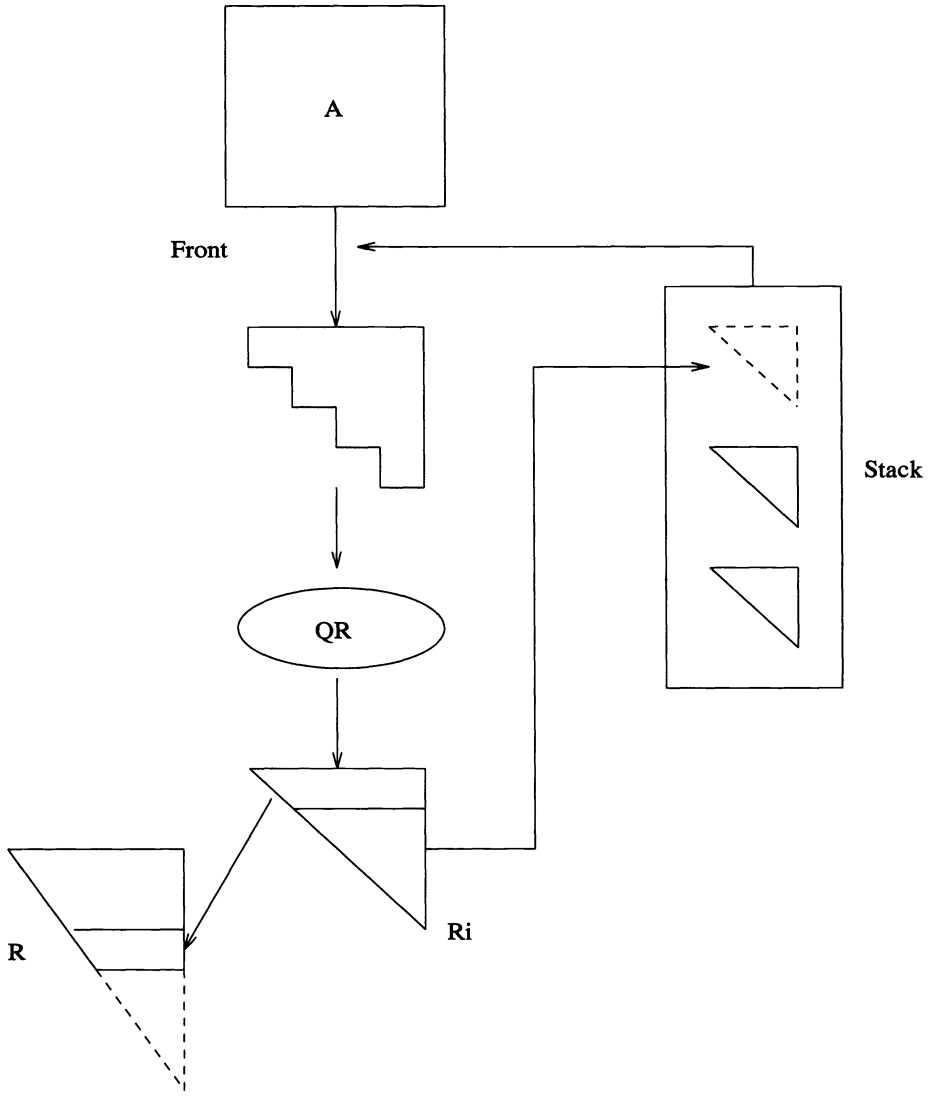


FIG. 7. *Data flow of multifrontal QR factorization.*

DEFINITION 2.4 (fundamental supernode). *A fundamental supernode, with respect to a postordering elimination tree, is a set of maximal number of contiguous vertices,* $S_j = \{i_{j_1}, i_{j_2}, \ldots, i_{j_{|S_j|}}\}$, *such that* i_{j_k} *is the only son of* $i_{j_{k+1}}$ *and the structure of row* $i_{j_{k+1}}$ *in the factor* R *is identical to the structure of the off-diagonal part of row* i_{j_k}, $k = 1, 2, \ldots, |S_j| - 1$. *Furthermore,* $|S_j|$ *is called the size of the supernode.*

Duff and Reid [8] explored the use of supernodes in the multifrontal method. They amalgamate vertices if one of the following conditions is satisfied:

1. if they form a *fundamental supernode*,
2. if the number of fully summed vertices in the parent and the child is less than a user-defined parameter NEMIN.

Details of the implementation and a complete study on the efficiency of the value of NEMIN on the performance of the multifrontal QR factorization are given by Matstoms [20] and Puglisi [23].

We build a supernodal elimination tree by substituting a single node for all the nodes belonging to the same supernode in the original elimination tree. We formally define this elimination tree below.

DEFINITION 2.5 (supernodal elimination tree). *Let A be as in Definition* 2.2. *Let the set $\{1,2,\ldots,n\}$ be partitioned into $\{1,2,\ldots,n\} = S_1 \cup \cdots \cup S_{ns}$, where $S_1,\ldots,S_{ns}$ are the supernodes of A by Definition* 2.4. *Then S_j is the parent of S_i in the supernodal elimination tree if for some vertex $v \in S_j$ and $w \in S_i$, v is the parent of w in the elimination tree of A.*

Associated with each supernode is an $m_j \times n_j$ frontal matrix F_j where n_j is the number of nonzero elements in the rows of $\{S_j\} \cup Tree(S_j)$ and $Tree(S_j)$ is the tree rooted at S_j. We can use the supernodal elimination tree as the representation of the order of the multifrontal factorization process as follows. The merge operation corresponds to computing the QR factorization of a frontal matrix composed of all the update matrices $U_{j_1}, U_{j_2}, \ldots, U_{j_t}$ and rows of A with leading nonzeros from the set $\{j_1, j_2, \ldots, j_t\}$, the indices of the supernode. With the QR factorization of a frontal matrix, we compute multiple rows of the factor R (i.e., those rows belonging to the set of indices of the supernode). Moreover, we spend less time manipulating the update matrices by reducing the data movement. The use of supernodes avoids the redundancy of separate merges by increasing the size of the frontal matrix and combining these merges into the application of one block Householder transformation. The amount of fill in the factor R remains unchanged.

3. The proposed method. In this section, we present an efficient method for storing frontal Householder vectors and, thus, computing $Q^T b$ by applying the multifrontal Householder QR factorization method. Instead of storing the Householder matrix H itself, we store the frontal Householder matrix Y_i of each multifrontal matrix F_i of A. Recall that Y_i is a lower trapezoidal matrix in which each column k is a Householder vector $w_k^{(i)}$ of F_i such that

$$Q_i^T = (I - w_{n_i}^{(i)} w_{n_i}^{(i)T}) \cdots (I - w_1^{(i)} w_1^{(i)T}) \tag{3}$$

and

$$Q_i^T F_i = R_i.$$

Here, Q_i^T is the orthogonal factor that is used to factor the $m_i \times n_i$ frontal matrix F_i. The first $|S_i|$ rows of R_i are used to fill the corresponding rows of R, where $|S_i|$ is the size of the supernode S_i. From the multifrontal process, we have

$$Q^T A \equiv Q_{ns}^T \circledast Q_{ns-1}^T \circledast \cdots \circledast Q_1^T \circledast A, \tag{4}$$

where ns is the number of supernodes in the supernodal elimination tree and Q_i^T is the orthogonal factor that is used to factor the m_i by n_i frontal matrix F_i. $\circledast$ is called "extended multiplication." The extended product in $Q_i^T \circledast A$ factors the part of A that contributes to forming frontal matrix F_i. It follows that

$$Q^T b \equiv Q_{ns}^T \circledast Q_{ns-1}^T \circledast \cdots \circledast Q_1^T \circledast b. \tag{5}$$

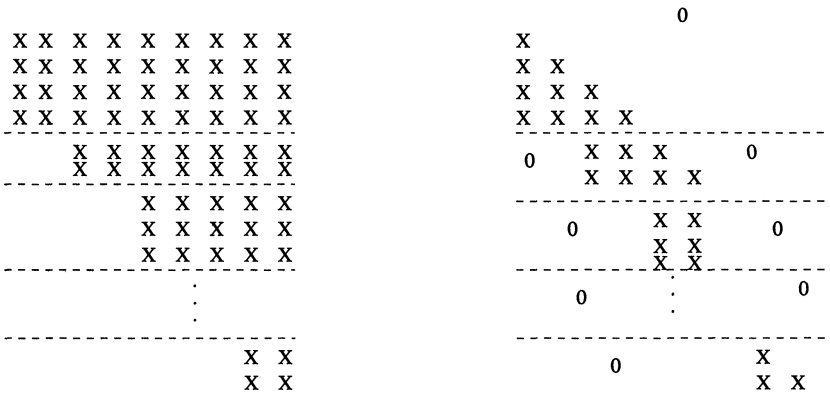


FIG. 8. *A frontal matrix of a supernode and the corresponding Y matrix.*

From (3) we have

$$Q^T b \equiv ((I - w_{n_{ns}}^{(ns)} w_{n_{ns}}^{(ns)T}) \cdots (I - w_1^{(ns)} w_1^{(ns)T})) \circledast \cdots \circledast ((I - w_{n_1}^{(1)} w_{n_1}^{(1)T}) \cdots (I - w_1^{(1)} w_1^{(1)T})) \circledast b. \tag{6}$$

Since R_i in the QR factorization of a frontal matrix F_i,

$$Q_i^T F_i = R_i,$$

is invariant under row orderings of F_i, we can therefore sort the rows of the frontal matrices by the column indices of their leading entries. Therefore, we achieve a block triangular structure for each frontal matrix. Block triangular matrices are efficiently factored by respecting the block structure. It follows that the corresponding frontal Householder matrix Y_i is also a block matrix. Figure 8 is a small example of a block triangular frontal matrix F_i and the corresponding frontal Householder matrix Y_i. The data structure for storing all Y_i's is not complicated; we need only a real data buffer to store the nonzeros of Y_i's and an integer data buffer to record the row indices of each frontal matrix. The required storage for computing $Q^T b$ using (6) is the same as the number of nonzeros of all the Y_i's. Since each Householder vector, w_i, in (6) is applied twice, the required number of multiplications is twice the number of nonzeros of all the Y_i's.

4. Theoretical results. In this section, we develop an upper bound on the number of nonzeros in all Y_i's for the model problem and extend the result to the $\sqrt{n}$ -separator problem. For notational convenience, we denote $|Y_i|$ as the number of nonzeros in Y_i and $|Y| = \sum_{i=1}^{ns} |Y_i|$, where ns is the number of supernodes. Thus $|Y|$ is the quantity we wish to bound.

Lipton, Rose, and Tarjan's "generalized nested dissection" ordering [17], which includes the separators in the recursive call, guarantees bounds of $O(n \log n)$ on fill-in and generates balanced elimination trees for the $\sqrt{n}$-separator problem. We assume all the matrices are ordered by that column ordering in our analysis for the $\sqrt{n}$-separator problem in §4.2. On the other hand, George's original, simpler form of nested dissection [9], which does not include the separators in the recursive call, is actually sufficient for some special classes of the $\sqrt{n}$-separable graphs: planar graphs,

graphs of bounded genus or bounded excluded minor, and two-dimensional finite element meshes of bounded aspect ratio [12]. As a result, our analysis in §4.1 for the model problem uses George's nested dissection ordering.

4.1. The model problem. If George's nested dissection ordering [9] is applied to the model problem, then each internal node has at most two children. To count the number of nonzeros in Y_i's, we begin with the initial frontal matrices. Let F_j be the frontal matrix associated with a leaf node of the elimination tree. Since F_j has at most $4s$ rows and nine columns, the number of nonzeros in Y_j is actually a constant. As a result, the sum of the number of nonzeros in all these Y_j's is $O(n)$ in total.

Now, we consider the internal nodes. The special structure of the model problem implies that the merge process for forming a frontal matrix involves only two trapezoidal matrices. To simplify the proof of the theorems and obtain an upper bound of the nonzero count, we use the two assumptions as in [16]:

1. the two update trapezoidal matrices are full triangular matrices,
2. the two update matrices are u by u and v by v, respectively. Also, they have t columns in common.

Let $C(u, v, t)$ denote the total number of nonzeros in all Householder vectors $w_k^{(i)}$'s such that $Q_i^T = (I - w_i^{(n_i)} w_i^{(n_i)T}) \cdots (I - w_i^{(1)} w_i^{(1)T})$ and $Q_i^T F_i = R_i$, where the first row of R_i is used to fill the ith row of the upper triangular factor R of A. The unreduced frontal matrix has the form given in Figure 9.

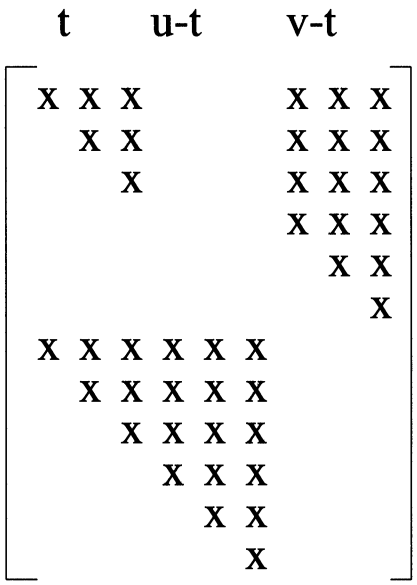


FIG. 9. *Unreduced frontal matrix.*

We then have

$$\begin{aligned} C(u, v, t) &= \sum_{i=1}^{t} (i+1) + \sum_{i=t+1}^{u+v-t} (t+1) \\ &= \frac{t(t+3)}{2} + (t+1)(u+v-2t). \end{aligned}$$

We can use the concept of "bordered K by K grids" [10] to perform the merge operation. Let $\Theta(k,i)$ be the number of nonzeros in the matrix Q, which is used to factor a K by K grid that is bordered on i sides. According to [16, 10], the following recurrence relations are valid.

$$\Theta(K,4) = 4\Theta\left(\frac{K}{2},4\right) + 2C\left(2K,2K,\frac{K}{2}\right) + C(3K,3K,K),$$

$$\Theta(K,3) = 2\Theta\left(\frac{K}{2},3\right) + 2\Theta\left(\frac{K}{2},4\right) + 2C\left(2K,\frac{3K}{2},\frac{K}{2}\right) + C\left(\frac{5K}{2},\frac{5K}{2},K\right),$$

$$\Theta(K,2) = \Theta\left(\frac{K}{2},2\right) + 2\Theta\left(\frac{K}{2},3\right) + \Theta\left(\frac{K}{2},4\right) + C\left(\frac{3K}{2},K,\frac{K}{2}\right) + C\left(2K,\frac{3K}{2},\frac{K}{2}\right) + C\left(\frac{5K}{2},\frac{3K}{2},K\right),$$

$$\Theta(K,0) = 4\Theta\left(\frac{K}{2},2\right) + 2C\left(K,K,\frac{K}{2}\right) + C(K,K,K).$$

Because we are interested in the K by K unbordered grid, $\Theta(K,0)$ is desired.

Using the approaches in [10] and our definition of $C(u,v,t)$, we have

$$|Y| = \Theta(K,0) = \frac{31}{4}K^2\log_2 K + \frac{29}{6}K^2 + 32K\log_2 K - 35K. \tag{7}$$

According to [10], the number of nonzeros of the R factor of a K by K unbordered grid matrix is $N_R = \frac{31}{4}K^2\log_2 K - \frac{73}{3}K^2 + O(K\log_2 K)$. It follows that $|Y| \approx N_R$ as $n \to \infty$. The following theorem summarizes the result in this section.

THEOREM 4.1. *For the model problem, the required storage and number of multiplications to compute Q^Tb as in* (6) *is $O(N_R)$, where N_R is the number of nonzeros of the upper triangular matrix R.* We note that since $|Y| \approx N_R + 30K^2$ for most practical values of n, $|Y| \approx cN_R$ for a constant $c > 1$.

4.2. The extended model problem. We prove the bound of $O(n\log n)$ on $|Y|$ for $\sqrt{n}$-separator matrices. Let A be a $\sqrt{n}$-separator matrix whose columns are ordered by "generalized nested dissection" ordering [17]. If A has no more than $n_0 = (\beta/(1-\alpha))^2$ columns, this recursive numbering algorithm numbers the unnumbered columns arbitrarily.

In order to limit the initial $|Y_i|$, we assume that each s_i, the number of rows of $A[i]$, is much smaller than n and could be treated as a constant. That is, there exists a constant s such that $s_i \le s$ for all i. Note that this is a reasonable assumption as long as $m = O(n)$ because of the fact that $\sum_{i=1}^{n} s_i = m$ and we are studying the matrices after column ordering, which generally permutes relatively fuller columns toward the end of the matrices to reduce fill-in.

LEMMA 4.2. *Let J denote the set of leaf nodes of the elimination tree of A. Then* $\sum_{j\in J}|Y_j| = O(n\log n)$.

Proof. Let j be a leaf node and F_j is the corresponding frontal matrix with m_j rows and n_j columns. From the definition of frontal matrices, F_j is identical to $A[j]$; thus $m_j = s_j$. Since $|Y_j| \le m_j \times n_j = s_j \times n_j$ we have

$$\begin{aligned}\sum_{j\in J}|Y_j| &< \sum_{j\in J} s_j \times n_j \\ &\le s \times \sum_{j\in J} n_j\end{aligned}$$

$$< s \times N_R.$$

According to [17], N_R is $O(n \log n)$. We then have $\sum_{j \in J} |Y_j| = O(n \log n)$. □

We now consider the internal nodes using the supernodal elimination tree. Since we apply the generalized nested dissection ordering and the special property of the $\sqrt{n}$-separable graphs, an internal supernode S_j of the supernodal elimination tree is actually a collection of the tree nodes corresponding to those vertices of C that are not previously numbered, where C is the separator of the subgraph corresponding to S_j and the subtree rooted at S_j.

From the process of multifrontal QR factorization, we have that n_j is the number of nonzeros in the i_1th row of the upper triangular factor R, where i_1 is the vertex in S_j with lowest number. That is, n_j is the number of fill-in edges whose lower numbered vertex is i_1. Suppose the recursive numbering algorithm is applied to an n-vertex graph G with ℓ vertices previously numbered. If G has n vertices, then by the definition of separator $|S_j| \leq \beta\sqrt{n}$ we thus have

$$\begin{aligned} n_j &= |S_j| + \ell \\ &\leq \beta\sqrt{n} + \ell. \end{aligned}$$

LEMMA 4.3. *Let I denote the set of internal supernodes of the supernodal elimination tree. Then $\sum_{j \in I} |Y_j| = O(n \log n)$.*

Proof. The proof is similar to the one for the fill-in bound of Lipton, Rose, and Tarjan's work in [17]. The construction of F_j involves only those $A[i]$'s, where i is a vertex in S_j, and the two upper triangular or trapezoidal update matrices from the two children of S_j in the supernodal elimination tree. Let the two update matrices be u by u and v by v, respectively. Then $u + v \leq \ell + 2\beta\sqrt{n}$. Note that the two update matrices have $\beta\sqrt{n}$ columns in common. That is the size of the separator. In order to get the maximum of $|Y_j|$, we assume the first $\beta\sqrt{n}$ columns are those common columns. A frontal matrix F_j with row reordering according to their leading nonzeros has the form given in Figure 10.

We then have

$$\begin{aligned} |Y_j| &\leq \sum_{i=1}^{\beta\sqrt{n}} (i + 1 + s\beta\sqrt{n}) + \sum_{i=\beta\sqrt{n}+1}^{u+v-\beta\sqrt{n}} ((s+1)\beta\sqrt{n} + 1) \\ &\quad + \frac{1}{2}((s+1)\beta\sqrt{n} + 1)^2 \\ &\leq \left(\left(s + \frac{1}{2}\right)\beta\sqrt{n} + \frac{3}{2}\right)\beta\sqrt{n} + (s+2)\beta\ell\sqrt{n} \\ &\quad + \frac{1}{2}((s+1)\beta\sqrt{n} + 1)^2 \\ &= \frac{1}{2}(s^2 + 4s + 2)\beta^2 n + (s+2)\beta\ell\sqrt{n} + \left(s + \frac{5}{2}\right)\beta\sqrt{n} + \frac{1}{2} \\ &\leq c_1 n + c_2 \ell\sqrt{n} + c_3\sqrt{n}, \end{aligned}$$

where $c_1 = \frac{1}{2}(s^2 + 4s + 2)\beta^2$, $c_2 = (s+2)\beta$, and $c_3 = (s+3)\beta$. Assume the subgraph corresponding to $Tree(S_j) \cup \{S_j\}$ has n vertices, of which ℓ are previously numbered. Let $f(\ell, n)$ be the maximum of $\sum_{i \in SN_j} |Y_i|$, where SN_j is a set whose members are the supernodes in $Tree(S_j) \cup \{S_j\}$. Then

$$f(\ell, n) \leq |Y_j| + \max\{f(\ell_1, k_1) + f(\ell_2, k_2)\} \tag{8}$$

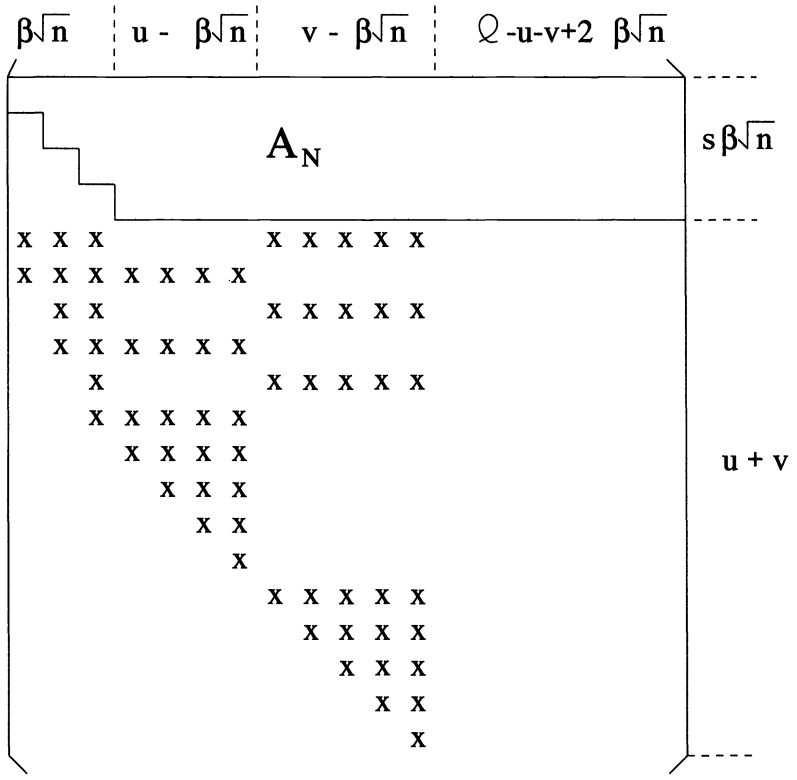


FIG. 10. *A sample frontal matrix with row reordering.*

$$\leq c_1 n + c_2 \ell\sqrt{n} + c_3\sqrt{n} + \max\{f(\ell_1, k_1) + f(\ell_2, k_2)\}, \tag{9}$$

where the maximum is taken over values satisfying

$$\ell_1 + \ell_2 \leq \ell + 2\beta\sqrt{n}, \tag{10}$$

$$n \leq k_1 + k_2 \leq n + \beta\sqrt{n}, \text{ and} \tag{11}$$

$$(1-\alpha)n \leq k_i \leq \alpha n + \beta\sqrt{n} \text{ for } i = 1, 2. \tag{12}$$

An analysis similar to [17, pp. 349–350, Thm. 2] shows that

$$f(\ell, n) \leq c_4(n+\ell)\log n + c_5\ell\sqrt{n},$$

where c_4 and c_5 are some suitably large constants.

Since $\sum_{S_j \in I} |Y_j| = f(0, n)$, the desired bound of $O(n \log n)$ on $\sum_{S_j \in I} |Y_j|$ then follows. □

From Lemmas 4.2 and 4.3 we have $|Y| = O(n \log n)$.

The following theorem summarizes the results of this section.

THEOREM 4.4. *Let A be an m by n matrix that is defined on a generalized nested dissection ordered $\sqrt{n}$-separable graph. If the number of rows of each $A[i]$ is bounded by a constant, then the proposed method for computing $Q^T b$ requires $O(n \log n)$ storage and multiplications.*

From Theorem 4.4, the proposed method for computing $Q^T b$ is more efficient than using the Householder matrix H or the orthogonal factor Q itself when $m - n$ is of the order of n.

5. Introducing BLAS-2 operations. In order to introduce BLAS-2 operations in (6), we use the YTY representation [24], also called the "storage-efficient WY representation." Here each Q_i^T can be written as

$$Q_i^T = I - Y_i T_i Y_i^T, \tag{13}$$

where Y_i is the frontal Householder matrix of F_i as defined before and T_i is an n_i by n_i lower triangular matrix which is computed by the following algorithm:

$$T_i^{(1)} = 1,$$

$$T_i^{(k)} = \begin{pmatrix} T_i^{(k-1)} & 0 \\ Z_i^{(k)} & 1 \end{pmatrix}, \quad k = 2, \ldots, n_i,$$

where $Z_i^{(k)} = -w_i^{(k)T} Y_i^{(k-1)} T_i^{(k-1)}$ and $Y_i = Y_i^{(n_i)}$, $T_i = T_i^{(n_i)}$.

From (6) and (13) we have

$$Q^T b = (I - Y_{ns} T_{ns} Y_{ns}^T) \circledast (I - Y_{ns-1} T_{ns-1} Y_{ns-1}^T) \circledast \cdots \circledast (I - Y_1 T_1 Y_1^T) \circledast b. \tag{14}$$

The structure in (14) is suitable for parallel computing; each Y_i and Y_j are independent blocks if node i and node j are not ancestor and descendant in the elimination tree. That is also true for the T matrices. As a result, the matrix-vector computations $(I - Y_i T_i Y_i^T) \circledast b$ and $(I - Y_j T_j Y_j^T) \circledast b$ can be done simultaneously and each matrix-vector computation can be performed in parallel. It follows that the computing time of (14) is based on the height of the supernodal elimination tree and the communication time among the processors. Note that the required number of multiplications to compute $Q^T b$ by (14) is $2|Y| + |T|$.

LEMMA 5.1. *The required storage and number of multiplications to compute $Q^T b$ for the K by K grid model problem using BLAS-2 operations is $O(N_R)$.*

Proof. By the same argument as in §4.1 and by redefining $C(u, v, t) = \frac{1}{2}(u + v - t)^2$, we can prove that $|T| = \frac{99}{4} K^2 \log_2 K + O(K^2) \approx 3N_R$. It follows that the storage requirement is $4N_R$ and the required number of multiplications is $5N_R$. □

THEOREM 5.2. *Let A be an m by n matrix that is defined on a generalized nested dissection ordered $\sqrt{n}$-separable graph. If the number of rows of each $A[i]$ is a constant, then the required storage and number of multiplications for computing $Q^T b$ using BLAS-2 operations is $O(n(\log n)^2) = O(N_R \log n)$.*

Proof. Use the same argument as in §4.2 and redefine $f(\ell, n)$ as the maximum of $\sum_{i \in SN_j} |T_j|$, where

$$\begin{aligned} |T_j| &= \frac{1}{2}(n_j)^2 \\ &= \frac{1}{2}(\beta\sqrt{n} + \ell)^2. \end{aligned}$$

Here again ℓ is the number of vertices that have already been labeled. Similar to (8)–(9) we define

$$f(\ell, n) \le \frac{1}{2} n^2 \le \frac{1}{2}(n_0) n, \quad n \le n_0,$$

$$f(\ell, n) \leq |T_j| + \max\{f(\ell_1, k_1) + f(\ell_2, k_2)\} \tag{15}$$

$$\leq \frac{1}{2}\beta^2 n + \beta\ell\sqrt{n} + \ell^2 + \max\{f(\ell_1, k_1) + f(\ell_2, k_2)\}, \quad n > n_0, \tag{16}$$

where the maximum is again taken over the set (10)–(12). Here $n_0 < n$ which is independent of n. We claim that the solution for all $n \geq 1$ is

$$f(\ell, n) \leq c_4 n(\log_2 n)^2 + c_5\ell^2 \log_2 n + c_6\ell\sqrt{n}\log n,$$

where c_4, c_5, and c_6 are constants. The value $f(0, n) = |T| = \sum_{i \in NS} |T_i|$. This claim can be shown by induction on n and by using the approach given by Lipton, Rose, and Tarjan in [17, pp. 349–350, Thm. 2]. The proof is as follows. Let n be large and suppose the claim is true for values smaller than n. Then the recurrences (15) and (16) give us

$$\begin{aligned} f(\ell, n) \leq\ & c_4 n(\log_2 n)^2 + c_5\ell^2 \log_2 n \\ & + (2c_4 \log_2(1-\epsilon) + 4c_5\beta^2 + 2c_6\beta\sqrt{\alpha})n \log_2 n + (4c_5\beta + \sqrt{\alpha}c_6)\ell\sqrt{n}\log n \\ & + \left(c_5 \log_2(1-\epsilon) + \frac{1}{2}\right)\ell^2 \\ & + \left\{c_4(\log_2(1-\epsilon))^2 + (4c_5\beta^2 + 2c_6\beta)\log_2(1-\epsilon) + \frac{1}{2}\beta^2\right\} n + h(n), \end{aligned}$$

where $h(n)$ is of order $O(\sqrt{n}(\log_2 n)^2)$ and $\epsilon = \alpha + \beta/\sqrt{1+n_0}$. Clearly, n_0 must be large enough so that $\epsilon < 1$. Suppose we choose c_5 such that $c_5 \log_2(1-\epsilon) + \frac{1}{2} \leq 0$, choose c_6 large enough such that $4c_5\beta + \sqrt{\alpha}c_6 \leq c_6$, and choose c_4 large enough such that $\frac{3}{2}c_4 \log_2(1-\epsilon) + 4c_5\beta^2 + 2c_6\beta\sqrt{\alpha} + \frac{1}{2}\beta^2 = 0$. Then $f(\ell, n) \leq c_4 n(\log_2 n)^2 + c_5\ell^2 \log_2 n + c_6\ell\sqrt{n}\log n + (-\frac{1}{2}c_4)(\log_2(1-\epsilon))^2 n + h(n)$. Since n is large, we have $f(\ell, n) \leq c_4 n(\log_2 n)^2 + c_5\ell^2 \log_2 n + c_6\ell\sqrt{n}\log n$, as desired. As a result, the bound on $|T|$ is $O(n(\log n)^2)$. It follows that the required storage and number of multiplications is $O(n(\log n)^2)$. □

Note that referring to Lemma 5.1, there is an extra $\log n$ term in the result of Theorem 5.2. This is because in the grid model problem case, the ℓ term in the boundary of $|T|$ is replaced by a lower order term based on the information from the separators.

Remark 5.3. The previous complexity results apply to Bischof and Van Loan's [4] WY representation. This would generate

$$Q_i^T = I - W_i Y_i^T,$$

where Y_i is the same as above, but W_i is computed according to

$$W_i^{(1)} = (w_i^{(1)}),$$

$$W_i^{(k)} = \begin{pmatrix} W_i^{(k-1)} & z^{(k)} \end{pmatrix}; \ z^{(k)} = (I - W_i^{(k-1)} Y_i^{(k-1)T}) w_i^{(k)};$$

thus $W_i = W_i^{(n_i)}$. To the best of our knowledge, this was never stated formally, but from an easy induction argument it is evident that $W_i = Y_i T_i$. Since T_i is full, it is reasonable to assume that W_i will be as well. Moreover, our bounds will apply to the Bischof–Van Loan representation, but with slightly different constants. The YTY representation always requires less storage. In our experiments, the WY representation tended to compute $Q^T b$ somewhat faster, but that result varies among architectures [24].

TABLE 1
Numerical operation count for the model problem.

K	$\|T\|/N_R$	$\|Y\|/N_R$	$(2\|Y\|+\|T\|)/N_R$
20	3.21	3.12	9.45
40	3.38	2.55	8.48
60	3.42	2.35	8.12
80	3.45	2.25	7.95
100	3.45	2.17	7.79

TABLE 2
Numerical operation count for the general problems.

Prob.	m	n	NZ	$\|T\|/N_R$	$\|Y\|/N_R$	$\frac{m}{n}$
ILLC1033	1033	320	4732	2.12	3.40	3.23
WELL1033	1033	320	4732	2.14	3.43	3.23
ILLC1850	1850	712	8758	2.73	3.27	2.60
WELL1850	1850	712	8758	2.73	3.27	2.60
CONVEC8	3362	484	13997	3.30	7.24	6.95
DUNES8	5514	771	24796	3.49	6.72	7.15
MIMBUS	23871	1325	181972	4.93	15.78	18.02
STRAT8	16640	2205	66192	3.43	6.32	7.55

6. Numerical results. In this section, we examine the performance of our method for computing $Q^T b$ and solve the linear least squares problem given in (1) by QR method using our method to compute $Q^T b$.

We check the ratio of the required number of multiplications for computing $Q^T b$ versus N_R for the model problem in Table 1. In Table 2 we do the same test for the general problems from the Harwell–Boeing test collection and Bramley's test matrices. The results show our method also performs well for those problems (i.e., the required number of multiplications is less than $\frac{3m}{n} N_R = O(N_R)$, where m and n are the number of rows and columns of the problem, respectively). Here, we use $\frac{m}{n}$ as an approximation of the average value of s in §4.2, since $\sum_i^n s_i = m$, and $s = \max_i s_i$ and is presumed to be constant.

We also solve the linear least squares problems given in (1) by the QR method by [14] using our method for computing $Q^T b$. We compare the QR method with the method of CSNE by [5]. The CSNE method is in fact the method of seminormal equations with one or more correction steps,

$$\begin{aligned} r &= b - Ax, \\ R^T R \delta x &= A^T r, \\ x &\leftarrow x + \delta x. \end{aligned}$$

The tests of CSNE were on QR27 routines by [20]. Two correction steps are used as suggested in QR27. In Tables 3 and 4, we list the residuals $||r||_2 = ||b - Ax||_2$ and the timing for solving (1) by both CSNE and the QR method on the model problem. Here, we assume the upper triangular factor R has been precomputed; the timing results do not include the factorization time. All the tests in this part were run on a Sun 4. We see from Table 3 that the residuals from both methods are about the same. However, the QR method is more time consuming than the CSNE method. This should be expected because the QR method requires more operations and, therefore,

TABLE 3
Residuals for the model problem.

K	QR	CSNE
10	4.08	4.08
20	9.21	9.21
30	13.99	13.99
40	19.54	19.19
50	24.27	24.27
60	24.98	28.98
70	31.32	33.94

TABLE 4
Timing for the model problem (in seconds).

K	QR(YTY BLAS-2)	QR(WY BLAS-2)	QR(BLAS-1)	CSNE
10	0.04	0.04	0.03	0.05
20	0.45	0.44	0.37	0.32
30	1.36	1.35	1.34	0.87
40	2.99	2.98	2.81	1.65
50	5.07	4.72	4.80	2.34
60	10.00	7.31	7.93	4.58
70	13.39	9.03	10.13	4.89
80	18.87	15.02	16.95	7.81

is slower. Also, we compare the performance of the BLAS-2 method (14) and the BLAS-1 method (6). Both WY and storage-efficient WY (YTY) representations are applied in the BLAS-2 method. The results are listed in Table 4. The WY BLAS-2 method is faster than the BLAS-1 method for larger problems. The BLAS-1 method seems faster than the YTY BLAS-2 method, but the gap between them narrows with larger problems. This suggests that the BLAS-1 method requires fewer floating point operations, but the BLAS-2 method takes more advantage of features of the architecture.

In [5], Björck stated that the CSNE method does not obtain good accuracy on "stiff" problems:

$$\begin{pmatrix} \omega A_1 \\ A_2 \end{pmatrix} x = \begin{pmatrix} \omega b_1 \\ b_2 \end{pmatrix},$$

where the rows are of widely differing norms.

It has been shown that the QR method performs well for "stiff" problems [1]. There is some comment on this problem in [2, 3]. In order to confirm that the QR method using the proposed method maintains this property, we apply the QR and CSNE methods on a sample "stiff" problem given in Figure 11. We take the exact solution x as

$$x = (10.0, 1.0, 10^{-1}, 10^{-2}, 10^{-3})^T$$

and set the right-hand side vector b to be Ax. We define error by

$$error = ||\tilde{x} - x||_2.$$

The results for the QR and CSNE methods using single precision with $\omega = 10^4, 10^5$, 10^6, and 10^7 are given in Table 5. The QR method performs consistently well and indeed gives better accuracy for increasing ω. These results are consistent with the work by Björck [5] and Matstoms [20].

$$\begin{bmatrix}
\omega & \omega & \omega & \omega & \\
\omega & 2\omega & 4\omega & 8\omega & \\
1. & 3. & 9. & 27. & \\
 & 4. & & 64. & 256. \\
 & 5. & & 125. & 625. \\
 & 6. & & 216. & 1296. \\
 & 7. & & 343. & 2401. \\
1. & 8. & & & 4096. \\
1. & 9. & & & 6561.
\end{bmatrix}$$

FIG. 11. *A sample "stiff" problem.*

TABLE 5
Errors for a sample "stiff" problem.

ω	No. of refinements	Errors QR	Errors CSNE
10^4	0	9.83E−5	1.78E+1
	1		1.11E−5
	2		1.07E−6
10^5	0	3.32E−4	1.56E+3
	1		8.78E−3
	2		2.32E−2
	3		9.74E−5
10^6	0	1.84E−4	1.81E+5
	1		1.64E+3
	2		1.06E+2
	3		2.45E+0
	4		9.62E−2
	5		9.68E−7
10^7	0	3.45E−4	5.60E+7
	1		1.02E+7
	2		3.38E+7
	3		2.84E+7
	4		2.33E+7
	5		1.96E+7

7. Conclusion. In this paper, we have provided a multifrontal-based method for storing Q and, thus, computing $Q^T b$ by using the frontal Householder matrices for an m by n large and sparse matrix with $m \geq n$. We have shown that the use of a multifrontal paradigm requires $O(N_R)$ storage and multiplications for the K by K grid model problem and problems defined on the $\sqrt{n}$-separable graphs, where $N_R = n \log n$ represents the number of nonzeros in the upper triangular factor R.

This method is more efficient than using the Householder matrix H or Q directly if the matrix is m by n with $m-n = O(n)$ and defined on a $\sqrt{n}$-separable graph. In that case both H and Q have $O(n\sqrt{n})$ nonzeros. Thus one can solve sparse linear least squares problems by the orthogonal method using the proposed method for computing $Q^T b$ efficiently. In order to introduce BLAS-2 operations, we also use the "storage-efficient WY representation" for the orthogonal factor of each frontal matrix. This representation brings the bound on storage and operation counts up to $O(n(\log n)^2)$ for matrices defined on $\sqrt{n}$-separable graphs. This is still more efficient than using Q itself or the "storage-efficient WY representation" of Q directly if the matrix is m by n with $m - n = O(n)$ and defined on a $\sqrt{n}$-separable graph, under which Q and its "storage-efficient WY representation" have $O(n\sqrt{n})$ and $O(n^2)$ nonzeros, respectively. The proposed method has possibilities for parallel computing as seen on the iPSC/2 [19]. In a future report, we will test different representations of the orthogonal factors on $\sqrt{n}$-separator matrices and various practical problems such as the geodesy problems [15] and the equilibrium systems problem [25] on advanced architectures which support BLAS-2 and BLAS-3 operations.

Acknowledgments. We would like to thank Åke Björck and Pontus Matstoms for making the QR27 software available to us and John Gilbert and Esmond Ng for introducing us to this problem. The second author thanks Don Beaver for some helpful discussion.

REFERENCES

[1] J. L. Barlow and S. L. Handy, *The direct solution of weighted and equality constrained least squares problems*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 704–716.

[2] J. L. Barlow, *Error analysis and implementation aspects of deferred correction for equality constrained least squares problems*, SIAM J. Numer. Anal., 25 (1988), pp. 1340–1358.

[3] J. L. Barlow and U. B. Vemulapati, *A note on deferred correction for equality constrained least squares problems*, SIAM J. Numer. Anal., 29 (1992), pp. 249–256.

[4] C. H. Bischof and C. F. Van Loan, *The WY representation for products of Householder matrices*, SIAM J. Sci. Stat. Comput., 8 (1987), pp. s2–s13.

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

[6] J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. Stewart, *LINPACK Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1979.

[7] E. Chu, *Orthogonal Decomposition of Dense and Sparse Matrices on Multiprocessors*, Tech. report CS-88-08, University of Waterloo, Waterloo, Ontario, Canada, 1988.

[8] I. S. Duff and J. K. Reid, *The multifrontal solution of indefinite sparse symmetric linear systems*, ACM Trans. Math. Software, 9 (1983), pp. 302–325.

[9] A. George, *Nested dissection of a regular finite element mesh*, SIAM J. Numer. Anal., 10 (1973), pp. 345–363.

[10] A. George and J. W. Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1981.

[11] A. George, J. W. Liu, and E. G. Ng, *A data structure for sparse QR and LU factorizations*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 100–121.

[12] J. R. Gilbert and R. E. Tarjan, *The analysis of a nested dissection algorithm*, Numer. Math., 50 (1987), pp. 377–404.

[13] J. R. Gilbert, E. G. Ng, and B. W. Peyton, *Separators and Structure Prediction in Sparse Orthogonal Factorization*, Tech. report, Xerox Palo Alto Research Center, Palo Alto, CA, 1993.

[14] G. H. Golub, *Numerical methods for solving linear least squares problems*, Numer. Math., 7 (1965), pp. 206–216.

[15] G. H. Golub, P. Manneback, and Ph. L. Toint, *A comparison between direct and iterative methods for certain large scale geodetic least squares problems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 799–816.

[16] J. G. LEWIS, D. J. PIERCE, AND D. K. WAH, *Multifrontal Householder QR Factorization*, Tech. report ECA-TR-127, Boeing Computer Services, Seattle, WA, 1989.

[17] R. J. LIPTON, D. J. ROSE, AND R. E. TARJAN, *Generalized nested dissection*, SIAM J. Numer. Anal., 16 (1979), pp. 346–358.

[18] J. W. LIU, *On general row merging schemes for sparse Givens transformations*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 1190–1211.

[19] S. M. LU AND J. L. BARLOW, *Parallel computation of orthogonal factors of sparse matrices*, in Proc. Sixth SIAM Conference on Parallel Processing for Scientific Computing, Norfolk, VA, 1993, pp. 486–490.

[20] P. MATSTOMS, *The Multifrontal Solution of Sparse Linear Least Squares Problems*, Thesis No. 293, LIU-TEK-LIC-1991:33, Department of Mathematics, Linköping University, Sweden, 1991.

[21] ———, *Parallel Sparse QR Factorization on Shared Memory Architectures*, LiTH-MAT-R-1993-18, Department of Mathematics, Linköping University, Sweden, 1993.

[22] E. G. NG AND B. W. PEYTON, *A Tight and Explicit Representation of Q in Sparse QR Factorization*, Tech. report ORNL/TM-12059, Oak Ridge National Laboratory, Argonne, IL, 1992.

[23] C. PUGLISI, *QR Factorization of Large Sparse Overdetermined and Square Matrices Using the Multifrontal Method in a Multiprocessor Environment*, Ph.D. thesis, De L'institut National Polytechnique de Toulouse, Toulouse, France, 1993.

[24] R. SCHREIBER AND C. VAN LOAN, *A storage-efficient WY representation for products of Householder transformations*, SIAM J. Sci. Stat. Comput., 10 (1989), pp. 53–57.

[25] S. A. VAVASIS, *Stable numerical algorithms for equilibrium systems*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1108–1131.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 3, pp. 680–690, July 1996

017

ON DOUBLY SYMMETRIC TRIDIAGONAL FORMS FOR COMPLEX MATRICES AND TRIDIAGONAL INVERSE EIGENVALUE PROBLEMS*

A. GEORGE†, Kh. IKRAMOV‡, W.-P. TANG†, AND V. N. TCHUGUNOV‡

Abstract. It is well known that for any distinct real numbers $\lambda_1, \ldots, \lambda_n$ there exists a doubly symmetric (i.e., symmetric and persymmetric or, equivalently, symmetric and centrosymmetric) tridiagonal real $n \times n$ matrix T with the λ's as its eigenvalues. Such a matrix can be constructed finitely using only arithmetic operations and square roots. We prove in this paper that the analogous assertions hold for any distinct complex numbers $\lambda_1, \ldots, \lambda_n$ with T being a complex matrix. It follows that any complex $n \times n$ matrix with distinct eigenvalues is similar to a doubly symmetric tridiagonal matrix. The condition that the eigenvalues be distinct is essential: we show that the tridiagonal form above does not exist or is trivial (depending on the Jordan structure imposed) if $\lambda_1 = \lambda_2 = \lambda_3 = 0$.

Key words. inverse eigenvalue problem (IEP), tridiagonalization, persymmetry, centrosymmetry

AMS subject classification. 65F10

1. Introduction.

An $n \times n$ matrix A is called persymmetric if

$$a_{ij} = a_{n+1-j,n+1-i} \quad \forall \ i,j$$

and centrosymmetric if

$$a_{ij} = a_{n+1-i,n+1-j} \quad \forall \ i,j \ .$$

Of the three matrix properties—symmetry, persymmetry, and centrosymmetry—any two imply the third one. We call A a *doubly symmetric matrix* if it is symmetric and persymmetric (or centrosymmetric) at the same time.

The following inverse eigenvalue problem (IEP) is well known (and can be found, for example, in [8, pp. 136–138]): given distinct real numbers $\lambda_1, \ldots, \lambda_n$, find a doubly symmetric real tridiagonal matrix T with the λ's as its eigenvalues. It is proved in [8] that this IEP is always solvable, and its (essentially unique) solution T could be constructed by a finite procedure involving only arithmetic operations and square roots. In fact, the use of the Lanczos algorithm is suggested in [8] (see also [2, 3]), although Householder transformations or rotations could be applied as well.

From the matrix theory point of view, the solvability of the IEP above implies that the following statement is valid.

THEOREM 1.1. *Any real symmetric $n \times n$ matrix A with distinct eigenvalues can be transformed via orthogonal similarity to a doubly symmetric tridiagonal form.*

* Received by the editors August 12, 1994; accepted for publication (in revised form) by A. Bunse-Gerstner October 4, 1995. 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 N2L 3G1, Canada (jageorge@sparse1.uwaterloo.ca, wptang@riacs.edu).

‡ Moscow State University, Faculty of Numerical Mathematics and Cybernetics, Moscow 119899, Russia (ikramov@cmc.msk.su).

If we do not insist on the transformations being orthogonal then we can state another matrix result.

THEOREM 1.2. *Any $n \times n$ matrix A, real or complex, with distinct real eigenvalues can be transformed via similarity to a doubly symmetric tridiagonal form.*

Now, the natural question is as follows: Are the statements analogous to Theorems 1.1 and 1.2 valid for matrices with complex spectra? We prove in this paper that they are, i.e., that the following assertions hold.

THEOREM 1.3. *Any complex $n \times n$ matrix A with distinct eigenvalues is similar to a (generally complex) doubly symmetric tridiagonal matrix.*

THEOREM 1.4. *Any complex symmetric $n \times n$ matrix A with distinct eigenvalues can be transformed via orthogonal similarity to a (generally complex) doubly symmetric tridiagonal form.*

Both assertions are almost immediate consequences of the main result we prove here.

THEOREM 1.5. *For any distinct complex numbers $\lambda_1, \ldots, \lambda_n$ there exists a (generally complex) doubly symmetric tridiagonal matrix T with the λ's as its eigenvalues. This IEP has generically $2^{k-1}\binom{n}{k}$ solutions where $k = \lfloor n/2 \rfloor$. For almost any n-tuple $(\lambda_1, \ldots, \lambda_n)$, all the solutions can be found by a finite procedure involving only arithmetic operations and square roots.*

Recall that a complex $n \times n$ matrix A is called symmetric if $A = A^T$, and orthogonal if $AA^T = I$. These matrices lack many of the desirable properties of their real counterparts, or Hermitian and unitary matrices. For example, there are no special properties of the spectrum or Jordan structure of such matrices. Complex orthogonal matrices, unlike real ones, may have entries exceeding one in modulus, and so on. For a somewhat more comprehensive discussion of these points and some related work the reader is referred to Scott [9, 10]. He also considers the problem of reducing complex symmetric matrices to a simpler form, although the approach and the resulting form are different from what is presented in this work. In particular, the question of the finiteness of the process is not considered, and the target form is pentadiagonal rather than tridiagonal, and not generally persymmetric. On the other hand, his is a canonical form—all complex symmetric matrices can be reduced to that form. In this paper we restrict our attention to matrices with distinct eigenvalues.

Theorem 1.5 and its proof are interesting in at least two respects. First, we do not use the special properties of eigenvectors of tridiagonal matrices; these are usually employed in the real case (see again [8]). Instead, we exploit the similarity that decomposes any centrosymmetric matrix into the direct sum of two blocks of (roughly) half the order. This similarity is well known but, to the best of the present authors' knowledge, has not been used in relation with the previous IEP. We review this transformation in §2. Second, our proof clearly demonstrates the origin of multiple solutions to this IEP. In particular, for real λ's, in addition to the classical real solution there exist many complex solutions as well. We give a small illustration of this point in §5. We also mention there that the condition that the prescribed eigenvalues be distinct is essential: we may not be able to find the tridiagonal form above for an $n \times n$ Jordan block. The proof of Theorem 1.5 is contained in §3. In §4 we show how Theorems 1.3 and 1.4 follow from Theorem 1.5.

2. Decomposing centrosymmetric tridiagonals. A doubly symmetric tridiagonal $n \times n$ matrix T can be depicted as follows:

$$T = \begin{bmatrix} x_1 & y_1 & & & & & & & \\ y_1 & x_2 & \ddots & & & & & & \\ & \ddots & \ddots & y_{k-1} & & & & & \\ & & y_{k-1} & x_k & y_k & & & & \\ & & & y_k & x_k & y_{k-1} & & & \\ & & & & y_{k-1} & \ddots & \ddots & & \\ & & & & & \ddots & \ddots & \ddots & \\ & & & & & & \ddots & x_2 & y_1 \\ & & & & & & & y_1 & x_1 \end{bmatrix} \tag{2.1}$$

if $n = 2k$, and

$$T = \begin{bmatrix} x_1 & y_1 & & & & & & & \\ y_1 & x_2 & \ddots & & & & & & \\ & \ddots & \ddots & \ddots & & & & & \\ & & \ddots & x_{k-1} & y_{k-1} & & & & \\ & & & y_{k-1} & x_k & y_{k-1} & & & \\ & & & & y_{k-1} & x_{k-1} & & & \\ & & & & & \ddots & \ddots & \ddots & \\ & & & & & & \ddots & x_2 & y_1 \\ & & & & & & & y_1 & x_1 \end{bmatrix} \tag{2.2}$$

if $n = 2k - 1$.

Centrosymmetry as a matrix property can be described by the commutativity relation

$$T\mathcal{P}_n = \mathcal{P}_n T,$$

where

$$\mathcal{P}_n = \begin{bmatrix} 0 & & & 1 \\ & & \cdot & \\ & \cdot & & \\ 1 & & & 0 \end{bmatrix}.$$

The matrix of the same form as $\mathcal{P}_n$ is used in decomposing the centrosymmetric matrix T. Namely, define Q as the block matrix of the form

$$Q = \frac{1}{\sqrt{2}} \begin{bmatrix} I_k & I_k \\ \mathcal{P}_k & -\mathcal{P}_k \end{bmatrix}$$

if $n = 2k$, and

$$Q = \frac{1}{\sqrt{2}} \begin{bmatrix} I_{k-1} & 0 & I_{k-1} \\ 0 & \sqrt{2} & 0 \\ \mathcal{P}_{k-1} & 0 & -\mathcal{P}_{k-1} \end{bmatrix}$$

if $n = 2k - 1$. Then the orthogonal similarity

$$T \to U = Q^T T Q \tag{2.3}$$

converts matrix (2.1) into the block diagonal matrix

$$U = U_1 \oplus U_2, \tag{2.4}$$

where

$$U_1 = \begin{bmatrix} x_1 & y_1 & & & \\ y_1 & x_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & x_{k-1} & y_{k-1} \\ & & & y_{k-1} & x_k + y_k \end{bmatrix} \tag{2.5}$$

and

$$U_2 = \begin{bmatrix} x_1 & y_1 & & & \\ y_1 & x_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & x_{k-1} & y_{k-1} \\ & & & y_{k-1} & x_k - y_k \end{bmatrix}. \tag{2.6}$$

For n odd, the same transformation (2.3) applied to matrix (2.2) gives the block diagonal matrix (2.4) with the blocks

$$U_1 = \begin{bmatrix} x_1 & y_1 & & & \\ y_1 & x_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & x_{k-1} & \sqrt{2}y_{k-1} \\ & & & \sqrt{2}y_{k-1} & x_k \end{bmatrix} \tag{2.7}$$

and

$$U_2 = \begin{bmatrix} x_1 & y_1 & & \\ y_1 & x_2 & \ddots & \\ & \ddots & \ddots & y_{k-2} \\ & & y_{k-2} & x_{k-1} \end{bmatrix}. \tag{2.8}$$

Remark. Being orthogonal, similarity (2.3) retains the symmetry of the matrix being transformed. We can see that blocks (2.5) and (2.6) differ only in position (k, k). For n odd, block (2.8) is a principal submatrix of block (2.7).

Of course, transformation (2.3) decomposes any centrosymmetric matrix, not only the tridiagonal one. But here we need it only for tridiagonals.

3. Proof of Theorem 1.5. Instead of dealing with T directly, we will look for the blocks U_1 and U_2 in its representation (2.4). After these blocks are constructed the inverse transformation to (2.3) gives us the tridiagonal matrix desired.

Below we thoroughly examine the case when n is even. Then, with fewer particulars, we consider the case when n is odd, which differs only in minor details. We show that the values of the unknowns $(y_k,)x_k, y_{k-1}, x_{k-1}, \ldots, x_2, y_1, x_1$ can be determined one by one for any n-tuple $(\lambda_1, \ldots, \lambda_n)$ where the λ's are distinct, with the possible exception of points of some algebraic manifold in $\mathbf{C}^n$. Finally, using some considerations relating to commutative algebra, we conclude that our IEP is solvable even for these exceptional n-tuples $(\lambda_1, \ldots, \lambda_n)$.

We begin by introducing some notation. Let σ_i be the ith symmetric function of the numbers $\lambda_1, \ldots, \lambda_n$ [6, p. 41]. Let S_i be the sum of all the principal minors of order i of the matrix T $(i = 1, 2, \ldots, n)$. Taking into account that both σ_i and S_i are equal, up to the same sign, to the coefficients of the characteristic polynomial of T, we conclude that our tridiagonal IEP is equivalent to the system of polynomial equations

$$S_i = \sigma_i, \quad i = 1, 2, \ldots, n, \tag{3.1}$$

with x's and y's as unknowns. We need therefore to show that system (3.1) is solvable over $\mathbf{C}$.

As was pointed out above, instead of dealing with (3.1) directly we prefer to use decomposition (2.4). We argue in the following way. Every solution of (3.1) (if any) defines the partition of the T's spectrum $\lambda_1, \ldots, \lambda_n$ into the spectra $\lambda'_1, \ldots, \lambda'_k$ and $\lambda''_1, \ldots, \lambda''_k$ (remember that n is even at the moment!) of the submatrices U_1 and U_2 in (2.4). Introducing symmetric functions σ'_i and σ''_i in numbers λ''s and λ'''s, respectively, and also the sums S'_i and S''_i of all the principal minors of order i for U_1 and U_2, we can replace (3.1) by the equivalent system

$$S'_i = \sigma'_i, \qquad i = 1, \ldots, k, \tag{3.2}$$

$$S''_i = \sigma''_i, \qquad i = 1, \ldots, k. \tag{3.3}$$

Note that there exist altogether $\binom{n}{k}$ ways of partitioning $\lambda_1, \ldots, \lambda_n$ into the subsets λ''s and λ'''s. We show next how, for any particular partition, the unknowns in system (3.2)–(3.3) can be determined. To describe this procedure some additional notation is useful; namely, we denote by S_i^m $(m = 1, 2, \ldots, k-1)$ the sum of all the principal minors of order i in the matrix formed by deleting the last $k - m$ rows and columns in U_1 (or, equivalently, in U_2). We also let

$$S_i^m = \begin{cases} 1, & i = 0, \\ 0, & i > m \end{cases} \qquad (m = 1, 2, \ldots, k-1). \tag{3.4}$$

It will be convenient to divide the procedure for computing the x's and the y's into $k - 1$ stages. The first stage and the last one are somewhat different from the rest and will be described separately. The main process is uniform, and it will be enough to describe Stage 2 as the typical one. Generally, at stage i of the main process, we compute x_{k+1-i} and y_{k-i}.

The main consideration much exploited in our procedure is the well-known rela-

tion between the tridiagonal determinant

$$\triangle_m = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \alpha_{m-1} & \beta_{m-1} \\ & & & \beta_{m-1} & \alpha_m \end{bmatrix}$$

and its principal minors $\triangle_{m-1}$ and $\triangle_{m-2}$ of order $m-1$ and $m-2$, respectively:

$$\triangle_m = \alpha_m \triangle_{m-1} - \beta_{m-1}^2 \triangle_{m-2}. \tag{3.5}$$

We begin Stage 1 by computing y_k. Taking traces of the submatrices U_1 and U_2 in (2.4), we can write (this corresponds to letting $i = 1$ in (3.2) and (3.3))

$$\sigma_1' = S_1' = S_1^{k-1} + (x_k + y_k) \tag{3.6}$$

and

$$\sigma_1'' = S_1'' = S_1^{k-1} + (x_k - y_k). \tag{3.7}$$

Subtracting (3.7) from (3.6) gives us the value of y_k:

$$y_k = (\sigma_1' - \sigma_1'')/2. \tag{3.8}$$

It is vital for the rest of the procedure that y_k be nonzero. This can be violated only in points of the nontrivial algebraic manifold (even linear subspace, in this case) described by the equation

$$\sigma_1' - \sigma_1'' = 0. \tag{3.9}$$

We assume that the condition

$$y_k \neq 0 \tag{3.10}$$

is satisfied and return to (3.9) at the end of this section.

Now, using (3.2) and (3.5) we can write

$$\sigma_i' = S_i' = S_i^{k-1} + (x_k + y_k) S_{i-1}^{k-1} - y_{k-1}^2 S_{i-2}^{k-2}, \qquad i = 2, 3, \ldots, k \tag{3.11}$$

and

$$\sigma_i'' = S_i'' = S_i^{k-1} + (x_k - y_k) S_{i-1}^{k-1} - y_{k-1}^2 S_{i-2}^{k-2}, \qquad i = 2, 3, \ldots, k. \tag{3.12}$$

Again subtracting (3.12) from (3.11), we obtain the relations

$$2 y_k S_{i-1}^{k-1} = \sigma_i' - \sigma_i'', \qquad i = 2, 3, \ldots, k. \tag{3.13}$$

Since $y_k \neq 0$, we can compute from (3.13) all the sums $S_j^{k-1} (j = 1, \ldots, k-1)$ for the principal submatrix $U^{(k-1)}$ of order $k-1$ in U_1(or U_2).

Returning to (3.6), we find that

$$x_k = \sigma_1' - S_1^{k-1} - y_k. \tag{3.14}$$

Now, (3.11), with $i = 2$, gives us

$$y_{k-1}^2 = S_2^{k-1} + (x_k + y_k)S_1^{k-1} - \sigma_2'. \tag{3.15}$$

It is again important for the rest of the computation to have

$$y_{k-1} \neq 0, \tag{3.16}$$

which could be violated only in points of the nontrivial algebraic (in fact, quadratic in this case) manifold

$$S_2^{k-1} + (x_k + y_k)S_1^{k-1} - \sigma_2' = 0. \tag{3.17}$$

Assuming that (3.16) is satisfied, take either of the two values of y_{k-1} defined by equation (3.15). Now, equation (3.11) with $i = 3, \ldots, k$ lets us determine the values for the sums S_j^{k-2}. This completes Stage 1.

Now, at the entry of Stage 2 we know all the sums $S_\ell^{k-1} (\ell = 1, 2, \ldots, k-1)$ and $S_j^{k-2} (j = 1, 2, \ldots, k-2)$. Since

$$S_1^{k-1} = S_1^{k-2} + x_{k-1}$$

we immediately find x_{k-1}. Further, the relation

$$S_2^{k-1} = S_2^{k-2} + x_{k-1}S_1^{k-2} - y_{k-2}^2 \tag{3.18}$$

defines y_{k-2}. Again we assume

$$y_{k-2} \neq 0, \tag{3.19}$$

which holds for any n-tuple $(\lambda_1, \ldots, \lambda_n)$, with the possible exception of points of the algebraic manifold (which is now a cubic one)

$$S_2^{k-1} - S_2^{k-2} - x_{k-1}S_1^{k-2} = 0. \tag{3.20}$$

We can choose either of the two values of y_{k-2} in equation (3.18). Then the relations

$$S_i^{k-1} = S_i^{k-2} + x_{k-1}S_{i-1}^{k-2} - y_{k-2}^2 S_{i-2}^{k-3}, \quad i = 3, 4, \ldots, k-1$$

determine the sums $S_m^{k-3}, m = 1, 2, \ldots, k-3$. It follows that on exit from Stage 2 (and at the entry of Stage 3) we know all the sums S_j^{k-2} and S_m^{k-3}.

At the entry of the final stage (Stage $k-1$) we know the quantities

$$S_1^2 = x_1 + x_2, \quad S_1^1 = x_1,$$

and

$$S_2^2 = x_1x_2 - y_1^2.$$

These relations determine x_1, x_2, and y_1 in an obvious way. Again we have a choice of the two values for y_1 (if $S_2^2 - x_1x_2 \neq 0$).

Now consider the case of n odd: $n = 2k - 1$. As has already been mentioned, in this case U_2 is a principal submatrix of U_1. This means that in the previous notation

$$S_i'' = S_i^{k-1}, \quad i = 1, 2, \ldots, k-1.$$

Stage 1 now proceeds as follows. First, we find x_k as

$$x_k = S_1' - S_1^{k-1} = \sigma_1' - \sigma_1''.$$

Then, using the relations

$$\sigma_2' = S_2' = S_2^{k-1} + x_k S_1^{k-1} - 2y_{k-1}^2 = \sigma_2'' + x_k \sigma_1'' - 2y_{k-1}^2, \tag{3.21}$$

we obtain the value of y_{k-1}. Once again, we should assume

$$y_{k-1} \neq 0$$

and the choice of either of the two values for y_{k-1} in (3.21) is possible. Finally, the equations

$$\sigma_i' = S_i' = S_i^{k-1} + x_k S_{i-1}^{k-1} - 2y_{k-1}^2 S_{i-2}^{k-2}, \quad i = 3, \ldots, k$$

give us the values for the sums $S_j^{k-2} (j = 1, \ldots, k-2)$. Now Stage 1 is complete. The rest of the procedure is the same as in the case of n even.

Summarizing, it has been shown that system (3.1) is consistent for any n-tuple $(\lambda_1, \ldots, \lambda_n)$, with the possible exception of points of the algebraic manifold comprised of manifolds (3.9), (3.17), (3.20), and so on. Moreover, this system has generically 2^{k-1} solutions for any $\binom{n}{k}$ partitionings of the set $\lambda_1, \ldots, \lambda_n$ into the two subsets with cardinalities $\lceil n \rceil$ and $\lfloor n \rfloor$, respectively (here $k = \lceil n \rceil$). Generically again, each of these solutions can be found by a finite sequence of arithmetic operations and square root extractions.

We can now prove that system (3.1) remains consistent for exceptional values of λ's as well. For this goal we will need some notions and results from commutative algebra. We refer the reader to [4, Chap. 3].

Interpreting the σ's (or more precisely, the λ's) in equations (3.1) as parameters and rewriting (3.1) in the form

$$f_i(x_1, \ldots, x_k, y_1, \ldots, y_{k-1}, (y_k)) = 0, \quad i = 1, 2, \ldots, n, \tag{3.22}$$

we can relate with system (3.22) the ideal I in the commutative ring of polynomials in the variables $x_1, \ldots, x_k, y_1, \ldots, y_{k-1}, (y_k)$ generated by the polynomials f_i, $i = 1, 2, \ldots, n$. Assume that some order over the monomials in the variables above is established, and a standard (or so-called Gröbner) basis of I with respect to this order is found.

THEOREM 3.1 (see [4, Thm. 3 on p. 114]). *For the particular values of $\lambda_1, \ldots, \lambda_n$ system (3.22) is inconsistent iff the corresponding standard basis contains a nonzero constant.*

For our purposes, we must modify Theorem 3.1 somewhat. The polynomials (3.22) depend on eigenvalues $\lambda_1, \ldots, \lambda_n$ prescribed (in fact, only the constant terms σ_i of these polynomials depend on λ's). Let us consider f_i as functions $\mathcal{F}_i$ not only in the former variables x's and y's but the variables $\lambda_1, \ldots, \lambda_n$ as well. Suppose any order over the monomials in the new variables is established such that any of the former variables precedes any of the λ's. Again, we assume that some standard basis $\{G_j\}$ of the ideal $\hat{I}$ generated by the polynomials $\mathcal{F}_i$ is computed. Now we can restate Theorem 3.1 as follows.

THEOREM 3.2. *System (3.22) is consistent for any values of λ's iff the standard basis $\{G_j\}$ does not contain a polynomial depending only on the λ's. If the standard*

basis does contain such polynomials $G_s(\lambda_1,\dots,\lambda_n),\dots,G_u(\lambda_1,\dots,\lambda_n)$ *then system* (3.22) *is consistent only for* n*-tuples* $(\lambda_1,\dots,\lambda_n)$ *belonging to the algebraic manifold described by the equations*

$$\begin{aligned} G_s(\lambda_1,\dots,\lambda_n) &= 0,\\ &\vdots\\ G_u(\lambda_1,\dots,\lambda_n) &= 0. \end{aligned} \tag{3.23}$$

Returning to Theorem 1.5, we see that the existence of the nontrivial polynomials $G_s,\dots,G_u$ in the corresponding standard basis would imply that system (3.1) is solvable only for λ's belonging to algebraic manifold (3.23). Meanwhile, we have already shown that this system is consistent for almost any n-tuple $(\lambda_1,\dots,\lambda_n)$. Therefore, system (3.1) is solvable for any λ's, which completes the proof of Theorem 1.5.

Remark. It has been vital for our reasoning here that system (3.1) be comprised of polynomial equations. Otherwise, our conclusion (i.e., the solvability almost everywhere implies the solvability everywhere) may not be correct. For example, the equation

$$e^z = \lambda$$

is consistent for any $\lambda \neq 0$ but not for $\lambda = 0$.

4. Proof of Theorems 1.3 and 1.4. Theorem 1.3 is an immediate consequence of Theorem 1.5 and the simple fact that any two complex matrices with identical and distinct eigenvalues are similar.

For the proof of Theorem 1.4, we refer to the following matrix theory result (see [5, Thm. 4 on p. 8, Vol. II]).

THEOREM 4.1. *If two complex symmetric matrices are similar then they are orthogonally similar.*

We mention that Theorem 4.1 has been generalized considerably in [7] by the second author of the present paper. A complex matrix A of order n is called persymmetric if $A^T = \mathcal{P}A\mathcal{P}$ and perorthogonal if $A^T = \mathcal{P}A^{-1}\mathcal{P}$. If in the relation $B = Q^{-1}AQ$ the matrix Q is perorthogonal then we say the A and B are perorthogonally similar. It follows from [7] that if two complex persymmetric matrices are similar then they are perorthogonally similar. Hence, along with Theorem 1.4 the following assertion holds.

THEOREM 4.2. *Any complex persymmetric matrix with distinct eigenvalues can be transformed via perorthogonal similarity to a doubly symmetric tridiagonal matrix.*

5. Concluding remarks. Let the prescribed eigenvalues $\lambda_1,\dots,\lambda_n$ be all real. Then our IEP admits the classical real solution T_+ with all the y's positive (and also $2^{k-1}-1$ other real solutions obtained from T_+ by changing the signs of any of the elements $y_1,\dots,y_{k-1}$). The solution T_+ corresponds to the well-defined partition of the set $\lambda_1,\dots,\lambda_n$ into the subsets λ''s and λ'''s (this fact has also been mentioned in [1]). It is most easily seen for n odd: according to (2.7)–(2.8), U_2 is a principal submatrix of U_1. Therefore, the numbers $\lambda''_1,\dots,\lambda''_{k-1}$ must interlace with $\lambda'_1,\dots,\lambda'_k$, i.e.,

$$\lambda'_1 > \lambda''_1 > \lambda'_2 > \lambda''_2 > \cdots > \lambda'_{k-1} > \lambda''_{k-1} > \lambda'_k. \tag{5.1}$$

We give a small illustration of this point. Assume that $\{\lambda\} = \{1, 2, 3\}$. For $n = 3$, the block U_2 in formula (2.8) is just the number x_1. Hence, the element x_1 in the matrix

$$T = \begin{bmatrix} x_1 & y & 0 \\ y & x_2 & y \\ 0 & y & x_1 \end{bmatrix} \tag{5.2}$$

must coincide with an eigenvalue of T. If we let $x_1 = 2$ then $\lambda_1' = 3$, $\lambda_2' = 1$, and relations (5.1) are satisfied:

$$3 > \lambda_1'' = 2 > 1.$$

We further obtain

$$x_2 = \sigma_1 - 2x_1 = 6 - 4 = 2$$

and

$$2y^2 = x_1^2 + 2x_1x_2 - \sigma_2 = 1,$$

which gives $y^2 = 1/2$. Therefore, we have

$$T_+ = \begin{bmatrix} 2 & 2^{-1/2} & 0 \\ 2^{-1/2} & 2 & 2^{-1/2} \\ 0 & 2^{-1/2} & 2 \end{bmatrix}.$$

Alternatively, if we take $x_1 = 3$ then $x_2 = 0$, and $y^2 = -1$. Hence, one of the two solutions corresponding to this choice for x_1 is

$$T_1 = \begin{bmatrix} 3 & i & 0 \\ i & 0 & i \\ 0 & i & 3 \end{bmatrix}.$$

Finally, with $x_1 = 1$ we have $x_2 = 4$, and again $y^2 = -1$. Consequently, one of the two solutions for $\lambda_1'' = 1$ is

$$T_2 = \begin{bmatrix} 1 & i & 0 \\ i & 4 & i \\ 0 & i & 1 \end{bmatrix}.$$

Our second remark concerns the situation when, in the algorithm of §3, some y_i is zero. Theorem 3.2 tells us that we can still find a doubly symmetric tridiagonal matrix T with the prescribed spectrum. On the other hand, with a zero y_i, T is bound to have some double eigenvalues. This is apparent when, for example, y_k is zero in the matrix (2.1), which implies that T is the direct sum of two permutationally similar tridiagonal matrices of half the order. It follows that an appearance of a zero y_i is possible only if some prescribed eigenvalues are repeated.

Our last remark relates to the condition in Theorem 1.5 that the λ's be distinct. As our next example shows, this condition is essential. Assume that $n = 3$, and $\{\lambda\} = \{0, 0, 0\}$. Then for matrix (5.2) we should have $x_1 = 0$, $x_2 = \sigma_1 - 2x_1 = 0$, and $y^2 = (x_1^2 + 2x_1x_2 - \sigma_2)/2 = 0$. It follows that $T = 0$, which means that we may not

obtain a nontrivial Jordan structure for our tridiagonal matrix. On the other hand, if we drop the requirement for T to be centrosymmetric then the matrix

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & i \\ 0 & i & 0 \end{bmatrix}$$

is similar to the Jordan block

$$J_3(0) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Acknowledgment. The second author wishes to thank Professor V.N. Latyshev with whom he consulted on matters relating to systems of polynomial equations. The authors would also like to thank the referees for suggestions that significantly improved the presentation.

REFERENCES

[1] A. L. ANDREW, *Eigenvectors of certain matrices*, Linear Algebra Appl., 7 (1973), pp. 151–162.
[2] D. BOLEY AND G. H. GOLUB, *A survey of matrix inverse eigenvalue problems*, Inverse Problems, 3 (1987), pp. 595–622.
[3] C. DE BOOR, *The numerically stable reconstruction of a Jacobi matrix from spectral data*, Linear Algebra Appl., 21 (1978), pp. 245–260.
[4] J. H. DAVENPORT, Y. SIRE, AND E. TOURNIER, *Computer Algebra*, 2nd ed., Academic Press, New York, 1993.
[5] F. R. GANTMACHER, *Matrix Theory*, Vol. II, Chelsea, New York, 1977.
[6] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, London, UK, 1985.
[7] Kh. D. IKRAMOV, *Singular numbers and the polar decomposition of an operator in a bilinear metric space*, U.S.S.R. Comput. Math. and Math. Phys., 28 (1988), pp. 85–87.
[8] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[9] N. H. SCOTT, *A theorem on isotropic null vectors and its application to thermoelasticity*, Proc. Roy. Soc. London A, 440 (1993), pp. 431–442.
[10] ———, *A new canonical form for complex symmetric matrices*, Proc. Roy. Soc. London A, 441 (1993), pp. 625–640.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 3, pp. 691–705, July 1996

018

UNIQUENESS OF SUM DECOMPOSITIONS OF SYMMETRIC MATRICES *

DANIEL HERSHKOWITZ†

Abstract. Let $\mathbf{F}$ be an arbitrary field with characteristic different from 2. A matrix B over $\mathbf{F}$, whose digraph is a subdigraph of a digraph D and whose sequence of row sums is equal to a sequence r of elements of $\mathbf{F}$, is said to be a (D,r)-decomposition of a symmetric matrix A if $A = B + B^T$. Necessary and sufficient conditions for the uniqueness of a (D,r)-decomposition of a given symmetric matrix A are proven. In our results we stress the case of nonnegative matrices.

Key words. symmetric matrices, sum decompositions, cycles

AMS subject classifications. 15, 05

1. Introduction.

The question of the existence of an entrywise nonnegative $m \times n$ matrix B with given row sums and column sums is of long standing. In [1] the authors study the case where B is not necessarily nonnegative but is over an arbitrary field $\mathbf{F}$ with characteristic different from 2 and where $B + B^T$ is given. Also, restrictions on the location of the nonzero entries are allowed. Generalizations of this problem to the case where only partial information is given are proved in [3] for real matrices. In this paper we investigate uniqueness of the solutions to the problems discussed in [1] and [3]. Our investigation is carried out for matrices over arbitrary fields with characteristic different from 2 and for nonnegative matrices.

Let $E(D)$ denote the arc set of a digraph D, whose vertex set we denote by $V(D)$. A digraph D' is said to be a *subdigraph* of a digraph D if $V(D') \subseteq V(D)$ and $E(D') \subseteq E(D)$. We write $D' \subseteq D$ to indicate that D' is a subdigraph of D. Let A be an $n \times n$ matrix. The *digraph* $D(A)$ *of* A is the digraph with vertex set $\{1,\ldots,n\}$ and where (i,j) is an arc in $D(A)$ if and only if $a_{ij} \neq 0$. Let A be a symmetric $n \times n$ matrix over $\mathbf{F}$, let D be a digraph with vertex set $\{1,\ldots,n\}$, and let $r = (r_1,\ldots,r_n)$ be a sequence of elements of $\mathbf{F}$. A matrix B with $D(B) \subseteq D$, with row sums $r_1,\ldots,r_n$, and such that $B + B^T = A$ is said to be a (D,r)-*decomposition* of A. If $\mathbf{F} = \mathbb{R}$ and B is nonnegative, then it is said to be a *nonnegative* (D,r)-*decomposition* of A.

In §2 we characterize the digraphs D for which there exists a unique (D,r)-decomposition of A for some sequence r in $\mathbf{F}$. We show that a (D,r)-decomposition of A is unique if and only if the graph D has a certain acyclicity property. Since the uniqueness depends entirely on D, it follows that either for *every* sequence r in $\mathbf{F}$ every (D,r)-decomposition of A is unique or *for every* sequence r in $\mathbf{F}$ every (D,r)-decomposition of A is nonunique. It also follows from our results that there always exist a digraph D and a sequence r in $\mathbf{F}$ such that A has a unique (D,r)-decomposition. We conclude §2 by showing that for *every* sequence r in $\mathbf{F}$ satisfying a certain necessary condition, there exists a digraph D such that A has a unique (D,r)-decomposition.

In §3 we prove that a matrix B may serve as a unique *nonnegative* (D,r)-decomposition for A even if D does not have the acyclicity property as long as the digraph of B has that property, and we characterize the cases in which nonnegative (D,r)-decompositions of A are unique.

*Received by the editors November 18, 1994; accepted for publication (in revised form) by R. Horn October 6, 1995.

†Mathematics Department, Technion–Israel Institute of Technology, Haifa 32000, Israel.

The discussion in §3 raises a few natural questions. Given a symmetric nonnegative matrix A, what are the digraphs D for which there exists *some* sequence r such that A has a unique nonnegative (D,r)-decomposition? What are the digraphs D for which for *every* sequence r of nonnegative numbers, a nonnegative (D,r)-decomposition of A is unique? What are the sequences r for which there exists *some* digraph D such that A has a unique nonnegative (D,r)-decomposition? What are the sequences r for which for *every* digraph D, a nonnegative (D,r)-decomposition of A is a unique nonnegative (D,r)-decomposition of A? All these questions are answered in §4. Our results are given in terms of acyclicity properties of certain graphs.

2. Uniqueness of decompositions of general symmetric matrices. Let A be a symmetric $n \times n$ matrix over an arbitrary field $\mathbf{F}$ with characteristic different from 2. A necessary and sufficient condition for the existence of a (D,r)-decomposition for A, for a *given* sequence r of n elements of $\mathbf{F}$ and a *given* digraph D with n vertices, is given in Theorem 2.5 of [1]. By Theorem 2.14 in [1] there exists a (D,r)-decomposition of A for *some* digraph D if and only if

$$\sum_{i=1}^{n} r_i = \frac{1}{2} \sum_{i,j=1}^{n} a_{ij}. \tag{2.1}$$

For a given digraph D it is easy to verify that a (D,r)-decomposition of A exists for *some* sequence r in $\mathbf{F}$ if and only if $D(A)$ is a subdigraph of the *symmetric closure* $\bar{D}$, that is, the digraph with $V(\bar{D}) = V(D)$, and where (i,j) is an arc in $\bar{D}$ whenever (i,j) and/or (j,i) is an arc in D. To see this note that if B is a (D,r)-decomposition of A, then $D(B) \subseteq D$ and, since $A = B + B^T$, we have that $D(A) \subseteq \overline{D(B)} \subseteq \bar{D}$. Conversely, if $D(A) \subseteq \bar{D}$, then the matrix B, defined by

$$b_{ij} = \begin{cases} a_{ij}, & (i,j) \in E(D), (j,i) \notin E(D); \\ \frac{a_{ij}}{2}, & (i,j), (j,i) \in E(D); \\ 0, & (i,j) \notin E(D), \end{cases}$$

satisfies $A = B + B^T$ as well as $D(B) \subseteq D$. Furthermore, since the matrix B is nonnegative entrywise whenever A is such, it follows that if A is *nonnegative*, then there exists a *nonnegative* (D,r)-decomposition of A for *some* sequence r in $\mathbf{F}$ if and only if $D(A) \subseteq \bar{D}$.

The following result characterizes the digraphs D for which a *unique* (D,r)-decomposition of A exists for some sequence r in $\mathbf{F}$. For this purpose, we denote by $\mathrm{row}(B)$ the sequence of row sums of a matrix B and by $R_i(B)$ the ith row sum of B. Also, for a digraph D we define the *symmetric part* $\mathrm{sym}(D)$ of D as the subdigraph of D whose vertex set is $V(D)$ and whose arc set consists of all arcs (i,j) of D such that (j,i) also is an arc of D. Note that if (i,i) is an arc of D, then it is also an arc in $\mathrm{sym}(D)$.

THEOREM 2.1. *Let A and B be $n \times n$ matrices over an arbitrary field $\mathbf{F}$, with characteristic different from 2, such that $A = B + B^T$; and let D be a digraph with vertex set $\{1,\ldots,n\}$ such that $D(B) \subseteq D$. The following are equivalent.*

(i) *B is a unique $(D, \mathrm{row}(B))$-decomposition of A.*

(ii) *The symmetric part of the digraph D has no cycle of length greater than 2.*

(iii) *For every sequence r in $\mathbf{F}$, a (D,r)-decomposition of A is a unique (D,r)-decomposition of A.*

Proof. (i) $\Rightarrow$ (ii). Assume that $\operatorname{sym}(D)$ has a cycle γ of length greater than 2. Define the $n \times n$ matrix E by

$$e_{ij} = \begin{cases} 0, & \text{neither } (i,j) \text{ nor } (j,i) \text{ is an arc in } \gamma, \\ \varepsilon, & (i,j) \text{ is an arc in } \gamma, \\ -\varepsilon, & (j,i) \text{ is an arc in } \gamma, \end{cases} \qquad i,j \in \{1,\dots,n\},$$

where $\varepsilon \neq 0$. Note that since γ is a cycle of length greater than 2, if (i,j) is an arc in γ, then (j,i) is not an arc in γ (although it is an arc in $\operatorname{sym}(D)$), and so the matrix E is well defined. Since γ is a cycle in $\operatorname{sym}(D)$, it follows that if (i,j) is an arc in γ, then both (i,j) and (j,i) are arcs in D. Thus, we have $D(E) \subseteq D$. Note that $E + E^T = 0$. Also, since γ is a cycle of length greater than 2 in D, it follows that each nonzero row of the matrix E has two nonzero elements, where one is ε and the other one is $-\varepsilon$. Therefore, E has zero row sums. It now follows that the matrix $B + E$ is also a $(D, \operatorname{row}(B))$-decomposition of A, in contradiction to (i).

(ii) $\Rightarrow$ (iii). Let B and C be (D,r)-decompositions of A for some sequence r in $\mathbf{F}$. It follows that $E = B - C$ is a skew-symmetric matrix with zero row sums. Hence, for every subset S of $\{1,\dots,n\}$ we have

$$\sum_{(i,j)\in S\times S^C} e_{ij} = \frac{1}{2}\sum_{i,j\in S}(e_{ij}+e_{ji}) + \sum_{(i,j)\in S\times S^C} e_{ij} = \sum_{i\in S, j\in\{1,\dots,n\}} e_{ij} = \sum_{i=1}^{n} R_i(E) = 0.$$

The matrix E thus satisfies

$$\begin{cases} D(E) \subseteq D, \\ \displaystyle\sum_{(i,j)\in S\times S^C} e_{ij} = 0, \qquad S \subseteq \{1,\dots,n\}. \end{cases} \tag{2.2}$$

Given that $\operatorname{sym}(D)$ has no cycle of length greater than 2, it follows from Theorem 3.21 in [2] that there exists a unique $n \times n$ matrix E satisfying (2.2). Since the zero matrix satisfies (2.2), it now follows that $E = 0$. Therefore, we have $B = C$.

(iii) $\Rightarrow$ (i) is trivial. □

Note that it follows from Theorem 2.1 that one can always find a digraph D and a sequence r in $\mathbf{F}$ such that a unique (D,r)-decomposition of A exists. To see this, choose D to be the digraph with vertex set $\{1,\dots,n\}$ and arc set $\{(i,j)\colon a_{ij} \neq 0, i \leq j\}$. Since $D(A) = \bar{D}$, it follows that a (D,r)-decomposition of A exists for some sequence r in $\mathbf{F}$. Since D contains no cycles other than loops, it follows from Theorem 2.1 that the (D,r)-decomposition of A is unique.

Let D be a digraph. It follows from Theorem 2.1 that either *for every* sequence r in $\mathbf{F}$ a (D,r)-decomposition of A is a unique (D,r)-decomposition of A, or *for every* sequence r in $\mathbf{F}$ a (D,r)-decomposition of A is not a unique (D,r)-decomposition of A. It also follows from Theorem 2.1 that for $n > 2$ there exist digraphs D for which (D,r)-compositions of A are not unique. For example, let D be the complete digraph D with n vertices and let $r = \operatorname{row}(\frac{1}{2}A)$. The matrix $\frac{1}{2}A$ is a (D,r)-decomposition of A and, since $\operatorname{sym}(D)$ has cycles of length greater than 2, $\frac{1}{2}A$ is not a unique (D,r)-decomposition of A. We now show that for *every* sequence $r = (r_1,\dots,r_n)$ in $\mathbf{F}$ satisfying (2.1) there exists a digraph D such that a unique (D,r)-decomposition of A exists.

THEOREM 2.2. *Let A be a symmetric $n \times n$ matrix over an arbitrary field $\mathbf{F}$ with characteristic different from 2, and let $r = (r_1, \ldots, r_n)$ be a sequence in $\mathbf{F}$ satisfying (2.1). Then there exists a digraph D with a unique (D, r)-decomposition of A.*

Proof. Let D' be the complete digraph with n vertices. By Theorem 2.14 in [1] there exists a (D', r)-decomposition B of A. Note that B is a $(D(B), r)$-decomposition of A. If $\mathrm{sym}(D(B))$ has no cycle of length greater than 2, then by Theorem 2.1 B is a unique $(D(B), r)$-decomposition of A. Else, there exists a cycle $\gamma = (i_1, i_2, \ldots, i_t), t \geq 3$, in $\mathrm{sym}(D(B))$. Let $c = b_{i_1 i_2}$ and let E be the $n \times n$ matrix defined by

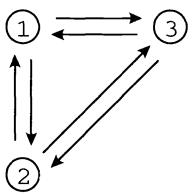
$$e_{ij} = \begin{cases} c, & (i,j) = (i_k, i_{k+1}), k = 1, \ldots, t; \\ -c, & (i,j) = (i_{k+1}, i_k), k = 1, \ldots, t; \\ 0, & \text{otherwise.} \end{cases}$$

Since $t \geq 3$, it follows that each nonzero row of E contains exactly two nonzero elements, one of which is equal to c and the other to $-c$. Therefore, the row sums of E are all zero. Since also $E + E^T = 0$, the matrix $B - E$ is a $(D(B - E), r)$-decomposition. Observe that the arc set of $D(B - E)$ is properly contained in the arc set of $D(B)$, as $(B - E)_{i_1 i_2} = 0$ while $b_{i_1 i_2} \neq 0$. We repeat this process of arc elimination from the decomposition of A until we reach a matrix C that is a $(D(C), r)$-decomposition of A and such that $\mathrm{sym}(D(C))$ has no cycle of length greater than 2. By Theorem 2.1, the digraph $D(C)$ has the required properties. □

Finally, we remark that for $n > 2$ there exists no sequence r in $\mathbf{F}$ satisfying (2.1) for which every (D, r)-decomposition of A for any digraph D is unique. To see this let D be the complete digraph with n vertices. By Theorem 2.14 in [1] there exists a (D, r)-decomposition of A that, by Theorem 2.1, is not unique.

3. Characterization of unique nonnegative decompositions. Let A be a symmetric $n \times n$ matrix. In the previous section we proved that the uniqueness of a (D, r)-decomposition of A depends entirely on the nature of the digraph D. This is not the case when *nonnegative* decompositions of nonnegative symmetric matrices are considered. Clearly, by Theorem 2.1, if B is a nonnegative (D, r)-decomposition of A and if D has no cycle of length greater than 2, then B is a unique nonnegative (D, r)-decomposition of A. The converse is, however, in general false, as is demonstrated by the following example.

Example 3.1. Let D be the digraph



,

let

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

and let $r = (r_1, r_2, r_3)$ be a sequence of nonnegative numbers. It is easy to verify that a nonnegative (D, r)-decomposition of A exists if and only if $r_1 = 0$ and $r_2 + r_3 = 1$

and that the only such decomposition is the matrix

$$B = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & r_2 \\ 0 & r_3 & 0 \end{pmatrix}.$$

Nevertheless, the digraph D contains cycles of length 3.

The analog of Theorem 2.1 in the nonnegative case is as follows.

THEOREM 3.1. *Let A and B be nonnegative $n \times n$ matrices such that $A = B + B^T$. The following are equivalent.*

(i) *B is a unique $(D(B), \text{row}(B))$-decomposition of A.*

(ii) *B is a unique nonnegative $(D(B), \text{row}(B))$-decomposition of A.*

(iii) *The symmetric part of the digraph $D(B)$ has no cycle of length greater than 2.*

(iv) *For every sequence r of real numbers, a $(D(B), r)$-decomposition of A is a unique $(D(B), r)$-decomposition of A.*

(v) *For every sequence r of nonnegative numbers, a nonnegative $(D(B), r)$-decomposition of A is a unique nonnegative $(D(B), r)$-decomposition of A.*

Proof. (i) $\Rightarrow$ (ii). The proof of this implication is trivial.

(ii) $\Rightarrow$ (iii). The proof of this implication is essentially the same as the proof of the implication (i) $\Rightarrow$ (ii) in Theorem 2.1, replacing the digraph D by $D(B)$ and observing that, since $D(E) \subseteq D(B)$, the matrix $B + E$ has the same sign pattern as B for ε sufficiently small in absolute value.

(iii) $\Rightarrow$ (iv). This implication is proven by Theorem 2.1.

(iv) $\Rightarrow$ (i). The proof of this implication is trivial.

(iv) $\Rightarrow$ (v) $\Rightarrow$ (ii). The proof of this implication also is trivial. □

Observe that Theorem 3.1 does not characterize digraphs D for which there exists a unique nonnegative (D, r)-decomposition of A for some sequence r, but it characterizes the digraphs of unique nonnegative decompositions. Clearly, a nonnegative matrix B satisfying $A = B + B^T$ is a $(D, \text{row}(B))$-decomposition of A for every digraph D such that $D(B) \subseteq D$. However, a unique nonnegative $(D(B), \text{row}(B))$-decomposition of A is not necessarily a unique nonnegative $(D, \text{row}(B))$-decomposition of A for $D(B) \subseteq D$, as is demonstrated by the following example.

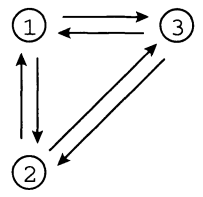
Example 3.2. Let

$$B = \begin{pmatrix} 0 & 2 & 2 \\ 2 & 0 & 2 \\ 0 & 2 & 0 \end{pmatrix}.$$

Since $\text{sym}(D(B))$ has no cycle of length greater than 2, it follows from Theorem 3.1 that B is a unique $(D(B), (4, 4, 2))$-decomposition of the matrix

$$A = \begin{pmatrix} 0 & 4 & 2 \\ 4 & 0 & 4 \\ 2 & 4 & 0 \end{pmatrix}.$$

Now, let D be the digraph



Clearly, $D(B) \subseteq D$. Now, while B is a nonnegative $(D,(4,4,2))$-decomposition of A, it is not a unique nonnegative $(D,(4,4,2))$-decomposition of A, as the matrix

$$C = \begin{pmatrix} 0 & 3 & 1 \\ 1 & 0 & 3 \\ 1 & 1 & 0 \end{pmatrix}$$

also forms such a decomposition.

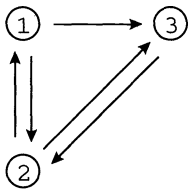
In the rest of this section we consider the case of a given nonnegative symmetric matrix A, a given digraph D, and a given sequence r of nonnegative numbers. A necessary and sufficient condition for the existence of nonnegative (D,r)-decomposition of A is given in Theorem 3.3 of [1]. Here we characterize the cases in which such a decomposition is a unique one.

DEFINITION 3.1. *Let D_1 and D_2 be two digraphs with the same vertex set. We defined $D_1 \cap D_2$ to be the digraph with the same vertex set, and whose arc set is $E(D_1) \cap E(D_2)$.*

DEFINITION 3.2. *Let D be a digraph with $V(D) = \{1,\ldots,n\}$, let A be an $n \times n$ matrix, and let $r = (r_1,\ldots,r_n)$ be a sequence of numbers. The digraph $\tilde{D}(D,A,r)$ is defined as the digraph obtained from $\mathrm{sym}(D) \cap D(A)$ by removing all arcs (u,v) for which there exists $S \subseteq \{1,\ldots,n\}$ such that $(u,v) \in S \times S^C$ or $(v,u) \in S \times S^C$ and*

$$\sum_{i \in S} \left(r_i - \sum_{\substack{j \in \{1,\ldots,n\} \\ (i,j) \in E(D),(j,i) \notin E(D)}} a_{ij} \right) = \frac{1}{2} \sum_{\substack{i,j \in S \\ (i,j) \in \mathrm{sym}(D)}} a_{ij}. \tag{3.1}$$

Example 3.3. Let D be the digraph

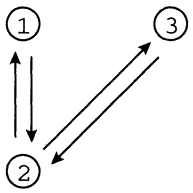


,

let

$$A = \begin{pmatrix} 0 & 2 & 2 \\ 2 & 0 & 2 \\ 2 & 2 & 0 \end{pmatrix},$$

and let $r = (0,2,4)$. The digraph $\mathrm{sym}\,(D) \cap D(A)$ is



.

It is immediate to verify that both sides of (3.1) are equal to 0 whenever $S = \{1\}$. Thus, (1, 2) and (2, 1) are not arcs in $\tilde{D}(D,A,r)$. Also, both sides of (3.1) are equal to 2 whenever $S = \{1,2\}$. Therefore, (2, 3) and (3, 2) are not arcs in $\tilde{D}(D,A,r)$, and it follows that $\tilde{D}(D,A,r)$ has no arcs.

DEFINITION 3.3. *A set S of vertices in a digraph D is said to be D-loose if for every $i \in S$ and every $j \in V(D)\backslash S$ at least one of the arcs (i,j) and (j,i) is not present in D. By convention, $\varnothing$ and $V(D)$ are D-loose sets. Observe that a set S of vertices in a digraph D is D-loose if and only if S is a union of strong components of* $\mathrm{sym}(D)$.

THEOREM 3.2. *Let A be a nonnegative symmetric $n \times n$ matrix, let D be a digraph with vertex set $\{1,\dots,n\}$, let $r = (r_1,\dots,r_n)$ be a sequence of nonnegative numbers, and assume that A has a nonnegative (D,r)-decomposition. The following are equivalent.*

(i) *A has a unique nonnegative (D,r)-decomposition.*

(ii) *The digraph $\tilde{D}(D,A,r)$ has no cycle of length greater than 2.*

Proof. (i) $\Rightarrow$ (ii). Assume that $\tilde{D}(D,A,r)$ has a cycle $\gamma = (i_1,i_2,\dots,i_t), t \geq 3$. We let $i_0 = i_t$ and $i_{t+1} = i_1$. Since A has a nonnegative (D,r)-decomposition, it follows from Theorem 3.3 in [1] that

$$\sum_{i\in S} r_i = \frac{1}{2}\sum_{i,j\in S} a_{ij} + \sum_{(i,j)\in S\times S^C\cap E(D)} a_{ij} \quad \text{for every } D\text{-loose set } S \tag{3.2}$$

and

$$\sum_{i\in S} r_i \geq \frac{1}{2}\sum_{i,j\in S} a_{ij} + \sum_{\substack{(i,j)\in S\times S^C\\ (j,i)\notin E(D)}} a_{ij} \quad \text{for every } S \subseteq \{1,\dots,n\}. \tag{3.3}$$

Note that γ is a cycle also in $D(A)$. Since A is a symmetric matrix, it follows that

$$a_{ij} > 0, \qquad (i,j) \text{ or } (j,i) \text{ is an arc in } \gamma. \tag{3.4}$$

We define an $n \times n$ matrix $\bar{A}$ by

$$\bar{a}_{ij} = \begin{cases} a_{ij}, & \text{neither } (i,j) \text{ nor } (j,i) \text{ is an arc in } \gamma,\\ a_{ij} - 2\varepsilon, & (i,j) \text{ or } (j,i) \text{ is an arc in } \gamma, \end{cases} \qquad i,j \in \{1,\dots,n\},$$

and a sequence $\bar{r} = (\bar{r}_1,\dots,\bar{r}_n)$ of numbers by

$$\bar{r}_i = \begin{cases} r_i - 2\varepsilon, & i \in \{i_1,\dots,i_t\},\\ r_i, & i \notin \{i_1,\dots,i_t\}, \end{cases} \qquad i \in \{1,\dots,n\}, \tag{3.5}$$

where $\varepsilon > 0$. By (3.4), $\bar{A}$ is a nonnegative matrix for ε sufficiently small. Let $k \in \{1,\dots,t\}$. Since (i_k, i_{k+1}) is an arc in $\tilde{D}(D,A,r)$, it follows by Definition 3.2 and by (3.3) that $r_{i_k} > 0$. Thus, by (3.5), the numbers $\bar{r}_1,\dots,\bar{r}_n$ are nonnegative for ε sufficiently small. Now, let S be a subset of $\{1,\dots,n\}$, let m be the cardinality of $\{i_1,\dots,i_t\} \cap S$, and let m' be the cardinality of the set $\{i_k \in S\colon\ i_{k+1} \in S\}$. Observe that

$$\sum_{i\in S} \bar{r}_i = \sum_{i\in S} r_i - 2m\varepsilon. \tag{3.6}$$

Let $i_k \in S$. Since $t \geq 3$, the i_kth row of $\bar{A}$ contains exactly two elements $\bar{a}_{ij}$ for which $\bar{a}_{ij} = a_{ij} - 2\varepsilon$, that is, $\bar{a}_{i_k i_{k+1}}$ and $\bar{a}_{i_k i_{k-1}}$. Therefore, we have

$$\sum_{i,j\in S} \bar{a}_{ij} = \sum_{i,j\in S} a_{ij} - 4m'\varepsilon \tag{3.7}$$

and

$$\sum_{(i,j)\in S\times S^C\cap E(D)} \bar{a}_{ij} = \sum_{(i,j)\in S\times S^C\cap E(D)} a_{ij} - 2(m-m')\varepsilon. \tag{3.8}$$

It now follows from (3.2), (3.6), (3.7), and (3.8) that

$$\sum_{i\in S} \bar{r}_i = \frac{1}{2}\sum_{i,j\in S} \bar{a}_{ij} + \sum_{(i,j)\in S\times S^C\cap E(D)} \bar{a}_{ij} \quad \text{for every } D\text{-loose set } S. \tag{3.9}$$

Also, if $m = m'$, then it follows from (3.3), (3.6), (3.7), and (3.8) that

$$\sum_{i\in S} \bar{r}_i \geq \frac{1}{2}\sum_{i,j\in S} \bar{a}_{ij} + \sum_{\substack{(i,j)\in S\times S^C \\ (j,i)\notin E(D)}} \bar{a}_{ij}. \tag{3.10}$$

If $m > m'$, then there must exist $k \in \{1,\dots,t\}$ such that $(i_k, i_{k+1}) \in S \times S^C$. Since (i_k, i_{k+1}) is an arc in $\tilde{D}(D, A, r)$, it follows from Definition 3.2 that

$$\sum_{i\in S} r_i > \frac{1}{2}\sum_{i,j\in S} a_{ij} + \sum_{\substack{(i,j)\in S\times S^C \\ (j,i)\notin E(D)}} a_{ij},$$

and again we have (3.10) for ε sufficiently small. Thus, we obtain

$$\sum_{i\in S} \bar{r}_i \geq \frac{1}{2}\sum_{i,j\in S} \bar{a}_{ij} + \sum_{\substack{(i,j)\in S\times S^C \\ (j,i)\notin E(D)}} \bar{a}_{ij} \quad \text{for every } S \subseteq \{1,\dots,n\}. \tag{3.11}$$

By Theorem 3.3 in [1], it follows from (3.9) and (3.11) that there exists a nonnegative $(D, \bar{r})$-decomposition $\bar{B}$ for $\bar{A}$. Let C be the $n \times n$ matrix defined by

$$c_{ij} = \begin{cases} \bar{b}_{ij}, & \text{neither } (i,j) \text{ nor } (j,i) \text{ is an arc in } \gamma, \\ \bar{b}_{ij} + \varepsilon, & (i,j) \text{ or } (j,i) \text{ is an arc in } \gamma, \end{cases} \qquad i,j \in \{1,\dots,n\}.$$

It is easy to verify that the C is a nonnegative (D, r)-decomposition of A. Also, $c_{ij} > \varepsilon$ whenever (i,j) or (j,i) is an arc in γ. So, $\mathrm{sym}(D(C))$ contains the cycle γ, which is of length greater than 2, and by Theorem 3.1, C is not a unique (D, r)-decomposition of A, in contradiction to the uniqueness assertion in (i).

(ii) $\Rightarrow$ (i). Let B and C be nonnegative (D, r)-decompositions of A. We have

$$C + C^T = B + B^T = A. \tag{3.12}$$

Let $D_1 = D \cap D(A)$ and let $T_1 = E(D_1)\backslash E(\mathrm{sym}(D_1))$. Observe that if for some i and j we have $b_{ij} \neq 0$ or $a_{ij} \neq 0$, then necessarily $(i,j) \in E(D)$ as well as $a_{ij} \neq 0$, and so $(i,j) \in D_1$. Therefore, if $(i,j) \in T_1$, then, since $(j,i) \notin E(D_1)$, we have $b_{ji} = c_{ji} = 0$. It now follows from (3.12) that

$$b_{ij} = c_{ij} = a_{ij}, \qquad (i,j) \in T_1. \tag{3.13}$$

We define the $n \times n$ matrices B^1 and C^1 by

$$b_{ij}^1 = \begin{cases} b_{ij}, & (i,j) \notin T_1, \\ 0, & (i,j) \in T_1; \end{cases} \qquad c_{ij}^1 = \begin{cases} c_{ij}, & (i,j) \notin T_1, \\ 0, & (i,j) \in T_1. \end{cases} \tag{3.14}$$

In view of (3.13) we have

$$R_i(B^1) = R_i(C^1) = r_i - \sum_{\substack{j \in \{1,\dots,n\} \\ (i,j) \in T_1}} a_{ij}, \qquad i \in \{1,\dots,n\}.$$

Also, it follows that

$$B = C \quad \text{if and only if } B^1 = C^1. \tag{3.15}$$

Let $A^1 = B^1 + (B^1)^T$. It follows from (3.12), (3.13), and (3.14) that $B^1 + (B^1)^T = C^1 + (C^1)^T = A^1$ and that $D(B^1), D(C^1) \subseteq \operatorname{sym}(D_1)$. Let T_2 be the set of all arcs (i,j) in $E(\operatorname{sym}(D_1))$ for which there exists $S \subseteq \{1,\dots,n\}$ such that $(i,j) \in S \times S^C$ and

$$\sum_{u \in S} r_u(B^1) = \frac{1}{2} \sum_{u,v \in S} a_{uv}^1. \tag{3.16}$$

Let $(i,j) \in T_2$ and let S be a subset of $\{1,\dots,n\}$ such that $(i,j) \in S \times S^C$ and (3.16) holds. Since $B^1 + (B^1)^T = A^1$, we have

$$\sum_{u \in S} r_u(B^1) = \sum_{u \in S,\, v \in \{1,\dots,n\}} b_{uv}^1 = \frac{1}{2} \sum_{u,v \in S} a_{uv}^1 + \sum_{(u,v) \in S \times S^C} b_{uv}^1. \tag{3.17}$$

Since A^1 is a nonnegative matrix, it follows from (3.16) and (3.17) that $b_{uv}^1 = 0$ whenever $(u,v) \in S \times S^C$. Thus, we obtain

$$b_{ij}^1 = 0, \qquad (i,j) \in T_2. \tag{3.18}$$

Similarly, we prove that

$$c_{ij}^1 = 0, \qquad (i,j) \in T_2. \tag{3.19}$$

Since $B^1 + (B^1)^T = C^1 + (C^1)^T = A^1$, it follows from (3.18) and (3.19) that

$$b_{ij}^1 = c_{ij}^1 = a_{ij}^1, \qquad (j,i) \in T_2. \tag{3.20}$$

We remove all the arcs (i,j) such that $(i,j) \in T_2$ or $(j,i) \in T_2$ from $\operatorname{sym}(D_1)$ and denote the resulting digraph by D_3. Observe that $D_3 = \tilde{D}(D, A, r)$. We define the $n \times n$ matrices B^2 and C^2 by

$$B_{ij}^2 = \begin{cases} b_{ij}^1, & (i,j),(j,i) \notin T_2, \\ 0, & \text{otherwise}; \end{cases} \qquad C_{ij}^2 = \begin{cases} c_{ij}^1, & (i,j),(j,i) \notin T_2, \\ 0, & \text{otherwise}. \end{cases}$$

In view of (3.18), (3.19), and (3.20) we have

$$R_i(B^2) = R_i(C^2) = R_i(B^1) - \sum_{\substack{j\in\{1,\ldots,n\}\\(j,i)\in T_2}} a^1_{ij}, \qquad i \in \{1,\ldots,n\}.$$

Also, it follows that we have $B^2 = C^2$ if and only if $B^1 = C^1$, and in view of (3.15) we obtain that

$$B = C \quad \text{if and only if} \quad B^2 = C^2. \tag{3.21}$$

Since $R_i(B^2) = R_i(C^2), i \in \{1,\ldots,n\}$, and since $B^2 + (B^2)^T = C^2 + (C^2)^T$, it follows that $E = B^2 - C^2$ is a skew-symmetric matrix with zero row sums. Hence, for every subset S of $\{1,\ldots,n\}$ we have

$$\sum_{(i,j)\in S\times S^C} e_{ij} = \frac{1}{2}\sum_{i,j\in S}(e_{ij}+e_{ji}) + \sum_{(i,j)\in S\times S^C} e_{ij} = \sum_{i\in S,\, j\in\{1,\ldots,n\}} e_{ij} = \sum_{i=1}^{n} R_i(E) = 0.$$

It now follows that the matrix E satisfies

$$\begin{cases} D(E) \subseteq \tilde{D}(D,A,r), \\ \displaystyle\sum_{(i,j)\in S\times S^C} e_{ij} = 0, \qquad S \subseteq \{1,\ldots,n\}. \end{cases} \tag{3.22}$$

Given that $\tilde{D}(D,A,r)$ has no cycle of length greater than 2, it follows from Theorem 3.21 in [2] that there exists a unique $n \times n$ matrix E satisfying (3.22). Since the zero matrix satisfies (3.22), it now follows that $E = 0$. Therefore, we have $B^2 = C^2$, implying by (3.21) that $B = C$. □

The following couple of examples demonstrate the claim of Theorem 3.2.

Example 3.4. Let D be the complete digraph with vertex set $\{1,2,3\}$, let

$$A = \begin{pmatrix} 0 & 2 & 2 \\ 2 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix},$$

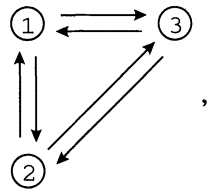
and let $r = (2,1,1)$. Observe that the nonnegative matrix

$$B = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

is a nonnegative (D,r)-decomposition of A. The digraph $\mathrm{sym}(D) \cap D(A)$, which is

has no cycle of length greater than 2. Therefore, $\tilde{D}(D,A,r)$ has no cycle of length greater than 2, and by Theorem 3.2, B is the unique nonnegative (D,r)-decomposition of A.

Example 3.5. Let D be the digraph



,

let

$$A = \begin{pmatrix} 0 & 2 & 2 \\ 2 & 0 & 2 \\ 2 & 2 & 0 \end{pmatrix},$$

and let $r = (0, 2, 4)$. Observe that the nonnegative matrix

$$B = \begin{pmatrix} 0 & 0 & 0 \\ 2 & 0 & 0 \\ 2 & 2 & 0 \end{pmatrix}$$

is a nonnegative (D, r)-decomposition of A. Since both sides of (3.1) are equal to 0 whenever $S = \{1\}$, it follows that (1, 2), (2, 1), (1, 3), and (3, 1) are not arcs in $\tilde{D}(D, A, r)$. Also, both sides of (3.1) are equal to 2 whenever $S = \{1, 2\}$. Therefore, (2, 3) and (3, 2) are not arcs in $\tilde{D}(D, A, r)$, and it follows that $\tilde{D}(D, A, r)$ has no arcs. By Theorem 3.2, the matrix B is the unique nonnegative (D, r)-decomposition of A.

4. More on unique nonnegative decompositions. Let A be a symmetric nonnegative matrix. In view of the discussion of the previous section, it is plausible to ask the following questions.

Question 4.1. (i) What are the digraphs D for which there exists a sequence r of nonnegative numbers such that a unique nonnegative (D, r)-decomposition of A exists?

(ii) What are the digraphs D for which for every sequence r of nonnegative numbers, a nonnegative (D, r)-decomposition of A is a unique nonnegative (D, r)-decomposition of A?

(iii) What are the nonnegative sequences r for which there exists a digraph D such that a unique nonnegative (D, r)-decomposition of A exists?

(iv) What are the nonnegative sequences r for which for every digraph D, a nonnegative (D, r)-decomposition of A is a unique nonnegative (D, r)-decomposition of A?

In this section we use our previous results to answer all these questions.

To answer our first question we prove the following lemma.

LEMMA 4.1. *Let A be a nonnegative symmetric $n \times n$ matrix, let D be a digraph with vertex set $\{1, \ldots, n\}$, let $r = (r_1, \ldots, r_n)$ be a sequence of nonnegative numbers, and assume that a nonnegative (D, r)-decomposition of A exists. If for a subset S of $\{1, \ldots, n\}$ we have*

$$\sum_{i \in S} r_i = \frac{1}{2} \sum_{i,j \in S} a_{ij}, \tag{4.1}$$

then $S \times S^C \cap E(\tilde{D}(D, A, r)) = \varnothing$.

Proof. Let B be a nonnegative (D,r)-decomposition of A. We define the $n \times n$ matrices $\tilde{A}$ and $\tilde{B}$ by

$$\tilde{a}_{ij} = \begin{cases} 0, & (i,j) \in E(D), (j,i) \notin E(D), \\ a_{ij}, & \text{otherwise;} \end{cases}$$

(4.2)

$$\tilde{b}_{ij} = \begin{cases} 0, & (i,j) \in E(D), (j,i) \notin E(D), \\ b_{ij}, & \text{otherwise.} \end{cases}$$

In addition we let

$$\tilde{r}_i = r_i - \sum_{\substack{j \in \{1,\ldots,n\} \\ (i,j) \in E(D), (j,i) \notin E(D)}} a_{ij}, \qquad i \in \{1,\ldots,n\}. \tag{4.3}$$

Observe that $\tilde{B} + \tilde{B}^T = \tilde{A}$ and that $R_i(\tilde{B}) = \tilde{r}_i, i \in \{1,\ldots,n\}$. Hence,

$$\sum_{i \in S} \tilde{r}_i = \sum_{i \in S,\, j \in \{1,\ldots,n\}} \tilde{b}_{ij} \geq \sum_{i,j \in S} \tilde{b}_{ij} = \frac{1}{2} \sum_{i,j \in S} \tilde{a}_{ij}. \tag{4.4}$$

On the other hand, it follows from (4.1), (4.2), and (4.3) and $\sum_{i \in S} \tilde{r}_i \leq \frac{1}{2} \sum_{i,j \in S} \tilde{a}_{ij}$, and in view of (4.4) we obtain $\sum_{i \in S} \tilde{r}_i = \frac{1}{2} \sum_{i,j \in S} \tilde{a}_{ij}$. Recall that since a (D,r)-decomposition of A exists, we have $D(A) \subseteq D$. Therefore, by (4.2) and (4.3) we now have (3.1) and by Definition 3.2 we obtain that $S \times S^C \cap E(\tilde{D}(D,A,r)) = \varnothing$. □

DEFINITION 4.1. *Let D be a digraph. A* path *in D is a sequence of distinct vertices $(i_1,\ldots,i_t)$ such that (i_k, i_{k+1}) is an arc in $D, k = 1,\ldots,t-1$. Every sequence that consists of one vertex is a path.*

LEMMA 4.2. *Let A be a nonnegative symmetric $n \times n$ matrix, let D be a digraph with vertex set $\{1,\ldots,n\}$, let $r = (r_1,\ldots,r_n)$ be a sequence of nonnegative numbers, and let B be a nonnegative (D,r)-decomposition of A. If there is no path from some vertex p to some vertex q in $D(B)$, then there is no path from p to q in $\tilde{D}(D,A,r)$.*

Proof. Let S be the set of all $k \in \{1,\ldots,n\}$ such that there exists a path from p to k in $D(B)$. It follows that $S \times S^C \cap E(D(B)) = \varnothing$. We now have

$$\sum_{i \in S} r_i = \sum_{i,j \in S} b_{ij} + \sum_{(i,j) \in S \times S^C} b_{ij} = \frac{1}{2} \sum_{i,j \in S} a_{ij} + \sum_{(i,j) \in S \times S^C \cap E(D(B))} b_{ij} = \frac{1}{2} \sum_{i,j \in S} a_{ij},$$

and it follows by Lemma 4.1 that $S \times S^C \cap E(\tilde{D}(D,A,r)) = \varnothing$. Since $p \in S$ and $q \in S^C$, it now follows that there is no path from a p to q in $\tilde{D}(D,A,r)$. □

The answer to Question 4.1.i now follows.

THEOREM 4.3. *Let A be a nonnegative symmetric $n \times n$ matrix, and let D be a digraph with vertex set $\{1,\ldots,n\}$. The following are equivalent.*

(i) *For some sequence r of nonnegative numbers there exists a unique nonnegative (D,r)-decomposition of A.*

(ii) *D has a subgraph D', with no cycle other than loops, such that $\bar{D}' = D(A)$.*

Proof. (i) ⇒ (ii). Assume that for some sequence r of nonnegative numbers there exists a unique nonnegative (D,r)-decomposition B of A. We have $D(B) \subseteq D$ and

$\overline{D(B)} = D(A)$. Also, by Theorem 3.1, $\mathrm{sym}(D(B))$ has no cycle of length greater than 2. We now remove from $D(B)$ all arcs (i,j) such that $i < j$ and $(j,i) \in E(D(B))$ and we denote the resulting graph by D'. Observe that D' has no cycle other than loops and that $\bar{D}' = \overline{D(B)} = D(A)$.

(ii) $\Rightarrow$ (i). Let D' be a subgraph of D, with no cycle other than loops, such that $\bar{D}' = D(A)$. We define a nonnegative $n \times n$ matrix B, with a digraph D', by $b_{ii} = a_{ii}/2, i \in \{1,\dots,n\}$, and

$$b_{ij} = \begin{cases} a_{ij}, & (i,j) \in E(D'), \\ 0, & (i,j) \notin E(D'), \end{cases} \qquad i,j \in \{1,\dots,n\}, i \neq j.$$

Since $D(A) = \bar{D}'$, it follows that whenever $a_{ij} \neq 0, i \neq j$, we have either $b_{ij} = a_{ij}$ or $b_{ji} = a_{ij}$ but not both. Also, if $a_{ij} = 0$, then $(i,j),(j,i) \notin E(D(A))$, and since $D(A) = \bar{D}'$, we have $(i,j),(j,i) \notin E(D')$. Thus, we have $B + B^T = A$. It now follows that B is a nonnegative (D,r)-decomposition of A, where $r = (R_1(B),\dots,R_n(B))$. Since D' has no cycle other than loops, it follows by Lemma 4.2 that $\tilde{D}(D,A,r)$ has no cycle other than loops, and by Theorem 3.2 A has a unique nonnegative (D,r)-decomposition. □

The answer to Question 4.1.ii is as follows.

THEOREM 4.4. *Let A be a nonnegative symmetric $n \times n$ matrix satisfying $D(A) \subseteq \bar{D}$, and let D be a digraph with vertex set $\{1,\dots,n\}$. The following are equivalent.*

(i) *For every sequence r of nonnegative numbers, a nonnegative (D,r)-decomposition of A is a unique nonnegative (D,r)-decomposition of A.*

(ii) *The digraph $\mathrm{sym}(D) \cap D(A)$ has no cycle of length greater than 2.*

Proof. (i) $\Rightarrow$ (ii). We define the nonnegative $n \times n$ matrix B with $D(B) = D \cap D(A)$ by

$$b_{ij} = \begin{cases} a_{ij}, & (i,j) \in E(D \cap D(A)), (j,i) \notin E(D), \\ \frac{1}{2}a_{ij}, & (i,j),(j,i) \in E(D \cap D(A)), \\ 0, & \text{otherwise}, \end{cases} \qquad i,j \in \{1,\dots,n\}.$$

Observe that since $D(A) \subseteq \bar{D}$, we have $A = B + B^T$, and so B is a nonnegative (D,r)-decomposition of A, where $r = (R_1(B),\dots,R_n(B))$. If $\mathrm{sym}(D(B)) = \mathrm{sym}(D) \cap D(A)$ has a cycle of length greater than 2, then by Theorem 3.1 B is not a unique nonnegative (D,r)-decomposition of A.

(ii) $\Rightarrow$ (i). Since $\mathrm{sym}(D) \cap D(A)$ has no cycle of length greater than 2, it follows by Definition 3.2 that for every sequence r of nonnegative numbers the digraph $\tilde{D}(D,A,r)$ has no cycle of length greater than 2. Our claim follows from Theorem 3.2. □

The following theorem is a related result.

THEOREM 4.5. *Let D be a digraph with vertex set $\{1,\dots,n\}$. The following are equivalent.*

(i) *For every nonnegative symmetric $n \times n$ matrix A satisfying $D(A) \subseteq \bar{D}$ and every sequence r of nonnegative numbers, a nonnegative (D,r)-decomposition of A is a unique nonnegative (D,r)-decomposition of A.*

(ii) *The symmetric part of the digraph D has no cycle of length greater than 2.*

Proof. (i) $\Rightarrow$ (ii). Assume that $\mathrm{sym}(D)$ has a cycle γ of length greater than 2. We define the nonnegative $n \times n$ matrix B with $D(B) = D$ by

$$b_{ij} = \begin{cases} 1, & (i,j) \in E(D), \\ 0, & (i,j) \notin E(D), \end{cases} \qquad i,j \in \{1,\dots,n\},$$

and let $r = (R_1(B), \ldots, R_n(B))$. B is a (D, r)-decomposition of $A = B + B^T$. Since γ is a cycle in $\mathrm{sym}(D)$, by Theorem 3.1 B is not a unique such decomposition.

(ii) $\Rightarrow$ (i). This implication follows from Theorem 2.1. □

Let A be a nonnegative symmetric $n \times n$ matrix, and let $r = (r_1, \ldots, r_n)$ be a sequence of nonnegative numbers. To answer to Question 4.1.iii we recall that, by Theorem 3.19 in [1], there exists a nonnegative (D, r)-decomposition of A for some digraph D if and only if the sequence r satisfies (2.1) as well as

$$\sum_{i \in S} r_i \geq \frac{1}{2} \sum_{i,j \in S} a_{ij} \quad \text{for every } S \subseteq \{1, \ldots, n\}. \tag{4.5}$$

THEOREM 4.6. *Let A be a nonnegative symmetric $n \times n$ matrix, and let $r = (r_1, \ldots, r_n)$ be a sequence of nonnegative numbers satisfying* (2.1) *and* (4.5). *Then there exists a digraph D with a unique nonnegative (D, r)-decomposition of A.*

Proof. Let D' be the complete digraph with n vertices. By Theorem 3.19 in [1] there exists a nonnegative (D', r)-decomposition B of A. Note that B is a $(D(B), r)$-decomposition of A. If $\mathrm{sym}(D(B))$ has no cycle of length greater than 2, then by Theorem 3.1 B is a unique nonnegative $(D(B), r)$-decomposition of A. Else, there exists a cycle $\gamma = (i_1, i_2, \ldots, i_t), t \geq 3$, in $\mathrm{sym}(D(B))$. Let $i_{t+1} = i_1$, and let $c = \min_{k \in \{1,\ldots,t\}} \{b_{i_k i_{k+1}}, b_{i_{k+1} i_k}\}$. Without loss of generality we may assume that $c = b_{i_1 i_2}$. Let E be the $n \times n$ matrix defined by

$$e_{ij} = \begin{cases} c, & (i,j) = (i_k, i_{k+1}), k = 1, \ldots, t; \\ -c, & (i,j) = (i_{k+1}, i_k), k = 1, \ldots, t; \\ 0, & \text{otherwise.} \end{cases}$$

Since $t \geq 3$, it follows that each nonzero row of E contains exactly two nonzero elements, one of which is equal to c and the other to $-c$. Therefore, the row sums of E are all zero. Also $B - E$ is a nonnegative matrix. Since $E + E^T = 0$, the matrix $B - E$ is a nonnegative $(D(B - E), r)$-decomposition. Observe that the arc set of $D(B - E)$ is properly contained in the arc set of $D(B)$, as $(B - E)_{i_1 i_2} = 0$ while $b_{i_1 i_2} \neq 0$. We repeat this process of arc elimination from the decomposition of A until we reach a matrix C that is a nonnegative $(D(C), r)$-decomposition of A and such that $\mathrm{sym}(D(C))$ has no cycle of length greater than 2. By Theorem 3.1, the digraph $D(C)$ has the required properties. □

Finally, we answer Question 4.1.iv.

DEFINITION 4.2. *Let A be an $n \times n$ matrix, and let $r = (r_1, \ldots, r_n)$ be a sequence of n numbers. The digraph $\tilde{D}(A, r)$ is defined as the digraph $\tilde{D}(D', A, r)$, where D' is the complete digraph with vertices $\{1, \ldots, n\}$. That is, $\tilde{D}(A, r)$ is obtained from D' by removing all arcs (u, v) for which there exists $S \subseteq \{1, \ldots, n\}$ such that $(u, v) \in S \times S^C$ or $(v, u) \in S \times S^C$ and*

$$\sum_{i \in S} r_i = \frac{1}{2} \sum_{i,j \in S} a_{ij}.$$

THEOREM 4.7. *Let A be a nonnegative symmetric $n \times n$ matrix, and let $r = (r_1, \ldots, r_n)$ be a sequence of nonnegative numbers satisfying* (2.1) *and* (4.5). *The following are equivalent.*

(i) *For every digraph D, a nonnegative (D, r)-decomposition of A is a unique nonnegative (D, r)-decomposition of A.*

(ii) *A has a unique nonnegative (D', r)-decomposition, where D' is the complete digraph with vertices $\{1, \ldots, n\}$.*

(iii) *The digraph $\tilde{D}(A, r)$ has no cycle of length greater than 2.*

Proof. (i) $\Rightarrow$ (ii). By Theorem 3.19 in [1] there exists a nonnegative (D', r)-decomposition B of A. Obviously, (ii) now follows from (i).

(ii) $\Rightarrow$ (iii). This implication is proven by Theorem 3.2.

(iii) $\Rightarrow$ (i). This implication is proven by Theorem 3.2 since for every digraph D a nonnegative (D, r)-decomposition of A is a nonnegative (D', r)-decomposition of A. □

REFERENCES

[1] J. A. Dias Da Silva, D. Hershkowitz, and H. Schneider, *Sum decompositions of symmetric matrices*, Linear Algebra Appl., 208/209 (1994), pp. 523–537.

[2] ———, *Existence of matrices with prescribed off-diagonal block element sums*, Linear and Multilinear Algebra, 40 (1995), pp. 15–28.

[3] D. Hershkowitz, A. J. Hoffman, and H. Schneider, *On the Existence of Matrices With Prescribed Partial Sums of Elements*, preprint.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 4, pp. 707–729, October 1996

MINIMAL RESIDUAL METHOD STRONGER THAN POLYNOMIAL PRECONDITIONING*

V. FABER†, W. JOUBERT†, E. KNILL†, AND T. MANTEUFFEL‡

Abstract. This paper compares the convergence behavior of two popular iterative methods for solving systems of linear equations: the s-step restarted minimal residual method (commonly implemented by algorithms such as GMRES(s)) and $(s-1)$-degree polynomial preconditioning. It is known that for normal matrices, and in particular for symmetric positive definite matrices, the convergence bounds for the two methods are the same. In this paper we demonstrate that for matrices unitarily equivalent to an upper triangular Toeplitz matrix, a similar result holds; namely, either both methods converge or both fail to converge. However, we show this result cannot be generalized to all matrices. Specifically, we develop a method, based on convexity properties of the generalized field of values of powers of the iteration matrix, to obtain examples of real matrices for which GMRES(s) converges for every initial vector, but every $(s-1)$-degree polynomial preconditioning stagnates or diverges for some initial vector.

Key words. linear systems, iterative methods, nonsymmetric, nonnormal matrix, GMRES, polynomial preconditioning, convergence, field of values

AMS subject classifications. 65F10, 65F15

1. Introduction.

A chief goal of numerical linear algebra is to solve linear systems of the form

$$Au = b \tag{1}$$

in a reliable and fast way. Here $A \in \mathbb{C}^{N\times N}$ is nonsingular and is possibly the result of a preconditioning operation such as $Q\hat{A}u = Q\hat{b}$.

The set of *polynomial methods* (sometimes loosely referred to as *Krylov subspace methods*) has proven to be extremely powerful for solving many types of linear systems. These are defined by

$$u^{(n)} = u^{(0)} + q_{n-1}(A)r^{(0)} \tag{2}$$

or

$$r^{(n)} = [I - Aq_{n-1}(A)]r^{(0)}, \tag{3}$$

where $u^{(0)}$ is the initial guess, $\{u^{(i)}\}_{i\geq 0}$ denote iterates, $r^{(i)} = b - Au^{(i)}$ are the associated residuals, and each q_{n-1} is a polynomial of degree no greater than $n-1$. Examples of such methods are the conjugate gradient method, the biconjugate gradient method, the minimal residual method, and polynomial preconditioned conjugate gradient methods (see [1], [12] for overviews of such methods).

Polynomial methods owe their strength to the fact that the properties of polynomials lend themselves to rapid convergence rates for many cases, in particular when A is Hermitian and positive definite (HPD). However, a comprehensive theory of convergence of polynomial methods for general matrices has remained elusive. The purpose of this paper is to address the issue of convergence rates of some of these methods.

* Received by the editors May 22, 1995; accepted for publication (in revised form) by A. Greenbaum October 16, 1995. This work was supported in part by Department of Energy grant W-7405-ENG-36, with Los Alamos National Laboratory.

† Los Alamos National Laboratory, Los Alamos, NM 87545 (vxf@lanl.gov, wdj@lanl.gov, knill@lanl.gov).

‡ University of Colorado at Boulder, Boulder, CO 80309 (tmanteuf@newton.colorado.edu).

A natural choice for a polynomial method is to require that q_{n-1} be "optimal" in some sense. For example, given A, b, and $u^{(0)}$, let

(4) q_{n-1} be a polynomial (cf. (3)) of degree at most $n-1$ which minimizes $||r^{(n)}||$,

where $||\cdot||$ is used here and throughout to refer to the standard 2-norm. This defines the *minimal residual method*, of which the GMRES algorithm is the best-known implementation [14]. To limit the average work per iteration, this method is typically restarted every s steps, leading to algorithms such as GMRES(s). The resulting method is

(5) $r^{(ms+s)} = [I - Aq_{s-1;m}(A)]r^{(ms)}$, $q_{s-1;m}$ selected by (4) based on $r^{(ms)}$.

The average work per iteration for such algorithms applied to general matrices is proportional to sN; larger values of s generally improve convergence but also increase the work per iteration.

A considerably cheaper algorithm is *polynomial preconditioning* coupled with the basic one-step iterative method; namely,

$$r^{(ms)} = [I - Aq_{s-1}(A)]^m r^{(0)}, \tag{6}$$

where the polynomial q_{s-1} is chosen in some appropriate fashion. (Of course, polynomial preconditioning can also be accelerated, for example, by applying GMRES to the preconditioned system $q_{s-1}(A)Au = q_{s-1}(A)b$.) Provided that a good polynomial q_{s-1} can be found, this algorithm requires only order N work per iteration, independent of s. Furthermore, the algorithm can be very successful on certain computer architectures for which inner product computations are particularly expensive, since GMRES requires inner product computations but polynomial preconditioning does not.

Good polynomials q_{s-1} are not always easy to find, so we consider here the *optimal polynomial preconditioning* of degree $s-1$ for a matrix A, defined as a polynomial q_{s-1} of degree no greater than $s-1$ which solves the minimization problem

$$\text{minimize } ||I - Aq_{s-1}(A)||. \tag{7}$$

It can be shown that such a minimizer exists and under reasonable assumptions in fact is unique [5].

The performance of this preconditioner is in some sense the best possible for a polynomial preconditioner. However, it should be noted that more sophisticated optimization procedures might be considered, such as

$$\text{minimize } \left\| \prod_{i=1}^{m} [I - Aq_{s-1;i}(A)]r^{(0)} \right\|, \tag{8}$$

$$\text{minimize } ||[I - Aq_{s-1}(A)]^m||, \tag{9}$$

which may in some cases yield faster convergence. In particular, (8) selects a set of polynomials to give optimality globally over all s-step cycles rather than locally for each cycle (as GMRES(s) does), and (9) selects a single polynomial preconditioner that performs well over an aggregated set of cycles without regard to its single-cycle

performance. The study of the convergence behavior of these methods is beyond the scope of this paper.

Methods (5) and (6), (7) are similar, but they differ in the following important respect: (6), (7) uses the same polynomial repeatedly, whereas (5) selects the best polynomial for each cycle. If an adequate polynomial can be found, then (6), (7) is much more economical than (5). This is true especially when s is large, which is usually desirable in order to increase the convergence rate [11]. However, it is not clear whether (6), (7) converges as fast as (5). The purpose of this study is to investigate the relative rates of convergence of (6), (7) compared to (5).

It can be shown that the convergence behavior of the restarted minimal residual method is bounded by

$$\frac{||r^{(ms)}||}{||r^{(0)}||} \leq \left[\max_{||r||=1} \min_{q_{s-1}} ||[I - Aq_{s-1}(A)]r||\right]^m, \tag{10}$$

whereas the convergence behavior of the basic iterative method applied to optimal polynomial preconditioning is bounded by

$$\frac{||r^{(ms)}||}{||r^{(0)}||} \leq \left[\min_{q_{s-1}} ||I - Aq_{s-1}(A)||\right]^m = \left[\min_{q_{s-1}} \max_{||r||=1} ||[I - Aq_{s-1}(A)]r||\right]^m. \tag{11}$$

This motivates the question of the relative behavior of

$$\psi_s(A) = \max_{||r||=1} \min_{q_{s-1}} ||[I - Aq_{s-1}(A)]r|| \quad \text{and} \quad \varphi_s(A) = \min_{q_{s-1}} \max_{||r||=1} ||[I - Aq_{s-1}(A)]r||. \tag{12}$$

The two functions $\psi_n(A)$ and $\varphi_n(A)$ will be used as measures of the convergence behavior of these two popular iterative methods.

We should say a few words about the tightness of bounds (10), (11). It is not clear that inequality (10) is sharp, in the sense that for every A, m, and s there is an $r^{(0)}$ for which (10) is an equality. This difficulty is due to the nonlinear nature of the minimization process (5). However, it does hold that $\psi_s(A) = 1$ if and only if there is an $r^{(0)}$ such that $r^{(0)} = r^{(s)} = r^{(2s)} \ldots$; i.e., the iterative method stagnates. Similarly, bound (11) may not be sharp, and in view of (9), polynomial preconditioners may exist which have better multicycle convergence than the (locally) optimal polynomial preconditioner described here. However, the (locally) optimal polynomial preconditioner is in some sense based on the best information known for a single cycle, and $\varphi_s(A) = 1$ if and only if this polynomial preconditioner coupled with the basic iterative method stagnates for some $r^{(0)}$. Furthermore, this assumes the optimal preconditioner can be economically found; more standard preconditioners may give an even worse performance.

The comparison of $\psi_n(A)$ and $\varphi_n(A)$ can tell us whether replacing the more strongly convergent GMRES with the faster polynomial preconditioning can be done without destroying convergence. For some classes of matrices, e.g., HPD matrices and normal matrices (i.e., matrices A for which $AA^* = A^*A$, where * denotes conjugate transpose—this includes Hermitian, skew-Hermitian, unitary, and circulant matrices, for example), it is known that $\psi_n(A) = \varphi_n(A)$, so both methods have the same convergence rate for such matrices [2], [10], [3]. In this paper we show further that the class of upper triangular Toeplitz matrices A satisfy $\psi_n(A) = 1$ iff $\varphi_n(A) = 1$; that is, replacing GMRES(s) with the optimal polynomial preconditioning of degree

$s-1$ cannot cause stagnation. On the other hand, we do give an example over the real numbers of a matrix for which restarted GMRES(s) converges but the optimal polynomial preconditioning of degree $s-1$ can stagnate. That is, GMRES is overall a more robust iterative method than the corresponding polynomial preconditioning.

Here is an outline of the remainder of the paper. In §2 a general theoretical framework for ψ_n and φ_n is established, various elementary results are obtained, and known results are summarized. In §3 the results for Toeplitz matrices are presented, and in §4 an example for which $\psi_n \neq \varphi_n$ is given. Implications of this result are discussed in §5.

2. General results on convergence. The following sections give the basic framework of tools used to analyze the convergence behavior of these iterative methods. Furthermore, a combination of existing and new results is given on the convergence behavior of the minimal residual method and optimal polynomial preconditioning.

2.1. Convergence bounds: Definitions and elementary results. The convergence bounds for the minimal residual method and for optimal polynomial preconditioning are given below. These definitions are slightly more general than the definitions given in §1 in that they differentiate between the solution of real and complex linear systems.

Let $\mathbb{K}$ denote either the field of real numbers $\mathbb{R}$ or the complex numbers $\mathbb{C}$. Let $\mathbb{K}_i[z]$ denote polynomials over $\mathbb{K}$ of degree no greater than i. Then for $A \in \mathbb{K}^{N\times N}$, let

$$\begin{aligned}\varphi_{n,\mathbb{K}}(A) &= \inf_{q\in\mathbb{K}_{n-1}[z]} \sup_{v\in\mathbb{K}^N:||v||=1} ||(I - Aq(A))v|| \\ &= \inf_{q\in\mathbb{K}_{n-1}[z]} ||I - Aq(A)||, \\ \psi_{n,\mathbb{K}}(A) &= \sup_{v\in\mathbb{K}^N:||v||=1} \inf_{q\in\mathbb{K}_{n-1}[z]} ||(I - Aq(A))v||.\end{aligned}$$

For both φ and ψ, when $\mathbb{K} = \mathbb{R}$, the infimum can be taken over either real or complex polynomials without affecting the values of φ and ψ [11].

Let us now confirm that in fact the convergence bound for the minimal residual method is at least as strong as that for optimal polynomial preconditioning. The proposition also sheds some light on what happens to the bounds when A is singular. Define the degree of a matrix $d(A)$ as $\min\{\deg(P) : P(A) = 0,\ P \text{ monic}\}$. Then we have the following proposition.

PROPOSITION 2.1. *Let $A \in \mathbb{K}^{N\times N}$. Then $0 \leq \psi_{n,\mathbb{K}}(A) \leq \varphi_{n,\mathbb{K}}(A) \leq 1$. If $d = d(A) \leq N$ is the degree of the minimal polynomial of A, then $0 < \psi_{n,\mathbb{K}}(A) \leq \varphi_{n,\mathbb{K}}(A)$ for any $n < d$. For $n \geq d$ and for A also nonsingular, $\psi_{n,\mathbb{K}}(A) = \varphi_{n,\mathbb{K}}(A) = 0$. If A is singular, then $\psi_{n,\mathbb{K}}(A) = \varphi_{n,\mathbb{K}}(A) = 1$ for any n.*

Proof. The first inequality is easily shown; see [11]. The result for $n < d$ is shown as follows. If $\psi_{n,\mathbb{K}}(A) = 0$, then for all $v \in \mathbb{K}^N$, $||v|| = 1$, $\inf_q ||(I - Aq(A))v|| = 0$. It is easily seen that if v is chosen to contain nonzero components of all generalized eigenvectors of A, this leads to a contradiction. For $n \geq d$, if A is invertible, the monic minimal polynomial for A can be renormalized so that the constant term is 1. If A is not invertible, then v can be chosen from the null space of A, and $(I - Aq(A))v = v$ for any q. □

Define $f_n : \mathbb{K}^N \times \mathbb{K}^{N\times N} \to \mathbb{K}$ by

$$f_n(v, A) = \inf_{q\in\mathbb{K}_{n-1}[x]} ||(I - Aq(A))v||^2.$$

This function defines the convergence of the minimal residual method applied to a specific vector: for $A \in \mathbb{K}^{N\times N}$, $\psi_{n,\mathbb{K}}(A)^2 = \sup_{v\in\mathbb{K}^N} f_n(v, A)/||v||^2$.

Let $\underline{K}_n(v, A) \in \mathbb{C}^{N\times n}$ be defined by $\underline{K}_n(v, A)e_i = A^{i-1}v$, where e_i is the standard unit basis vector. Also define the degree of a vector $d(v, A)$ as $\min\{\deg(P) : P(A)v = 0, P \text{ monic}\}$. Note that $A\underline{K}_n(v, A) = \underline{K}_n(Av, A)$ is full rank if and only if $d(Av, A) \geq n$, and when A is nonsingular $d(Av, A) = d(v, A)$. Then for $A\underline{K}_n(v, A)$ full rank,

(13)
$$f_n(v, A) = \hat{f}_n(v, A) \equiv v^*v - v^*A\underline{K}_n(v, A)[\underline{K}_n(v, A)^*A^*A\underline{K}_n(v, A)]^{-1}\underline{K}_n(v, A)^*A^*v.$$

PROPOSITION 2.2. *For fixed A, $f_n(\cdot, A)$, considered as a function of $2N$ real variables under the identification $\mathbb{R}^{2N} \cong \mathbb{C}^N$, is C^∞ on the complement of the closed set $S = \{v : d(Av, A) < n\}$. If A is nonsingular then $f_n(\cdot, A)$ is continuous everywhere, and if furthermore $n \leq d(A)$ then S has measure zero.*

Proof. The first statement follows from the above remarks and the fact that $\hat{f}_n(v, A)$ is a rational function of the real and imaginary parts of the elements of v and A. When A is nonsingular and $n \geq d(A)$, $f_n(\cdot, A)$ is uniformly zero. Otherwise, it suffices to show that for $v_i \notin S$ and $v_i \to v \in S$, $f_n(v_i, A) \to f_n(v, A)$. Let $p(A)$ be the minimal polynomial for v with respect to A; note that $\deg p < n$. Let $\tilde{p}(z) = p(z)/p(0)$. Then $0 \leq f_n(v_i, A) \leq ||\tilde{p}(A)v_i||^2 \to 0 = f_n(v, A)$ □

2.2. Continuity of the bound functions. To understand the convergence of polynomial iterative methods, it is desirable to get a better understanding of how the bound functions $\psi_{n,\mathbb{K}}$ and $\varphi_{n,\mathbb{K}}$ behave. See [11] and [10] for elementary results on these functions. In what follows we demonstrate in particular that $\psi_{n,\mathbb{K}}$ and $\varphi_{n,\mathbb{K}}$ are continuous functions on the open set of nonsingular matrices. It will be noted, however, that they are not generally continuous everywhere; for example, $\psi_{N,\mathbb{K}} = \varphi_{N,\mathbb{K}} = 1$ for A singular but $\psi_{N,\mathbb{K}} = \varphi_{N,\mathbb{K}} = 0$ for A nonsingular.

We begin with the following lemma.

LEMMA 2.3. *For A, $A_i \in \mathbb{C}^{N\times N}$ with $A_i \to A$ and for r, $r_i \in \mathbb{C}^N$ with $r_i \to r$, $\liminf_{i\to\infty} d(r_i, A_i) \geq d(r, A)$.*

Proof. Let P_i be the minimal polynomial for r_i with respect to A_i, and let P be the minimal polynomial for r with respect to A. Note first that the eigenvalues of all $\{A_i\}$ form a bounded set: if $\{\lambda, v\}$ is an eigenpair of A_i, then $|\lambda| = ||A_iv||/||v|| \leq ||A_i||$, but since $||A_i||$ is bounded near $||A||$, $|\lambda|$ must be bounded. Thus the polynomials P_i must reside within a bounded set because the coefficients of P_i are products and sums of the eigenvalues of A_i.

Suppose there exists a subsequence i_j such that $\deg(P_{i_j}) = d(r_{i_j}, A_{i_j}) < d(r, A)$ for all j. By boundedness, this subsequence has a convergent subsequence P_{i_k} with $P_{i_k} \to \hat{P}$ for some polynomial $\hat{P}$. We note that necessarily $\deg(\hat{P}) < d(r, A)$ by the choice of P_{i_j}. Furthermore,

$$||\hat{P}(A)r|| \leq ||\hat{P}(A)r - P_{i_k}(A)r|| + ||P_{i_k}(A)r - P_{i_k}(A_{i_k})r_{i_k}||$$

since $P_{i_k}(A_{i_k})r_{i_k} = 0$. Since $P_{i_k} \to \hat{P}$, $||\hat{P}(A)r - P_{i_k}(A)r|| \to 0$. Furthermore, since the $\{P_{i_k}\}$ are bounded, $A_{i_k} \to A$ and $r_{i_k} \to r$, we have for $P_{i_k}(z) = \sum c_{i_kj}z^j$,

$||P_{i_k}(A)r - P_{i_k}(A_{i_k})r_{i_k}|| \leq \sum |c_{i_k,j}| \cdot ||A^j r - A^j_{i,k} r_{i_k}|| \to 0$. Thus $\hat{P}(A)r = 0$, which is a contradiction. □

Note that in fact we may have $d(r_i, A_i) \to d > d(r, A)$; e.g., $A_i = \text{diag}[1+1/i, 1+2/i]$, $r_i = [1,1]^T$.

LEMMA 2.4. *Let $f : S_1 \times S_2 \to \mathbb{R}$ be continuous, and let $S_1 \subseteq \mathbb{C}^{n_1}$ compact and $S_2 \subseteq \mathbb{C}^{n_2}$ open. Then $F(u) = \sup_{v \in S} f(v,u)$ and $G(u) = \inf_{v \in S} f(v,u)$ are continuous.*

Proof. We prove the result for F. Note that $F(u) < \infty$ for any u. Let $u_i \to u$, with u_i, $u \in S_2$. We first show that $\sup_{v \in S} |f(v,u_i) - f(v,u)| \to 0$. Otherwise, there exists $\epsilon > 0$, a subsequence i_j, and vectors $v_{i_j} \in S$ such that $|f(v_{i_j}, u_{i_j}) - f(v_{i_j}, u)| \geq \epsilon$ for all j. By the compactness of S_1, there exists i_k, a subsequence of i_j, such that $v_{i_k} \to \hat{v} \in S$. Then

$$|f(v_{i_k}, u_{i_k}) - f(v_{i_k}, u)| \leq |f(v_{i_k}, u_{i_k}) - f(\hat{v}, u)| + |f(\hat{v}, u) - f(v_{i_k}, u)| \to 0$$

by the continuity of f, which is a contradiction.

Thus for any $\epsilon > 0$ there is some m such that for all $i > m$,

$$f(v,u) - \epsilon < f(v,u_i) < f(v,u) + \epsilon$$

for any $v \in S_1$. Taking suprema over S_1 yields

$$F(u) - \epsilon \leq F(u_i) \leq F(u) + \epsilon,$$

implying $|F(u) - F(u_i)| \leq \epsilon$, giving the result. □

THEOREM 2.5. *The function $\psi_{n,\mathbb{K}}(\cdot)$ is continuous on the open set of nonsingular matrices in $\mathbb{K}^{N \times N}$.*

Proof. Note $\psi_{n,\mathbb{K}}(A)^2 = \sup_{v \in \mathbb{K}^N, ||v||=1} f_n(v,A)$, where f_n is as defined earlier. If $n \geq d(A)$, then $\psi_{n,\mathbb{K}}(A) = 0$ and the result follows by letting P_n be a scaling of the minimal polynomial for A: $0 \leq \psi_{n,\mathbb{K}}(A_i) \leq \varphi_{n,\mathbb{K}}(A_i) \leq ||P_n(A_i)|| \to 0$ for $A_i \to A$. Otherwise we proceed as follows.

Since the set of nonsingular matrices constitutes an open set in $\mathbb{K}^{N \times N}$, the previous lemma will give the result if it can be shown that the map $f_n(\cdot,\cdot)$ is continuous on $\{r : ||r|| = 1\} \times \{A : A \text{ nonsingular}\}$. Select A, A_i nonsingular and $||r|| = ||r_i|| = 1$, and let $A_i \to A$ and $r_i \to r$. If $d(r,A) \leq n$, then for P_n, a scaling of the minimal polynomial of r with respect to A, $0 \leq f_n(r_i, A_i) \leq ||P_n(A_i)r_i|| \to 0$. Otherwise, by the previous lemma, for some m, $d(r_i, A_i) \geq d(r,A) > n$ for all $i > m$. Then $f_n(r_i, A_i) = \hat{f}_n(r_i, A_i)$ for all $i > m$, and $f_n(r,A) = \hat{f}_n(r,A)$. Since the rational function $\hat{f}_n(\cdot,\cdot)$ is continuous on the open set of elements for which it is defined, the result follows. □

THEOREM 2.6. *The function $\varphi_{n,\mathbb{K}}(\cdot)$ is continuous on the open set of nonsingular matrices in $\mathbb{K}^{N \times N}$.*

Proof. Since $\varphi_{n,\mathbb{R}}(A) = \varphi_{n,\mathbb{C}}(A)$ for $A \in \mathbb{R}^{N \times N}$, the result need only be shown for $\mathbb{K} = \mathbb{C}$. We have

$$\varphi_{n,\mathbb{C}}(A) = \inf_{P_n(0)=1} ||P_n(A)||.$$

As in the previous theorem, assume that $n < d(A)$. Note that $\varphi_{n,\mathbb{C}}(A) = \inf_{P_n \in S} ||P_n(A)||$, where $S = \{P_n \in \mathbb{C}_n[z] : P_n(0) = 1, ||P_n(A)|| \leq 2\}$, and as usual $\deg P_n \leq n$. Note that S is nonempty: $P(z) = 1$ defines a polynomial found in S. Importantly, by the linear independence of $\{A^i\}_{i=0}^n$, the set S is bounded. Also note

that S is closed. Thus, using the notation of Lemma 2.4, we let $f(P_n, A) = ||P_n(A)||$ to obtain the result via that lemma. □

The previous two theorems confirm the continuity of the bound functions. This is an expected result—a small perturbation of the matrix A should cause only a small perturbation in the behavior of the iterative method.

2.3. The generalized field of values. In the study of conjugate gradient methods, the field of values $F(A) = \{x^*Ax : x \in \mathbb{C}^N, \ ||x|| = 1\}$ [7] plays a prominent role in determining convergence behavior of the methods. In this study we will make use of the concept of the *generalized* field of values (see, e.g., [6]) to develop a more powerful set of results.

Let us first establish some notation. For any matrices $A = \{a_{i,j}\} \in \mathbb{K}^{r_a \times c_a}$ and $B = \{b_{i,j}\} \in \mathbb{K}^{r_b \times c_b}$, define the standard *tensor product of matrices* $A \otimes B \in \mathbb{K}^{r_a r_b \times c_a c_b}$ by $e^*_{j+r_b(i-1)}(A \otimes B)e_{l+c_b(k-1)} = a_{i,k}b_{j,l}$ [13]. Note that when A, B, C, and D are matrices of dimension such that AC and BD are well defined, then $(A \otimes B)(C \otimes D) = AC \otimes BD$.

For a set of matrices $\{A_i\}_{i=1}^n \subseteq \mathbb{C}^{N \times N}$, the matrix $R = \sum_{i=1}^n e_i \otimes A_i$ can be considered as a vector of quadratic forms $v \mapsto (I \otimes v^*)R(1 \otimes v) \in \mathbb{C}^n$. Thus R may be thought of as representing a map from $\mathbb{C}^N$ to $\mathbb{C}^n$.

Let the *generalized field of values* over a field $\mathbb{K}$ for a set of matrices $\{A_i\}_{i=1}^n \subseteq \mathbb{C}^{N \times N}$ be defined by

$$F_{\mathbb{K}}(\{A_i\}_{i=1}^n) = \left\{ (I \otimes v^*) \left(\sum_{i=1}^n e_i \otimes A_i \right) (1 \otimes v) : \ v \in \mathbb{K}^N, \ ||v|| = 1 \right\}$$
$$= \left\{ \sum_{i=1}^n e_i v^* A_i v : \ v \in \mathbb{K}^N, \ ||v|| = 1 \right\} \subseteq \mathbb{C}^n.$$

Note that the quantity $F_{\mathbb{C}}(\{A\})$ coincides with the standard field of values of a matrix [7].

Also define the *conical* generalized field of values,

$$\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n) = \left\{ \sum_{i=1}^n e_i v^* A_i v : \ v \in \mathbb{K}^N \right\} \subseteq \mathbb{C}^n.$$

It is clear that this object is a cone; i.e., for real $\alpha > 0$, $f \in \check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n) \Rightarrow \alpha f \in \check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n)$. Note that $1 \oplus F_{\mathbb{K}}(\{A_i\}_{i=1}^n) = H_{\mathbb{K}}(e_1) \cap \check{F}_{\mathbb{K}}(\{I\} \cup \{A_i\}_{i=1}^n)$, where $H_{\mathbb{K}}(v)$ denotes the hyperplane $\{u \in \mathbb{K}^N | v^*u = 1\}$ for a vector $v \in \mathbb{K}^N$, and more generally $H_{\mathbb{K}}(v, r_0) = \{u \in \mathbb{K}^N : v^*u = r_0\}$. Note also that the conical field of values is preserved by simultaneous congruence transformation: for $P \in \mathbb{K}^{N \times N}$ nonsingular, $\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n) = \check{F}_{\mathbb{K}}(\{P^*A_iP\}_{i=1}^n)$.

We may now use the concept of generalized field of values to find characterizations of when $\psi_{n,\mathbb{K}}(A)$ or $\varphi_{n,\mathbb{K}}(A)$ is equal to 1, i.e., when the corresponding iterative methods can stagnate. This allows the performance of these methods to be studied in terms of the geometric properties of these objects, in particular, their convexity properties, as in the case of the standard field of values.

The following two theorems characterize stagnation of the methods in terms of properties of the generalized field of values.

THEOREM 2.7. *For nonsingular matrices $A \in \mathbb{K}^{N \times N}$, $\psi_{n,\mathbb{K}}(A) = 1$ if and only if $0 \in F_{\mathbb{K}}(\{A^i\}_{i=1}^n)$.*

Proof. Suppose that $\psi_{n,\mathbb{K}}(A) = 1$. By continuity of $f_n(\cdot, A)$ and by compactness, there exists a vector $v \in \mathbb{K}^N$, $||v|| = 1$, such that $f_n(v, A) = \inf_{q \in \mathbb{K}_{n-1}[z]} ||(I - Aq(A))v|| = 1$. Note that for such v, $d(v, A) > n$. In view of the definition of $\hat{f}_n$ (13), this can hold only if v is perpendicular to the space generated by $Av, A^2v, \ldots, A^nv$, which implies that $v^*A^iv = 0$ for $1 \le i \le n$. Thus $0 \in F_{\mathbb{K}}(\{A^i\}_{i=1}^n)$.

Conversely, if $0 \in F_{\mathbb{K}}(\{A^i\}_{i=1}^n)$, then for some nonzero $v \in \mathbb{K}^N$ with $||v|| = 1$, $v^*A^iv = 0$ for $1 \le i \le n$. This implies that v is perpendicular to each A^iv for $1 \le i \le n$ and for any q,

$$||(I - Aq(A))v||^2 = ||v - q_0Av - q_1A^2v - \cdots - q_{n-1}A^2v||^2 = ||v||^2 + ||Aq(A)v||^2 \ge 1.$$

The result follows. □

THEOREM 2.8. *For nonsingular matrices $A \in \mathbb{K}^{N\times N}$, $\varphi_{n,\mathbb{K}}(A) = 1$ if and only if $0 \in \mathrm{cvx}(F_{\mathbb{K}}(\{A^i\}_{i=1}^n))$, the convex hull of the generalized field of values.*

Proof. By the Hahn–Banach theorem, $0 \notin \mathrm{cvx}(F_{\mathbb{K}}(\{A^i\}_{i=1}^n))$ iff there exists a hyperplane separating 0 from $F_{\mathbb{K}}(\{A^i\}_{i=1}^n)$; i.e., for some $c \in \mathbb{K}^n$, $\mathrm{Re}\, c^*w > 0$ for all $w \in F_{\mathbb{K}}(\{A^i\}_{i=1}^n)$. (In the complex case, let $c = c_r + ic_i$ and $w = w_r + iw_i \in F_{\mathbb{K}}(\{A^i\}_{i=1}^n)$, so that $\mathrm{Re}\, c^*w = c_r^*w_r + c_i^*w_i$. Let $\hat{c} = c_r \oplus c_i$ and $\hat{w} = w_r \oplus w_i$. Then $\mathrm{Re}\, c^*w = \hat{c}^*\hat{w}$, and the result follows from the Hahn–Banach theorem in $\mathbb{R}^{2n}$.)

Let $C(c, z) = \sum_{i=1}^n (c^*e_i)z^{i-1}$. Then $\mathrm{Re}(c^*F_{\mathbb{K}}(\{A^i\}_{i=1}^n)) = \mathrm{Re}(F_{\mathbb{K}}(AC(c, A)))$.

Suppose $\varphi_{n,\mathbb{K}}(A) < 1$. Then there exist $P(z) = 1 - zq(z)$ and ρ such that $||P(A)v||^2 = 1 - 2\,\mathrm{Re}\, v^*Aq(A)v + ||Aq(A)v||^2 \le \rho < 1$ for every $v \in \mathbb{K}^N$ with $||v|| = 1$. Thus $1 - 2\,\mathrm{Re}\, v^*Aq(A)v \le \rho$ for such v, or $\mathrm{Re}\, v^*Aq(A)v \ge (1 - \rho)/2$. Defining c by $q(z) = C(c, z)$, we obtain $\mathrm{Re}\, c^*F_{\mathbb{K}}(\{A^i\}_{i=1}^n) \ge (1 - \rho)/2 > 0$, giving $0 \notin \mathrm{cvx}(F_{\mathbb{K}}(\{A^i\}_{i=1}^n))$.

Suppose there exists $C(c, z) = \sum (c^*e_i)z^{i-1}$ with $\mathrm{Re}(F_{\mathbb{K}}(AC(c, A))) > 0$. We have $||[I - \epsilon AC(c, A)]v||^2 = ||v||^2 - 2\epsilon\,\mathrm{Re}\, v^*AC(c, A)v + \epsilon^2||AC(c, A)v||^2$ for $\epsilon > 0$. By compactness, there exists $\delta > 0$ such that $\mathrm{Re}(F_{\mathbb{K}}(AC(c, A))) > \delta$. Hence, for a sufficiently small ϵ, $||[I - \epsilon AC(c, A)]v||^2 < 1 - \epsilon\delta$ for all v with $||v|| = 1$. Thus $\varphi_{n,\mathbb{K}}(A) < 1$. □

The principles behind these two theorems will be used heavily in §4 to construct the counterexample. In particular, note that if the generalized field of values associated with the powers of A is convex, then either both methods converge or both diverge. On the other hand, if the generalized field of values is nonconvex at the origin, in the sense of Theorem 2.8, then restarted GMRES will necessarily converge but the associated polynomial preconditioning may diverge.

2.4. Bounds based on the generalized field of values. Theorems 2.7 and 2.8 can be quantified in terms of the distances between 0 and either of the sets $F_{\mathbb{K}}(\{A^i\}_{i=1}^n)$ and $\mathrm{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^n)]$. In particular, the convergence rates of the two methods can be bounded in terms of these distances.

First consider the map $\Gamma_A : \mathbb{K}^n \to \mathbb{K}^{N\times N}$ defined by $\Gamma_A(c) = AQ(A)$, where $Q(z) = \sum_{i=1}^n (e_i^*c)z^{i-1}$. Then Γ_A is a bounded map under the induced norm

$$||\Gamma_A|| = \sup_{c\in\mathbb{K}^n, ||c||=1} ||AQ(A)|| = \sup_{c\in\mathbb{K}^n, ||c||=1} \sup_{v\in\mathbb{K}^N, ||v||=1} ||AQ(A)v||.$$

Observe that $||\Gamma_A|| \le \sum_{i=1}^n ||A^i||$.

The following two theorems give bounds on $\psi_{n,\mathbb{K}}(A)$ and $\varphi_{n,\mathbb{K}}(A)$. Here, let $A \in \mathbb{K}^{N\times N}$, and let $S_{n,\mathbb{K}}$ denote the sphere in $\mathbb{K}^n$.

THEOREM 2.9. *Let $\eta = \sup\{\rho \ge 0 : (\rho \cdot S_{n,\mathbb{K}}) \cap F_{\mathbb{K}}(\{A^i\}_{i=1}^n) = \{\}\}$, the distance from $F_{\mathbb{K}}(\{A^i\}_{i=1}^n)$ to the origin. Then $\psi_{n,\mathbb{K}}(A) \le (1 - (\eta/||\Gamma_A||)^2)^{1/2}$.*

Proof. Given $v \in \mathbb{K}^N$, $||v|| = 1$, let $w = v^* A\underline{K}_n(v, A)$ and let c satisfy $\operatorname{Re} c^* w \geq \eta$, $||c|| = 1$, for example, by $c = w/||w||$. Let $Q(z) = \sum_{i=1}^{n} (e_i^* c) z^{i-1}$. Then let $\hat{Q} = \epsilon Q$ for $\epsilon > 0$.

$$||(I - A\hat{Q}(A))v||^2 = v^*v - 2\epsilon \operatorname{Re} v^* AQ(A)v + \epsilon^2 ||AQ(A)v||^2 \leq v^*v - 2\epsilon\eta + \epsilon^2 ||\Gamma_A||^2.$$

Setting $\epsilon = \eta/||\Gamma_A||^2$ gives the result. □

THEOREM 2.10. *Let* $\eta = \sup\{\rho \geq 0 : \rho \cdot S_{n,\mathbb{K}} \cap \operatorname{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^n)] = \{\}\}$, *the distance from* $\operatorname{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^n)]$ *to the origin. Then* $\varphi_{n,\mathbb{K}}(A) \leq (1 - (\eta/||\Gamma_A||)^2)^{1/2}$.

Proof. The result is trivial for $\eta = 0$. Otherwise, note that

$$\eta \cdot S_{n,\mathbb{K}} \cap \operatorname{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^n)]$$

must contain exactly one point. It has at least one point since $f(v) = ||v^* A\underline{K}_n(v, A)||$ must attain its infimum on $\{v : ||v|| = 1\}$. On the other hand, if the intersection contains two distinct points γ_1 and γ_2, then $(\gamma_1 + \gamma_2)/2$ is in the interior of $\eta \cdot S_{n,\mathbb{K}}$ and also in $\operatorname{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^n)]$ due to convexity, which is a contradiction due to closedness. Now, let γ be the point in this intersection.

We claim that for $c = \gamma/||\gamma||$, $c^* w \geq \eta$ for every $v \in \mathbb{K}^N$, $||v|| = 1$, with $w = v^* A\underline{K}_n(v, A)$. This follows if we can show the result for all $w \in \operatorname{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^n)]$. Otherwise there exists $w \in \operatorname{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^n)]$ with $\operatorname{Re} \gamma^* w < ||\gamma||^2$, where $||\gamma|| = \eta$. Let $w_\epsilon = \epsilon w + (1 - \epsilon)\gamma$, $0 \leq \epsilon \leq 1$. Note that $w_\epsilon \in \operatorname{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^n)]$ for any such ϵ. Furthermore,

$$||w_\epsilon||^2 = ||\gamma||^2 - 2\epsilon||\gamma||^2 + 2\epsilon \operatorname{Re} \gamma^* w + \mathcal{O}(\epsilon^2) < ||\gamma||^2$$

for sufficiently small ϵ. For this ϵ, w_ϵ is in the interior of $\eta \cdot S_{n,\mathbb{K}}$ and also in $\operatorname{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^n)]$, which is a contradiction.

Defining Q and $\hat{Q}$ as in the proof of the previous theorem we see that

$$||(I - A\hat{Q}(A))v||^2 = v^*v - 2\epsilon \operatorname{Re} v^* AQ(A)v + \epsilon^2 ||AQ(A)v||^2 \leq v^*v - 2\epsilon\eta + \epsilon^2 ||\Gamma_A||^2$$

for every $v \in \mathbb{K}^N$. Setting $\epsilon = \eta/||\Gamma_A||^2$ gives the result. □

These bounds will become useful in the numerical example given later.

2.5. Deriving results for other matrices. This subsection gives a collection of results that add further insight into the behavior of the bound functions and will be useful later for extending results to wider classes of matrices.

The first result shows that the generalized field of values is convex and the bound functions are equal for normal matrices.

THEOREM 2.11. *For* $A \in \mathbb{K}^{N \times N}$ *nonsingular and normal (e.g., Hermitian, real symmetric, skew Hermitian, real skew symmetric, unitary, or circulant),* $\check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^n)$ *is convex, and, in fact,* $\psi_{n,\mathbb{K}}(A) = \varphi_{n,\mathbb{K}}(A)$.

Proof. See [10], [3], [8]. □

The following result shows that for a block diagonal matrix, the convergence rate of either of the methods is no better than the convergence rate of that method for any of the diagonal submatrices of the block diagonal matrix.

THEOREM 2.12. *For* $A_i \in \mathbb{K}^{N_i \times N_i}$, $i = 1, 2$, $\psi_{n,\mathbb{K}}(A_1) \leq \psi_{n,\mathbb{K}}(\operatorname{diag}[A_1, A_2])$ *and* $\varphi_{n,\mathbb{K}}(A_1) \leq \varphi_{n,\mathbb{K}}(\operatorname{diag}[A_1, A_2])$. *Furthermore,* $\check{F}_{\mathbb{K}}(\{A_1^i\}_{i=0}^n) \subseteq \check{F}_{\mathbb{K}}(\{\operatorname{diag}[A_1, A_2]^i\}_{i=0}^n)$.

Proof. The first result follows easily from

$$\psi_{n,\mathbb{K}}(A_1) = \sup_{v\in\mathbb{K}^{N_1}:||v||=1} \inf_{P\in\mathbb{C}_n[x]:P(0)=1} ||P(A)v||$$

$$= \sup_{v\in\mathbb{K}^{N_1}:||v||=1} \inf_{P\in\mathbb{C}_n[x]:P(0)=1} \left\| P\left(\begin{bmatrix} A_1 & \\ & A_2 \end{bmatrix}\right)\begin{bmatrix} v \\ 0 \end{bmatrix}\right\|$$

$$\leq \sup_{v\in\mathbb{K}^{N_1+N_2}:||v||=1} \inf_{P\in\mathbb{C}_n[x]:P(0)=1} ||P(\mathrm{diag}[A_1,A_2])v|| = \psi_{n,\mathbb{K}}(\mathrm{diag}[A_1,A_2]),$$

$$\varphi_{n,\mathbb{K}}(A_1) = \inf_{P\in\mathbb{C}_n[x]:P(0)=1} \sup_{v\in\mathbb{K}^{N_1}:||v||=1} ||P(A)v||$$

$$= \inf_{P\in\mathbb{C}_n[x]:P(0)=1} \sup_{v\in\mathbb{K}^{N_1}:||v||=1} \left\| P\left(\begin{bmatrix} A_1 & \\ & A_2 \end{bmatrix}\right)\begin{bmatrix} v \\ 0 \end{bmatrix}\right\|$$

$$\leq \inf_{P\in\mathbb{C}_n[x]:P(0)=1} \sup_{v\in\mathbb{K}^{N_1+N_2}:||v||=1} ||P(\mathrm{diag}[A_1,A_2])v|| = \varphi_{n,\mathbb{K}}(\mathrm{diag}[A_1,A_2]).$$

The subset inclusion result follows from a similar line of argument. □

The next result shows that for the special case of a submatrix replicated down the main diagonal of a block diagonal matrix, the convergence rate for polynomial preconditioning is unchanged by the replication.

THEOREM 2.13. *For $A \in \mathbb{K}^{N\times N}$ and I an identity matrix of any size, $\varphi_{n,\mathbb{K}}(A\otimes I) = \varphi_{n,\mathbb{K}}(A)$. Furthermore, for $\{A_i\}_{i=1}^n \subseteq \mathbb{K}^{N\times N}$, $\mathrm{cvx}[\check{F}_{\mathbb{K}}(\{A_i\otimes I\}_{i=1}^n)] = \mathrm{cvx}[\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n)]$.*

Proof. Using the results on tensor products from [13],

$$\varphi_{n,\mathbb{K}}(A\otimes I) = \inf_P ||P(A\otimes I)|| = \inf_P \lambda_{\max}[[P(A)^*P(A)]\otimes I]^{1/2}$$

$$= \inf_P \lambda_{\max}[P(A)^*P(A)]^{1/2} = \varphi_{n,\mathbb{K}}(A).$$

To show the set equality, it is sufficient to note that

$$\sum_i e_i \left[\sum_j v_j\otimes e_j\right]^* [A_i\otimes I]\left[\sum_j v_j\otimes e_j\right] = \sum_i e_i \sum_j [v_j\otimes e_j]^*[A_i\otimes I][v_j\otimes e_j]$$

$$= \sum_j \sum_i e_i v_j^* A_i v_j = \sum_j (1/n)\sum_i e_i(\sqrt{n}v_j)^* A_i(\sqrt{n}v_j) \in \mathrm{cvx}[\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n)]. \quad \square$$

The following weaker result holds for the minimal residual method.

THEOREM 2.14. *For $A_i \in \mathbb{K}^{N\times N}$ and I_N an identity matrix of size N, the set $\check{F}_{\mathbb{K}}(\{A_i\otimes I_N\}_{i=1}^n)$ is convex.*

Proof. We view $\check{F}_{\mathbb{K}}(\{A_i\otimes I_N\}_{i=1}^n)$ as the image of the composition of two maps, where the first one has a convex range and the second one is linear.

First let $v = \sum_i v_i\otimes e_i$ for vectors $\{v_i\}_{i=1}^N \subseteq \mathbb{K}^N$. The first map ρ takes v to $\rho(v) = \sum_{i=1}^N v_i v_i^* \in \mathbb{K}^{N\times N}$. Note that the range of ρ is exactly the convex cone of Hermitian nonnegative definite matrices in $\mathbb{K}^{N\times N}$.

The second map σ takes a matrix $P \in \mathbb{K}^{N\times N}$ to $\sigma(P) = \sum_{i=1}^n e_i \mathrm{trace}(PA_i) \in \mathbb{K}^n$. This map is linear.

Let R be the map associated with $\check{F}_{\mathbb{K}}(\{A_i\otimes I_N\}_{i=1}^n)$. To complete the proof, we show that $\sigma(\rho(v)) = R(v)$:

$$R(v) = \sum_{i=1}^n e_i v^*(A_i\otimes I_N)v = \sum_{i=1}^n e_i \sum_{j=1}^N v_j^* A_i v_j$$

$$= \sum_{i=1}^{n} e_i \operatorname{trace}\left(\left[\sum_{j=1}^{N} v_j v_j^*\right] A_i\right) = \sigma(\rho(v)). \qquad \square$$

We may also characterize situations in which $\psi_{n,\mathbb{K}}(A \otimes I_k) = \varphi_{n,\mathbb{K}}(A \otimes I_k)$. For this result, let us state the following definitions. Consider $\mathbf{A} = \sum_{i=1}^{n} e_i \otimes A_i$, $A_i \in \mathbb{K}^{N \times N}$. For our purposes, $A_i = A^i$. For $v \in \mathbb{K}^n$, define $v \cdot \mathbf{A} = [v^T \otimes I]\mathbf{A} = \sum_i [e_i^* v] A_i$. Define $\varphi_{\mathbb{K}}(\mathbf{A}) = \inf_v \sup_x ||(I + v \cdot \mathbf{A})x||$, $\psi_{\mathbb{K}}(\mathbf{A}) = \sup_x \inf_v ||(I + v \cdot \mathbf{A})x||$, where x and v are vectors over $\mathbb{K}$ and the suprema are over $||x|| = 1$.

Let us define the *restricted* generalized field of values: for such $\mathbf{A}$ and for M a subspace of $\mathbb{C}^N$, $F_M(\mathbf{A}) = \{\sum_i e_i [x^* A_i x] : ||x|| = 1, x \in M\}$. Also, let $\Sigma_{\mathbb{K}}(B) \subseteq \mathbb{K}^N$ be the subspace over $\mathbb{K}$ spanned by the (right) singular vectors associated with the maximal singular value of $B \in \mathbb{K}^{N \times N}$.

The following result generalizes Theorem 2.8.

LEMMA 2.15. *Given* $\mathbf{A} = \sum_i e_i \otimes A_i$, $A_i \in \mathbb{K}^{N \times N}$, v *is a minimizer of* $\varphi_{\mathbb{K}}(\mathbf{A})$ *if and only if for* $B = v \cdot \mathbf{A}$, $0 \in \operatorname{cvx} F_{\Sigma_{\mathbb{K}}(I+B)}([I \otimes [I + B]]^* \mathbf{A})$.

Proof. Following the proof of Theorem 2.8, let us suppose first that $0 \in \operatorname{cvx} F_{\Sigma_{\mathbb{K}}(I+B)}([I \otimes [I+B]]^* \mathbf{A})$. Then for any w there exists $x_w \in \Sigma_{\mathbb{K}}(I+B)$, $||x_w|| = 1$, such that $\operatorname{Re} x_w^* (I + B)^* [w \cdot \mathbf{A}] x_w = 0$. Thus for any w and for $\epsilon > 0$,

$$||(I + (v + \epsilon w) \cdot \mathbf{A})||^2 \geq ||(I + (v + \epsilon w) \cdot \mathbf{A}) x_w||^2$$
$$= ||(I + v \cdot \mathbf{A}) x_w||^2 + \epsilon^2 ||(w \cdot \mathbf{A}) x_w||^2 \geq ||(I + v \cdot \mathbf{A}) x_w||^2 = ||I + v \cdot \mathbf{A}||^2,$$

since the cross term is zero. Thus v is a local minimizer. By a convexity argument it can be shown that v is thus a global minimizer.

If $0 \notin \operatorname{cvx} F_{\Sigma_{\mathbb{K}}(I+B)}([I \otimes [I + B]]^* \mathbf{A})$, then as before there is w such that for all $x \in \Sigma_{\mathbb{K}}(I + B)$ such that $||x|| = 1$, $\operatorname{Re} x^* (I + B)^* [w \cdot \mathbf{A}] x < 0$. Then

$$||(I + (v + \epsilon w) \cdot \mathbf{A}) x||^2 = ||(I + B) x||^2$$
$$+ 2\epsilon \operatorname{Re} x^* (I + B)^* [w \cdot \mathbf{A}] x + \epsilon^2 ||(w \cdot \mathbf{A}) x||^2 < ||(I + B) x||^2$$

uniformly for ϵ sufficiently small. For more general x such that $||x|| = 1$, since x has components of right singular vectors associated with smaller singular values of $I + B$, it can also be shown that $||(I + (v + \epsilon w) \cdot \mathbf{A}) x||^2 < ||I + B||^2$ uniformly for sufficiently small ϵ. By this line of argument we arrive at a contradiction to v being a minimizer. $\square$

LEMMA 2.16. *Given* $\mathbf{A} = \sum_i e_i \otimes A_i$, $A_i \in \mathbb{K}^{N \times N}$, $\psi_{\mathbb{K}}(\mathbf{A}) = \varphi_{\mathbb{K}}(\mathbf{A})$ *if and only if for some (equivalently, every) minimizer* v *of* $\varphi_{\mathbb{K}}(\mathbf{A})$ *and for* $B = v \cdot \mathbf{A}$, $0 \in F_{\Sigma_{\mathbb{K}}(I+B)}([I \otimes [I + B]]^* \mathbf{A})$.

Proof. If $0 \in F_{\Sigma_{\mathbb{K}}(I+B)}([I \otimes [I+B]]^* \mathbf{A})$, then there exists $x \in \Sigma_{\mathbb{K}}(I+B)$, $||x|| = 1$, such that $(I + B)x \perp A_i x$ for all i. Then such B solves the least squares problem for x; i.e., $||(I+B)x|| = \inf_w ||(I+w \cdot \mathbf{A})x||$, so in fact $\varphi_{\mathbb{K}}(\mathbf{A}) \leq ||I+B|| = ||(I+B)x|| \leq \psi_{\mathbb{K}}(\mathbf{A}) \leq \varphi_{\mathbb{K}}(\mathbf{A})$.

Now suppose that $\psi_{\mathbb{K}}(\mathbf{A}) = \varphi_{\mathbb{K}}(\mathbf{A})$. Let v be a minimizer for φ, and let $B = v \cdot \mathbf{A}$. Let x be a maximizer for ψ, $||x|| = 1$. Then by the definition of ψ, letting B' solve the least squares problem for x, $\varphi_{\mathbb{K}}(\mathbf{A}) = \psi_{\mathbb{K}}(\mathbf{A}) = ||(I + B')x|| \leq ||(I + v \cdot \mathbf{A})x|| = ||(I + B)x|| \leq \varphi_{\mathbb{K}}(\mathbf{A})$. But then x must be a maximal right singular vector of $I + B$, and in fact this inequality is an equality. Furthermore, v solves the minimization problem $\inf_w ||(I + w \cdot \mathbf{A})x||$, so $(I + B)x \perp A_i x$, giving the result. $\square$

COROLLARY 2.17. *Given* $\mathbf{A} = \sum_i e_i \otimes A_i$, $A_i \in \mathbb{K}^{N \times N}$, *if for some minimizer* v *of* $\varphi_{\mathbb{K}}(\mathbf{A})$ *and for* $B = v \cdot \mathbf{A}$, $F_{\Sigma_{\mathbb{K}}(I+B)}([I \otimes [I + B]]^* \mathbf{A})$ *is convex, then* $\psi_{\mathbb{K}}(\mathbf{A}) = \varphi_{\mathbb{K}}(\mathbf{A})$.

These results lead to the following theorem, which shows that for k sufficiently large, ψ and φ are equal for $\mathbf{A}\otimes I_k$.

THEOREM 2.18. *Given* $\mathbf{A} = \sum_{i=1}^{n} e_i \otimes A_i$, $A_i \in \mathbb{K}^{N\times N}$, *if* $k \le N$ *is the smallest dimension for the maximal singular vector space of* $I + B = I + v \cdot \mathbf{A}$ *for* v *any minimizer for* $\varphi_{\mathbb{K}}(\mathbf{A})$, *then for* I_k *the identity matrix of dimension* $k\times k$, $\psi_{\mathbb{K}}(\mathbf{A}\otimes I_k) = \varphi_{\mathbb{K}}(\mathbf{A}\otimes I_k)$.

Proof. Let B' be a minimizer for $\varphi_{\mathbb{K}}(\mathbf{A}\otimes I_k)$. Then $B' = v\cdot\mathbf{A}\otimes I_k = (v\cdot\mathbf{A})\otimes I_k = B\otimes I_k$, where $B = v\cdot\mathbf{A}$. It is clear from Theorem 2.13 that B' is a minimizer for $\varphi(\mathbf{A}\otimes I_k)$ if and only if B is a minimizer for $\varphi(\mathbf{A})$. Note also that for $I + B = U\Sigma V^*$ a singular value decomposition, $[I + B]\otimes I_k = (U\otimes I_k)(\Sigma\otimes I_k)(V\otimes I_k)^*$ is also a singular value decomposition. Thus $\Sigma_{\mathbb{K}}(I + B') = \bigoplus_{i=1}^{k}\Sigma_{\mathbb{K}}(I + B)$. Also, $[I\otimes[I + B']]^*[\mathbf{A}\otimes I_k] = ([I\otimes[I + B]]^*\mathbf{A})\otimes I_k = \mathbf{C}\otimes I_k$ for $\mathbf{C} = \sum_i e_i\otimes C_i$, $C_i = [I + B]^*A_i$. We conclude that $F_{\Sigma_{\mathbb{K}}(I+B')}([I\otimes[I + B']]^*\mathbf{A}') = \{\sum_i\sum_j e_i v_j^* C_i v_j : \sum_j ||v_j||^2 = 1,\ v_j \in \Sigma_{\mathbb{K}}(I + B)\}$. It is enough to show this is convex.

For $v_i \in \Sigma_{\mathbb{K}}(I + B)$ let $v = \sum v_i\otimes e_i$ and $\rho(v) = \sum_i v_i v_i^*$. Furthermore, let $\sigma(P) = \sum_{i=0}^{n} e_i \mathrm{trace}(PC_i)$, letting $C_0 = I$. Note that this sum begins at $i = 0$. Since, without loss of generality, $k \ge \dim\Sigma_{\mathbb{K}}(I + B)$, it is easily seen that the range of ρ is convex. Since σ is linear, the range $\check{F}$ of $\sigma\circ\rho$ is convex. But then $H_{\mathbb{K}}(e_1)\cap\check{F} = F_{\Sigma_{\mathbb{K}}(I+B')}([I\otimes[I + B']]^*\mathbf{A}')$ is convex. ∎

It should be added that when $A_i = A^i$, under appropriate conditions the minimizer for φ is unique (see [5]). We finally conclude that, given A and n, for some k no greater than N, $\psi_{n,\mathbb{K}}(A\otimes I_k) = \varphi_{n,\mathbb{K}}(A\otimes I_k)$. Note also that $\psi_{n,\mathbb{K}}(A\otimes I_k)$ is nondecreasing in k (Theorem 2.12), whereas $\varphi_{n,\mathbb{K}}(A\otimes I_k)$ is constant as a function of k (Theorem 2.13).

The result that follows allows us to characterize convergence of the two methods solely in terms of the *conical* generalized field of values rather than the standard generalized field of values. These results will be useful later in the paper.

PROPOSITION 2.19. *For* $A \in \mathbb{K}^{N\times N}$, $0 \in F_{\mathbb{K}}(\{A^i\}_{i=1}^{n})$ *holds if and only if* $e_1 \in \check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^{n})$. *Also,* $0 \in \mathrm{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^{n})]$ *if and only if* $e_1 \in \mathrm{cvx}[\check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^{n})]$.

Proof. Recalling that $H_{\mathbb{K}}(e_1) = \{u \in \mathbb{K}^N : e_1^* u = 1\}$, note that

$$0 \in F_{\mathbb{K}}(\{A^i\}_{i=1}^{n}) \iff e_1 \in 1\oplus F_{\mathbb{K}}(\{A^i\}_{i=1}^{n}) = H_{\mathbb{K}}(e_1)\cap\check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^{n}) \subseteq \check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^{n}),$$
$$0 \in \mathrm{cvx}\, F_{\mathbb{K}}(\{A^i\}_{i=1}^{n}) \iff e_1 \in 1\oplus \mathrm{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^{n})] = \mathrm{cvx}[1\oplus F_{\mathbb{K}}(\{A^i\}_{i=1}^{n})]$$
$$= \mathrm{cvx}[H_{\mathbb{K}}(e_1)\cap\check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^{n})] \subseteq \mathrm{cvx}[\check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^{n})].$$

Note also that if $e_1 \in \check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^{n})$ then since $e_1 \in H_{\mathbb{K}}(e_1)$, $e_1 \in H_{\mathbb{K}}(e_1)\cap\check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^{n})$. Now suppose $e_1 \in \mathrm{cvx}(\check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^{n}))$; i.e., there exist $t_i \ge 0$, $\sum_i t_i = 1$, $e_1 = \sum_i t_i[\sum_{j=0}^{n} e_j v_i^* A^j v_i]$ for some $v_i \in \mathbb{K}^N$. That is, $1 = \sum_i t_i v_i^* v_i$, and for $j \ge 1$, $0 = \sum_i t_i v_i^* A^j v_i$. Now, let $t_i' = t_i v_i^* v_i$, and let $v_i' = v_i/||v_i||$ when $v_i \ne 0$ and otherwise let v_i' be any vector of norm 1. Then, for $S = \{i : v_i \ne 0\}$, $e_1 = \sum_{i\in S} t_i'[\sum_j e_j v_i'^* A^j v_i'] = \sum_i t_i'[\sum_j e_j v_i'^* A^j v_i']$ since $t_i' = t_i v_i^* v_i = 0$ for $i \notin S$, $\sum_i t_i' = \sum_i t_i v_i^* v_i = 1$, and $||v_i'|| = 1$ for all i, so $0 \in \mathrm{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^{n})]$. ∎

The simple result below states that if the conical field of values of a set of matrices is convex, then it is also convex for a subset of the matrices. This can be useful for transferring a result on equivalence of convergence of iterative methods to a lower iteration number.

PROPOSITION 2.20. *For* $A_i \in \mathbb{C}^{N\times N}$ *if* $\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^{n})$ *is convex, then for* $m \le n$, $\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^{m})$ *is convex. More generally, if for* $A_i \in \mathbb{C}^{N\times N}$, $\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^{n})$ *is convex, and* $V \in \mathbb{C}^{m\times n}$, *then* $V\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^{n}) = \check{F}_{\mathbb{K}}(\{\sum v_{i,j}A_j\}_{i=1}^{n})$ *is convex.*

Proof. A linear operator applied to a convex set preserves convexity. ∎

The following are two further results on combining multiple matrices into a block diagonal matrix.

PROPOSITION 2.21. *Let* $A_i \in \mathbb{C}^{N_1 \times N_1}$ *and* $B_i \in \mathbb{C}^{N_2 \times N_2}$, *and also suppose that* $\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n)$ *and* $\check{F}_{\mathbb{K}}(\{B_i\}_{i=1}^n)$ *are convex. Then* $\check{F}_{\mathbb{K}}(\{\mathrm{diag}[A_i, B_i]\}_{i=1}^n)$ *is convex.*

Proof. It is enough to show that for $t_i \geq 0$, $\sum_i t_i = 1$, $v_{i,1} \in \mathbb{K}^{N_1}$, $v_{i,2} \in \mathbb{K}^{N_2}$, there exist $v_1 \in \mathbb{K}^{N_1}$, $v_2 \in \mathbb{K}^{N_2}$ such that for $1 \leq k \leq n$, $\sum_i t_i[\sum_k e_k[v_{i,1}^* A_k v_{i,1} + v_{i,2}^* B_k v_{i,2}]] = \sum_k [e_k v_1^* A_k v_1 + v_2^* B_k v_2]$. This may be done simply by letting

$$\sum_k e_k[v_1^* A_k v_1] = \sum_i t_i \left[\sum_k e_k[v_{i,1}^* A_k v_{i,1}] \right],$$
$$\sum_k e_k[v_2^* B_k v_2] = \sum_i t_i \left[\sum_k e_k[v_{i,2}^* B_k v_{i,2}] \right]. \qquad \blacksquare$$

COROLLARY 2.22. *Let* $A_i \in \mathbb{C}^{N \times N}$, *and suppose that* $\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n)$ *is convex. Then* $\check{F}_{\mathbb{K}}(\{A_i \otimes D_m\}_{i=1}^n)$ *is convex for* $D_m \in \mathbb{C}^{m \times m}$ *any diagonal matrix.*

The next two results further extend the above convexity results to the tensor product of a matrix with an appropriate normal matrix.

PROPOSITION 2.23. *Let* $A_i \in \mathbb{C}^{N \times N}$, *and suppose that* $\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n)$ *is convex. Also let* $B \in \mathbb{K}^{M \times M}$ *be any normal matrix with eigenvalues in* $\mathbb{K}$, *and let* $p_i \in \mathbb{C}[x]$. *Then* $\check{F}_{\mathbb{K}}(\{A_i \otimes p_i(B)\}_{i=1}^n)$ *is convex.*

Proof. Let $B = U \Lambda U^*$, $U \in \mathbb{K}^{N \times N}$ unitary, $\Lambda = \{\lambda_i\} \in \mathbb{K}^{N \times N}$ diagonal. Note that

$$\check{F}_{\mathbb{K}}(\{A_i \otimes p_i(B)\}_{i=1}^n) = \check{F}_{\mathbb{K}}(\{A_i \otimes p_i(\Lambda)\}_{i=1}^n) = \left\{ \sum_i e_i \sum_j (v_j \otimes e_j)^* [A_i \otimes p_i(\Lambda)](v_j \otimes e_j) \right\}$$
$$= \left\{ \sum_i e_i \sum_j p_i(\lambda_j)(v_j \otimes e_j)^* [A_i \otimes e_j e_j^*](v_j \otimes e_j) \right\}$$
$$= \left[\sum_{i,j} p_i(\lambda_j) e_i [e_i \otimes e_j]^* \right] \left\{ \sum_{i,j} e_i \otimes e_j [(v_j \otimes e_j)^* [A_i \otimes e_j e_j^*](v_j \otimes e_j)] \right\}$$
$$= \left[\sum_{i,j} p_i(\lambda_j) e_i [e_i \otimes e_j]^* \right] \check{F}_{\mathbb{K}}(\{A_i \otimes e_j e_j^*\}_{i=1\, j=1}^{n\quad M})$$
$$= \left[\sum_{i,j} p_i(\lambda_j) e_i [e_i \otimes e_j]^* \right] \oplus_{j=1}^M \check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n).$$

The result follows from noting that a direct sum of convex sets is convex and the linear transformation of a convex set is convex. ∎

COROLLARY 2.24. *For* A, $B \in \mathbb{K}^{N \times N}$, *for* B *normal with eigenvalues in* $\mathbb{K}$, *if* $\check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^n)$ *is convex, then so is* $\check{F}_{\mathbb{K}}(\{(A \otimes B)^i\}_{i=0}^n)$.

The next proposition shows that it is possible to replace matrices with their transpose or conjugate transpose.

PROPOSITION 2.25. *For* $A_i \in \mathbb{C}^{N \times N}$, *suppose that* $\check{F}_{\mathbb{K}}(\{A_i\})$ *is convex, and for each* i *let* $\tilde{A}_i$ *be either* A_i, A_i^*, *or* A_i^T. *Then* $\check{F}_{\mathbb{K}}(\{\tilde{A}_i\})$ *is convex.*

Proof. The result follows from $x^* A_i^* x = \overline{x^* A_i x}$ and $x^* A_i^T x = \overline{x^* \overline{A_i} x} = \overline{x}^* A_i \overline{x}$. ∎

Similarly, each matrix can be replaced with its Hermitian and skew-Hermitian parts. Here, $(M)_H$ denotes the Hermitian part of a matrix, $(M)_H = (M + M^*)/2$.

PROPOSITION 2.26. *For* $A_j \in \mathbb{R}^{N\times N}$, $\check{F}_{\mathbb{R}}(\{A_j\}_{j=1}^n) = \check{F}_{\mathbb{R}}(\{(A_j)_H\}_{j=1}^n)$, *and for* $A_j \in \mathbb{C}^{N\times N}$, $\check{F}_{\mathbb{C}}(\{A_j\}_{j=1}^n)$ *is isomorphic to* $\check{F}_{\mathbb{C}}(\{(A_j)_H, (iA_j)_H/i\}_{j=1}^n)$ *under the identification of* $\mathbb{C}^n$ *with* $\mathbb{R}^{2n}$.

Proof. For the real case, note that $v^*Av = v^*(A)_H v$; for the complex case, the real and imaginary parts of v^*Av are given by $v^*(A)_H v$ and $-v^*(iA)_H v$, respectively. □

The following result indicates that shifting a matrix by a constant multiple of the identity does not change convexity of the conical field of values.

PROPOSITION 2.27. *For* $A \in \mathbb{C}^{N\times N}$ *and* $c \in \mathbb{C}$, *if* $\check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^n)$ *is convex then* $\check{F}_{\mathbb{K}}(\{(A+cI)^i\}_{i=0}^n)$ *is convex.*

Proof. The result follows from the fact that a linear transformation T over $\mathbb{C}$ exists such that $\check{F}_{\mathbb{K}}(\{(A+cI)^i\}_{i=0}^n) = T\check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^n)$. □

3. Results for Toeplitz matrices. In this section we demonstrate that for the special class of upper triangular Toeplitz matrices, the generalized field of values of the powers of any such matrix is convex. The implication of this result, based on the results of the previous section, is that for such matrices, if GMRES(s) converges for given s, then the optimal polynomial preconditioning of corresponding degree must also converge.

Let us begin with notation. Let $D_k \in \mathbb{R}^{N\times N}$ be defined by $e_i^* D_k e_j = \delta_{i,j-k}$. These matrices form a basis for the Toeplitz matrices. Note that $D_0 = I$, $D_{-k} = D_k^*$, and for $kl \geq 0$, $D_k D_l = D_{kl}$. A Toeplitz matrix over $\mathbb{K}$ is defined to be $\sum_{i=1-N}^{N-1} t_i D_i$ for $t_i \in \mathbb{K}$.

We begin by proving the convexity of $\check{F}_{\mathbb{K}}(\{D_i\}_{i=1-N}^{N-1})$. Note that each $r = [r_{1-N}, \ldots, r_{N-1}] \in \check{F}_{\mathbb{K}}(\{D_i\})$ can be mapped by a linear injective map to the space of rational functions of the form $p(z) = \sum_{i=1-N}^{N-1} r_i z^i$, the z-transform. Letting $r_i = x^* D_i x$, with $x \in \mathbb{K}^N$, $r_i = 0$ for $i \notin [1-N, N-1]$, and $x_i = 0$ for $i \notin [1, N]$, we can write

$$p(z) = \sum_{i=1-N}^{N-1} \left(\sum_{j=1}^{N} \bar{x}_j x_{j+i} \right) z^i = \sum_{i=0}^{N-1} x_{i+1} z^i \sum_{i=0}^{N-1} \bar{x}_{i+1} z^{-i}.$$

Thus $p(z) = q(z)\bar{q}(1/z)$, where $q(z) = \sum_{i=0}^{N-1} x_{i+1} z^i$. Here the convention $\bar{q}(z) = \sum_{i=0}^{N-1} \bar{q}_i z^i$ is assumed, and similarly for rational functions. Note then that $r \in \check{F}_{\mathbb{K}}(\{D_i\})$ if and only if the corresponding p may be factored as $q(z)\bar{q}(1/z)$ for $q \in \mathbb{K}_{n-1}[x]$.

Let $\mathcal{P}'_{\mathbb{K}}$ be the set of symmetric rational functions over $\mathbb{K}$; i.e., $p(z) = \bar{p}(1/z)$, such that $z^{N-1}p(z) = \tilde{p}(z) \in \mathbb{K}[z]$. Note that $\mathcal{P}'_{\mathbb{K}}$ is a linear space over the reals, in the sense that $p_i \in \mathcal{P}'_{\mathbb{K}}$, $a_i \in \mathbb{R}$ imply that $\sum_i a_i p_i \in \mathcal{P}'_{\mathbb{K}}$. Furthermore, let $\mathcal{P}_{\mathbb{K}}$ denote the set of rational functions over $\mathbb{K}$ factorable as $q(z)\bar{q}(1/z)$, $q \in \mathbb{K}_{N-1}[x]$. Clearly $\mathcal{P}_{\mathbb{K}} \subseteq \mathcal{P}'_{\mathbb{K}}$. The convexity of $\check{F}_{\mathbb{K}}(\{D_i\})$ will follow if $\mathcal{P}_{\mathbb{K}}$ can be shown to be a convex subset of the linear space $\mathcal{P}'_{\mathbb{K}}$.

Let us consider the properties of polynomials in $z^{N-1}\mathcal{P}'_{\mathbb{K}}$. Note that if α is a root of $s(z) \in \mathcal{P}'_{\mathbb{K}}$, then so is $1/\bar{\alpha}$, and furthermore both of these are roots of the polynomial $z^{N-1}s(z) \in z^{N-1}\mathcal{P}'_{\mathbb{K}}$. Also, for nonzero α, $|\alpha| \neq 1$ if and only if α and $1/\bar{\alpha}$ are distinct, so for $\alpha \in \mathbb{K}$, $|\alpha| \neq 1$ implies that $s(z)/[(z-\alpha)(1/z-\bar{\alpha})] \in \mathcal{P}'_{\mathbb{K}}$. Similarly for $\mathbb{K} = \mathbb{R}$, $\alpha \in \mathbb{C} \setminus \mathbb{R}$, and $|\alpha| \neq 1$, the roots α, $\bar{\alpha}$, $1/\bar{\alpha}$, and $1/\alpha$ are all distinct, so $s(z)/[(z-\alpha)(1/z-\bar{\alpha})(z-\bar{\alpha})(1/z-\alpha)] \in \mathcal{P}'_{\mathbb{K}}$. This demonstrates that for $|\alpha| \neq 1$, the roots α and $1/\bar{\alpha}$ must have the same multiplicity in $z^{N-1}s(z) \in z^{N-1}\mathcal{P}'_{\mathbb{K}}$.

Note also the following lemma.

LEMMA 3.1. *For z on the unit circle, $s(z) \in \mathcal{P}'_{\mathbb{K}}$ takes on real values.*

Proof.

$$s(z) = \bar{s}(1/z) = \bar{s}(\bar{z}) = \overline{s(z)}. \qquad \square$$

The following result characterizes polynomials in $z^{N-1}\mathcal{P}_{\mathbb{K}}$.

LEMMA 3.2. *For nonzero rational functions $s(z)$, the polynomial $z^{N-1}s(z) \in z^{N-1}\mathcal{P}_{\mathbb{K}}$ if and only if the roots of $z^{N-1}s(z) \in \mathbb{K}[z]$, counted with multiplicity, consist of k zero roots and $(N-k-1)$ pairs of roots of the form $\{\alpha_i, 1/\bar{\alpha}_i\}$ and furthermore $s(z) \geq 0$ for all z on the unit circle.*

Proof. For $z^{N-1}s(z) \in z^{N-1}\mathcal{P}_{\mathbb{K}}$ we have

$$\begin{aligned} z^{N-1}s(z) = z^{N-1}q(z)\bar{q}(1/z) &= (\alpha\bar{\alpha})z^{N-1}\prod_{i=1}^{N-k-1}(z-\alpha_i)\prod_{i=1}^{N-k-1}(1/z-\bar{\alpha}_i) \\ &= (\alpha\bar{\alpha})z^k\left[\prod_{i=1}^{N-k-1}(-\bar{\alpha}_i)\right]\prod_{i=1}^{N-k-1}(z-\alpha_i)\prod_{i=1}^{N-k-1}(z-1/\bar{\alpha}_i) \end{aligned}$$

for some $k \geq 0$ and for $q(z) = \alpha\prod(z-\alpha_i)$. Note that $q(z)$ has strict degree $(N-k-1)$ and $z^{N-1}s(z)$ has strict degree $2(N-1)-k$. Also note that if $s(z) \in \mathcal{P}_{\mathbb{K}}$, then for $|z|=1$, $s(z) = q(z)\bar{q}(1/z) = q(z)\bar{q}(\bar{z}) = |q(z)|^2 \geq 0$. To show the converse, let

$$\begin{aligned} z^{N-1}s(z) &= cz^k\left[\prod_{i=1}^{N-k-1}(-\bar{\alpha}_i)\right]\prod_{i=1}^{N-k-1}(z-\alpha_i)\prod_{i=1}^{N-k-1}(z-1/\bar{\alpha}_i) \\ &= cz^{N-1}\prod_{i=1}^{N-k-1}(z-\alpha_i)\prod_{i=1}^{N-k-1}(1/z-\bar{\alpha}_i). \end{aligned}$$

By substituting values of z for which $1/z = \bar{z}$, we obtain necessarily that $c \geq 0$, so c may be written $c = \alpha\bar{\alpha}$, giving the result. $\square$

The previous result shows that roots α of the polynomial $z^{N-1}s(z) \in z^{N-1}\mathcal{P}_{\mathbb{K}}$ with $|\alpha| = 1$ must necessarily have even multiplicity. Thus, we have the following corollary.

COROLLARY 3.3. *For nonzero $s(z) \in \mathcal{P}'_{\mathbb{K}}$, $s(z) \notin \mathcal{P}_{\mathbb{K}}$ if and only if $s(z)$ is negative for some z on the unit circle or $s(z)$ has a root on the unit circle of odd multiplicity.*

We then conclude the following.

LEMMA 3.4. *For $s(z) \in \mathcal{P}'_{\mathbb{K}}$, $s(z) \in \mathcal{P}_{\mathbb{K}}$ if and only if $s(z)$ is nonnegative on the unit circle.*

Proof. The case when $s(z)$ is zero is readily dispensed with. If $s(z) \in \mathcal{P}_{\mathbb{K}}$, then, as shown earlier, $s(z)$ is nonnegative on the unit circle. For the converse, it suffices to show that if $s(z) \in \mathcal{P}'_{\mathbb{K}}$ has a root $e^{i\theta}$ of odd multiplicity on the unit circle, then $s(z)$ is negative somewhere on the unit circle. In this case $s(z) = (z-e^{i\theta})^{2m+1}t(z)$, where t is nonzero at $e^{i\theta}$. Let $z = e^{i(\theta+\delta)}$.

$$s(e^{i(\theta+\delta)}) = (e^{i\theta}(e^{i\delta}-1))^{2m+1}t(z) = e^{i(2m+1)\theta}[(-1)^m i\delta^{2m+1}]t(e^{i\theta}) + \mathcal{O}(\delta^{2m+2}).$$

Thus for sufficiently small δ, $s(e^{i(\theta+\delta)})$ and $s(e^{i(\theta-\delta)})$ must have different signs. $\square$

Thus, we have the following.

THEOREM 3.5. *$\check{F}_{\mathbb{K}}(\{D_i\}_{i=1-N}^{N-1})$ is convex.*

Proof. Both $\mathcal{P}_{\mathbb{K}}$ and the set of rational functions which are real and nonnegative on the unit circle are convex sets, so their intersection is convex, giving the result. □

COROLLARY 3.6. *For $T_i \in \mathbb{K}^{N\times N}$ Toeplitz, $\check{F}_{\mathbb{K}}(\{T_i\})$ is convex.*

COROLLARY 3.7. *For $T \in \mathbb{K}^{N\times N}$ upper triangular and Toeplitz, $\check{F}_{\mathbb{K}}(\{T^i\}_{i=0}^n)$ is convex.*

The results of §2 can be used to extend these results further.

COROLLARY 3.8. *For $T \in \mathbb{K}^{N\times N}$ upper triangular and Toeplitz and $B \in \mathbb{K}^{N\times N}$ normal with eigenvalues in $\mathbb{K}$, $\check{F}_{\mathbb{K}}(\{(T\otimes B)^i\}_{i=0}^n)$ is convex.*

COROLLARY 3.9. *For $T_i \in \mathbb{K}^{N_i\times N_i}$ upper triangular Toeplitz, $\check{F}_{\mathbb{K}}(\{(\mathrm{diag}\{T_i\})^j\}_{j=0}^n)$ is convex.*

COROLLARY 3.10. *For $T_i \in \mathbb{K}^{N_i\times N_i}$ upper triangular Toeplitz, and B and B_i normal matrices over $\mathbb{K}$ with eigenvalues in $\mathbb{K}$, $\check{F}_{\mathbb{K}}(\{((\mathrm{diag}\{T_i\}\otimes B_i\})\otimes B)^j\}_{j=0}^n)$ is convex.*

Thus, for a fairly large number of matrices, including normal matrices and direct sums of upper triangular Toeplitz matrices, if GMRES(s) converges, then the optimal polynomial preconditioner of corresponding degree must also converge, although in the latter case it is not clear that the convergence rate is necessarily the same.

One might be led to believe that the same result holds for all matrices. However, the next section shows that this is not the case.

4. Counterexamples for general matrices. In this section a method is given for generating matrices $A \in \mathbb{K}^{N\times N}$ for which $\psi_{n,\mathbb{K}}(A) < 1$ but $\varphi_{n,\mathbb{K}}(A) = 1$ for certain values of n and N.

In particular, we will construct a real nonsingular matrix A of dimension $N = 4$ such that $\psi_{N-1,\mathbb{R}}(A) < \varphi_{N-1,\mathbb{R}}(A) = 1$.

The following step-by-step process is used to construct the counterexample. It should be noted that the method given here may be used to generate other counterexamples $A \in \mathbb{R}^{N\times N}$, possibly for larger N, for which $\psi_{N-1,\mathbb{R}}(A) < \varphi_{N-1,\mathbb{R}}(A) = 1$; however, the construction is not necessarily always guaranteed to work, and each potential counterexample must be checked to confirm its validity.

Step 1: Construct HPD $M \in \mathbb{R}^{N\times N}$ and $w \in \mathbb{R}^N$ such that $H \equiv (D(w)M)_H$ is nonnegative definite with kernel of dimension 2. Here we define $D(w) = \sum_i (e_i^* w) e_i e_i^*$.

Let us note that for Hermitian M, $H = (D(w)M)_H = \Delta \circ M$, where $\Delta = \mathbf{1}w^* + w\mathbf{1}^*$, where $\mathbf{1}$ denotes the vector of all 1's, and $B\circ C$ denotes the Hadamard product of matrices, $\{b_{ij}c_{ij}\}$. We thus seek $H = \Delta\circ M$ with kernel of dimension 2. If w has no zero entries, then this may be written $M = \Delta^{\circ -1}\circ H$, where $B^{\circ -1}$ denotes the Hadamard inverse, $\{1/b_{ij}\}$.

We define the map $\mu : \mathbb{R}^N \times \prod_{i=1}^{N-2} \mathbb{R}^N \to \mathbb{R}^{N\times N}$ by

$$\mu(w, \{x_i\}_{i=1}^{N-2}) = (\mathbf{1}^* w + w^* \mathbf{1})^{\circ -1} \circ \sum_{i=1}^{N-2} x_i x_i^*.$$

Since $\Delta^{\circ -1}$ is nonnegative definite when $w_i > 0$ for all i [6, p. 348] and nonnegative definiteness is preserved by Hadamard products [6, p. 309], matrices in the image of μ for such w are nonnegative definite.

To obtain a matrix M in the image of μ that is not only nonnegative definite but also positive definite, we seek w and $\{x_i\}$ such that all symmetric matrices in a neighborhood of $\mu(w, \{x_i\})$ are in the image of μ. That is, we seek a point $(w, \{x_i\})$ where the Jacobian $J(\mu)$ is of rank $N(N+1)/2$, the dimension of the space of symmetric

matrices. This condition also assures a set of matrices M of positive measure which satisfy the desired condition.

It can be shown directly that for $N = 3$, $J(\mu)$, a matrix-valued function of size $N(N-1) = 6$, never has rank $N(N+1)/2 = 6$. However, when $N = 4$, the vectors w, x_1, x_2 yield $J(\mu)$ of size $N(N-1) = 12$ and of rank $N(N+1)/2 = 10$ and corresponding $M = \mu(w, x_1, x_2)$ of rank $(N-2) = 2$:

$$w = (1 \quad 2 \quad 2 \quad 3)^T, \quad x_1 = (8 \quad 8 \quad 3 \quad 9)^T, \quad x_2 = (5 \quad 8 \quad 2 \quad 8)^T.$$

This yields

$$M = \begin{pmatrix} 89/2 & 104/3 & 34/3 & 28 \\ 104/3 & 32 & 10 & 136/5 \\ 34/3 & 10 & 13/4 & 43/5 \\ 28 & 136/5 & 43/5 & 145/6 \end{pmatrix}.$$

The nullspace V of $H = \Delta \circ M = \sum_{i=1}^{N-2} x_i x_i^*$ consists of vectors perpendicular to both x_1 and x_2. A basis of V for our example is given by

$$y_1 = (0 \quad -3 \quad -4 \quad 4)^T, \quad y_2 = (-8 \quad -1 \quad 24 \quad 0)^T.$$

Step 2: Examine the image of $\{v \in \mathbb{K}^N : v^* M v = 1\} \cap \mathrm{Ker}_{\mathbb{K}}[(D(w)M)_H]$ by the vector of quadratic maps $\mathcal{M} = \sum_{i=1}^{N} e_i \otimes (e_i e_i^* M)$. For $\mathbb{K} = \mathbb{R}$ and $n = N - 1 = 3$, this is an ellipsoidal curve, being the image of an ellipse, which under appropriate conditions is nondegenerate.

We now demonstrate that for $\mathbb{K} = \mathbb{R}$ this image is the intersection of the supporting hyperplane $H_{\mathbb{K}}(w, 0)$ with $\check{F}_{\mathbb{K}}(\{e_i e_i^* M\}_{i=1}^N) \cap H_{\mathbb{K}}(\mathbf{1})$; the object $\check{F}_{\mathbb{K}}(\{e_i e_i^* M\}_{i=1}^N) \cap H_{\mathbb{K}}(\mathbf{1})$ is chosen here in anticipation of being transformed into $F_{\mathbb{K}}(\{A^i\}_{i=1}^{N-1})$ in a later stage of this construction. This result follows easily from noting that $\mathrm{Ker}_{\mathbb{K}}(D(w)M)_H = \{v \in \mathbb{K} : \mathrm{Re}[v^* D(w) M v] = 0\} = \{v \in \mathbb{K} : v^* D(w) M v = 0\}$ and

$$\begin{aligned} &H_{\mathbb{K}}(w, 0) \cap \check{F}_{\mathbb{K}}(\{e_i e_i^* M\}) \cap H_{\mathbb{K}}(\mathbf{1}) \\ &= \left\{ \sum_i e_i v^* (e_i e_i^* M) v : \sum_i w_i v^* (e_i e_i^* M) v = 0, \ \sum_i v^* (e_i e_i^* M) v = 1, \ v \in \mathbb{K}^N \right\} \\ &= \left\{ \sum_i e_i v^* (e_i e_i^* M) v : v^* D(w) M v = 0, \ v^* M v = 1, \ v \in \mathbb{K}^N \right\}. \end{aligned}$$

Note also that $\mathrm{Re}[w^* [\check{F}_{\mathbb{K}}(\{e_i e_i^* M\}_{i=1}^N) \cap H_{\mathbb{K}}(\mathbf{1})]] \geq 0$, since $(D(w)M)_H$ is nonnegative definite. Furthermore, this verifies that the image of the ellipse through the quadratic maps is contained in the two-dimensional plane $H_{\mathbb{R}}(w, 0) \cap H_{\mathbb{R}}(\mathbf{1})$.

For our example, let $y = a y_1 + b y_2$ represent an arbitrary element of $\mathrm{Ker}_{\mathbb{K}}[(D(w)M)_H]$. Then

$$\begin{aligned} y^* \mathcal{M} y = [a \quad b] \Bigg(&\begin{bmatrix} 0 & 448/3 \\ 448/3 & 2848/3 \end{bmatrix}, \begin{bmatrix} 408/5 & 588/5 \\ 588/5 & 208/3 \end{bmatrix}, \begin{bmatrix} 172/5 & -868/15 \\ -868/15 & -544 \end{bmatrix}, \\ &\begin{bmatrix} -232/3 & -448/5 \\ -448/5 & 0 \end{bmatrix} \Bigg) \begin{bmatrix} a \\ b \end{bmatrix} \equiv [a \quad b] \, \mathcal{Q} \begin{bmatrix} a \\ b \end{bmatrix}. \end{aligned}$$

We wish to verify that the curve generated by $(a, b)\mathcal{Q}(a, b)^T$, under the condition that (a, b) satisfies $(a, b)([y_1 \ y_2]^* M [y_1 \ y_2])(a, b)^T = 1$, is nondegenerate, i.e., is not

contained in a single line. This is true if three values of (a, b) can be given for which the three points $(a, b)\mathcal{Q}(a, b)^T$ are not collinear. In particular,

$$(1,0)\mathcal{Q}(1,0)^T = (\,0 \quad 408/5 \quad 172/5 \quad -232/3\,)^T,$$

$$(0,1)\mathcal{Q}(0,1)^T = (\,2848/3 \quad 208/3 \quad -544 \quad 0\,)^T,$$

$$(1,1)\mathcal{Q}(1,1)^T = (\,1248 \quad 5792/15 \quad -1876/3 \quad -3848/15\,)^T,$$

which can be shown not to be collinear.

What we have found then is a generalized field of values of matrices that has a supporting hyperplane whose intersection with the body is an ellipsoidal curve and thus is not convex.

Step 3: Select $x \in \text{cvx}[\check{F}_{\mathbb{K}}(\{e_i e_i^* M\}) \cap H_{\mathbb{K}}(\mathbf{1}) \cap H_{\mathbb{K}}(w, 0)] \setminus [\check{F}_{\mathbb{K}}(\{e_i e_i^* M\}) \cap H_{\mathbb{K}}(\mathbf{1}) \cap H_{\mathbb{K}}(w, 0)]$; that is, x is in the convex hull of the set just determined, but not in the set itself. Let us now confirm that such x also satisfies $x \in \text{cvx}[\check{F}_{\mathbb{K}}(\{e_i e_i^* M\})] \setminus [\check{F}_{\mathbb{K}}(\{e_i e_i^* M\})]$. Note that such x satisfies $x \in \text{cvx}[\check{F}_{\mathbb{K}}(\{e_i e_i^* M\}) \cap H_{\mathbb{K}}(\mathbf{1})] \setminus [\check{F}_{\mathbb{K}}(\{e_i e_i^* M\}) \cap H_{\mathbb{K}}(\mathbf{1})]$; specifically, $x \in \text{cvx}[\check{F}_{\mathbb{K}}(\{e_i e_i^* M\}) \cap H_{\mathbb{K}}(\mathbf{1}) \cap H_{\mathbb{K}}(w, 0)] \subseteq \text{cvx}[\check{F}_{\mathbb{K}}(\{e_i e_i^* M\}) \cap H_{\mathbb{K}}(\mathbf{1})]$, and $x \notin \check{F}_{\mathbb{K}}(\{e_i e_i^* M\}) \cap H_{\mathbb{K}}(\mathbf{1})$ since $x \in \text{cvx}[\check{F}_{\mathbb{K}}(\{e_i e_i^* M\}) \cap H_{\mathbb{K}}(\mathbf{1}) \cap H_{\mathbb{K}}(w, 0)] \subseteq \text{cvx}[H_{\mathbb{K}}(w, 0)] = H_{\mathbb{K}}(w, 0)$. Note also that $x \in \text{cvx}[\check{F}_{\mathbb{K}}(\{e_i e_i^* M\}) \cap H_{\mathbb{K}}(\mathbf{1})] \subseteq \text{cvx}[H_{\mathbb{K}}(\mathbf{1})] = H_{\mathbb{K}}(\mathbf{1})$. Thus $x \notin \check{F}_{\mathbb{K}}(\{e_i e_i^* M\})$.

For our example, let us select f_1, f_2 as values of $(a, b)\mathcal{Q}(a, b)^T$ and let $x = \alpha f_1 + \beta f_2$ for appropriate positive values of α, β. In particular, let $f_1 = (1,0)\mathcal{Q}(1,0)^T = (0, 408/5, 172/5, -232/3)$, $f_2 = (3,-1)\mathcal{Q}(3,-1)^T = (160/3, 1472/15, 564/5, -792/5)$, $\alpha = 15/8$, and $\beta = 1/8$, yielding $x = (100, 337, 276, -442) \in \text{cvx}[\check{F}_{\mathbb{K}}(\{e_i e_i^* M\})] \setminus [\check{F}_{\mathbb{K}}(\{e_i e_i^* M\})]$. Note that since $\check{F}_{\mathbb{K}}(\{e_i e_i^* M\})$ is a cone, such x may be easily scaled to ensure that $x \in \text{cvx}[\check{F}_{\mathbb{K}}(\{e_i e_i^* M\}) \cap H_{\mathbb{K}}(\mathbf{1}) \cap H_{\mathbb{K}}(w, 0)]$. However, this scaling is not important to the argument which follows.

Step 4: For such $x \in \mathbb{K}^N$ select $d \in \mathbb{K}^N$ with distinct positive entries such that $d^{\circ i} \perp x$ for $1 \le i \le N-1$, where $d^{\circ i}$ denote the Hadamard powers d, $d \circ d$, $d \circ d \circ d, \ldots$.

Let V be the Vandermonde matrix determined by d, $(V)_{i,j} = (d_j)^{i-1}$. Note that for such d, $e_i^* V x = 0$ for $i > 1$, and furthermore since $x \in H_{\mathbb{K}}(\mathbf{1})$, $Vx = e_1$.

For our example, the existence of a real solution for the entries of d can be proven formally by reducing the constraints through elimination of variables to a one-variable equation. The resulting one-variable equation can be seen to have a solution by the intermediate value theorem.

Writing $d = [d_1, d_2, d_3, 1]$, we have three equations:

$$100d_1 + 337d_2 + 276d_3 - 442 = 0, \tag{14}$$

$$100d_1^2 + 337d_2^2 + 276d_3^2 - 442 = 0, \tag{15}$$

$$100d_1^3 + 337d_2^3 + 276d_3^3 - 442 = 0. \tag{16}$$

We proceed by first solving equation (14) for d_3 and substituting in equations (15) and (16) to eliminate d_3. Next we formally solve equation (15) for d_2 (using a computer algebra system) and substitute one of the two results in equation (16) to eliminate d_2 (we found that which result was used did not affect the roots of the new equation (16)). We then proceed to solve equation (16). Although both computer algebra systems we tried returned several incorrect solutions, the one corresponding to $d_1 = -0.26769$

could be verified formally, simply by checking that the left-hand side $L(d_1)$ of equation (16) satisfies

$$L(-1/2) = -\frac{63362781}{751538} - \frac{183\sqrt{4519}^3}{751538\sqrt{7751}},$$
$$L(0) = \frac{20482722}{375769} - \frac{61\sqrt{35981055003114}}{\sqrt{7751}\ 375769}.$$

It can be checked that the signs of these two expressions are negative and positive, respectively.

We finally obtain

$$d = (-0.267698\ldots \quad 1.084117\ldots \quad 0.374718\ldots \quad 1).$$

Step 5: For such $M \in \mathbb{K}^{N\times N}$ let $M = P^{-1}P^{-*}$, $P \in \mathbb{K}^{N\times N}$ by, for example, singular value decomposition.

For our example, $M = UDU^*$, where $D = \text{diag}[(98.1079, 5.44092, 0.339205, 0.0286123)]$ and

$$U = \begin{bmatrix} -0.650955 & 0.732478 & -0.152607 & 0.12824 \\ -0.564055 & -0.330403 & 0.752453 & -0.0805637 \\ -0.180298 & -0.047945 & -0.257713 & -0.948039 \\ -0.474966 & -0.593306 & -0.586608 & 0.279797 \end{bmatrix}.$$

Then let $P^{-1} = U\sqrt{D}$.

Step 6: We now verify that for $A = PD(d)P^{-1}$, $\varphi_{N-1,\mathbb{R}}(A) = 1$ but $\psi_{N-1,\mathbb{R}}(A) < 1$. This will be done by using previous results to transform the general field of values of $\{e_i e_i M\}$ to that of $\{A^i\}$. This transformation will map the point x, located in the "hole" in the body, to the origin.

First note that for $A \in \mathbb{K}^{N\times N}$, $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, we may transfer from the standard to the conical generalized field of values:

$$0 \in F_{\mathbb{K}}(\{A^i\}_{i=1}^n) \Longleftrightarrow e_1 \in 1 \oplus F_{\mathbb{K}}(\{A^i\}_{i=1}^n) = H_{\mathbb{K}}(e_1) \cap \check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^n),$$

$$0 \in \text{cvx}\, F_{\mathbb{K}}(\{A^i\}_{i=1}^n) \Longleftrightarrow e_1 \in 1 \oplus \text{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^n)]$$

$$= \text{cvx}[1 \oplus F_{\mathbb{K}}(\{A^i\}_{i=1}^n)] = \text{cvx}[H_{\mathbb{K}}(e_1) \cap \check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^n)].$$

Note that for $n = N - 1$ and V the Vandermonde matrix defined above, since $V(I \otimes v^*) = (V \otimes 1)(I \otimes v^*) = V \otimes v^* = (I \otimes v^*)(V \otimes I)$, we have

$$V\left[(I \otimes v^*)\left(\sum_{i=1}^N e_i \otimes (Pe_i e_i^* P^{-1})\right)(1 \otimes v)\right]$$
$$= (I \otimes v^*)(V \otimes I)\left(\sum_{i=1}^N e_i \otimes (Pe_i e_i^* P^{-1})\right)(1 \otimes v)$$
$$= (I \otimes v^*)\left(\sum_{i,j} d_i^{j-1} e_j \otimes (Pe_i e_i^* P^{-1})\right)(1 \otimes v) = (I \otimes v^*)\left(\sum_{i=j}^N e_j \otimes A^{j-1}\right)(1 \otimes v),$$

and so for V, $P \in \mathbb{K}^{N\times N}$,

$$\begin{aligned}\check{F}_{\mathbb{K}}(\{A^{i-1}\}_{i=1}^{N}) &= V\check{F}_{\mathbb{K}}(\{Pe_ie_i^*P^{-1}\}_{i=1}^{N})\\ &= V\check{F}_{\mathbb{K}}(\{e_ie_i^*P^{-1}P^{-*}\}_{i=1}^{N}) = V\check{F}_{\mathbb{K}}(\{e_ie_i^*M\}_{i=1}^{N}),\end{aligned}$$

using the invariance of the conical field of values over the simultaneous congruence transformation. Note also that $VH_{\mathbb{K}}(\mathbf{1}) = H_{\mathbb{K}}(V^{-*}\mathbf{1}) = H_{\mathbb{K}}(e_1)$. We conclude that $H_{\mathbb{K}}(e_1) \cap \check{F}_{\mathbb{K}}(\{A^{i-1}\}_{i=1}^{N}) = V[H_{\mathbb{K}}(\mathbf{1}) \cap \check{F}_{\mathbb{K}}(\{e_ie_i^*M\}_{i=1}^{N})]$. As a result, since

$$x \in \operatorname{cvx}[\check{F}_{\mathbb{K}}(\{e_ie_i^*M\}_{i=1}^{N}) \cap H_{\mathbb{K}}(\mathbf{1})] \setminus [\check{F}_{\mathbb{K}}(\{e_ie_i^*M\}_{i=1}^{N}) \cap H_{\mathbb{K}}(\mathbf{1})],$$

we have

$$\begin{aligned}Vx = e_1 &\in \operatorname{cvx}[V\check{F}_{\mathbb{K}}(\{e_ie_i^*M\}_{i=1}^{N}) \cap H_{\mathbb{K}}(\mathbf{1})] \setminus [V\check{F}_{\mathbb{K}}(\{e_ie_i^*M\}_{i=1}^{N}) \cap H_{\mathbb{K}}(\mathbf{1})]\\ &= \operatorname{cvx}[\check{F}_{\mathbb{K}}(\{A^{i-1}\}_{i=1}^{N}) \cap H_{\mathbb{K}}(e_1)] \setminus [\check{F}_{\mathbb{K}}(\{A^{i-1}\}_{i=1}^{N}) \cap H_{\mathbb{K}}(e_1)],\end{aligned}$$

establishing the desired result.

The final result of this construction is the matrix

$$A = \begin{pmatrix} .469258949671 & .144764925686 & -.011212551044 & .000047280410\\ 2.610326570493 & .327595085371 & .028231187826 & -.010533891160\\ -3.242997268120 & .452834814744 & .976573801662 & -.038644914105\\ .162118329163 & -2.003125765058 & -.458143025300 & .417709385890 \end{pmatrix}.$$

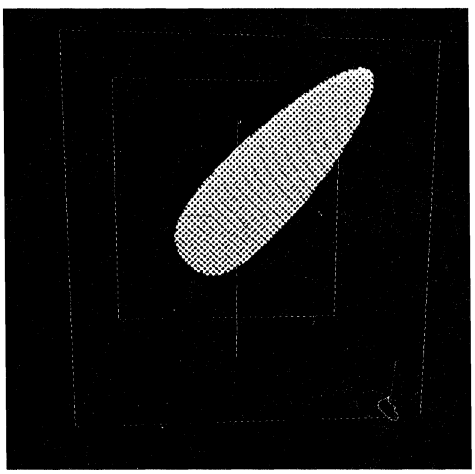


FIG. 1. *Generalized field of values.*

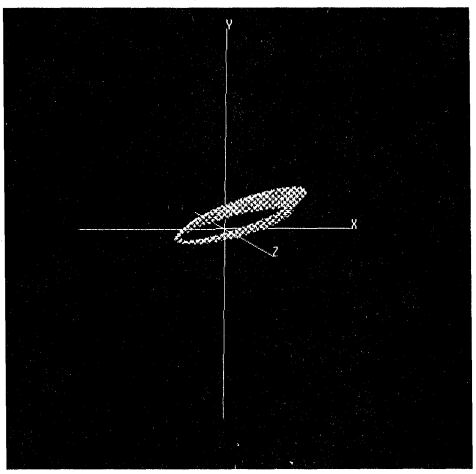


FIG. 2. *Slice of generalized field of values.*

In Figure 1 the generalized field of values $F_{\mathbb{R}}(\{A^i\}_{i=1}^{3})$ is rendered graphically. The generalized field of values is plotted in a box of extents $[-4, 4]$ along each axis, with the rightward x-axis corresponding to A, the upward y-axis to A^2, and the frontward z-axis to A^3.

The calculated distance of the generalized field of values to the origin is .001274. The calculated value of $\psi_{3,\mathbb{R}}(A)$ is approximately 0.99988.

The nonconvexity of this generalized field of values is supported by the illustration in Figure 2. This plot shows points in the generalized field of values that are on one side of the plane defined to be normal to the vector $(-.0691426, -.741006, .667929)$ and .00025 from the origin. The boundary of the convex hull of the body passes

exactly through the origin. Although the field of values in Figure 1 appears to be flat near the origin, in fact a small concavity exists near the origin, indicated by the "hole" in the graph of Figure 2, and furthermore $0 \in \text{cvx}[F_{\mathbb{R}}(\{A^i\}_{i=1}^3)] \setminus [F_{\mathbb{R}}(\{A^i\}_{i=1}^3)]$, so $\psi_{3,\mathbb{R}}(A) < \varphi_{3,\mathbb{R}}(A) = 1$.

5. Observations and open questions. Polynomial preconditioning is a popular and useful technique, insofar as it increases solution speed by reducing the requirements for inner product calculations, which is useful in its own right but has yet more advantage on certain advanced computer architectures for which inner products are particularly expensive (for a study of this issue, see, for example, [9]).

For HPD problems, polynomial preconditioning is robust. As shown in [2], not only do convergent preconditioners exist, but preconditioners with the same convergence rate as the conjugate gradient method exist. Thus, the main goal is to calculate preconditioners that give these good convergence rates.

On the other hand, the results of this paper indicate that for nonsymmetric problems, using polynomial preconditioning for the sake of increased speed may mean sacrificing robustness, the ability of the method to converge reliably to the solution of a given problem. Furthermore, this is a limitation, in principle, of the applicability of polynomial preconditioning as a technique. This problem is particularly critical for highly indefinite matrices, which commonly arise in practice and may require very many GMRES iterations before restarting in order to converge.

Let us summarize some particular facts we know regarding this issue.

1. Due to the counterexample given above, it is at least known that for $N = 4$ and $n = 3$, there exists $A \in \mathbb{R}^{N\times N}$ such that $\psi_{n,\mathbb{R}}(A) < \varphi_{n,\mathbb{K}} = 1$.

2. Since the counterexample matrix is nonsingular, it is known by the continuity theorems of §§2.2–2.5 that there is in fact a set of matrices of positive measure for which $\psi_{n,\mathbb{R}}(A) < \varphi_{n,\mathbb{K}}$ ($N = 4$, $n = 3$). In other words, there is a nonzero probability of an arbitrary matrix being such that restarted GMRES converges but the associated polynomial preconditioning does not. Exactly how large the set is or how to characterize the matrices is not known.

3. As shown in the theorem below, a set of matrices of positive measure exists for which both methods stagnate. Thus, $\psi_{n,\mathbb{R}}(A) = \varphi_{n,\mathbb{R}}(A)$ on a set of matrices of positive measure. This affirms the experience that a significant number of matrices result in slow convergence or stagnation for restarted GMRES (and thus slow convergence for other iterative methods such as biconjugate gradient or QMR as well). How to ascertain easily whether this will happen for a given matrix is not known.

THEOREM 5.1. *For every $N \geq 1$ and for every $n < N$, there exists a set of matrices $A \in \mathbb{R}^{N\times N}$ of positive measure in $\mathbb{R}^{N\times N}$ satisfying $1 = \psi_{n,\mathbb{R}}(A) = \varphi_{n,\mathbb{R}}(A)$.*

Proof. Without loss of generality, let $n = N - 1$. Let $\hat{A} = e_1 e_N^*/2 + \sum_{i=2}^N e_i e_{i-1}^*$. Note that for $\hat{r} = e_1$ and $g_A(r) \equiv r^* \underline{K}_N(r, A)$, $g_{\hat{A}}(\hat{r}) = e_1$. It is enough to show that for every sufficiently small real perturbation A of $\hat{A}$, the corresponding function g_A has a real solution r to the equation $g_A(r) = e_1$.

The Jacobian function $J(g_{\hat{A}})(r)$ for $g_{\hat{A}}$ with respect to r is the matrix-valued function $2[\,r \quad (\hat{A})_H r \quad \ldots \quad (\hat{A}^{N-1})_H r\,]$. Then

$$J(g_{\hat{A}})(\hat{r}) = [\,2e_1 \quad e_2 + e_N/2 \quad e_3 + e_{N-1}/2 \quad \ldots \quad e_N + e_2/2\,].$$

This matrix is of full rank by Gershgorin's theorem. Thus by the inverse function theorem there exist open sets $V \ni \hat{r}$ and $W \ni g_{\hat{A}}(\hat{r})$ such that $g_{\hat{A}} : V \to W$ is bijective.

There exists an open disk $\hat{W}_\epsilon \subseteq W$ of radius ϵ centered at $g_{\hat{A}}(\hat{r})$. By continuity, there exists $\delta > 0$ such that for every $A = \hat{A} + \delta E$, $||E|| = 1$, the image $g_A(V)$ contains $\hat{W}_{\epsilon/2}$, the disk centered at $g_{\hat{A}}(\hat{r})$ of radius $\epsilon/2$. Thus, for each such A, $g_{\hat{A}}(\hat{r})$ is in the image of g_A, and thus $g_A(r) = g_{\hat{A}}(\hat{r}) = e_1$ has a solution r. □

4. It is known that $\psi_{n,\mathbb{R}}(A) = \varphi_{n,\mathbb{R}}(A)$ on some important measure-zero sets of matrices such as Hermitian matrices. Furthermore, for the measure-zero set of upper triangular Toeplitz matrices, at least $\psi_{n,\mathbb{R}}(A) < \varphi_{n,\mathbb{R}}(A) = 1$ cannot occur. It is not clear whether a positive measure set exists for which $\psi_{n,\mathbb{R}}(A) = \varphi_{n,\mathbb{R}}(A) < 1$. However, it is known that positive measure sets exist for which $\psi_{n,\mathbb{R}}(A) \leq \varphi_{n,\mathbb{R}}(A) < 1$, due to continuity of the bound functions (small perturbations of an HPD matrix, for example).

5. One might ask how large the gap $\varphi_{n,\mathbb{R}}(A) - \psi_{n,\mathbb{R}}(A)$ can be. In the example given above, the gap is calculated to be approximately .00012. However, in a recent paper [15], a class of matrices is given for which $\varphi_{n,\mathbb{R}}(A) - \psi_{n,\mathbb{R}}(A)$ can be arbitrarily close to 1. Note that the gap cannot equal 1, since $0 = \psi_{n,\mathbb{R}}(A) < \varphi_{n,\mathbb{R}}(A)$ cannot occur. It is not known how to calculate this gap for a matrix in a simple and reliable way.

6. Conclusions. This paper has demonstrated several new results on the convergence rate of GMRES and polynomial preconditionings, including the fact that matrices exist for which restarted GMRES converges but every polynomial preconditioning of corresponding degree does not. Further research is required in order to devise practical tests for determining the convergence rates for these methods for matrices encountered in practice.

Acknowledgments. The authors would like to thank Anne Greenbaum for her comments and suggestions. The authors are also indebted to Roger Horn for pointing out some key results from his book with C. Johnson

REFERENCES

[1] 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.

[2] A. Greenbaum, *Comparison of splittings used with the conjugate gradient algorithm*, Numer. Math., 33 (1979), pp. 181–194.

[3] A. Greenbaum and L. Gurvits, *Max-min properties of matrix factor norms*, SIAM J. Sci. Comput., 15 (1994), pp. 427–439.

[4] A. Greenbaum and Z. Strakos, *Matrices that generate the same Krylov residual spaces*, in Recent Advances in Iterative Methods, G. Golub, A. Greenbaum, and M. Luskin, eds., Springer-Verlag, New York, 1994, pp. 95–118.

[5] A. Greenbaum and L. N. Trefethen, *GMRES/CR and Arnoldi/Lanczos as matrix approximation problems*, SIAM J. Sci. Comput., 15 (1994), pp. 427–439.

[6] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[7] A. S. Householder, *The Theory of Matrices in Numerical Analysis*, Dover, New York, 1964.

[8] W. Joubert, *Iterative Methods for the Solution of Nonsymmetric Systems of Linear Equations*, Report CNA-242, the University of Texas at Austin, Center for Numerical Analysis, 1990.

[9] W. D. Joubert and G. F. Carey, *Parallelizable restarted iterative methods for nonsymmetric linear systems. Part* I: *Theory,* Internat. J. Comput. Math., 44 (1992), pp. 243–267.

[10] W. Joubert, *A robust GMRES-based adaptive polynomial preconditioning algorithm for nonsymmetric linear systems*, SIAM J. Sci. Comput., 15 (1994), pp. 427–439.

[11] ———, *On the convergence behavior of the restarted GMRES algorithm for solving nonsymmetric linear systems*, Numer. Linear Algebra Appl., 1 (1994), pp. 427–447.

[12] W. D. JOUBERT AND T. A. MANTEUFFEL, *Iterative methods for nonsymmetric linear systems,* in Iterative Methods for Large Linear Systems, D. R. Kincaid and L. J. Hayes, eds., Academic Press, Boston, MA, 1990, pp. 149–171.

[13] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices*, 2nd ed., Academic Press, Boston, 1985.

[14] Y. SAAD AND M. H. SCHULTZ, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems,* SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.

[15] K.-C. TOH, *GMRES vs. ideal GMRES,* SIAM J. Matrix Anal. Appl., 18 (1997), to appear.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 4, pp. 730–743, October 1996

STABILITY THEORY FOR LINEAR INEQUALITY SYSTEMS*

M. A. GOBERNA†, M. A. LOPEZ†, AND M. TODOROV‡

Abstract. This paper develops a stability theory for (possibly infinite) linear inequality systems defined on a finite-dimensional space, analyzing certain continuity properties of the solution set mapping. It also provides conditions under which sufficiently small perturbations of the data in a consistent (inconsistent) system produce systems belonging to the same class.

Key words. linear inequality systems, stability, semicontinuity

AMS subject classifications. 65F99, 15A39, 49D39, 52A40

1. Introduction.

This paper deals with the stability of systems, in $\mathbb{R}^n$ (n fixed), of the form $\sigma_0 = \{a_t'x \geq b_t, t \in T\}$, where T represents a fixed arbitrary index set, $a_t = a(t) \in \mathbb{R}^n$, and $b_t = b(t) \in \mathbb{R}$ for all $t \in T$. Systems of this kind arise in fields such as functional approximation and linear semi-infinite programming. Observe that T is required to be neither a finite set (as in [5, 6]) nor endowed with a topology (as in [2, 3, 4, 8, 13, 17]). Since the right-hand side (RHS) functions do not necessarily range on a Banach space, our theory is not a special case of Robinson's [14, 15], whereas Tuy's stability theory [18] can be applied. In the present paper we prove the equivalence, for such a general class of systems, of different continuity properties of the solution set map for consistent systems, including the main stability concepts involved in the aforementioned papers. The attainment of conditions for the upper semicontinuity of this solution set map will be addressed in a forthcoming paper. In this paper we also study the stability properties of inconsistent systems, which are frequently neglected even though they arise in practical situations, not only due to the inaccuracy of the information involved in the models, but also because of the existence of antagonisms inherent to the modeled situations. In fact, authors such as Eremin (see [7] and the references therein) have considered the problem of perturbing a given inconsistent system in order to obtain a suitable (in some sense) consistent approximation. Therefore, in addition to the classical questions (i.e., which are the consistent systems such that sufficiently small perturbations provide consistent systems? and, if so, when the solution set changes gradually?) we consider others like the following one: When can an inconsistent system provide consistent systems through arbitrarily small perturbations? It will be useful to distinguish those systems that contain at least a finite-inconsistent subsystem (which will be called *strongly inconsistent*) from the remaining inconsistent system, the so-called *weakly inconsistent.*

Finally, we study the stability properties of a particular class of systems, not only for the aim of comparison with previous works, but also because of their wide range of applications. In fact, the so-called *continuous linear semi-infinite programming* (LSIP) (see [1] and the references therein) deals with the optimization of a linear

* Received by the editors October 16, 1995; accepted for publication by G. Cybenko October 17, 1995. Research supported by DGICYT of Spain grant PB93-0943, Generalitat Valenciana grant GV-2219/94, Bulgarian Ministry of Education and Science grant MM-21, and EC Commission grant CIPA3510PL929132.

† Department of Statistics and Operations Research, Faculty of Sciences, University of Alicante, 03071 Alicante, Spain.

‡ Bulgarian Academy of Sciences, Institute of Mathematics, 4002 Plovdiv, Bulgaria.

function over the solution set of a system σ_0 such that T is a compact Hausdorff space, $a(\cdot) \in \mathcal{C}(T)^n$, and $b(\cdot) \in \mathcal{C}(T)$. For this kind of system, which will be called *continuous* in what follows, we prove that any weakly inconsistent system can always be sufficiently approached through either consistent or strongly inconsistent systems, and we show some discontinuity properties of the solution set map. Therefore, the class of weakly inconsistent systems is intrinsically unstable in the context of continuous systems subject to continuous perturbations. It constitutes the expected behavior for a transition class.

2. Preliminaries. Given a nonempty set $X, X \subset \mathbb{R}^p, p \in \mathbb{N}$, we denote by conv X, cone X, and $\dim X$ the convex hull of X, the convex cone spanned by X, and the dimension of X, respectively. The Euclidean norm and distance in $\mathbb{R}^p$ will be denoted by $\|\cdot\|_2$ and ρ, respectively, whereas $\|\cdot\|$ represents the Chebyshev norm in $\mathbb{R}^p$ as well as in $\mathcal{C}(T)$. From the topological side, for $X \neq \varnothing$ contained in some topological space, int X, cl X, and bd X denote the interior, the closure, and the boundary of X, respectively.

The first part of the paper is devoted to (possibly noncontinuous) systems, where arbitrary perturbations of all the coefficients in all the constraints will be allowed. In order to define the size of a perturbation, recall that, given a metric space (X, δ), if X^T is the space of functions defined on T with values in X, the function $d\colon X^T \times X^T \to [0, +\infty]$, such that for f and g in $X^T, d(f,g) = \sup_{t\in T} \delta(f(t), g(t))$, provides a pseudometric space (X^T, d). The corresponding topology is Hausdorff, satisfies the first axiom of countability, and describes the uniform convergence on X^T. We shall consider, in particular, $X = \mathbb{R}^{n+1}$ endowed with the metric associated with $\|\cdot\|$, which yields the pseudometric *space of parameters* (Θ, d).

If $\sigma_1 := \{c_t' x \geq d_t, t \in T\}$, the pseudodistance between σ_1 and σ_0 is given (as in [12]) by

$$d(\sigma_1, \sigma_0) = \sup_{t\in T} \max\{|c_1(t) - a_1(t)|, \ldots, |c_n(t) - a_n(t)|, |d(t) - b(t)|\}.$$

Since we are not assuming, except in the last section, any particular property for the functional dependence between the coefficients and the associated indices, we prefer the notation a_t, b_t instead of $a(t), b(t)$.

We associate with a given system $\sigma_0 = \{a_t' x \geq b_t, t \in T\} \in \Theta$ its *solution set* F, two *moment cones* $M := \text{cone}\{a_t, t \in T\}$ and $\hat{M} := \text{cone}\{\binom{a_t}{b_t}, t \in T\}$, as well as its *characteristic cone*

$$K := \text{cone}\left\{\begin{pmatrix} a_t \\ b_t \end{pmatrix}, t \in T; \begin{pmatrix} 0_n \\ -1 \end{pmatrix}\right\}.$$

When at least two systems are simultaneously considered, they will be distinguished by subindices, as their corresponding solution sets, and their moment and characteristic cones (e.g., F_k will denote the solution set of $\sigma_k \in \Theta$).

Let us denote by LC, LW, and LS the subsets of Θ corresponding to the *linear consistent systems*, the *linear weakly inconsistent systems*, and the *linear strongly inconsistent systems*, respectively. According to [10, Thm. 4.1], $\sigma \in LC(LW, LS)$ if and only if $\binom{0_n}{1} \notin \text{cl } \hat{M}(\binom{0_n}{1} \in (\text{cl } \hat{M})\backslash\hat{M}, \binom{0_n}{1} \in \hat{M}$, respectively). The set of *linear inconsistent systems* will be represented by LI, and $LI = LW \cup LS$.

One of the purposes of the paper is the characterization of the interior parameters in the sets above. The approach given in this paper is closely related to the main aim—the study of the continuity properties of the solution set mapping $F : \Theta \to 2^{\mathbb{R}^n}$. Observe that given two sequences $\{x^r\}_{r=1}^{\infty} \subset \mathbb{R}^n$ and $\{\sigma_r\}_{r=1}^{\infty} \subset \Theta$, such that $x^r \in F_r$,

for $r = 1, 2, \ldots, \lim_{r\to\infty} x^r = x^0$, and $\lim_{r\to\infty} \sigma_r = \sigma_0$, since the functions describing the coefficients in $\{\sigma_r\}_{r=1}^{\infty}$ are pointwise convergent to the coefficients of σ_0, one has $x^0 \in F_0$. Hence F is a closed mapping.

Our attention will be focused on a number of continuity properties locally established at a parameter $\sigma_0 = \{a_t'x \geq b_t, t \in T\}$. First, let us recall the concept of lower semicontinuity: F is said to be *lower semicontinuous* (l.s.c.) at σ_0 if for each open set W in $\mathbb{R}^n$ such that $W \cap F_0 \neq \varnothing$ there exists an open set $\mathcal{V}$, $\sigma_0 \in \mathcal{V} \subset \Theta$, such that $W \cap F \neq \varnothing$ for all $\sigma \in \mathcal{V}$.

The next two conditions are, respectively, the classical Slater constraint qualification and a more restrictive version due to Helbig [13].

σ_0 satisfies the *Slater condition* if there exists a point $\bar{x} \in \mathbb{R}^n$ such that $a_t'\bar{x} > b_t$ for all $t \in T$.

σ_0 satisfies the *strong Slater condition* if there exist a positive scalar ε and a point $\bar{x} \in \mathbb{R}^n$ (called a strong Slater (SS) element) such that $a_t'\bar{x} \geq b_t + \varepsilon$ for all $t \in T$.

σ_0 is *noncritical* (in the sense of Tuy [18]) if the zero function satisfies $0 \notin \mathrm{bd}\{G_0(\mathbb{R}^n) - \mathbb{R}_+^T\}$, where $G_0 : \mathbb{R}^n \to \mathbb{R}^T$ is $G_0(x) := a_t'x - b_t$ and $\mathbb{R}_+^T$ denotes the positive cone in $\mathbb{R}^T$.

σ_0 is *regular* (in the sense of Robinson [15]) if $b(\cdot) \in \mathrm{int}\{A_0(\mathbb{R}^n) - \mathbb{R}_+^T\}$, where $A_0 : \mathbb{R}^n \to \mathbb{R}^T$ is $A_0(x) := a_t'x$, i.e., if sufficiently small perturbations of the RHS function do not affect the consistency.

F is *R-stable* (in Robinson's sense [15]) at σ_0 if it is consistent and for each $x^0 \in F_0$ two positive numbers exist, β and ε, such that the Hoffman-type inequality $\rho(x^0, F) \leq \beta v(x^0, \sigma)$ holds for any $\sigma = \{c_t'x \geq d_t, t \in T\}$, verifying $d(\sigma, \sigma_0) < \varepsilon$, where $\rho(x^0, F)$ $(= +\infty$ if $F = \varnothing)$ represents the Euclidean distance from x^0 to F (the solution set of σ), whereas

$$v(x^0, \sigma) = \max\{0, \sup_{t\in T}(d_t - c_t'x^0)\} \in [0, +\infty]$$

is a measure of the infeasibility, for σ, of the point x^0.

3. Consistent systems. Let $\sigma_0 = \{a_t'x \geq b_t, t \in T\} \in LC$ be given.

Next, we state the main result in this section.

THEOREM 3.1. *The following statements are equivalent.*

(i) F *is* l.s.c. *at* σ_0.

(ii) $\sigma_0 \in \mathrm{int}\ LC$.

(iii) σ_0 *is noncritical.*

(iv) σ_0 *is regular.*

(v) $0_{n+1} \notin \mathrm{cl\ conv}\{\binom{a_t}{b_t}, t \in T\}$.

(vi) σ_0 *satisfies the strong Slater condition.*

(vii) F *is R-stable at* σ_0.

Proof. After proving the chain of implications (i) $\to$ (ii) $\to$ (iii) $\to$ (iv) $\to$ (v) $\to$ (vi) $\to$ (i), we will show the equivalence between (vii) and the previous statements.

(i) $\to$ (ii) It is trivial.

(ii) $\to$ (iii) Let $\varepsilon > 0$ such that $\sigma \in LC$ provided that $d(\sigma, \sigma_0) \leq \varepsilon$. Let $f : T \to \mathbb{R}$ such that $|f(t)| \leq \varepsilon$ for all $t \in T$. Since $\sigma_1 := \{a_t'x \geq b_t + f(t), t \in T\} \in LC$, we can take some $x^1 \in F_1$. Then $p(t) := a_t'x^1 - b(t) - f(t) \geq 0$ for all $t \in T$; i.e., $f(\cdot) \in G_0(\mathbb{R}^n) - \mathbb{R}_+^T$, so that $0 \in \mathrm{int}\{G_0(\mathbb{R}^n) - \mathbb{R}_+^T\}$.

(iii) $\to$ (iv) Since σ_0 is a consistent noncritical system, it holds that $0 \in \mathrm{int}\{G_0(\mathbb{R}^n) - \mathbb{R}_+^T\}$. Then there will exist $\varepsilon > 0$ such that $f(\cdot) \in G_0(\mathbb{R}^n) - \mathbb{R}_+^T$ if

$|f(t)| \leq \varepsilon$ for all $t \in T$. For such a function $f, b(t) + f(t) \in A_0(\mathbb{R}^n) - \mathbb{R}_+^T$, so that $b(\cdot) \in \text{int}\{A_0(\mathbb{R}^n) - \mathbb{R}_+^T\}$.

(iv) $\rightarrow$ (v) Let us assume the contrary: $0_{n+1} \in \text{cl conv}\{\binom{a_t}{b_t}, t \in T\}$. Then a sequence $\{\lambda^r\}_{r=1}^{\infty}$ exists in the convex cone $\mathbb{R}_+^{(T)}$ of the generalized positive finite sequences (nonnegative real functions on T that vanish everywhere except on a finite subset of T), such that $\sum_{t\in T} \lambda_t^r = 1, r = 1, 2, \ldots$, and $0_{n+1} = \lim_{r\to\infty} \sum_{t\in T} \lambda_t^r \binom{a_t}{b_t}$.

Now, let $\varepsilon > 0$ such that $b(\cdot) + f(\cdot) \in A_0(\mathbb{R}^n) - \mathbb{R}_+^T$ for any $f : T \rightarrow \mathbb{R}$ such that $\sup_{t\in T} |f(t)| \leq \varepsilon$. In particular, if $f(t) = \varepsilon$ for all $t \in T, \sigma_\varepsilon := \{a_t' x \geq b_t + \varepsilon, t \in T\} \in LC$.

But $\lim_{r\to\infty} \sum_{t\in T} \lambda_t^r \binom{a_t}{b_t+\varepsilon} = \binom{0_n}{\varepsilon}$, so that $\binom{0_n}{\varepsilon} \in \text{cl } \hat{M}_\varepsilon$, which is a contradiction.

(v) $\rightarrow$ (vi) The consistency of σ_0 allows us to separate $\binom{0_n}{1}$ from $\text{cl}\hat{M}_0$, and because of (v) we can strongly separate 0_{n+1} from $\text{conv}\{\binom{a_t}{b_t}, t \in T\}$ (cf. [16, Thm. 11.4]).

Let $\hat{w} = \binom{w}{w_{n+1}}, w \in \mathbb{R}^n, w_{n+1} \in \mathbb{R}$, such that $\hat{w}'\binom{0_n}{1} < 0$ and $\hat{w}'\binom{a_t}{b_t} \geq 0$ for all $t \in T$. Analogously, there exist $\hat{u} \in \mathbb{R}^{n+1}$ and a $\delta > 0$ such that $\hat{u}'\binom{a_t}{b_t} \geq \delta$ for all $t \in T$.

Now consider a point $\hat{z} = \binom{z}{z_{n+1}} \in \{\hat{u} + \alpha\hat{w} | \alpha \geq 0\}$ such that $z_{n+1} < 0$.

Then for $\bar{x} := -\frac{1}{z_{n+1}} z$ one has

$$a_t'\bar{x} - b_t = -\frac{1}{z_{n+1}} \hat{z}' \begin{pmatrix} a_t \\ b_t \end{pmatrix} \geq -\frac{\delta}{z_{n+1}} > 0,$$

so that $\bar{x}$ is an SS element for σ_0.

(vi) $\rightarrow$ (i) Let $\bar{x} \in \mathbb{R}^n$ such that $a_t'x \geq b_t + \varepsilon$ for all $t \in T$ and for a certain $\varepsilon > 0$. Let $W \subset \mathbb{R}^n$ be an open set such that $W \cap F_0 \neq \varnothing$.

We shall show that $W \cap F_0$ contains an SS element for σ_0. Take an arbitrary $y \in W \cap F_0$ and consider $z := (1 - \lambda)y + \lambda\bar{x}$ for some $\lambda \in]0, 1]$ such that $z \in W \cap F_0$. One has

$$a_t'z = (1 - \lambda)a_t'y + \lambda a_t'\bar{x} \geq b_t + \lambda\varepsilon,$$

so that z is an SS element for σ_0. Now consider an arbitrary system $\sigma = \{c_t'x \geq d_t, t \in T\}$ such that $d(\sigma, \sigma_0) < \frac{\lambda\varepsilon}{2} \min\{1, n^{-1/2}\|z\|_2^{-1}\}$ (with $\|z\|_2^{-1} = +\infty$ if $z = 0_n$). The Cauchy– Schwartz inequality leads us to

$$c_t'z \geq a_t'z - \|z\|_2\|a_t - c_t\|_2 \geq a_t'z - \frac{\lambda\varepsilon}{2} \geq b_t + \frac{\lambda\varepsilon}{2} \geq d_t;$$

i.e., $z \in W \cap F$.

(vii) $\rightarrow$ (iv) Assume that F is R-stable at σ_0, and suppose that $b(\cdot) \in \text{bd}\{A_0(\mathbb{R}^n) - \mathbb{R}_+^T\}$. Take $x^0 \in F_0$ and consider the positive numbers β and ε for which $\rho(x^0, F) \leq \beta v(x^0, \sigma)$, provided that $d(\sigma, \sigma^0) < \varepsilon$. We can find $f : T \rightarrow \mathbb{R}$ such that $|f(t)| \leq \varepsilon/2$, for all $t \in T$, for which $b(\cdot) + f(\cdot) \notin A_0(\mathbb{R}^n) - \mathbb{R}_+^T$.

If $\sigma := \{a_t'x \geq b_t + f_t, t \in T\}$, we have $d(\sigma, \sigma_0) \leq \varepsilon/2 < \varepsilon$ and σ is inconsistent. Thus,

$$v(x^0, \sigma) = \sup_{t\in T}(f_t + b_t - a_t'x^0) \leq \sup_{t\in T} f_t + \sup_{t\in T}(b_t - a_t'x^0) < \varepsilon,$$

while $\rho(x^0, F) = +\infty$ and we get the contradiction $+\infty < \beta\varepsilon$.

Now assume that σ_0 satisfies the (equivalent) statements from (i) to (vi).

Let $x^0 \in F_0$ be given. Since (ii) holds, $F \neq \varnothing$ for $\sigma = \{c_t'x \geq d_t, t \in T\}$ close enough to σ_0, and there exists a point $x^F \in F$ such that $\rho(x^0, F) = \|x^F - x^0\|_2$.

We can confine ourselves to those systems σ such that $x^F \neq x^0$ and $v(x^0, \sigma) < +\infty$. (Otherwise, the aimed inequality holds trivially.)

It can easily be realized that $(x^F - x^0)'x \geq (x^F - x^0)'x^F$ is a consequent inequality of σ. Then, according to the Farkas lemma (see, e.g., [11, Thm. 2.1]), there exist sequences $\{\lambda^r\}_{r=1}^{\infty} \subset \mathbb{R}_+^{(T)}$ and $\{\mu_r\}_{r=1}^{\infty} \subset \mathbb{R}_+$ for which

$$\begin{pmatrix} x^F - x^0 \\ (x^F - x^0)'x^F \end{pmatrix} = \lim_{r\to\infty} \left\{ \sum_{t\in T} \lambda_t^r \begin{pmatrix} c_t \\ d_t \end{pmatrix} + \mu_r \begin{pmatrix} 0_n \\ -1 \end{pmatrix} \right\}; \tag{1}$$

i.e., the vector of coefficients lies in the closure of the characteristic cone. Multiplying both members of (1) by $\binom{x^F}{-1}'$, we obtain

$$\lim_{r\to\infty} \left\{ \sum_{t\in T} \lambda_t^r (c_t' x^F - d_t) + \mu_r \right\} = 0, \tag{2}$$

which itself implies $\lim_{r\to\infty} \mu_r = 0$ and

$$\lim_{r\to\infty} \sum_{t\in T} \lambda_t^r (c_t' x^F - d_t) = 0 \tag{3}$$

because of the nonnegativity of the general terms in both sequences.

Analogously, the scalar product by $\binom{x^F - x^0}{0}'$ yields

$$\|x^F - x^0\|_2^2 = \lim_{r\to\infty} \sum_{t\in T} \lambda_t^r [(c_t' x^F - d_t) + (d_t - c_t' x^0)].$$

Then, recalling (3), one has

$$\|x^F - x^0\|_2^2 = \lim_{r\to\infty} \sum_{t\in T} \lambda_t^r (d_t - c_t' x^0) \leq v(x^0, \sigma) \limsup_{r\to\infty} \sum_{t\in T} \lambda_t^r. \tag{4}$$

Since (vi) also holds, consider an SS element for $\sigma_0, \bar{x}$, and $\varepsilon > 0$ such that $a_t'\bar{x} - b_t \geq \varepsilon$ for all $t \in T$. Appealing once again to (1), we get

$$(x^F - x^0)'(\bar{x} - x^F) = \lim_{r\to\infty} \sum_{t\in T} \lambda_t^r (c_t'\bar{x} - d_t). \tag{5}$$

Let $\varepsilon_1 := \varepsilon(2 + 2n\|\bar{x}\|)^{-1}$.

If $d(\sigma, \sigma_0) < \varepsilon_1$, writing $c_t = a_t + f_t$ and $d_t = b_t + g_t$, one has, for all $t \in T$,

$$c_t'\bar{x} - d_t = (a_t'\bar{x} - b_t) + (f_t'\bar{x} - g_t) \geq \varepsilon - \varepsilon_1(n\|\bar{x}\| + 1) = \frac{\varepsilon}{2}.$$

Therefore, from (5), $(x^F - x^0)'(\bar{x} - x^F) \geq \frac{\varepsilon}{2} \limsup_{r\to\infty} \sum_{t\in T} \lambda_t^r$, so that

$$\limsup_{r\to\infty} \sum_{t\in T} \lambda_t^r \leq \frac{2}{\varepsilon} \|x^F - x^0\|_2 \|\bar{x} - x^F\|_2. \tag{6}$$

Now, consider the open set $W := \{x \in \mathbb{R}^n | \|x - x^0\|_2 < \varepsilon\}$, which obviously satisfies $W \cap F_0 \neq \varnothing$. Since (i) also applies, there will exist $\varepsilon_2 > 0, \varepsilon_2 \leq \varepsilon_1$, such that $W \cap F \neq \varnothing$ for any σ such that $d(\sigma, \sigma_0) < \varepsilon_2$. In this case $\rho(x^0, F) < \varepsilon$ and we have

$$\|\bar{x} - x^F\|_2 \leq \|\bar{x} - x^0\|_2 + \|x^0 - x^F\|_2 < \varepsilon + \|\bar{x} - x^0\|_2.$$

Replacing in (6) and taking $\beta := 2(1 + \varepsilon^{-1}\|\bar{x} - x^0\|_2)$, we get

$$\limsup_{r\to\infty} \sum_{t\in T} \lambda_t^r \leq \beta \|x^F - x^0\|_2. \tag{7}$$

Multiplying both members of (7) by $v(x^0, \sigma)$, recalling (4), and simplifying, we obtain the aimed conclusion:

$$\rho(x^0, F) = \|x^F - x^0\|_2 \leq \beta v(x^0, \sigma). \qquad \square$$

For consistent systems, the full dimensionality of F_0 is independent of the l.s.c. of F at σ_0. The following example illustrates such an assertion.

Example 1. Let $\sigma_i = \{r^i x_1 + x_2 \geq 0, r \in \mathbb{Z}\}, i = 1, 2.$

Obviously, $F_1 = \{0\} \times \mathbb{R}_+$ and $F_2 = \mathbb{R}^2_+$, so that $\dim F_1 < \dim F_2$. However, both systems are stable.

Let us make some additional remarks concerning Theorem 3.1 which shows, roughly speaking, that stability of linear inequality systems has the same meaning for most of the authors. Tuy proved, in a more general setting, the equivalence between (iii) and other stability concepts [18, Def. 1], pointing out that "results very near (...) have been obtained earlier by S. M. Robinson" [18, p. 33], without giving a precise statement. In fact, Robinson proved (iv) $\leftrightarrow$ (vii) for a class of linear systems which does not include Θ, giving an explicit bound for $\rho(x^0, F)$ [15, Thm. 1]. Our Theorem 3.1 also gives an error bound, $\beta v(x^0, \sigma)$, which depends on the SS element that we take, $\bar{x}$, and its associated positive scalar ε:

$$\beta = 2(1 + \varepsilon^{-1}\|\bar{x} - x^0\|_2).$$

Let us observe that stability, in the sense of lower semicontinuity, reduces itself to stability with respect to the RHS, so that Theorem 3.1 can be directly applied even to models where only the perturbation of $b(\cdot)$ makes sense. This is the case, e.g., for the following approximation problem: given a function $f : [\alpha, \beta] \to \mathbb{R}$, find a polynomial P of degree lower than a given $n \in \mathbb{N}$, such that $|f(t) - P(t)| \leq \varepsilon\, (\varepsilon > 0)$ for all $t \in [\alpha, \beta]$.

4. Inconsistent systems. Throughout this section $\sigma_0 = \{a_t' x \geq b_t, t \in T\}$ will be a given inconsistent system. First of all, observe that $F(\cdot)$ is trivially l.s.c. at σ_0, but it is neither R-stable nor regular, and it does not satisfy the Slater condition. Hence, our attention will be focused on the other remaining properties. Concerning $\sigma_0 \in \text{int } LI$, we are also interested in the sufficient conditions $\sigma_0 \in \text{int } LS$ and $\sigma_0 \in \text{int } LW$.

THEOREM 4.1. *If $\sigma_0 \in \text{int } LI$, then σ_0 is noncritical.*

Proof. Assume the contrary: $0 \in \text{bd}\{G_0(\mathbb{R}^n) - \mathbb{R}^T_+\}$. Then a sequence $\{f_r\}_{r=1}^{\infty} \subset G_0(\mathbb{R}^n) - \mathbb{R}^T_+$ exists such that $\lim_{r\to\infty} f_r(t) = 0$ uniformly on T. Consider the sequence $\sigma_r := \{a_t' x \geq b_t + f_r(t), t \in T\}, r = 1, 2, \ldots$. Obviously, $\sigma_r \in LC, r = 1, 2, \ldots$ and $\lim_{r\to\infty} \sigma_r = \sigma_0$, so that $\sigma_0 \in \text{bd } LC$. $\square$

Observe that the condition in Theorem 4.1 is not sufficient for $\sigma_0 \in \text{int } LI$ (consider $\{0x \geq 1\}$).

In order to characterize int LS we need the following lemma, whose proof is left to the reader.

LEMMA 4.2. *Let $\{a_s, s \in S\} \subset \mathbb{R}^n$ and $a \in \text{int cone}\{a_s, s \in S\}$. Then there exist some $\varepsilon > 0$ such that $a \in \text{int cone}\{c_s, s \in S\}$ for any function $c : S \to \mathbb{R}^n$ such that $\|a_s - c_s\| < \varepsilon$ for all $s \in S$.*

Let us introduce the family of systems $LB := \{\sigma \in \Theta | M = \mathbb{R}^n\}$. It has been proven [9, Thm. 2.1] that $LC \cap LB$ is the class of consistent systems whose solution set is bounded. As a consequence of Lemma 4.2, LB is an open set. In fact, applying Lemma 4.2 to $a = 0_n$, since $0_n \in \text{int cone}\{a_t, t \in T\} = \text{int } M$, if we consider $\sigma := \{c_t' x \geq b_t, t \in T\}$ with $\|a_t - c_t\| < \varepsilon$, for all $t \in T$, it holds that $0_n \in \text{int cone}\{c_t, t \in T\}$.

THEOREM 4.3. *The following statements are equivalent.*

(i) $\sigma_0 \in \text{int } LS$.

(ii) $\binom{0_n}{1} \in \text{int } \hat{M}_0$.

(iii) $\sigma_0 \in LB$.

Proof. The implication (i) $\rightarrow$ (ii) will be proven by assuming the contrary: $\binom{0_n}{1} \notin \text{int } \hat{M}_0$. Since $\sigma_0 \in LS, \binom{0_n}{1} \in \hat{M}_0$ and, therefore, $\binom{0_n}{1} \in \text{bd } \hat{M}_0$. The last assertion implies the existence of a $\hat{w} \in \mathbb{R}^{n+1}, \hat{w} \neq 0_{n+1}$, such that $\hat{w}'\binom{0_n}{1} = 0$ and $\hat{w}'\hat{y} \geq 0$ for all $\hat{y} \in \hat{M}_0$. Then we have $\hat{w} = \binom{w}{0}$ for a certain $w \neq 0_n$ and $w' a_t \geq 0$ for every $t \in T$.

Let $\sigma_\varepsilon := \{(a_t + \varepsilon w)' x \geq b_t, t \in T\}$ for $\varepsilon > 0$. If we had $\sigma_\varepsilon \in LS$, there would exist a $\lambda \in \mathbb{R}_+^{(T)}$ such that

$$\begin{pmatrix} 0_n \\ 1 \end{pmatrix} = \sum_{t \in T} \lambda_t \begin{pmatrix} a_t + \varepsilon w \\ b_t \end{pmatrix}.$$

Premultiplying both sides by $\hat{w}'$, we obtain

$$0 = \sum_{t \in T} \lambda_t (w' a_t + \varepsilon w' w) \geq (\varepsilon w' w) \sum_{t \in T} \lambda_t.$$

Since $\varepsilon w^T w > 0$, it should be $\lambda_t = 0$ for all $t \in T$, which yields a contradiction.

Therefore, for all $\varepsilon > 0, \sigma_\varepsilon \notin LS$; meanwhile, $d(\sigma_\varepsilon, \sigma_0) = \varepsilon \|w\|$ and we get the contradictory conclusion $\sigma_0 \notin \text{int } LS$.

(ii) $\rightarrow$ (i) Assume $\binom{0_n}{1} \in \text{int } \hat{M}_0$. Applying Lemma 4.2 to $\binom{0_n}{1}$ and $\{\binom{a_t}{b_t}, t \in T\}$, we know that there exists an $\varepsilon > 0$ such that $\binom{0_n}{1} \in \text{int cone}\{\binom{c_t}{d_t}, t \in T\}$ for any mapping $\binom{c_t}{d_t} : T \rightarrow \mathbb{R}^{n+1}$ such that $\|\binom{a_t}{b_t} - \binom{c_t}{d_t}\| < \varepsilon$ for all $t \in T$. Therefore, if $\sigma \in \Theta$ satisfies $d(\sigma, \sigma_0) < \varepsilon$, then $\binom{0_n}{1} \in \text{int } \hat{M} \subset \hat{M}$; i.e., $\sigma \in LS$.

(ii) $\rightarrow$ (iii) The projection map $\pi : \hat{x} = \binom{x}{x_{n+1}} \rightarrow x$ transforms open sets in $\mathbb{R}^{n+1}$ into open sets in $\mathbb{R}^n$. Hence, $0_n \in \pi(\text{int } \hat{M}_0) \subset \text{int } M_0$ and this implies $M_0 = \mathbb{R}^n$.

(iii) $\rightarrow$ (ii) Assuming the contrary, $\binom{0_n}{1} \in \text{bd } \hat{M}_0$ and (as in the proof of (i) $\rightarrow$ (ii)) there exists $w \in \mathbb{R}^n, w \neq 0_n$, such that $w' a_t \geq 0$ for all $t \in T$, but this is impossible because $\text{cone}\{a_t, t \in T\} = M_0 = \mathbb{R}^n$. □

Next, we shall provide a sufficient condition (Theorem 4.4) and a necessary condition (Theorem 4.7) for $\sigma_0 \in \text{int } LW$.

THEOREM 4.4. $\sigma_0 \in \text{int } LW$ *provided that the following two conditions hold:*

(i) $0_n \notin \text{cl conv}\{a_t, t \in T\}$ *and*

(ii) $\{b_t / \|a_t\|, t \in T\}$ *is unbounded from above.*

Proof. (i) guarantees the existence of an $\varepsilon > 0$ such that $\|x\| \geq \varepsilon$ for every $x \in \text{conv}\{a_t, t \in T\}$. Then $0_n \notin \text{cl conv}\{a_t, t \in T\} + \frac{\varepsilon}{2} \text{cl } B$. Let $\sigma := \{c_t' x \geq d_t, t \in T\}$ such that $d(\sigma, \sigma_0) \leq \frac{\varepsilon}{2}$. Since $c_t \in \{a_t\} + \frac{\varepsilon}{2} \text{cl } B$ for all $t \in T$,

$$\text{conv}\{c_t, t \in T\} \subset \text{cl conv}\{a_t, t \in T\} + \frac{\varepsilon}{2} \text{cl } B,$$

so that $0_n \notin \text{cl conv}\{c_t, t \in T\}$.

If we had $\binom{0_n}{1} \in \hat{M}$, there would exist $\lambda \in \mathbb{R}_+^{(T)}$ such that

$$\begin{pmatrix} 0_n \\ 1 \end{pmatrix} = \sum_{t \in T} \lambda_t \begin{pmatrix} c_t \\ d_t \end{pmatrix}.$$

Since $\sum_{t \in T} \lambda_t > 0$, it should be $0_n \in \operatorname{conv}\{c_t, t \in T\}$, and this is impossible. Therefore, $\binom{0_n}{1} \notin \hat{M}$.

On the other hand, according with (ii), for some sequence $\{t_k\}_{k=1}^{\infty} \subset T$, one has $\lim_{k \to \infty} \|a(t_k)\|^{-1} b(t_k) = +\infty$. Since $\|a(t_k)\| \geq \varepsilon$, necessarily $\lim_{k \to \infty} b(t_k) = +\infty$, and we can assume then that $b(t_k) > 0, k = 1, 2, \ldots$.

Since $\|c(t_k) - a(t_k)\| \leq \frac{\varepsilon}{2}$ and $|d(t_k) - b(t_k)| \leq \frac{\varepsilon}{2}, k = 1, 2, \ldots$, we can write

$$\|c(t_k)\| \leq \|a(t_k)\| + \frac{\varepsilon}{2} \quad \text{and} \quad \left| \frac{d(t_k)}{b(t_k)} - 1 \right| \leq \frac{\varepsilon}{2b(t_k)}.$$

Since $\lim_{k \to \infty} (d(t_k)/b(t_k)) = 1$, we also assume that $d(t_k) > 0, k = 1, 2, \ldots$.

Therefore,

$$\frac{\|c(t_k)\|}{d(t_k)} \leq \left\{ \frac{\|a(t_k)\|}{b(t_k)} + \frac{\varepsilon}{2b(t_k)} \right\} \frac{b(t_k)}{d(t_k)}$$

and, taking limits for $k \to \infty$, we get $\lim_{k \to \infty} (\|c(t_k)\|/d(t_k)) = 0$.

Lastly,

$$\lim_{k \to \infty} d(t_k)^{-1} \left(\frac{c(t_k)}{d(t_k)} \right) = \begin{pmatrix} 0_n \\ 1 \end{pmatrix},$$

which shows that $\binom{0_n}{1}$ belongs to $(\operatorname{cl} \hat{M}) \setminus \hat{M}$; i.e., $\sigma \in LW$. The proof is complete. □

LEMMA 4.5. *If $\sigma_0 \in LI$ and $b(\cdot)$ is bounded, then*

$$0_n \in \operatorname{cl} \operatorname{conv}\{a_t, t \in T\}.$$

Proof. Let $k > 0$ such that $|b(t)| \leq k$ for all $t \in T$.

Since $\binom{0_n}{1} \in \operatorname{cl} \hat{M}_0$, we can find $\hat{z}^r = \binom{z^r}{z^r_{n+1}} = \sum_{t \in T} \lambda_t^r \binom{a_t}{b_t}, \lambda^r \in \mathbb{R}_+^{(T)}, r = 1, 2, \ldots$, such that $\binom{0_n}{1} = \lim_{r \to \infty} \hat{z}^r$. We shall denote $\mu_r := \sum_{t \in T} \lambda_t^r, r = 1, 2, \ldots$.

For the sequence $\{\mu_r\}_{r=1}^{\infty} \subset \mathbb{R}_+$ three cases can arise.

If $\{\mu_r\}_{r=1}^{\infty}$ is unbounded, there exists a subsequence, denoted in the same way, such that $\lim_{r \to \infty} \mu_r = +\infty$. Since $\mu_r^{-1} z^r \in \operatorname{conv}\{a_t, t \in T\}$ and $\lim_{r \to \infty} \mu_r^{-1} z^r = 0_n$, we get $0_n \in \operatorname{cl} \operatorname{conv}\{a_t, t \in T\}$.

If $\{\mu_r\}_{r=1}^{\infty}$ is bounded, then it contains a convergent subsequence. Without loss of generality, we assume $\mu = \lim_{r \to \infty} \mu_r$. If $\mu > 0$ we can repeat the reasoning above to get the aimed conclusion.

If $\mu = 0$, since $z^r_{n+1} = \sum_{t \in T} \lambda_t^r b_t$, one has $|z^r_{n+1}| \leq k \mu_r$ and, taking limits once again, it holds that $\lim_{r \to \infty} |z^r_{n+1}| = 0$, which is a contradiction, showing that actually this case is impossible. □

LEMMA 4.6. *If $\sigma_0 \in (\operatorname{int} LI) \setminus (\operatorname{int} LS)$, then the RHS function $b(\cdot)$ is unbounded on T.*

Proof. We have $\sigma_0 \notin LB$ according to Theorem 4.3, and let $\varepsilon > 0$ for which $\sigma \in LI$ for any $\sigma \in \Theta$ such that $d(\sigma, \sigma_0) \leq \varepsilon$. The statement will be proven by assuming the contrary, i.e., the boundedness of $b(\cdot)$ on T, and obtaining a contradiction. First, observe that $0_n \in \operatorname{cl} \operatorname{conv}\{a_t, t \in T\}$, following Lemma 4.5.

If it were $0_n \in \text{bd conv}\{a_t, t \in T\}$, a nonzero vector w would exist such that $w'a_t \geq 0$ for every $t \in T$.

Now, consider $\sigma_1 := \{(a_t + \varepsilon\|w\|^{-1}w)'x \geq b_t, t \in T\}$. Observe that $w'(a_t + \varepsilon\|w\|^{-1}w) \geq \varepsilon\|w\|^{-1}w'w > 0$, for all $t \in T$, so that $0_n \notin \text{cl conv}\{a_t + \varepsilon\|w\|^{-1}w\}$, whereas $\sigma_1 \in LI$ since $d(\sigma_1, \sigma_0) = \varepsilon$. Applying Lemma 4.5 to σ_1, we conclude that $b(\cdot)$ should be unbounded, and this is not possible.

Hence, $0_n \in \text{int conv}\{a_t, t \in T\} \subset \text{int } M_0$ and $M_0 = \mathbb{R}^n$; i.e., $\sigma_0 \in LB$, which constitutes the aimed contradiction. □

THEOREM 4.7. *If $\sigma_0 \in \text{int } LW$, then $b(\cdot)$ is unbounded on T.*

Proof. It is a straightforward consequence of Lemma 4.6. □

A weakly inconsistent system with an unbounded RHS function does not necessarily belong to int LW (consider, e.g., $\{t^2x \geq t, t \in \mathbb{R}\}$).

5. Properties of the main parameter sets. This section deals with the properties of the sets LC, LS, LW, and LI, especially the existence of stable (int $LC \neq \varnothing$) and unstable (bd $LC \cap$ bd $LI \neq \varnothing$) linear inequality systems in $\mathbb{R}^n$ with a given index set T.

THEOREM 5.1. *The following propositions hold.*

(i) $\varnothing \neq \text{int } LC \subsetneqq LC$. *Moreover,* $LC \backslash \text{int } LC \subset \text{cl } LS$.

(ii) int $LS \subsetneqq LS$. *Moreover,* int $LS \neq \varnothing$ *if and only if* $|T| \geq n+1$.

(iii) $\varnothing \neq \text{int } LW \subsetneqq LW$ *if* $|T| = \infty$. *Otherwise,* $LW = \varnothing$.

(iv) int $LI \subsetneqq LI$. *Moreover,* int $LI \neq \varnothing$ *if and only if* $|T| \geq n+1$.

(v) int $LS \cup$ int $LW \subsetneqq$ int LI *if and only if* $|T| = \infty$.

Proof. (i) Consider $\sigma_i = \{0_n'x \geq b_t^i, t \in T\}, i = 1, 2$, where $b_t^1 = -1$ for all $t \in T$ and $b_t^2 = 0$ for all $t \in T$. It can easily be realized that $\sigma_1 \in \text{int } LC$ (since 0_n is an SS element) and $\sigma_2 \in LC \backslash \text{int } LC$ (because the Slater condition fails).

Now, let $\sigma_3 = \{a_t'x \geq b_t, t \in T\} \in LC, \sigma_3 \notin \text{int } LC$. According to Theorem 3.1, since (v) fails, there exists a sequence $\hat{u}^r = \binom{u^r}{\mu_r} \in \mathbb{R}^{n+1}, r = 4, 5, \ldots$, such that we can write $\hat{u}^r = \sum_{t \in T} \lambda_t^r \binom{a_t}{b_t}, \lambda^r \in \mathbb{R}_+^{(T)}, \sum_{t \in T} \lambda_t^r = 1$, and $\|\hat{u}^r\| \leq \frac{1}{2r}, r = 4, 5, \ldots$.

Let $\sigma_r := \{(a_t - u^r)'x \geq b_t - \mu_r + \frac{1}{2r}, t \in T\}, r = 4, 5, \ldots$.

Obviously, for any $r = 4, 5, \ldots$, one has $d(\sigma_r, \sigma_0) \leq \frac{1}{r}$, whereas

$$\sum_{t \in T} \lambda_t^r \begin{pmatrix} a_t - u^r \\ b_t - \mu_r + \frac{1}{2r} \end{pmatrix} = \frac{1}{2r} \begin{pmatrix} 0_n \\ 1 \end{pmatrix} \in \hat{M}_r,$$

so that $\sigma_r \in LS$.

(ii) Let $\sigma_0 = \{0_n'x \geq b_t, t \in T\}$ with $b_t = 1$ for all $t \in T$. According to Theorem 4.3, $\sigma_0 \in LS \backslash \text{int } LS$. On the other hand, if $|T| \leq n, \dim \hat{M} \leq n$ for all $\sigma \in \Theta$ and int $LS = \varnothing$ (also by Theorem 4.3). Conversely, if $|T| \geq n+1$, the functions $a_t : T \to \mathbb{R}^n$ and $b_t : T \to \mathbb{R}$ can be chosen in a variety of ways in order to fulfill condition (ii) in Theorem 4.3.

(iii) If $|T| < \infty$, then any inconsistent system is strongly inconsistent, so that $\varnothing = \text{int } LW = LW$.

If $|T| = \infty$, we can take a sequence $\{t_r\}_{r=1}^{\infty} \subset T$, such that $t_i \neq t_j$ for $i \neq j$. Let $\sigma_0 = \{x_n \geq b_t, t \in T\}$, where

$$b_t = \begin{cases} r \text{ if } t = t_r, r = 1, 2, \ldots, \\ 0 \text{ if } t \neq t_r \text{ for all } r. \end{cases}$$

Since cl conv$\{a_t, t \in T\} = \{\binom{0_{n-1}}{1}\}$ and $\lim_{r \to \infty}(b(t_r)/\|a(t_r)\|) = +\infty, \sigma_0 \in \text{int } LW$ as a consequence of Theorem 4.4.

The system

$$\sigma_1 = \left\{ \begin{array}{ll} r^{-2}x_n \geq r^{-1}, & t = t_r, r = 1, 2, \ldots, \\ 0_n'x \geq 0, & t \neq t_r \text{ for all } r \end{array} \right\}$$

is the limit, when $k \to \infty$, of the sequence

$$\sigma_k = \left\{ \begin{array}{c} (r^{-2} - k^{-2})x_n \geq r^{-1} - k^{-2}, t = t_r, r = 1, 2, \ldots, \\ 0_n'x \geq 0, t \neq t_r \text{ for all } r \end{array} \right\}, k = 2, 3, \ldots.$$

Since the inequality corresponding to t_k in σ_k is itself inconsistent, $\sigma_k \in LS$ for all $k = 2, 3, \ldots$, whereas $\sigma_1 \in LW$.

(iv) The first assertion is a consequence of $\sigma_0 \in LI \backslash \text{int } LI$ for σ_0 as in (ii), whereas the second one comes from (ii) and (iii).

(v) The direct statement is trivial since $LI = LS$ when $|T| < \infty$. Now, assume $|T| = \infty$. Let $\{t_r\}_{r=1}^{\infty} \subset T$, with $t_i \neq t_j$ for $i \neq j$, and consider $\sigma_0 = \{0_n'x \geq r, t = t_r, r = 1, 2, \ldots; 0_n'x \geq 0, t \neq t_r \text{ for all } r\}$. We shall prove that for any $\sigma \in \Theta$ such that $d(\sigma, \sigma_0) < \infty$, it holds that $F = \varnothing$. In fact, let $\sigma = \{a_t'x \geq b_t, t \in T\}$ with $d(\sigma, \sigma_0) \leq \delta, \delta > 0$. Assuming the contrary, i.e., $F \neq \varnothing$, and taking any pair $x \in F$ and $t \in T$, the Cauchy–Schwarz inequality gives

$$n^{1/2}\delta \|x\|_2 \geq \|a_t\|_2 \|x\|_2 \geq |a_t'x| \geq b_t.$$

Hence, $n^{1/2}\delta \|x\|_2 \geq \lim_{r \to \infty} b(t_r) = +\infty$, which is a contradiction. Consequently, $\sigma_0 \in \text{int } LI$.

On the other hand, let us consider, for any $\varepsilon > 0$, the following systems:

$$\sigma_\varepsilon := \{\varepsilon x_n \geq r, t = t_r, r = 1, 2, \ldots; 0_n'x \geq 0, t \neq t_r \text{ for all } r = 1, 2, \ldots\}.$$

Since $d(\sigma_\varepsilon, \sigma_0) = \varepsilon$ and $\sigma_\varepsilon \in LW$ for every $\varepsilon > 0$ whereas $\sigma_0 \in LS$, we conclude that $\sigma_0 \notin (\text{int } LW) \cup (\text{int } LS)$. □

As a consequence of Theorem 5.1, if $|T| = \infty$, the four (proper) subsets of Θ considered here are neither open nor closed. Consider σ_0 as in (v). One has $\sigma_0 \in LS \backslash \text{int } LS$, whereas $\sigma_0 \notin \text{cl } LC$. Hence, the symmetric inclusion of $LC \backslash \text{int } LC \subset \text{cl } LS$ fails.

The next example shows the existence of (nontrivial) highly unstable systems.

Example 2. Assume $|T| = \infty$ and take $\{t_r\}_{r=1}^{\infty} \subset T$ with $t_i \neq t_j$ if $i \neq j$. Let $s_1, s_2, \ldots$ be an enumeration of $\mathbb{Q} \cap [0, 1]$ and consider, for $\gamma \in \mathbb{R}$, the system $\sigma_\gamma := \{(1 + \cos 2\pi s_r)x_n \geq \gamma + \sin 2\pi s_r, t = t_r, r = 1, 2, \ldots; 0_n'x \geq 0, t \neq t_r \text{ for all } r = 1, 2, \ldots\}$.

Since $\binom{0_n}{1} \notin \text{cl } \hat{M}_\gamma$ for $\gamma < 0, \binom{0_n}{1} \in (\text{cl } \hat{M}_0) \backslash \hat{M}_0$ and $\binom{0_n}{1} \in \hat{M}_\gamma$ for $\gamma > 0$; then $\sigma_0 \in (\text{bd } LC) \cap LW \cap (\text{bd } LS)$ and $\sigma_0 \in (\text{bd } LC) \cap (\text{bd } LS) \cap (\text{bd } LW)$.

6. Continuous systems. In what follows T is assumed to be a compact Hausdorff space. Let us denote by Θ_c the (parameter) set of all the systems in $\mathbb{R}^n$ with (fixed) index set T and continuous coefficient functions. Consequently, we consider throughout this section only continuous perturbations of the coefficients. As can easily be seen, d induces the uniform distance in Θ_c, which becomes a complete metric space.

Let us denote by CC (CW, CS, CI) the set of continuous consistent (weakly inconsistent, strongly inconsistent, inconsistent) systems. The main purpose of this section

is to investigate the continuity properties of the solution set mapping $F\colon \Theta_c \to 2^{\mathbb{R}^n}$. Such properties will be the same as those that were defined in §2, except with minor changes. Specifically, in the definition of l.s.c. $\mathcal{V}$ will be an open set in Θ_c, whereas the topological operators involved in the concepts of noncritical and regular systems are those corresponding to $\mathcal{C}(T)$, whose positive cone, $\mathcal{C}_+(T)$, replaces $\mathbb{R}^T_+$. More in detail, our aim is to review and complete well-known results on the stability of continuous systems (including the finite systems, for which $\Theta_c = \Theta$) by exploiting the general theory developed in the previous sections. To achieve such a reduction we frequently appeal to Urisohn's lemma, which applies here since any compact Hausdorff space is a normal topological space.

In order to avoid confusion, we shall distinguish the toplogical operators in Θ_c by means of the subindex c. Thus, $\operatorname{int}_c CC$ denotes the interior of CC in Θ_c, which does not coincide with its interior in Θ, $\operatorname{int} CC$.

LEMMA 6.1. $\operatorname{int}_c CC = \Theta_c \cap \operatorname{int} LC$.

Proof. We have only to prove that $\operatorname{int}_c CC \subset \Theta_c \cap \operatorname{int} LC$. Assume that $\sigma_0 := \{a_t' x \geq b_t, t \in T\} \in \operatorname{int}_c CC$ and $\sigma_0 \notin \operatorname{int} LC$, and let $\varepsilon > 0$. Since $\sigma_0 \in LC \backslash \operatorname{int} LC$, applying Theorem 5.1(i), we conclude the existence of $\sigma := \{c_t' x \geq d_t, t \in T\} \in LS(\sigma$ probably noncontinuous) such that $d(\sigma, \sigma_0) < \frac{\varepsilon}{n+1}$.

Then, for some $\lambda \in \mathbb{R}_+^{(T)}$, it holds that $\binom{0_n}{1} = \sum_{t \in T} \lambda_t \binom{c_t}{d_t}$ and, according to Carathèodory's theorem, a set $\{t_1, \ldots, t_k\} \subset T$ exists, with $k \leq n+1$, such that $\lambda_t = 0$ for $t \notin \{t_1, \ldots, t_k\}$. Now, by Urisohn's lemma, let $f_i \in \mathcal{C}(T), f_i\colon T \to [0,1]$, such that $f_i(t_j) = 1$, if $j = i$, and $f_i(t_j) = 0$ for all $j \neq i, i = 1, 2, \ldots, k$. Consider the functions $g(t) := \sum_{i=1}^k [c(t_i) - a(t_i)] f_i(t)$ and $h(t) := \sum_{i=1}^k [d(t_i) - b(t_i)] f_i(t)$, which belong to $\mathcal{C}(T), \|g(\cdot)\| < \varepsilon, \|h(\cdot)\| < \varepsilon, g(t_i) = c(t_i) - a(t_i)$, and $h(t_i) = d(t_i) - b(t_i)$, $i = 1, 2, \ldots, k$.

Finally, consider $\sigma_\varepsilon := \{(a(t) + g(t))' x \geq b(t) + h(t), t \in T\}$. Obviously, $\sigma_\varepsilon \in \Theta_c$ and $d(\sigma_\varepsilon, \sigma_0) < \varepsilon$. Moreover,

$$\sum_{t \in T} \lambda_t \begin{pmatrix} a(t) + g(t) \\ b(t) + h(t) \end{pmatrix} = \sum_{i=1}^k \lambda_{t_i} \begin{pmatrix} c_{t_i} \\ d_{t_i} \end{pmatrix} = \begin{pmatrix} 0_n \\ 1 \end{pmatrix},$$

so that $\sigma_\varepsilon \in CS$. Therefore $\sigma_0 \notin \operatorname{int}_c CC$, which contradicts the assumption. □

THEOREM 6.2. *Given* $\sigma_0 := \{a_t' x \geq b_t, t \in T\} \in CC$, *the following statements are equivalent.*

(i) F *is* l.s.c. *at* σ_0.
(ii) $\sigma_0 \in \operatorname{int}_c CC$.
(iii) σ_0 *is noncritical.*
(iv) σ_0 *is regular.*
(v) $0_{n+1} \notin \operatorname{conv}\{\binom{a_t}{b_t}, t \in T\}$.
(vi) σ_0 *satisfies the strong Slater condition.*
(vii) σ_0 *satisfies the Slater condition.*
(viii) F *is R-stable at* σ_0.
(ix) $\dim F_0 = n$ *and* σ_0 *does not contain the trivial inequality* $0_n' x \geq 0$.

Proof. By Lemma 6.1, $\sigma_0 \in \operatorname{int}_c CC$ if and only if $\sigma_0 \in \operatorname{int} LC$. This condition can be replaced by any of the other six equivalent conditions listed in Theorem 3.1, which in some cases can be reformulated.

In fact, since $\{\binom{a_t}{b_t}, t \in T\}$ is a compact set,

$$\operatorname{cl}\operatorname{conv}\left\{\begin{pmatrix} a_t \\ b_t \end{pmatrix}, t \in T\right\} = \operatorname{conv}\left\{\begin{pmatrix} a_t \\ b_t \end{pmatrix}, t \in T\right\},$$

whereas Slater and strong Slater conditions coincide in Θ_c. Hence, (ii) is equivalent to statements (v) through (viii) and it implies (i). On the other hand, the proofs of (i) $\rightarrow$ (ii) $\rightarrow$ (iii) $\rightarrow$ (iv) $\rightarrow$ (v) in Theorem 3.1 remain valid here with the exception of small changes (the details are left to the reader). Therefore, statements (i) through (viii) are equivalent. Finally, (vii) is equivalent to (ix) by the straightforward application of [9, Cor. 3.2.1]. □

Some of the equivalences established in Theorem 6.2 appeared in previous papers: (ii) $\leftrightarrow$ (v) was proven in [17, Prop. 1.3], (v) $\leftrightarrow$ (vii) was proven in [4, Thm. 3.5], and (iv) $\leftrightarrow$ (viii) was established, in a more general frame, in [15]. Concerning the l.s.c. of F at σ_0, the classical sentence establishes its equivalence with the following statement: either σ_0 satisfies Slater condition or $|F_0| = 1$ [8, Thm. 4.1], [3, Thm. 2.1(c)], and [13, Cor. 4.3]. The following example shows that $|F_0| = 1$ is not a sufficient condition, even for finite systems.

Example 3. Consider, for $n = 2$ and $T = \{1, 2, 3, 4\}$, the systems, for $\varepsilon \geq 0, \sigma_\varepsilon := \{x_1 \geq \varepsilon, -x_1 \geq 0, x_2 \geq 0, -x_2 \geq 0\}$. Obviously, $F_0 = \{0_2\}$, whereas $F_\varepsilon = \varnothing$ for any $\varepsilon > 0$. Since $d(\sigma_\varepsilon, \sigma_0) = \varepsilon, F$ is not l.s.c. at σ_0.

COROLLARY 6.2.1. *If $\sigma_0 = \{a_t' x \geq b_t, t \in T\} \in CC$ and its corresponding homogeneous system, $\{a_t' x \geq 0, t \in T\}$, has some strict solution, then F is* l.s.c. *at σ_0.*

Proof. If $a_t' x^1 > 0$ for all $t \in T$, then some $\alpha > 0$ exists such that $\max_{t \in T} b_t < \alpha \min_{t \in T} a_t' x^1$, so that $x^0 := \alpha x^1$ is a strict solution of σ_0. □

THEOREM 6.3. *Given $\sigma_0 \in \Theta_c$, the following statements are equivalent.*

(i) $\sigma_0 \in \operatorname{int}_c CS$.

(ii) $\binom{0_n}{1} \in \operatorname{int} \hat{M}_0$.

(iii) $\sigma_0 \in LB$.

Proof. It is similar to the proof of Theorem 4.3. □

Concerning $\operatorname{int}_c CW$, we prove next that it is the empty set, so that its characterization is meaningless.

THEOREM 6.4. *The following propositions hold.*

(i) $\varnothing \neq \operatorname{int}_c CC \subsetneq CC$. *Moreover,* $CC \backslash \operatorname{int}_c CC \subset \operatorname{cl} CS$.

(ii) $\operatorname{int}_c CS \subsetneq CS$. *Moreover,* $\operatorname{int}_c CS \neq \varnothing$ *if and only if* $|T| \geq n + 1$.

(iii) $\operatorname{int}_c CW = \varnothing$. *Moreover,* $CW \neq \varnothing$ *if and only if* $|T| = \infty$.

(iv) $\operatorname{int}_c CI = \operatorname{int}_c CS$. *Therefore,* $\operatorname{int}_c CI \neq \varnothing$ *if and only if* $|T| \geq n + 1$.

Proof. (i) can be proven as Theorem 5.1(i) was, realizing that the limiting process, in the second part, can be avoided.

(ii) The first part is the same as the proof of Theorem 5.1(ii). On the other hand, if $|T| \leq n$, $\operatorname{int}_c CS = \varnothing$ as a consequence of Theorem 6.3.

Now, assume $|T| \geq n + 1$. Let $t_1, \ldots, t_{n+1}$ be different elements of T and take functions $f_1, \ldots, f_{n+1}$ (as in Lemma 6.1) such that $f_i \colon T \rightarrow [0, 1], f_i \in C(T)$, and $f_i(t_j) = 1$ if $j = i$, whereas $f_i(t_j) = 0$ for all $j \neq i, i = 1, 2, \ldots, n + 1$.

Let $\hat{z}^1, \ldots, \hat{z}^{n+1}$ be arbitrarily chosen points in $\mathbb{R}^{n+1}$ such that $\binom{0_n}{1} \in \operatorname{int} \operatorname{cone}\{\hat{z}^1, \ldots, \hat{z}^{n+1}\}$. Defining $\binom{a(t)}{b(t)} := \sum_{i=1}^{n+1} f_i(t) \hat{z}^i$, one has

$$\binom{0_n}{1} \in \operatorname{int} \operatorname{cone} \left\{ \binom{a(t_i)}{b(t_i)}, i = 1, 2, \ldots, n + 1 \right\} \subset \operatorname{int} \hat{M}_0$$

for $\sigma_0 := \{a_t' x \geq b_t, t \in T\}$. Therefore, $\sigma_0 \in \operatorname{int}_c CS$ (again by Theorem 6.3).

(iii) Since $\operatorname{int}_c CW \cap \operatorname{int}_c CS = \varnothing$ and $\operatorname{int}_c CW \subset \operatorname{int}_c CI$, the first statement will be a straightforward consequence of (iv).

Now, assume $|T| = \infty$. At least an ω-accumulation point exists, say $\bar{t} \in T$. For an arbitrary point $t_1 \in T, t_1 \neq \bar{t}$, two open disjoint sets exist, V_1 and W_1, such that $t_1 \in V_1$ and $\bar{t} \in W_1$. Since $T \backslash V_1 (\supset W_1)$ is a compact infinite set, we can take, in the same way, $t_2 \in T \backslash V_1, t_2 \neq \bar{t}$, and two open sets which do not contain t_1, V_2, and W_2, such that $t_2 \in V_2$ and $\bar{t} \in W_2$. We inductively obtain a sequence $\{t_r\}_{r=1}^{\infty} \subset T$ such that $V_k \cap \{t_1, t_2, \ldots\} = \{t_k\}, k = 1, 2, \ldots$. Given $k \in \mathbb{N}$, consider the closed disjoint sets $\{t_1, t_2, \ldots, t_k\}$ and $\mathrm{cl}\{t_i, i = k+1, \ldots\} \subset T \backslash \cup_{i=1}^{k} V_i$, which is also closed.

Urysohn's lemma can be applied to the above disjoint closed sets. Let $f_k \colon T \to [0,1], f_k \in \mathcal{C}(T)$, such that

$$\begin{cases} f_k(t_i) = 0, i = 1, 2, \ldots, k, \\ f_k(t) = 1 \text{ for all } t \in \mathrm{cl}\{t_i, i = k+1, \ldots\}. \end{cases}$$

Hence, the function $\varphi(t) := \sum_{k=1}^{\infty} 2^{-k} f_k(t)$ satisfies $\varphi \colon T \to [0,1], \varphi \in \mathcal{C}(T)$ and $\varphi(t_r) = \sum_{k=1}^{\infty} 2^{-k} f_k(t_r) = \sum_{k=r}^{\infty} 2^{-k} = 2^{1-r}$ for $r = 1, 2, \ldots$.

Finally, let us verify that $\sigma := \{\varphi^2(t) x_n \geq \varphi(t), t \in T\} \in CW$. In fact,

$$\lim_{r \to \infty} 2^{r-1} \begin{pmatrix} 0_{n-1} \\ \varphi^2(t_r) \\ \varphi(t_r) \end{pmatrix} = \begin{pmatrix} 0_n \\ 1 \end{pmatrix} \in \mathrm{cl}\ \hat{M}.$$

On the other hand, if it were

$$\begin{pmatrix} 0_n \\ 1 \end{pmatrix} = \sum_{t \in T} \lambda_t \begin{pmatrix} 0_{n-1} \\ \varphi^2(t) \\ \varphi(t) \end{pmatrix}$$

for some $\lambda \in \mathbb{R}_+^{(T)}$, it would be $\sum_{t \in T} \lambda_t \varphi^2(t) = 0$, whereas $\sum_{t \in T} \lambda_t \varphi(t) = 1$, which is a contradiction. Therefore $\binom{0_n}{1} \in (\mathrm{cl}\ \hat{M}) \backslash \hat{M}$ and $\sigma \in CW$.

(iv) We have to prove only that $\mathrm{int}_c CI \subset \mathrm{int}_c CS$. Let $\sigma_0 := \{a_t' x \geq b_t, t \in T\} \in \Theta_c \backslash \mathrm{int}_c CS$. According to Theorem 6.3, $\sigma_0 \notin LB$, and there exists a vector $w \neq 0_n$ such that $w' a_t \geq 0$ for all $t \in T$. Consider that $\sigma_\varepsilon := \{(a_t + \varepsilon w)' x \geq b_t, t \in T\}$ for $\varepsilon > 0$. Since $\lim_{r \to \infty} (a_t + \varepsilon w)' r w = +\infty$, whereas $b(\cdot)$ is bounded on $T, \sigma_\varepsilon \in CC$ for all $\varepsilon > 0$, so that $\sigma_0 \notin \mathrm{int}_c CI$. This completes the proof. □

Finally, let us show through a suitable example the existence of nontrivial systems in $(\mathrm{bd}_c CC) \cap (\mathrm{bd}_c CS) \cap (\mathrm{bd}_c CW)$.

Example 4. Assume $|T| = \infty$ and take $\varphi \in \mathcal{C}(T)$ and $\{t_r\}_{r=1}^{\infty} \subset T$ such that $\varphi(t_r) = 2^{1-r}, r = 1, 2, \ldots$, (as in the proof of Theorem 6.4(iii)). Obviously, $\varphi^{-1}(0) \neq \varnothing$. Define, for $\gamma \in \mathbb{R}$,

$$\sigma_\gamma := \{[1 - \cos \varphi(t)] x_n \geq \gamma + \sin \varphi(t), t \in T\}.$$

Reasoning as in Example 2, $\sigma_0 \in CW$, whereas $\sigma_\gamma \in CC$ for $\gamma < 0$ and $\sigma_\gamma \in CS$ for $\gamma > 0$. The conclusion comes straightforwardly.

Acknowledgments. The authors are indebted to the anonymous referees for their valuable suggestions.

REFERENCES

[1] E. J. Anderson and P. Nash, *Linear Programming in Infinite Dimensional Spaces*, Wiley, New York, 1987.

[2] B. BROSOWSKI, *Parametric Semi-Infinite Optimization*, Verlag Peter Lang, Frankfurt am Main, 1982.

[3] ———, *Parametric semi-infinite linear programming* I. *Continuity of the feasible set and the optimal value*, Math. Programming Study, 21 (1984), pp. 18–42.

[4] G. CHRISTOV AND M. TODOROV, *Semi-infinite optimization: Existence and uniqueness of the solution*, Math. Balkanica, 2 (1988), pp. 182–191.

[5] J. W. DANIEL, *On perturbations in systems of linear inequalities*, SIAM J. Numer. Anal., 10 (1973), pp. 299–307.

[6] ———, *Remarks on perturbations in linear inequalities*, SIAM J. Numer. Anal., 12 (1975), pp. 770–772.

[7] I. I. EREMIN, *Methods of parametrization in the analysis of improper mathematical programming problems*, in Mathematical Research No. 35, Akademie-Verlag, Berlin, 1987, pp. 82–94.

[8] T. FISCHER, *Contributions to semi-infinite linear optimization*, in Methoden und Verfahren der mathematischen Physik, Band 27, Verlag Peter Lang, Berlin, 1983, pp. 175–199.

[9] M. A. GOBERNA AND M. A. LÓPEZ, *A theory of the linear inequality systems*, Linear Algebra Appl., 106 (1988), pp. 77–115.

[10] M. A. GOBERNA, M. A. LÓPEZ, J. A. MIRA, AND J. VALLS, *On the existence of solutions for linear inequality systems*, J. Math. Anal. Appl., 192 (1995), pp. 133–150.

[11] M. A. GOBERNA, M. A. LÓPEZ, AND J. T. PASTOR, *Farkas–Minkowski systems in semi-infinite programming*, Appl. Math. Optim., 7 (1981), pp. 295–308.

[12] H. J. GREENBERG AND W. P. PIERSKALLA, *Stability theorems for infinitely constrained mathematical programs*, J. Opt. Theory Appl., 16 (1975), pp. 409–428.

[13] S. HELBIG, *Stability in disjunctive linear optimization* I: *Continuity of the feasible set*, Optimization, 21 (1990), pp. 855–869.

[14] S. M. ROBINSON, *Normed convex processes*, Trans. Amer. Math. Soc., 174 (1972), pp. 127– 140.

[15] ———, *Stability theory for systems of inequalities. Part* I: *Linear systems*, SIAM J. Numer. Anal., 12 (1975), pp. 754–769.

[16] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[17] M. TODOROV, *Generic existence and uniqueness of the solution set to linear semi-infinite optimization problems*, Numer. Funct. Anal. Optim., 8 (1985–1986), pp. 541–556.

[18] H. TUY, *Stability property of a system of inequalities*, Math. Operationsforsch. Statist. Optim., 8 (1977), pp. 27–39.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 4, pp. 744–747, October 1996

003

A CHARACTERIZATION AND REPRESENTATION OF THE DRAZIN INVERSE*

WEI YIMIN†

Abstract. We establish characterization and representation for the Drazin inverse of an arbitrary square matrix which reduce to the well-known result if the matrix is nonsingular.

Key words. Drazin inverse, index, rank, restriction

AMS subject classifications. 15A09, 65F20

1. Introduction. If A is an $n \times n$ complex matrix, then the Drazin inverse [4] of A, denoted by A^D, is the unique matrix X satisfying the relations

$$A^{k+1}X = A^k, \quad XAX = X, \quad AX = XA, \tag{1.1}$$

where $k = Ind(A)$, the index of A, is the smallest nonnegative integer for which $\text{rank}(A^k) = \text{rank}(A^{k+1})$.

Particularly when $Ind(A) = 1$, the matrix X satisfying (1.1) is called the group inverse of A and is denoted by $X = A^{\#}$. If A is nonsingular, then it is easily seen that $Ind(A) = 0$ and A^{-1} satisfies (1.1); i.e., $A^D = A^{-1}$.

The Drazin inverse is very useful since various applications (for example, applications in singular differential difference equations, Markov chains, cryptography, iterative methods, and multibody system dynamics) were found in the literature [2, 6, 9–11, 14–16], respectively.

In this paper, we present a characterization and representation for the Drazin inverse which reduce to the well-known result if the matrix is nonsingular.

As usual, let $R(A)$ be the range of A, $N(A)$ the null space of A.

2. A characterization for the Drazin inverse. It is a well-known fact that if A is a nonsingular matrix of order n, then the inverse of A, A^{-1}, is the unique matrix X for which

$$\text{rank}\begin{pmatrix} A & I \\ I & X \end{pmatrix} = \text{rank}(A). \tag{2.1}$$

In this section,we present a generalization of this fact to singular matrix A to obtain an analogous result for the Drazin inverse A^D of A.

The following lemma is needed in what follows.

LEMMA 1. *Let M be a $2n \times 2n$ matrix partitioned as*

$$M = \begin{pmatrix} A & AQ \\ PA & B \end{pmatrix}.$$

Then

$$\text{rank}(M) = \text{rank}(A) + \text{rank}(B - PAQ).$$

* Received by the editors March 6, 1995; accepted for publication (in revised form) by G. P. Styan October 21, 1995.

† Institute of Mathematics, Fudan University, Shanghai 200433, People's Republic of China (yzyang@ms.fudan.sh.cn).

Proof. The proof is immediate from [7, Thm. 19].

We mention here that related ideas appear in [1, 3], and [5].

In the following theorem, a characterization for the Drazin inverse A^D is presented.

THEOREM 1. *Suppose* $A \in C^{n\times n}$ *with* $Ind(A) = k$ *and* $\operatorname{rank}(A^k) = r$. *Then there exists a unique matrix* Y *such that*

$$A^kY = 0, \quad YA^k = 0, \quad Y^2 = Y, \quad \operatorname{rank}(Y) = n - r \tag{2.2}$$

and a unique matrix X *such that*

$$\operatorname{rank}\begin{pmatrix} A & I-Y \\ I-Y & X \end{pmatrix} = \operatorname{rank}(A). \tag{2.3}$$

The matrix X *is the Drazin inverse* A^D *of* A. *Further, we have*

$$Y = I - A^DA = I - AA^D. \tag{2.4}$$

Proof. To prove the first statement, let U be a nonsingular matrix for which

$$A^k = U\begin{pmatrix} J & 0 \\ 0 & 0 \end{pmatrix}U^{-1},$$

where J is a nonsingular matrix of order r. It is easy to verify that

$$Y = U\begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}U^{-1}$$

satisfies condition (2.2). To show uniqueness, let Y_0 be a matrix which satisfies (2.2). Let $Y_1 = U^{-1}Y_0U$, and let Y_1 be partitioned as

$$Y_1 = \begin{pmatrix} E & F \\ G & H \end{pmatrix},$$

with E being $r \times r$. By (2.2),

$$\begin{pmatrix} J & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} E & F \\ G & H \end{pmatrix} = 0$$

and

$$\begin{pmatrix} E & F \\ G & H \end{pmatrix}\begin{pmatrix} J & 0 \\ 0 & 0 \end{pmatrix} = 0.$$

So $E = 0$, $F = 0$, and $G = 0$. It follows that $H = I$, since Y_1 again satisfies $Y_1^2 = Y_1$ and has to have rank $n - r$. Thus, we obtain $Y_0 = Y$.

Let A^D be the Drazin inverse of A. Observe that then (2.4) holds. For this Y and

$$\begin{pmatrix} A & I-Y \\ I-Y & X \end{pmatrix} = \begin{pmatrix} A & AA^D \\ A^DA & X \end{pmatrix}.$$

Thus, by Lemma 1 and condition (2.3), we have

$$X - A^DAA^D = 0, \tag{2.5}$$

which by (1.1) implies $X = A^D$. This completes the proof. □

Notice. The projection $Y = I - A^DA$ was called the eigenprojection of A. (It is a projection on $N(A^k)$ along $R(A^k)$; cf. [11, Chap. 2].) Methods to compute the eigenprojection of a square matrix A are discussed in [12].

3. A representation of the Drazin inverse. For a nonsingular matrix A, A^{-1} can be characterized in terms of a well-known limit process,

$$A^{-1} = \lim_{\epsilon \to 0} (\epsilon I + A)^{-1}, \tag{3.1}$$

where in the limit, as $\epsilon \to 0$, of the above expression involving $(\epsilon I + A)^{-1}$, we assume that $-\epsilon \notin \sigma(A)$, the set of all eigenvalues of A. The same assumption will be used in the following.

One extension of the limit expression (3.1) to the Drazin inverse was established in [8], and the Drazin inverse A^D can be similarly characterized by

$$A^D = \lim_{\epsilon \to 0} (\epsilon I + A^{l+1})^{-1} A^l \quad \text{for every } l \geq k = Ind(A). \tag{3.2}$$

The representation of the Drazin inverse in terms of its eigenprojection $Y = I - A^D A$ was given in [13]:

$$A^D = (A - Y)^{-1}(I - Y) = (I - Y)(A - Y)^{-1}. \tag{3.3}$$

In this section, we shall give another representation of A without its eigenprojection Y.

THEOREM 2. *Let $A \in C^{n \times n}$ with $Ind(A) = k$. Then*

$$A^D = \tilde{A}^{-1} A^k, \tag{3.4}$$

where $\tilde{A} = A^{k+1}|_{R(A^k)}$ is the restriction of A^{k+1} to $R(A^k)$.

Proof. Notice that $\tilde{A} = A^{k+1}|_{R(A^k)}$ is a one-to-one map of $R(A^k)$ onto $R(A^k)$. Since $\tilde{A}x = 0$ where $x \in R(A^k)$, there exists a $y \in C^n$ such that $x = A^k y$. However,

$$A^{2k+1} y = \tilde{A} x = 0;$$

i.e.,

$$y \in N(A^{2k+1}) = N(A^k).$$

Thus,

$$x = A^k y = 0.$$

On the other hand, for every $y \in R(A^k)$ there exists an $x \in C^n$ such that $y = A^{k+1}(A^k x) \in \tilde{A} R(A^k)$ since $R(A^k) = R(A^{2k+1})$.

These indicate the nonsingularity of $\tilde{A}$.

We let $X = \tilde{A}^{-1} A^k$ and consider the decomposition of any $z \in C^n$ as $z = z_1 + z_2$ with $z_1 \in N(A^k)$ and $z_2 \in R(A^k)$. It follows that there exists a $t \in C^n$ such that $z_2 = A^{k+1} t$ since $R(A^k) = R(A^{k+1})$.

Next, we will verify that $X = \tilde{A}^{-1} A^k$ satisfies the three equations in (1.1):

$$\begin{aligned} A^{k+1} X z &= A^{k+1} \tilde{A}^{-1} A^k z = A^{k+1} \tilde{A}^{-1} A^k z_2 \\ &= A^{k+1} \tilde{A}^{-1} A^{k+1} (A^k t) = A^{k+1} A^k t \\ &= A^k z_2 = A^k z \end{aligned}$$

and

$$\begin{aligned} XAXz &= XA\tilde{A}^{-1} A^k z = XA\tilde{A}^{-1} A^k z_2 \\ &= XA\tilde{A}^{-1} A^{k+1} (A^k t) = XA^{k+1} t \\ &= \tilde{A}^{-1} A^k z_2 = \tilde{A}^{-1} A^k z = Xz. \end{aligned}$$

Finally,

$$\begin{aligned} AXz &= A\tilde{A}^{-1}A^k z = A\tilde{A}^{-1}A^k z_2 \\ &= A\tilde{A}^{-1}A^{k+1}(A^k t) = A^{k+1}t = z_2, \\ XAz &= \tilde{A}^{-1}A^{k+1}z = \tilde{A}^{-1}A^{k+1}z_2 = z_2; \end{aligned}$$

i.e.,

$$AXz = XAz.$$

The above-mentioned fact is true for the arbitrary $z \in C^n$; thus we have

$$A^{k+1}X = A^k, \quad XAX = X, \quad AX = XA,$$

which completes the proof. □

Remark. Since it is always valid that $k = Ind(A) \leq n$, we can take $k = n$ in (2.2) and (3.4).

Acknowledgments. The author would like to thank Cao Zhihao and the referee for their helpful comments on the original version of this paper.

REFERENCES

[1] A. Albert, *Conditions for positive and nonnegative definiteness in terms of pseudoinverses,* SIAM J. Appl. Math., 17 (1969), pp. 434–440.

[2] S. L. Campbell, C. D. Meyer, Jr., and N. J. Rose, *Applications of the Drazin inverse to linear systems of differential equations with singular constant coefficients,* SIAM J. Appl. Math., 31 (1976), pp. 411–425.

[3] D. Carlson, E. Haynsworth, and T. L. Markham, *A generalization of the Schur complement by means of the Moore-Penrose inverse,* SIAM J. Appl. Math., 26 (1974), pp. 169–175.

[4] M. P. Drazin, *Pseudoinverses in associative rings and semigroups,* Amer. Math. Monthly, 65 (1958), pp. 506–514.

[5] M. Fiedler and T. L. Markham, *A characterization of the Moore-Penrose inverse,* Linear Algebra Appl., 179 (1993), pp. 129–133.

[6] R. E. Hartwig and J. Levine, *Applications of the Drazin inverse to the hill cryptographic system, Part* III, Cryptologia, 5 (1981), pp. 67–77.

[7] G. Marsaglia and G. P. H. Styan, *Equalities and inequalities for ranks of matrices,* Linear and Multilinear Algebra, 2 (1974), pp. 269–292.

[8] C. D. Meyer, Jr., *Limits and the index of a square matrix,* SIAM J. Appl. Math., 26 (1974), pp. 469–478.

[9] ———, *The role of the group generalized inverse in the theory of finite Markov chains,* SIAM Rev., 17 (1975), pp. 443–464.

[10] C. D. Meyer, Jr. and R. J. Plemmons, *Convergent powers of a matrix with applications to iterative methods for singular linear systems,* SIAM J. Numer. Anal., 14 (1977), pp. 699–705.

[11] U. G. Rothblum, *Multiplicative Markov Decision Chains,* Ph.D. dissertation, Stanford University, Stanford, CA, 1974.

[12] ———, *Computation of the eigenprojection of a nonnegative matrix at its spectral radius,* Math. Programming Stud., 6 (1976), pp. 188–201.

[13] ———, *A representation of the Drazin inverse and characterization of the index,* SIAM J. Appl. Math., 31 (1976), pp. 646–648.

[14] B. Simeon, C. Fuhrer, and P. Rentrop, *The Drazin inverse in multibody system dynamics,* Numer. Math., 64 (1993), pp. 521–539.

[15] G. R. Wang, *A Cramer rule for finding the solution of a class of singular equations,* Linear Algebra Appl., 116 (1989), pp. 27–34.

[16] Y. M. Wei and G. R. Wang, *The perturbation theory for the Drazin inverse and its applications,* Linear Algebra Appl., to appear.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 4, pp. 748–762, October 1996

004

COMPUTING THE SMALLEST EIGENVALUE OF AN M-MATRIX*

XUE JUNGONG†

Abstract. A computation of the smallest eigenvalue and the corresponding eigenvector of an irreducible nonsingular M-matrix A is considered. It is shown that if the entries of A are known with high relative accuracy, the smallest eigenvalue and each component of the corresponding eigenvector will be determined to high relative accuracy. A known inverse iteration algorithm with new stopping criterion is presented to compute them. Under certain assumptions, the algorithm will have a small componentwise backward error, which is consistent with the perturbation results.

Key words. irreducible nonsingular M-matrix, backward error, componentwise perturbation

AMS subject classifications. 65F70, 15A06, 15A18

1. Introduction. In this paper we consider the problem of how to accurately compute the smallest eigenvalue and the corresponding eigenvector of an irreducible nonsingular M-matrix. Here, "smallest" is in modules sense. It is well known [4, p. 135] that this eigenvalue is a positive simple eigenvalue to which there corresponds a positive eigenvector.

First, we discuss the new perturbation theory for the smallest eigenvalue. Let A be an $n \times n$ irreducible nonsingular M-matrix and δA be a small perturbation matrix to A with $|\delta A| \leq \eta |A|$. Here, $|Q|$ denotes the matrix of entries $|Q_{i,j}|$, and $Q \leq P (Q < P)$ means $Q_{i,j} \leq P_{i,j} (Q_{i,j} < P_{i,j})$ for all i and j. For vectors, $|y|$ and $y \leq x (y < x)$ are defined in an analogous way. Let λ and λ' be the smallest eigenvalues of A and $A + \delta A$, respectively. From the standard first-order perturbation theory [12, p. 69], we have

$$\frac{|\lambda - \lambda'|}{\lambda} \leq \frac{u^T |\delta A| v}{u^T A v} \eta + O(\eta^2), \tag{1.1}$$

where $u > 0$ and $v > 0$ are the normalized left and right eigenvectors of A corresponding to λ. In this paper, we prove the following stronger result: write $A = D - N$, where D is diagonal and N has zero diagonal. Let γ denote the spectral radius of $D^{-1}N$.

Then

$$\frac{|\lambda - \lambda'|}{\lambda} \leq \frac{1 + \gamma}{1 - \gamma} \eta. \tag{1.2}$$

This error bound is independent of the angle between u and v. It is possible that $(1 + \gamma)/(1 - \gamma) \ll \kappa(A)$, where $\kappa(A) = \|A\|_2 \|A^{-1}\|_2$.

Now we consider the eigenvector. Let $v' > 0$ be the normalized right eigenvector of $A + \delta A$ corresponding to λ'. Under the assumption that A has distinct eigenvalues $\lambda, \lambda_1, \ldots, \lambda_{n-1}$, the standard perturbation theory [8, p. 346] says that

$$v' \approx v + \sum_{i=1}^{n-1} \frac{y_i^H \delta A v}{(\lambda - \lambda_i) y_i^H x_i} x_i + O(\eta^2); \tag{1.3}$$

* Received by the editors November 9, 1994; accepted for publication (in revised form) by N. J. Higham October 24, 1995.

† Institute of Mathematics, Fudan University, Shanghai 200433, People's Republic of China.

here y_i and x_i are the normalized left and right eigenvectors of A corresponding to λ_i. Formula (1.3) implies

$$\|v' - v\|_2 \leq \frac{n(n-1)\eta}{s \cdot \text{absgap}_\lambda} + O(\eta^2), \tag{1.4}$$

where

$$\text{absgap}_\lambda = \min_i |\lambda - \lambda_i| / \|A\|_2 \quad \textit{and} \quad s = \min_i |y_i^H x_i|. \tag{1.5}$$

We call absgap_λ the *absolute gap* for λ as in [3]. The error is measured in normwise sense and the bound depends on absgap_λ and the angles between the right and left eigenvectors. In this paper, even without the assumption that A has distinct eigenvalues, we prove a generally stronger result:

$$|v - v'| \leq \left[2n\left(1 + \frac{\gamma + \text{relgap}_\lambda}{(1-\gamma)\text{relgap}_\lambda}\right)\left(1 + \frac{2}{(1-\gamma)\text{relgap}_\lambda}\right)\eta + O(\eta^2)\right] v. \tag{1.6}$$

The *relative gap* relgap_λ is defined as follows: let μ_i be the smallest eigenvalue of the principal submatrix of A obtained by deleting the ith row and column. Obviously [4, p. 156] $\mu_i > \lambda$. Define

$$\text{relgap}_\lambda \equiv \max_i \frac{\mu_i - \lambda}{\lambda}. \tag{1.7}$$

Formula (1.6) gives the relative error bound for each component of the eigenvector. This bound is independent of the angle between the left and right eigenvectors, and absgap_λ is replaced by *relative gap* relgap_λ. The point is that if A has two or more tiny eigenvalues, their *absolute gap* is necessarily small, but their *relative gap* may be large.

To illustrate, consider the irreducible nonsingular M-matrix $A = D - N$, where

$$A = \begin{pmatrix} 10^{10}+1 & 0 & -1 \\ 0 & 2 & -10^3 \\ -10^{10} & -10^{-3} & 3 \end{pmatrix}, \quad D = \text{diag}(10^{10}+1, 2, 3).$$

The smallest eigenvalue is 1 and the second smallest is about 3. It is easy to show that $\kappa(A) > \sqrt{2} \cdot 10^{10}$ and $\text{absgap}_\lambda \leq \sqrt{2} \cdot 10^{-10}$. But $\text{relgap}_\lambda \approx 2$ and $\gamma \approx 0.7$. According to the new perturbation theory, this problem of computing the smallest eigenvalue is well conditioned.

In view of (1.2) and (1.6), it is desirable to have an algorithm to compute the smallest eigenvalue and each component of the corresponding eigenvector to guarantee high relative accuracy. Our algorithm is a variation of that in [6], which computes the Perron roots of irreducible and nonnegative matrices. It is based on a variant of the inverse iteration due to Noda [10], which was shown by Elsner [5] to be quadratically convergent.

For $x = (x_1, \ldots, x_n)^T > 0$, define

$$S_x = \frac{\max_i(x_i)}{\min_i(x_i)}.$$

Let y be the right Perron vector of nonnegative matrix $N(D - \lambda I)^{-1}$. If S_y is not seriously large, with proper stopping criteria our algorithm computes the eigenvalue

and the eigenvector with small componentwise backward error; i.e., the computed eigenvalue and eigenvector are exactly those of $A + E$, with E here satisfying

$$|E| \leq ((n+1)\epsilon + tol)|A|. \tag{1.8}$$

Here ϵ is the machine precision and *tol* is a small threshold. Usually we choose $tol = \epsilon^{\frac{1}{2}}$.

Any algorithm that first reduces the matrix to Hessenberg form may not compute the smallest eigenvalue to high relative accuracy, since reducing a dense matrix to Hessenberg form can't guarantee small componentwise backward error. In §6, we present an example for which the standard QR algorithm breaks down, whereas our algorithm computes the smallest eigenvalue with high accuracy.

This paper is organized as follows. Section 2 contains lemmas for M-matrices. Section 3 discusses perturbation theory for the eigenvalue. We present the perturbation theory for each component of the eigenvector and explain the observation in [6] in §4. Section 5 discusses the algorithm and gives the componentwise backward error. We present a numerical example in §6 to illustrate the stability of the proposed algorithm.

2. Lemmas for M-matrices. An $n \times n$ matrix A is called an M-matrix if it can be expressed in the form

$$A = sI - B, \quad s \geq 0,\ B \geq 0$$

with $s \geq \rho(B)$, the spectral radius of B.

First we present a basic lemma for the M-matrix. The proof can be found in [4, p. 156].

LEMMA 1. *Let B be a singular irreducible M-matrix. Then each principal submatrix of B other than B itself is a nonsingular M-matrix.*

The following result is due to Fiedler et al. and can be found in [7].

LEMMA 2. *If P is a nonnegative irreducible matrix with Perron root ρ and Perron vectors $x > 0$, $y > 0$, $Px = \rho x$, $P^T y = \rho y$, then for all positive diagonal matrices Δ, the inequality*

$$y^T P x \leq y^T \Delta P \Delta^{-1} x$$

holds, and the equality is attained if and only if Δ is a scalar matrix.

The following lemma is a consequence of Lemma 2, and it is an important tool for establishing relative perturbation theorems in the remainder of this paper.

LEMMA 3. *Let $A = D - N$ be an $n \times n$ irreducible nonsingular M-matrix, where D is diagonal and N has zero diagonal. Let $u > 0$, $v > 0$ denote left and right eigenvectors corresponding to smallest eigenvalue λ, and $\gamma = \rho(D^{-1}N)$. Then*

$$\frac{u^T N v}{u^T D v} \leq \gamma.$$

The equality is attained if and only if D is a scalar matrix.

Proof. We apply Lemma 2 to $P = tI - D + N$, which for sufficiently large t satisfies the assumptions. Hence for any diagonal Δ,

$$u^T(tI - D + N)v \leq u^T \Delta (tI - D + N)\Delta^{-1} v,$$

from which we infer that

$$u^T N v \leq u^T \Delta N \Delta^{-1} v.$$

Now choose Δ such that $v = \Delta z$, where $z > 0$ satisfies $D^{-1}Nz = \rho(D^{-1}N)z$; then

$$u^T N v \leq u^T \Delta D D^{-1} N z = \gamma u^T \Delta D z = \gamma u^T D \Delta z = \gamma u^T D v.$$

The equality is attained if and only if Δ is a scalar matrix. This means that v is also a Perron vector of $D^{-1}N$. From the equation

$$(D - N)v = \lambda v,$$

i.e.,

$$(I - D^{-1}N)v = \lambda D^{-1} v,$$

we get $D^{-1}v = \tau v$ for some constant τ. Thus D is a scalar matrix. □

Now we give the componentwise error bound for the inverse of a nonsingular irreducible M-matrix. The following lemma can be found in [13].

LEMMA 4. *Let A and γ be as in Lemma 3. Let δA be a perturbation matrix with $|\delta A| \leq \eta |A|$. If $0 < \eta < (1-\gamma)/(1+\gamma)$, then $A + \delta A$ is still a nonsingular M-matrix and*

$$|(A + \delta A)^{-1} - A^{-1}| \leq \left(\frac{n}{1-\gamma}\eta + (n-1)\eta + O(\eta^2) \right) A^{-1}.$$

In the next two sections we will use these lemmas to derive perturbation theorems for the smallest eigenvalue and the corresponding eigenvector.

3. A perturbation theorem for eigenvalues. In this section, we prove the following theorem, which gives the relative error bound for the smallest eigenvalue of an irreducible nonsingular M-matrix.

THEOREM 1. *Let A, γ be as in Lemma 3. Let δA be a perturbation of A with $|\delta A| \leq \eta |A|$ and $0 < \eta < (1-\gamma)/(1+\gamma)$. If λ and λ' are the smallest eigenvalues of A and $A + \delta A$, then*

$$\frac{|\lambda - \lambda'|}{\lambda} \leq \frac{1+\gamma}{1-\gamma}\eta.$$

Proof. Obviously

$$A - \eta|A| \leq A + \delta A \leq A + \eta|A|.$$

Since $0 < \eta < (1-\gamma)/(1+\gamma)$, $A + \delta A$ is also an irreducible nonsingular M-matrix. Let $\lambda_1(\eta)$ and $\lambda_2(\eta)$ be the smallest eigenvalues of $A + \eta|A|$ and $A - \eta|A|$. We can easily get

$$\lambda_2(\eta) \leq \lambda' \leq \lambda_1(\eta).$$

For $0 \leq t \leq \eta$, let $u(t) > 0$ and $v(t) > 0$ be left and right eigenvectors of $A + t|A|$ corresponding to $\lambda_1(t)$. We have

$$(A + t|A|)v(t) = \lambda_1(t)v(t).$$

By differentiating this equation with respect to t, we obtain

$$|A|v(t) + (A + t|A|)\dot{v}(t) = \dot{\lambda}_1(t)v(t) + \lambda_1(t)\dot{v}(t).$$

Applying $u^T(t)$ to both sides of this equation, dividing by $u^T(t)v(t)$ gives

$$\dot{\lambda}_1(t) = \frac{u^T(t)|A|v(t)}{u^T(t)v(t)}.$$

Thus

$$\frac{d}{dt}(\ln \lambda_1(t)) = \frac{\dot{\lambda}_1(t)}{\lambda_1(t)} = \frac{u^T(t)(D+N)v(t)}{u^T(t)[(1+t)D - (1-t)N]v(t)}.$$

From Lemma 3,

$$\frac{u^T(t)((1-t)N)v(t)}{u^T(t)((1+t)D)v(t)} \leq \rho(((1+t)D)^{-1}((1-t)N));$$

i.e.,

$$\frac{u^T(t)Nv(t)}{u^T(t)Dv(t)} \leq \rho(D^{-1}N) = \gamma.$$

Thus

$$\frac{d}{dt}\ln(\lambda_1(t)) \leq \frac{(1+\gamma)}{(1-\gamma) + (1+\gamma)t}. \tag{3.1}$$

Noting that $\lambda_1(0) = \lambda$ and integrating (3.1) from 0 to η yields

$$\ln \lambda_1(\eta) - \ln \lambda \leq \ln((1-\gamma) + (1+\gamma)\eta) - \ln(1-\gamma);$$

i.e.,

$$\frac{\lambda_1(\eta) - \lambda}{\lambda} \leq \frac{1+\gamma}{1-\gamma}\eta.$$

Similarly,

$$\frac{\lambda - \lambda_2(\eta)}{\lambda} \leq \frac{1+\gamma}{1-\gamma}\eta.$$

Thus

$$\frac{|\lambda' - \lambda|}{\lambda} \leq \frac{1+\gamma}{1-\gamma}\eta. \qquad \square$$

Thus the sensitivity of λ to relative perturbations in entries of A is governed by γ, independent of the condition number $\kappa(A)$ and the angle between the left and right eigenvectors.

4. Perturbation theorems for eigenvectors. In this section we discuss the sensitivity of each component of the eigenvector under the same small perturbation as in §3.

LEMMA 5. *Let $B = \Lambda - F$ be an $n \times n$ nonsingular M-matrix with smallest eigenvalue μ. Here Λ is diagonal and F has zero diagonal. Let ν be the spectral radius of $\Lambda^{-1}F$. If $\nu(\alpha)$ is the spectral radius of $(\Lambda - \alpha\mu I)^{-1}F$ with $0 \leq \alpha \leq 1$, then*

$$\nu(\alpha) \leq \frac{\nu}{1 - \alpha(1-\nu)}. \tag{4.1}$$

Proof. We only consider B irreducible. Otherwise there exists a permutation matrix P such that P^TBP is a block triangular matrix with irreducible diagonal blocks.

Let

$$f(\alpha) = \frac{\nu}{1 - \alpha(1-\nu)} - \nu(\alpha).$$

Since $\nu(0) = \nu$ and $\nu(1) = 1$, we have

$$f(0) = f(1) = 0.$$

If Λ is a scalar matrix, $f(\alpha) \equiv 0$. Otherwise $f(\alpha) = 0$ has at most n roots. Let $y(\alpha)$ and $x(\alpha)$ be the normalized positive left and right eigenvectors of $(\Lambda - \alpha\mu I)^{-1}F$. We have

$$(\Lambda - \alpha\mu I)^{-1}Fx(\alpha) = \nu(\alpha)x(\alpha). \tag{4.2}$$

We rewrite (4.2) as

$$\left(\Lambda - \frac{1}{\nu(\alpha)}F\right)x(\alpha) = \alpha\mu x(\alpha). \tag{4.3}$$

Thus $x(\alpha)$ is the right eigenvector of $\Lambda - \frac{1}{\nu(\alpha)}F$ corresponding to the smallest eigenvalue $\alpha\mu$. Similarly, we can prove $(\Lambda - \alpha\mu I)^{-1}y(\alpha)$ is the left eigenvector. From Lemma 3, we have

$$\frac{y^T(\alpha)(\Lambda - \alpha\mu I)^{-1}Fx(\alpha)}{y^T(\alpha)(\Lambda - \alpha\mu I)^{-1}\Lambda x(\alpha)} < \nu. \tag{4.4}$$

Substituting (4.2) into (4.4), we get

$$\frac{y^T(\alpha)(\Lambda - \alpha\mu I)^{-1}\Lambda x(\alpha)}{y^T(\alpha)x(\alpha)} > \frac{\nu(\alpha)}{\nu}. \tag{4.5}$$

We are now in a position to show that if there exists some $\xi \in (0, 1]$ such that $f(\xi) = 0$, i.e., $\nu(\xi) = \frac{\nu}{1-\xi(1-\nu)}$, then $\dot{f}(\xi) < 0$.

By differentiating (4.2) with respect to α and applying $y^T(\alpha)$ to both sides, we have

$$\dot{\nu}(\alpha) = \mu \cdot \nu(\alpha)\frac{y^T(\alpha)(\Lambda - \alpha\nu I)^{-1}x(\alpha)}{y^T(\alpha)x(\alpha)}.$$

Thus

$$\begin{aligned}\dot{f}(\xi) &= \frac{\nu(1-\nu)}{(1-\xi(1-\nu))^2} - \dot{\nu}(\xi) \\ &= \nu(\xi)\left(\frac{1-\nu}{1-\xi(1-\nu)} + \frac{1}{\xi} - \frac{1}{\xi} - \frac{\mu y^T(\xi)(\Lambda-\xi\mu I)^{-1}x(\xi)}{y^T(\xi)x(\xi)}\right) \\ &= \frac{v(\xi)}{\xi}\left(\frac{\nu(\xi)}{\nu} - \frac{y^T(\xi)(\Lambda-\xi\mu I)^{-1}\Lambda x(\xi)}{y^T(\xi)x(\xi)}\right) \\ &< 0.\end{aligned}$$

Since $f(1) = 0$, from the above proof $\dot{f}(1) < 0$. Assume there exists $0 < \xi_1 < 1$ such that $f(\xi_1) < 0$. From continuity and $\dot{f}(1) < 0$ there exists $\xi_1 < \xi_2 < 1$ such that $f(\xi_2) = 0$. Without loss of generality, we assume there is no other root in $(\xi_2, 1)$. From the above proof, $\dot{f}(\xi_2) < 0$. Thus $f(\xi) < 0$ for $\xi_2 \le \xi < 1$, which implies $\dot{f}(1) \ge 0$. This is in contradiction with $\dot{f}(1) < 0$. So there is no root in $(0, 1)$. From $\dot{f}(1) < 0$, we have $f(\alpha) > 0$ for all $\alpha \in (0, 1)$, which completes the proof. □

Now we present the componentwise relative error bound for the eigenvector. The *absolute gap* absgap_λ in standard perturbation theory is replaced by *relative gap* relgap_λ, defined in (1.7).

THEOREM 2. *Let $A = D - N$ be an $n \times n$ irreducible nonsingular M-matrix and δA be a small perturbation of A with $|\delta A| \le \eta |A|$. Let $v > 0$, $v' > 0$ be the normalized eigenvectors of A and $A + \delta A$ corresponding to the smallest eigenvalues λ and λ', repectively. If*

$$0 < \eta < \frac{((1-\gamma)\text{relgap}_\lambda)^2}{(2+(1-\gamma)\text{relgap}_\lambda)(2\gamma+(1+\gamma)\text{relgap}_\lambda)}, \tag{4.6}$$

then

$$|v' - v| \le \left[2n\left(1 + \frac{\gamma + \text{relgap}_\lambda}{(1-\gamma)\text{relgap}_\lambda}\right) \cdot \left(1 + \frac{2}{(1-\gamma)\text{relgap}_\lambda}\right)\eta + O(\eta^2)\right] v. \tag{4.7}$$

Proof. Partition

$$A = \begin{pmatrix} A_1 & -a \\ -b^T & \alpha \end{pmatrix} \quad \text{and} \quad A + \delta A = \begin{pmatrix} B_1 & -c \\ -d^T & \beta \end{pmatrix},$$

where A_1 and B_1 are $(n-1) \times (n-1)$. Let

$$w = \begin{pmatrix} (A_1 - \lambda I)^{-1}a \\ 1 \end{pmatrix} \quad \text{and} \quad w' = \begin{pmatrix} (B_1 - \lambda' I)^{-1}c \\ 1 \end{pmatrix}.$$

Without loss of generality, we assume the smallest eigenvalue μ of A_1 satisfies $\text{relgap}_\lambda = (\mu - \lambda)/\lambda$.

From Theorem 1,

$$|\lambda - \lambda'| \le \frac{1+\gamma}{1-\gamma}\eta\lambda.$$

The ith diagonal entries of $A_1 - \lambda I$ and $B_1 - \lambda' I$ are $A_{i,i} - \lambda$ and $A_{i,i} + \delta A_{i,i} - \lambda'$. Noting that $A_{i,i} \geq \mu$, we have

$$\frac{|(A_{i,i} - \lambda) - (A_{i,i} + \delta A_{i,i} - \lambda')|}{A_{i,i} - \lambda} \leq \frac{\eta(A_{i,i} + \frac{1+\gamma}{1-\gamma}\lambda)}{A_{i,i} - \lambda} \leq \left(1 + \frac{2}{(1-\gamma)\text{relgap}_\lambda}\right)\eta.$$

Thus

$$|(A_1 - \lambda I) - (B_1 - \lambda' I)| \leq \left(1 + \frac{2}{(1-\gamma)\text{relgap}_\lambda}\right)\eta |A_1 - \lambda I|.$$

Write $A_1 = D_1 - N_1$, where D_1 is diagonal and N_1 has zero diagonal. Obviously $\rho(D_1^{-1}N_1) \leq \gamma$.

From Lemma 5,

$$\begin{aligned}\rho[(D_1 - \lambda I)^{-1}N_1] &\leq \frac{\gamma}{1 - \lambda(1-\gamma)/\mu} \\ &= \frac{(1 + \text{relgap}_\lambda)\gamma}{\gamma + \text{relgap}_\lambda}.\end{aligned}$$

We rewrite (4.6) as

$$0 < \left(1 + \frac{2}{(1-\gamma)\text{relgap}_\lambda}\right)\eta < \frac{(1-\gamma)\text{relgap}_\lambda}{2\gamma + (1+\gamma)\text{relgap}_\lambda} = \frac{1 - \rho((D_1 - \lambda I)^{-1}N_1)}{1 + \rho((D_1 - \lambda I)^{-1}N_1)},$$

which implies that $B_1 - \lambda' I$ is still a nonsingular M-matrix. Using Lemma 4 and noting that a and c are nonnegative vectors, we have

$$|w - w'| \leq \left(n \cdot \frac{2 - \rho((D_1 - \lambda I)^{-1}N_1)}{1 - \rho((D_1 - \lambda I)^{-1}N_1)} \cdot \left(1 + \frac{2}{(1-\gamma)\text{relgap}}\right)\eta + O(\eta^2)\right) w.$$

Since $v = w/\|w\|_2$ and $v' = w'/\|w'\|_2$, it is straightforward to get (4.7).

Now we present an analogous result for the Perron vector of an irreducible nonnegative matrix. It is observed in [6] that the Perron vector can be computed with high componentwise relative accuracy, but with a lack of theoretical analysis. Here we explain this observation.

LEMMA 6. *Let $P = \Lambda + F$ be an $n \times n$ nonnegative matrix, where Λ is diagonal and F has zero diagonal. Let ρ be the spectral radius of P and $\alpha > 0$. Then*

$$\rho(((1+\alpha)\rho I - \Lambda)^{-1}F) \leq \frac{1}{1+\alpha}. \tag{4.8}$$

Proof. For $\epsilon > 0$, the matrix $(1+\alpha\epsilon)\rho I - \Lambda - F$ is a nonsingular M-matrix. Thus $(1+\alpha\epsilon)\rho I - \Lambda$ is nonsingular and

$$\rho(((1+\alpha\epsilon)\rho I - \Lambda)^{-1}F) < 1. \tag{4.9}$$

If $0 < \epsilon < 1$, we have

$$\begin{aligned}((1+\alpha)\rho I - \Lambda)^{-1}F &= \left(\frac{1+\alpha}{1+\alpha\epsilon} \cdot (1+\alpha\epsilon)\rho I - \Lambda\right)^{-1} F \\ &\leq \frac{1+\alpha\epsilon}{1+\alpha}((1+\alpha\epsilon)\rho I - \Lambda)^{-1}F.\end{aligned}$$

Combining with (4.9), we have

$$\rho(((1+\alpha)\rho I - \Lambda)^{-1}F) < \frac{1+\alpha\epsilon}{1+\alpha}.$$

Letting $\epsilon \to 0$, we obtain (4.8). □

Let $P = \Lambda + F$ be an $n \times n$ irreducible nonnegative matrix, where Λ is diagonal and F has zero diagonal. Let ρ be the spectral radius of P and ρ_i be the spectral radius of the submatrix obtained by deleting the ith row and column. Obviously $\rho > \rho_i$. Define the *relative gap*

$$\text{relgap}_\rho = \max_i \frac{\rho - \rho_i}{\rho}.$$

The following result gives the error bound for each component of the Perron vector.

THEOREM 3. *Let P be an $n \times n$ irreducible nonnegative matrix as above and let δP be a perturbation matrix to P with $|\delta P| \le \eta P$. Let $u > 0$ and $u' > 0$ be the right normalized Perron vectors of P and $P + \delta P$, respectively. If*

$$0 < \eta < \frac{(\text{relgap}_\rho)^2}{2(2 - \text{relgap}_\rho)},$$

then

$$|u - u'| \le \left[4n \cdot \frac{1 + \text{relgap}_\rho}{(\text{relgap}_\rho)^2}\eta + O(\eta^2)\right] u.$$

Proof. Let ρ' be the spectral radius of $P + \delta P$. It is shown in [6] that

$$|\rho - \rho'| \le \eta\rho.$$

Partition

$$P = \begin{pmatrix} P_1 & a \\ b^T & \alpha \end{pmatrix} \quad \text{and} \quad P + \delta P = \begin{pmatrix} Q_1 & c \\ d^T & \beta \end{pmatrix},$$

where P_1 and Q_1 are $(n-1) \times (n-1)$. Let

$$w = \begin{pmatrix} (\rho I - P_1)^{-1}a \\ 1 \end{pmatrix} \quad \text{and} \quad w' = \begin{pmatrix} (\rho' I - Q_1)^{-1}c \\ 1 \end{pmatrix}.$$

Without loss of generality, we assume that the spectral radius ρ_1 of P_1 satisfies $\text{relgap}_\rho = (\rho - \rho_1)/\rho$.

Write $P_1 = D_1 + N_1$, where D_1 is diagonal and N_1 has zero diagonal. From Lemma 6,

$$\rho((\rho I - D_1)^{-1}N_1) \le \rho_1/\rho.$$

The ith diagonal entries of $\rho I - P_1$ and $\rho' I - Q_1$ are $\rho - P_{i,i}$ and $\rho' - P_{i,i} - \delta P_{i,i}$. Noting that $\rho > \rho_1 \ge P_{i,i}$, we have

$$\frac{|(\rho - P_{i,i}) - (\rho' - P_{i,i} - \delta P_{i,i})|}{\rho - P_{i,i}} \le \frac{\eta(P_{i,i} + \rho)}{\rho - P_{i,i}} \le \frac{2}{\text{relgap}_\rho}\eta.$$

Thus

$$|(\rho I - P_1) - (\rho' I - Q_1)| \le \frac{2}{\text{relgap}_\rho}\eta|\rho I - P_1|.$$

The rest of the proof is similar to that of Theorem 2. □

Because this error bound depends only on relgap_ρ, it is stronger than that for an M-matrix.

5. The algorithm. Let A be a nonsingular irreducible M-matrix, λ be the smallest eigenvalue of A, and $v > 0$ be the corresponding eigenvector with $\|v\|_\infty = 1$. In this section we present an algorithm that computes λ and v accurately.

For any two n vectors $x = (x_1, \ldots, x_n)^T$ and $y = (y_1, \ldots, y_n)^T$, $y > 0$, we define

$$\max\left(\frac{x}{y}\right) = \max_{1\le i\le n} \frac{x_i}{y_i}, \quad \min\left(\frac{x}{y}\right) = \min_{1\le i\le n} \frac{x_i}{y_i}$$

and

$$\operatorname{osc}\left(\frac{x}{y}\right) = \max\left(\frac{x}{y}\right) - \min\left(\frac{x}{y}\right).$$

In [10], Noda provided an inverse iteration for computing the Perron root of an irreducible nonnegative matrix. This algorithm was shown to be quadratically convergent by Elsner [5]. Now we modify it to compute the smallest eigenvalue of a nonsingular irreducible M-matrix.

INVERSE ITERATION ALGORITHM. For a given $v_0 > 0$ iteratively define

$$\begin{aligned} \lambda_s &= \min\left(\frac{Av_s}{v_s}\right), \\ w_s &= (A - \lambda_s I)^{-1} v_s, \\ v_{s+1} &= \frac{v_s}{\|v_s\|_\infty}. \end{aligned}$$

Then $\lambda_s \le \lambda_{s+1} \le \lambda$ and $\lambda - \lambda_{s+1} \le c(\lambda - \lambda_s)^2$, where c is a constant depending on v_0 and A.

This inverse iteration algorithm is the basis for our algorithm. The main task in each step is to solve the linear equations $(A-\lambda_s I)w_s = v_s$. If v_0 is not a scalar multiple of v, $A - \lambda_s I$ is an M-matrix. The algorithm due to Ahac and Olesky [1] can be used to compute the solution. It is shown in [9] that if A is tridiagonal, this algorithm has small componentwise backward error. If A is general, as recommended by Skeel [11], we add one step of iterative refinement to get small componentwise backward error. To prevent cancellation in computing λ_s, we get λ_s from λ_{s-1} following the relation

$$\lambda_s = \min\left(\frac{Av_s}{v_s}\right) = \min\left(\frac{Aw_{s-1}}{w_{s-1}}\right) = \lambda_{s-1} + \min\left(\frac{v_{s-1}}{w_{s-1}}\right).$$

Thus we can formulate our algorithm as follows.

ALGORITHM. Let *tol* be a small threshold and ϵ be the machine precision. Start with $v_0 > 0$ and $\lambda_0 = \min(Av_0/v_0)$. For $s = 0, 1, \ldots$

1. Compute the LU factorization

$$(A - \lambda_s I) = L_s U_s$$

and solve for w_s,

$$L_s U_s w_s = v_s,$$

by the Ahac and Olesky algorithm; save the LU factors.

2. Compute $r = (A - \lambda_s I)w_s - v_s$.

3. Solve $(A-\lambda_s I)d = r$ using the saved LU factors L_s and U_s.
4. Update $\overline{w}_s = w_s - d$.
5. Compute

$$\lambda_{s+1} = \lambda_s + \min\left(\frac{v_s}{\overline{w}_s}\right) \quad \text{and} \quad v_{s+1} = \frac{\overline{w}_s}{\|\overline{w}_s\|_\infty}.$$

6. Proceed until

$$\operatorname{osc}\left(\frac{v_s}{v_{s+1}}\right)\frac{1}{\|\overline{w}_s\|_\infty}\frac{1}{\min A_{i,i}} \le tol.$$

Denote $\widehat{\kappa}(A) = \||A||A^{-1}|\|_\infty$ and $\sigma(A,x) = \max(|A||x|)/\min(|A||x|)$. According to Skeel's theorem [11], which is stated in a simpler form in [2], there exists a function $f(A-\lambda_s I, \overline{w}_s)$. Under the assumption

$$\widehat{\kappa}(A-\lambda_s I)\sigma(A-\lambda_s I, \overline{w}_s) \le (f(A-\lambda_s I, \overline{w}_s)\epsilon)^{-1}, \tag{5.1}$$

the iterative refinement in Steps 2, 3, and 4 can solve $(A-\lambda_s I)w_s = v_s$ with the componentwise backward error no more than $(n+1)\epsilon$. In [11], Skeel states that although $f(A-\lambda_s I, \overline{w}_s)$ typically behaves as $O(n)$, it can grow exponentially. However, in our algorithm $f(A-\lambda_s I, \overline{w}_s)$ must grow modestly since we use the Ahac and Olesky algorithm to compute the LU factorization. Now we explain this.

From Skeel's theorem, the function $f(A-\lambda_s I, \overline{w}_s)$ depends on $\|C\|_\infty$, where C is a nonnegative matrix satisfying

$$|r| \le \epsilon \cdot C|A-\lambda_s I||w_s|. \tag{5.2}$$

From the error analysis in [8, p. 115], the computed solution w_s satisfies

$$(A-\lambda_s I + E)w_s = v_s,$$

where

$$|E| \le n\epsilon(3|A-\lambda_s I| + 5\widehat{P}|\widehat{L}_s||\widehat{U}_s|\widehat{P}^T) + O(\epsilon^2). \tag{5.3}$$

Here, $\widehat{P}$, $\widehat{L}_s$, and $\widehat{U}_s$ are the computed analogs of P, L_s, and U_s. We denote by $f^{(j)}$ and $g^{(j)}$ the jth columns of the matrices $A-\lambda_s I$ and $|\widehat{U}_s|\widehat{P}$, respectively. The algorithm of Ahac and Olesky guarantees that

$$\frac{\|f^{(j)}\|_\infty}{\|g^{(j)}\|_\infty} \le n-1. \tag{5.4}$$

Noting that $\widehat{P}$ is a permutation matrix and each entry of $|\widehat{L}_s|$ is no more than 1, we can easily construct a matrix C_1 with $\|C_1\|_\infty = O(n^3)$ to rewrite (5.3) as

$$|E| \le \epsilon C_1|A-\lambda_s I|, \tag{5.5}$$

which implies that

$$|r| \le \epsilon C_1|A-\lambda_s I||w_s|. \tag{5.6}$$

Thus the function $f(A-\lambda_s I, \overline{w}_s)$ grows modestly.

Now we explain the stopping criteria in Step 6. In fact, this stopping criteria can be written as

$$\frac{\lambda_{s+1} - \lambda_s}{\min A_{i,i}} \leq tol.$$

In the following, variables computed by the algorithm are denoted by hats.

THEOREM 4. *Let* $\epsilon_1 = \|\widehat{\overline{w}}_s\|_\infty^{-1}$ *and* $\epsilon_2 = \mathrm{osc}(\widehat{v}_s/\widehat{v}_{s+1})$. *Suppose that the algorithm terminates when* $\epsilon_1\epsilon_2/\min A_{i,i} \leq tol$. *Suppose further that* $\epsilon_2 > (n+1)\epsilon$, $(n+1)tol < 1$, *and* $\widehat{\kappa}(A - \widehat{\lambda}_s I)\sigma(A - \widehat{\lambda}_s I, \widehat{\overline{w}}_s) \leq (f(A - \widehat{\lambda}_s I, \widehat{\overline{w}}_s)\epsilon)^{-1}$. *Then* $\widehat{\lambda}_s + \epsilon_1$ *and* $\widehat{v}_{s+1}$ *are, respectively, the exact smallest eigenvalue and its corresponding eigenvector of matrix* $\overline{A}$, *where*

$$|\overline{A} - A| \leq [(n+2)\epsilon + 2tol]|A|. \tag{5.7}$$

Proof. It follows from Skeel's theorem that

$$(A - \widehat{\lambda}_s I - E)\widehat{\overline{w}}_s = \widehat{v}_s + \Delta v, \tag{5.8}$$

where

$$|E| \leq (n+1)\epsilon|A - \widehat{\lambda}_s I| \quad \text{and} \quad |\Delta v| \leq (n+1)\epsilon \cdot \widehat{v}_s.$$

Dividing by $\|\widehat{\overline{w}}_s\|_\infty$, we can write (5.3) as

$$(A - \widehat{\lambda}_s I - E)\widehat{v}_{s+1} = \epsilon_1(I + D_\epsilon)\widehat{v}_s, \tag{5.9}$$

where D_ϵ is a diagonal matrix such that $|f_i| \leq (n+1)\epsilon$ with $D_\epsilon(i,i) = f_i$. Noting that the infinity norms of $\widehat{v}_{s+1}$ and $\widehat{v}_s$ equal 1 and that they are both positive, we can get the following inequality as in [6]:

$$(1 - \epsilon_2)\widehat{v}_{s+1} \leq \widehat{v}_s \leq (1 + \epsilon_2)\widehat{v}_{s+1}.$$

Hence we can write $\widehat{v}_s = (I + D_{\epsilon_2})\widehat{v}_{s+1}$, where D_{ϵ_2} is a diagonal matrix such that $|h_i| \leq \epsilon_2$ with $D_{\epsilon_2}(i,i) = h_i$.

Substituting into (5.4) yields

$$(A - \widehat{\lambda}_s I - E)\widehat{v}_{s+1} = \epsilon_1(I + D_\epsilon)(I + D_{\epsilon_2})\widehat{v}_{s+1}.$$

This can be rewritten as

$$(A_{i,i} - E_{i,i} - \epsilon_1(1+f_i)(1+h_i) - \widehat{\lambda}_s)\widehat{v}_{s+1}(i) = \sum_{j=1, j\neq i}^{n} (-A_{i,j} + E_{i,j})\widehat{v}_{s+1}(j).$$

Using the condition $\epsilon_2 > (n+1)\epsilon$ and $(n+1)tol < 1$, we get

$$\begin{aligned} \frac{|E_{i,i} + \epsilon_1(f_i + h_i) + \epsilon_1 f_i h_i|}{A_{i,i}} &\leq \frac{1}{A_{i,i}}(|E_{i,i}| + 2\epsilon_1\epsilon_2 + (n+1)\epsilon_1\epsilon_2\epsilon) \\ &\leq (n+2)\epsilon + 2tol, \end{aligned}$$

which completes the proof. □

From Theorem 1, we have

$$\frac{\lambda - (\widehat{\lambda}_s + \epsilon_1)}{\lambda} \leq \frac{1+\gamma}{1-\gamma}((n+1)\epsilon + 2tol).$$

The relative precision of the computed eigenvalue depends on *tol*.

Now we comment on condition (5.1). First we give the upper bounds for $\widehat{\kappa}(A-\widehat{\lambda}_s I)$ and $\sigma(A - \widehat{\lambda}_s I, \widehat{v}_{s+1})$.

THEOREM 5. *Let A, D, N, γ, and λ be as in Theorem* 1. *Let $\widehat{v}_{s+1}$ and $\widehat{\lambda}_s$ denote the quantities computed in the algorithm for some value of s. For $x > 0$, define*

$$S_x = \frac{\max(x)}{\min(x)}.$$

Let $y_s > 0$ with $\|y_s\|_\infty = 1$ be the right Perron vector of $N(D - \widehat{\lambda}_s I)^{-1}$. Let $x_s = (D - \widehat{\lambda}_s I)\widehat{v}_{s+1}/\|(D - \widehat{\lambda}_s I)\widehat{v}_{s+1}\|_\infty$. Suppose $|x_s - y_s| \leq \eta y_s$. If $0 < \eta < 1$ and $\widehat{\lambda_s} > 0$, then

$$\frac{\max(|A - \widehat{\lambda}_s I|\widehat{v}_{s+1})}{\min(|A - \widehat{\lambda}_s I|\widehat{v}_{s+1})} \leq \frac{1+\eta}{1-\eta} S_{y_s}, \tag{5.10}$$

$$\||A - \widehat{\lambda}_s I|(A - \widehat{\lambda}_s I)^{-1}\|_\infty \leq \left(1 + \frac{2\lambda}{\lambda - \widehat{\lambda}_s} \cdot \frac{\gamma}{1-\gamma}\right) S_{y_s}. \tag{5.11}$$

Proof. Observing that $A_{i,i} > \lambda > \widehat{\lambda}_s$, we can get

$$|A - \widehat{\lambda}_s I| = D - \widehat{\lambda}_s I + N.$$

Let ρ_1 be the spectral radius of $N(D - \widehat{\lambda}_s I)^{-1}$. From Lemma 5,

$$\rho_1 \leq \frac{\gamma}{1 - (1-\gamma)\widehat{\lambda}_s/\lambda}.$$

Thus

$$|A - \widehat{\lambda}_s I|\widehat{v}_{s+1} = (I + N(D - \widehat{\lambda}_s I)^{-1})(D - \widehat{\lambda}_s I)\widehat{v}_{s+1} \leq c(1+\eta)(1+\rho_1)y_s,$$

where

$$c = \|(D - \widehat{\lambda}_s I)\widehat{v}_{s+1}\|_\infty.$$

Similarly,

$$|A - \widehat{\lambda}_s I|\widehat{v}_{s+1} \geq c(1-\eta)(1+\rho_1)y_s.$$

Combining the above two inequalities, we prove (5.7). We have

$$|A - \widehat{\lambda}_s I|(A - \widehat{\lambda}_s I)^{-1} = I + 2N(D - \widehat{\lambda}_s I)^{-1}(I - N(D - \widehat{\lambda}_s I)^{-1})^{-1}.$$

Let ρ_2 be the spectral radius of $|A - \widehat{\lambda}_s I|(A - \widehat{\lambda}_s I)^{-1}$. Then

$$\rho_2 = \frac{1+\rho_1}{1-\rho_1} \leq 1 + \frac{2\lambda}{\lambda - \widehat{\lambda}_s}\frac{\gamma}{1-\gamma}.$$

Obviously y_s is the right Perron vector of $|A - \widehat{\lambda}_s I|(A - \widehat{\lambda}_s I)^{-1}$. Let e be the vector of all ones and $D_{y_s} = \mathrm{diag}(y_s(1), \ldots, y_s(n))$. We have

$$D_{y_s}^{-1}|A - \widehat{\lambda}_s I|(A - \widehat{\lambda}_s I)^{-1} D_{y_s} e \leq \rho_2 e.$$

Using the definition of S_{y_s}, we have

$$\||A - \widehat{\lambda}_s I|(A - \widehat{\lambda}_s I)^{-1}\|_1 \leq \left(1 + \frac{2\lambda}{\lambda - \widehat{\lambda}_s} \cdot \frac{\gamma}{1-\gamma}\right) y_s.$$

Using the definition of S_{y_s}, we prove (5.11). □

If the algorithm converges, then $\widehat{\lambda}_s$ converges to λ and $\widehat{v}_{s+1}$ converges to v in componentwise sense. Thus x_s and y_s converge to y, the Perron vector of $N(D-\lambda I)^{-1}$, in componentwise sense. For sufficiently large s, the assumptions in Theorem 5 can be satisfied.

It is stated in [6] that whether condition (5.1) holds depends mainly on S_v and the absolute accuracy of the computed eigenvalue $\lambda - \widehat{\lambda}_s$, whereas in our analysis it depends on S_y and $(\lambda - \widehat{\lambda}_s)/\lambda$, the relative accuracy of the computed eigenvalue. It is possible that $S_y << S_v$ and $\lambda - \widehat{\lambda}_s << (\lambda - \widehat{\lambda}_s)/\lambda$. The following 2×2 matrix illustrates this:

$$A = \begin{pmatrix} 10^8 + 1 & -10^2 \\ -10^6 & 2 \end{pmatrix}.$$

If S_v is sufficiently large, or $\lambda - \widehat{\lambda}_s$ is small but still does not achieve the required accuracy, (5.1) doesn't hold according to the analysis in [6], while it may still hold according to our analysis.

Finally, we now discuss how to choose *tol*. The smaller *tol* is, the more accurate the computed solution is. On the other hand, *tol* should be chosen a little large to make the algorithm converge in finite precision. We recommend choosing $tol = O(\epsilon^{\frac{1}{2}})$. We have tested this threshold on many matrices and have found that the algorithm always converges. The interesting thing is that the relative accuracy of $\widehat{\lambda}_s + \epsilon$ and $\widehat{v}_{s+1}$ is near to the machine precision ϵ, whereas Theorem 4 predicts the relative accuracy to be $O(\epsilon^{\frac{1}{2}})$. The observation is akin to that in [6]. Unfortunately, we can't explain this phenomenon.

6. A numerical example. In this section, we give a typical numerical example from our experiment to illustrate the stability of our algorithm. Let A be the 7×7 matrix

$$A = \begin{pmatrix} 2 & & & & & & -10^{-6} \\ -10 & 10^3+1 & & & & & \\ & -10^4 & 10^6+1 & & & & \\ & & -10^7 & 10^9+1 & & & \\ & & & -10^{10} & 10^{12}+1 & & \\ & & & & -10^{13} & 10^{15}+1 & \\ & & & & & -10^{16} & 2 \end{pmatrix}.$$

It is easy to see that $\lambda = 1$ and $v = (10^{-6}, 10^{-8}, 10^{-10}, 10^{-12}, 10^{-14}, 10^{-16}, 1)^T$. We can estimate that $\kappa(A) > 10^{16}$ and $\mathrm{absgap}_\lambda < 5 \cdot 10^{-15}$. But $1 - \gamma \approx 0.18$ and $\mathrm{relgap}_\lambda \geq 1$, and according to the new perturbation theory, this eigenproblem is not

TABLE 1

tol	λ_s	$\lambda_s + \epsilon_1$	$\max \lvert v_s(i) - v(i) \rvert / v(i)$
10^{-1}	0.604737	0.944033	5.6×10^{-2}
10^{-3}	0.998518	0.999999	1.6×10^{-6}
10^{-5}	0.998518	0.999999	1.0×10^{-6}
10^{-7}	0.999999	1.000000	1.0×10^{-7}

ill conditioned. Choosing $v_0 = (1, 1, 1, 1, 1, 1, 1)$ in Table 1, we give the numerical results. The experiment is performed with computer precision $\epsilon \approx 10^{-7}$ using an IBM 486. The iteration is stopped when $\epsilon_1 \epsilon_2 / \min A_{i,i} \leq tol$. It takes 33 steps to get the result on the first line, 35 steps on the second line and third line, and 36 steps on the fourth line. Our algorithm computes λ accurately, while the standard QR algorithm of MATLAB (the function eig(A)) gives the result 0.7567.

Acknowledgment. The author thanks his advisor, Prof. Jiang Erxiong, for his careful reading of this paper and helpful comments. The original proof of Lemma 3 in the draft was long and complicated, and the short one in this paper is due to Prof. Elsner. The author expresses his sincere gratitude to him.

REFERENCES

[1] A. A. AHAC AND D. D. OLESKY, *A stable method for the LU factorization of M-matrices*, SIAM J. Alg. Disc. Meth., 7 (1986), pp. 368–378.

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

[3] J. BARLOW AND J. DEMMEL, *Computing accurate eigensystems of scaled diagonally dominant matrices*, SIAM J. Numer. Anal., 27 (1990), pp. 762–791.

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

[5] L. ELSNER, *Inverse iteration for calculating the spectral radius of a nonnegative irreducible matrix*, Linear Algebra Appl., 15 (1976), pp. 235–242.

[6] L. ELSNER, I. KOLTRACHT, M. NEUMANN, AND D. XIAO, *On accurate computations of the Perron root*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 456–467.

[7] M. FIEDLER, C. R. JOHNSON, T. L. MARKHAM, AND M. NEUMANN, *A trace inequality for M-matrices and the symmetriczability of a real matrix by a positive diagonal matrix*, Linear Algebra Appl., 71 (1985), pp. 81–94.

[8] G. GOLUB AND C. VAN LOAN, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

[9] N. J. HIGHAM, *Bounding the error in Gaussian elimination for tridiagonal systems*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 521–530.

[10] T. NODA, *Note on the computation of the maximal eigenvalue of a nonnegative irreducible matrix*, Numer. Math., 17 (1971), pp. 382–386.

[11] R. D. SKEEL, *Iterative refinememt implies numerical stability for Gaussian elimination*, Math. Comput., 35 (1980), pp. 817–832.

[12] J. H. WILKINSON, *The Algebraic Eigenvalue Problems*, Oxford University Press, London, UK, 1965.

[13] X. JUNGONG AND J. ERXIONG, *Entrywise relative perturbation theory for nonsingular M-matrices*, BIT, 35 (1995), pp. 417–427.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 4, pp. 763–788, October 1996

005

ON LINEAR LEAST-SQUARES PROBLEMS WITH DIAGONALLY DOMINANT WEIGHT MATRICES*

ANDERS FORSGREN†

Abstract. The solution of the unconstrained weighted linear least-squares problem is known to be a convex combination of the *basic solutions* formed by the nonsingular subsystems if the weight matrix is diagonal and positive definite. In particular, this implies that the norm of this solution is uniformly bounded for any diagonal and positive definite weight matrix. In addition, the solution set is known to be the relative interior of a finite set of polytopes if the weight matrix varies over the set of positive definite diagonal matrices. In this paper, these results are reviewed and generalized to the set of weight matrices that are symmetric, positive semidefinite, and diagonally dominant and that give unique solution to the least-squares problem. This is done by means of a particular symmetric diagonal decomposition of the weight matrix, giving a finite number of diagonally weighted problems but in a space of higher dimension. Extensions to equality-constrained weighted linear least-squares problems are given. A discussion of why the boundedness properties do not hold for general symmetric positive definite weight matrices is given. The motivation for this research is from interior methods for optimization.

Key words. unconstrained linear least-squares problem, weighted linear least-squares problem, equality-constrained linear least-squares problem

AMS subject classifications. 65F20, 65F35, 65K05

1. Introduction.

A fundamental problem in linear algebra is the linear least-squares problem; see, e.g., Lawson and Hanson [19], Golub and Van Loan [13, Chap. 5], and Gill, Murray, and Wright [11, Chap. 6]. In this paper, we consider the *weighted* linear least-squares problem

$$\underset{\pi\in\mathbb{R}^m}{\text{minimize}} \qquad \|W^{1/2}(A^T\pi - g)\|_2^2, \tag{1.1}$$

where A is an $m \times n$ matrix of full row rank and W is a positive definite symmetric $n \times n$ matrix. (Here, $W^{1/2}$ denotes the matrix square root of W; see, e.g., Golub and Van Loan [13, p. 554].) In many cases, W is diagonal, but it is also of interest to consider the case in which W is a general symmetric positive definite matrix. A motivation for this is given in §1.1. An individual problem of the form (1.1) can be converted to an unweighted problem by substituting $\tilde{A} = AW^{1/2}$ and $\tilde{g} = W^{1/2}g$. However, our interest is in *sequences* of weighted problems, where the weight matrix W changes and A is constant (see §1.1). In this situation, the weight matrix is of importance.

The solution of (1.1) is given by the *normal equation*

$$AWA^T\pi = AWg \tag{1.2}$$

or alternatively as the solution to the *augmented system* (or *Karush–Kuhn–Tucker KKT system*)

$$\begin{pmatrix} W^{-1} & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} r \\ \pi \end{pmatrix} = \begin{pmatrix} g \\ 0 \end{pmatrix}. \tag{1.3}$$

* Received by the editors April 3, 1995; accepted for publication (in revised form) by D. P. O'Leary November 3, 1995. This research was supported by the Swedish Natural Science Research Council (NFR).

† Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology, S-100 44 Stockholm, Sweden (andersf@math.kth.se).

There are a large number of papers giving reasons for solving systems of the type (1.2) or (1.3), starting with Bartels, Golub, and Saunders [1]. Although our intention is ultimately to focus on computational aspects, in this paper we deal with the linear algebra. For recent papers with discussions on computational aspects, see, e.g., Duff et al. [9], Björck [5], Gulliksson and Wedin [16], Wright [27, 28], Björck and Paige [6], Vavasis [24], Forsgren, Gill, and Shinnerl [10], and Gill, Saunders, and Shinnerl [12].

This paper concerns properties of the solution to (1.1) when W varies over a particular set of symmetric matrices. Note that if W is positive definite and symmetric, the solution of (1.1) given by (1.2) is unique. If W is positive semidefinite and symmetric, then the solution is unique if and only if AWA^T is nonsingular, i.e., positive definite. We shall consider sets of matrices of these two types, and, in either case, the solution of (1.1) is unique and given by

$$\pi = (AWA^T)^{-1}AWg. \tag{1.4}$$

The characterization of π when W varies over the set of diagonal and positive definite matrices is known, and these results are reviewed in §2, together with the closely related case when W varies over the set of matrices for which W is diagonal and positive semidefinite and AWA^T is positive definite. Section 3 gives an expression for π of (1.4) when W is a general symmetric matrix such that AWA^T is nonsingular by means of diagonal decomposition. In §4, a particular diagonal decomposition, called the *signature decomposition*, is presented. This decomposition is the key to the analysis of §5, where the characterization of π for W belonging to the set of diagonal positive definite weight matrices is extended to the case in which W belongs to the set of matrices for which W is symmetric, positive semidefinite, and diagonally dominant, and AWA^T is positive definite. In essence, the analysis for diagonal weight matrices can still be applied but in a space of higher dimension. In §6, the signature decomposition is used to generalize the results to the case of *equality-constrained* linear least-squares problems, i.e., the case in which infinite diagonal weight is put on some constraints. Finally, a concluding discussion is given in §7. The signature decomposition here gives some insight into why the properties of π for positive definite diagonal weight matrices do not hold for general positive definite symmetric weight matrices.

1.1. Motivation. Our interest in this problem is from interior methods for optimization. There are a vast number of papers on interior methods; here we give only a brief motivation for the weighted linear least-squares problems that arise. For detailed discussions of the summary given here and an overview of interior methods, see, e.g., Gonzaga [14] and Wright [26]. We derive the *primal* barrier equations associated with a linear programming problem in standard form

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{minimize}} & c^Tx \\ \text{subject to} & Ax = b, \\ & x \geq 0. \end{array} \tag{1.5}$$

We assume that A is an $m \times n$ matrix of full row rank and that the *relative interior* of the feasible region is nonempty; i.e., there is an $\bar{x} \in \mathbb{R}^n$ such that $A\bar{x} = b$ and $\bar{x} > 0$. For a positive *barrier parameter* μ, the associated barrier subproblem is

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{minimize}} & c^Tx - \mu \sum_{i=1}^n \ln x_i \\ \text{subject to} & Ax = b. \end{array} \tag{1.6}$$

The optimality conditions for (1.6) are given by

$$c - \mu X^{-1}e - A^T\pi = 0, \tag{1.7a}$$

$$Ax - b = 0, \tag{1.7b}$$

with $X = \operatorname{diag}(x)$ and e the n-dimensional vector with all components one. If $x(\mu)$ denotes the minimizer of (1.6), then under mild conditions $\lim_{\mu\to 0^+} x(\mu) = x^*$, where x^* is a minimizer of (1.5). The Newton equations associated with (1.7) are given by

$$\begin{pmatrix} \mu X^{-2} & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} -\Delta x \\ \Delta\pi \end{pmatrix} = \begin{pmatrix} c - \mu X^{-1}e - A^T\pi \\ Ax - b \end{pmatrix}. \tag{1.8}$$

If x is strictly feasible (i.e., x is positive and satisfies $Ax = b$), then a comparison of (1.3) and (1.8) shows that the Newton equations (1.8) can be associated with a weighted linear least-squares problem with a diagonal and positive definite weight matrix $(1/\mu)X^2$. For the case of convex quadratic programming, where the objective function of (1.5) is changed to

$$\tfrac{1}{2}x^THx + c^Tx,$$

for a positive semidefinite and symmetric H, the weight matrix in (1.8) is altered to $(H+\mu X^{-2})^{-1}$, and it is not diagonal in general. A sequence of strictly feasible iterates $\{x_k\}_{k=0}^{\infty}$ gives rise to a sequence of weighted linear least-squares problems where the weight matrix changes but A is constant. Moreover, if $\lim_{k\to\infty} x_k = x^*$, the condition numbers of the corresponding weight matrices tend to infinity in general.

Our motivation for considering extensions to nondiagonal weight matrices is primarily twofold: (i) to investigate what perturbations can be made to diagonal weight matrices and the properties of the solution still remain the same and (ii) to see if more general weight matrices can be considered using the same analysis.

1.2. Notation. When referring to vector norms and matrix norms, when we make no explicit reference to what type of norm is considered, it can be any vector norm and associated subordinate matrix norm such that $\|(x^T\ 0)^T\| = \|x\|$ holds for any vector x. This condition on the norm is made to make the notation convenient, and it is not strictly necessary. Any frequently used matrix norm satisfies this condition; see, e.g., Golub and Van Loan [13, p. 57]. For a matrix M, we denote by $|M|$ the matrix whose components are the absolute values of the components of M. The identity matrix of appropriate dimension is denoted by I, and its ith column is denoted by e_i. For a vector r, we denote by r_i its ith component, and for a matrix U, we denote by u_i its ith column. For a diagonal matrix D, a slightly different notation is used, and d_i denotes the ith diagonal element. In §6, partitioned vectors and matrices are considered. A partition of a vector r as $r^T = (r_1^T\ r_2^T)$, and similarly for matrices, means the partition of r into two blocks of a specified size. The meaning will be clear from the context.

The signature operator is used frequently. For a vector r, when we write $\operatorname{sign}(r)$, we mean the vector s of the same dimension as r with components $s_i = 1$ if $r_i > 0$, $s_i = 0$ if $r_i = 0$, and $s_i = -1$ if $r_i < 0$. We refer to a vector s with all components -1, 0, or 1 as a *signature vector*, and if all components of s are nonzero, i.e., -1 or 1, we refer to it as a *dense signature vector*. Similarly, for a vector r, when we write $\operatorname{sign}^+(r)$, we mean the vector s of the same dimension as r with components $s_i = 1$

if $r_i \geq 0$ and $s_i = -1$ if $r_i < 0$. Hence, for any vector r, $\text{sign}^+(r)$ is a dense signature vector.

For an $m \times n$ matrix A of full row rank, we shall be interested in its nonsingular $m \times m$ submatrices. Let $\mathcal{J}$ denote the $\binom{n}{m}$ subsets of $\{1, 2, \ldots, n\}$ that have cardinality m. For $J \in \mathcal{J}$, we denote by A_J the $m \times m$ submatrix of A comprising the columns of A with indices in J. We denote by $\mathcal{J}(A)$ the family of such sets of column indices associated with the nonsingular $m \times m$ submatrices of A; i.e., $\mathcal{J}(A) = \{J \in \mathcal{J} : A_J \text{ is nonsingular}\}$. For example, if

$$A = \begin{pmatrix} 1 & -1 & 2 \\ 0 & 1 & -2 \end{pmatrix},$$

then $\mathcal{J}(A) = \{\{1,2\}, \{1,3\}\}$, and associated with $\mathcal{J}(A)$ we have

$$A_{\{1,2\}} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad A_{\{1,3\}} = \begin{pmatrix} 1 & 2 \\ 0 & -2 \end{pmatrix}.$$

Associated with $J \in \mathcal{J}(A)$, for a diagonal $n \times n$ matrix D, we denote by D_J the $m \times m$ diagonal matrix formed by the elements of D that have row and column indices in J. Similarly, for a vector g of dimension n, we denote by g_J the vector of dimension m with the components of g that have indices in J. For example, with $\mathcal{J}(A)$ as above, if

$$D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \quad \text{and} \quad g = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix},$$

then

$$D_{\{1,2\}} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \; D_{\{1,3\}} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}, \; g_{\{1,2\}} = \begin{pmatrix} 4 \\ 5 \end{pmatrix}, \quad \text{and} \quad g_{\{1,3\}} = \begin{pmatrix} 4 \\ 6 \end{pmatrix}.$$

The slightly different meanings of A_J, D_J, and g_J are used in order not to make the notation more complicated than necessary. The analogous notation is used for an $m \times n$ matrix A of full row rank and an $n \times r$ matrix U of full row rank in that we associate $\mathcal{J}(AU)$ with the set of column indices corresponding to nonsingular $m \times m$ submatrices of AU. Associated with $J \in \mathcal{J}(AU)$, for a diagonal $r \times r$ matrix D, we denote by D_J the $m \times m$ diagonal matrix formed by the elements of D that have row and column indices in J. Similarly, for a vector g of dimension r, we denote by g_J the vector of dimension m with the components of g that have indices in J. Since column indices of AU are also column indices of U, for $J \in \mathcal{J}(AU)$, we denote by U_J the $n \times m$ submatrix of full column rank formed by the columns of U with indices in J. Note that each element of $\mathcal{J}(A)$ as well as each element of $\mathcal{J}(AU)$ is a collection of m indices.

2. Review of results for diagonal weight matrices. The case in which W is a diagonal positive definite matrix has been considered independently by several authors. In this section, some of these results are reviewed. To stress that the weight matrix is diagonal, we replace W by D. We consider two sets of $n \times n$ diagonal weight

matrices associated with an $m \times n$ matrix A of full row rank. The first set is $\mathcal{D}_+(A)$, defined by

$$\mathcal{D}_+(A) = \{D \in \mathbb{R}^{n\times n} : D \text{ is diagonal and positive definite}\}, \tag{2.1}$$

and the second is $\mathcal{D}_0(A)$, defined by

$$\mathcal{D}_0(A) = \left\{ D \in \mathbb{R}^{n\times n} : \begin{array}{l} D \text{ is diagonal and positive semidefinite,} \\ ADA^T \text{ is positive definite} \end{array} \right\}. \tag{2.2}$$

Note that $\mathcal{D}_+(A) \subseteq \mathcal{D}_0(A)$, and both $\mathcal{D}_+(A)$ and $\mathcal{D}_0(A)$ are convex sets in which ADA^T is positive definite. The reason we stress that $\mathcal{D}_+(A)$ and $\mathcal{D}_0(A)$ are associated with A is that the dimensions of the diagonal matrices in $\mathcal{D}_+(A)$ and $\mathcal{D}_0(A)$ are given by the number of columns of A. In §5, the results reviewed in this section are used, but with the matrix A replaced by $AU(s)$, where $U(s)$ is a matrix of full row rank with dimension $n \times \frac{n(n+1)}{2}$. In this circumstance, the number of columns of $AU(s)$ determines the size of matrices in $\mathcal{D}_+(AU(s))$ and $\mathcal{D}_0(AU(s))$, meaning that they are diagonal matrices of dimension $\frac{n(n+1)}{2} \times \frac{n(n+1)}{2}$.

The following theorem, which states that the diagonally weighted linear least-squares solution can be written as a certain convex combination, has been given independently by several authors. To the best of our knowledge, Dikin [8, p. 55] was the first to state this result in the convergence analysis of the interior method for linear programming he proposed [7]; see Vanderbei and Lagarias [23, p. 118]. The proof is based on the Cauchy–Binet formula and Cramer's rule; see, e.g., Horn and Johnson [18, pp. 21–22]. The same result is given by Ben-Tal and Teboulle [3, Cor. 2.1]. Also, the closely related result for unweighted linear least-squares problems where A may not have full row rank is given by Berg [4, p. 67]. Extending this analysis, Ben-Israel [2, p. 108] shows that Theorem 2.1 can be generalized to the case in which A does not have full row rank.

THEOREM 2.1 (see Dikin [8]). *Let A be an $m \times n$ matrix of full row rank, let g be a vector of dimension n, and let $\mathcal{D}_+(A)$ be defined by* (2.1). *If $D \in \mathcal{D}_+(A)$, then*

$$(ADA^T)^{-1}ADg = \sum_{J\in\mathcal{J}(A)} \left(\frac{\det(D_J)\det(A_J)^2}{\sum_{K\in\mathcal{J}(A)} \det(D_K)\det(A_K)^2} \right) A_J^{-T} g_J,$$

where $\mathcal{J}(A)$ is the set of column indices associated with nonsingular $m\times m$ submatrices of A.

Proof. See, e.g., Ben-Tal and Teboulle [3, Cor. 2.1]. (See also Theorem 3.1 below.) □

Theorem 2.1 implies that if the weight matrix is diagonal and positive definite, then the solution to the weighted least-squares problem (1.1) lies in the convex hull of the *basic solutions* formed by satisfying m linearly independent equations. Hence, this theorem provides an expression on the supremum of $\|(ADA^T)^{-1}ADg\|$ and $\|(ADA^T)^{-1}AD\|$ for D diagonal and positive definite, as the following corollary shows.

COROLLARY 2.2. *Let A be an $m \times n$ matrix of full row rank, let g be a vector of dimension n, and let $\mathcal{D}_+(A)$ be defined by* (2.1). *Then*

$$\sup_{D\in\mathcal{D}_+(A)} \|(ADA^T)^{-1}ADg\| = \max_{J\in\mathcal{J}(A)} \|A_J^{-T} g_J\| \quad \textit{and}$$

$$\sup_{D\in\mathcal{D}_+(A)} \|(ADA^T)^{-1}AD\| = \max_{J\in\mathcal{J}(A)} \|A_J^{-T}\|,$$

where $\mathcal{J}(A)$ is the set of column indices associated with nonsingular $m \times m$ submatrices of A.

Proof. Let $\mathcal{J}(A)$ denote the set of column indices associated with nonsingular $m \times m$ submatrices of A. If D is diagonal and positive definite, then $\det(D_J) > 0$ for all $J \in \mathcal{J}(A)$. Theorem 2.1 and norm properties immediately give

$$\|(ADA^T)^{-1}ADg\| \leq \max_{J \in \mathcal{J}(A)} \|A_J^{-T} g_J\|,$$

and hence

$$\sup_{D \in \mathcal{D}_+(A)} \|(ADA^T)^{-1}ADg\| \leq \max_{J \in \mathcal{J}(A)} \|A_J^{-T} g_J\|. \tag{2.3}$$

Let $K \in \mathcal{J}(A)$ be such that the maximum in the right-hand side of (2.3) is achieved; i.e.,

$$\|A_K^{-T} g_K\| = \max_{J \in \mathcal{J}(A)} \|A_J^{-T} g_J\|. \tag{2.4}$$

Assume that A can be partitioned as $A = (A_K \ A_\epsilon)$ without loss of generality. For a positive ϵ, let $D(\epsilon)$ be partitioned conformally with A as $D(\epsilon) = \operatorname{diag}(I, \epsilon I)$. Then $D(\epsilon) \in \mathcal{D}_+(A)$ for $\epsilon > 0$ with

$$\lim_{\epsilon \to 0^+} \|(AD(\epsilon)A^T)^{-1}AD(\epsilon)g\| = \|A_K^{-T} g_K\|. \tag{2.5}$$

A combination of (2.3), (2.4), and (2.5) gives

$$\sup_{D \in \mathcal{D}_+(A)} \|(ADA^T)^{-1}ADg\| = \max_{J \in \mathcal{J}(A)} \|A_J^{-T} g_J\|, \tag{2.6}$$

proving the first statement.

Since (2.6) is an identity for every $g \in \mathbb{R}^n$, we obtain

$$\max_{\|g\|=1} \sup_{D \in \mathcal{D}_+(A)} \|(ADA^T)^{-1}ADg\| = \max_{\|g\|=1} \max_{J \in \mathcal{J}(A)} \|A_J^{-T} g_J\|. \tag{2.7}$$

Reversing the order of the maximizations in (2.7) gives

$$\sup_{D \in \mathcal{D}_+(A)} \|(ADA^T)^{-1}AD\| = \max_{J \in \mathcal{J}(A)} \|A_J^{-T}\|,$$

proving the second statement. □

In this paper, we only consider the case in which A has full row rank, but we are interested in replacing the condition D diagonal and positive definite by D diagonal and positive semidefinite such that ADA^T is positive definite. Adding zero diagonals is of no significance to the above results, as the following corollary shows.

COROLLARY 2.3. *Theorem 2.1 and Corollary 2.2 still hold if $\mathcal{D}_+(A)$ is replaced by $\mathcal{D}_0(A)$, with $\mathcal{D}_0(A)$ defined by* (2.2).

Proof. Let $D \in \mathcal{D}_0(A)$. Assume that D can be partitioned as $D = \operatorname{diag}(D_+, 0)$ without loss of generality, where D_+ is diagonal and positive definite. Let A and g be partitioned conformally with D as $A = (\ A_+ \ A_0\ )$ and $g = (\ g_+^T \ g_0^T\ )^T$. Then $ADA^T = A_+ D_+ A_+^T$ and

$$(ADA^T)^{-1}ADg = (A_+ D_+ A_+^T)^{-1} A_+ D_+ g_+. \tag{2.8}$$

Since D_+ is positive definite, Theorem 2.1 applies to $(A_+D_+A_+^T)^{-1}A_+D_+g_+$. If $\mathcal{J}(A)$ denotes the set of column indices associated with nonsingular $m \times m$ submatrices of A, for each $J \in \mathcal{J}(A)$, any D_J having at least one zero diagonal has $\det(D_J) = 0$. Hence, they do not affect the convex combination of Theorem 2.1. □

Hanke and Neumann [17] give the geometry of the set $(ADA^T)^{-1}ADg$ as D varies over the set of positive definite diagonal matrices. The result is that this set is the union of the relative interiors of a finite number of polytopes and in general a proper, and possibly nonconvex, subset of the convex hull of the basic solutions. The trick used by Hanke and Neumann [17] is to divide into several cases, depending on the signature of the residual vector r of (1.3). We review this result below and give a direct proof. For a more elaborate discussion, see Hanke and Neumann [17].

THEOREM 2.4 (see Hanke and Neumann [17]). *Let A be an $m \times n$ matrix of full row rank, let g be a vector of dimension n, and let*

$$\Pi^{\mathcal{D}_+(A)}(A,g) = \{(ADA^T)^{-1}ADg : D \in \mathcal{D}_+(A)\},$$

with $\mathcal{D}_+(A)$ defined by (2.1). *If $\mathcal{S}$ denotes the set of n-dimensional signature vectors, then*

$$\Pi^{\mathcal{D}_+(A)}(A,g) = \cup_{s\in\mathcal{S}}\Pi^s(A,g),$$

where

$$\Pi^s(A,g) = \{\pi : Su + A^T\pi = g,\ ASv = 0,\ u > 0,\ v > 0\},$$

with $S = \operatorname{diag}(s)$.

Proof. Suppose $\pi \in \Pi^{\mathcal{D}_+(A)}(A,g)$. Then $ADA^T\pi = ADg$ for some $D \in \mathcal{D}_+(A)$. Hence, with $r = D(g - A^T\pi)$, these r and π solve the augmented system (1.3) given by

$$\begin{pmatrix} D^{-1} & A^T \\ A & 0 \end{pmatrix}\begin{pmatrix} r \\ \pi \end{pmatrix} = \begin{pmatrix} g \\ 0 \end{pmatrix}. \tag{2.9}$$

Let $s = \operatorname{sign}(r)$ and $S = \operatorname{diag}(s)$. Define v with components $v_i = |r_i|$ if $r_i \neq 0$ and $v_i = 1$ if $r_i = 0$. Then, $Sv = r$ and $v > 0$. Let $u = D^{-1}v$. Then $Su = SD^{-1}v = D^{-1}Sv = D^{-1}r$ and $u > 0$. Hence, we have

$$\begin{pmatrix} S & 0 & A^T \\ 0 & AS & 0 \end{pmatrix}\begin{pmatrix} u \\ v \\ \pi \end{pmatrix} = \begin{pmatrix} g \\ 0 \end{pmatrix}, \tag{2.10}$$

with $u > 0$ and $v > 0$. Consequently, $\pi \in \Pi^s(A,g)$ for $s = \operatorname{sign}(r)$, and it follows that $\Pi^{\mathcal{D}_+(A)}(A,g) \subseteq \cup_{s\in\mathcal{S}}\Pi^s(A,g)$.

Conversely, suppose that $\pi \in \Pi^s(A,g)$ for some signature vector s. Then there are positive vectors u and v such that π, u, and v satisfy (2.10) with $S = \operatorname{diag}(s)$. Let $r = Sv$ and $D = VU^{-1}$, with $U = \operatorname{diag}(u)$ and $V = \operatorname{diag}(v)$. Since v and u are positive vectors, D is well defined and positive definite, and we have $Su = SUV^{-1}v = UV^{-1}Sv = D^{-1}r$. Hence, this choice of D gives π and r as the solution of (2.9). Since s is an arbitrary element in $\mathcal{S}$, we conclude that $\cup_{s\in\mathcal{S}}\Pi^s(A,g) \subseteq \Pi^{\mathcal{D}_+(A)}(A,g)$. □

In the later analysis, we shall also be concerned with the geometry of the set $(ADA^T)^{-1}ADg$ as D varies over the set of positive semidefinite diagonal matrices

such that ADA^T is positive definite. In principle, this can be dealt with by considering subsets of columns of A such that the submatrix created that way has full row rank. However, in order not to make the notation too complex, we shall only associate this set with $\Pi^{\mathcal{D}_+(A)}(A,g)$ and its closure $\bar{\Pi}^{\mathcal{D}_+(A)}(A,g)$, as stated in the following corollary.

COROLLARY 2.5. *Let A be an $m \times n$ matrix of full row rank, let g be a vector of dimension n, and let*

$$\Pi^{\mathcal{D}_0(A)}(A,g) = \{(ADA^T)^{-1}ADg : D \in \mathcal{D}_0(A)\},$$

with $\mathcal{D}_0(A)$ defined by (2.2). Then

$$\Pi^{\mathcal{D}_+(A)}(A,g) \subseteq \Pi^{\mathcal{D}_0(A)}(A,g) \subseteq \bar{\Pi}^{\mathcal{D}_+(A)}(A,g),$$

where $\Pi^{\mathcal{D}_+(A)}(A,g)$ is defined by Theorem 2.4 and $\bar{\Pi}^{\mathcal{D}_+(A)}(A,g)$ denotes the closure of $\Pi^{\mathcal{D}_+(A)}(A,g)$.

Proof. Since $\mathcal{D}_+(A) \subseteq \mathcal{D}_0(A)$, we obtain $\Pi^{\mathcal{D}_+(A)}(A,g) \subseteq \Pi^{\mathcal{D}_0(A)}(A,g)$.

To show that $\Pi^{\mathcal{D}_0(A)}(A,g) \subseteq \bar{\Pi}^{\mathcal{D}_+(A)}(A,g)$, suppose that $\pi \in \Pi^{\mathcal{D}_0(A)}(A,g)$. Then $\pi = (ADA^T)^{-1}ADg$ for some positive semidefinite and diagonal D such that ADA^T is positive definite. For $\epsilon > 0$ let

$$\pi(\epsilon) = (A(D+\epsilon I)A^T)^{-1}A(D+\epsilon I)g. \tag{2.11}$$

Then $\pi(\epsilon) \in \Pi^{\mathcal{D}_+(A)}(A,g)$ for $\epsilon > 0$, and, since ADA^T is nonsingular, (2.11) gives

$$\lim_{\epsilon \to 0^+} \pi(\epsilon) = \pi,$$

and hence $\pi \in \bar{\Pi}^{\mathcal{D}_+(A)}(A,g)$. □

The essence of Corollary 2.5 is that the difference between the sets $\Pi^{\mathcal{D}_+(A)}(A,g)$ and $\Pi^{\mathcal{D}_0(A)}(A,g)$ is insignificant in that they have the same closure.

Stewart [21] gives a bound on the supremum of $\|A^T(ADA^T)^{-1}AD\|_2$ for D diagonal positive definite. This bound has subsequently been shown to be sharp by O'Leary [20], and the analysis has also been generalized to the case in which A does not have full row rank by Wei [25]. For an orthonormal matrix Q whose columns span the range space of A^T, the bound is given as the inverse of the smallest positive singular value of any submatrix of Q that has m columns. Since

$$\|A^T(ADA^T)^{-1}AD\|_2 = \|Q(Q^TDQ)^{-1}Q^TD\|_2 = \|(Q^TDQ)^{-1}Q^TD\|_2$$

and since for Q_J, a nonsingular $m \times m$ submatrix of Q, it holds that

$$\|Q_J^{-1}\|_2 = \frac{1}{\sigma_{\min}(Q_J)},$$

where $\sigma_{\min}(Q_J)$ denotes the smallest singular value of Q_J, this result can also be obtained from Corollary 2.2 with the additional information that the smallest positive singular value of any submatrix of Q that has m columns can be found by minimizing the smallest singular value of the nonsingular $m \times m$ submatrices of Q. Todd [22, pp. 1011–1012] also derives the boundedness of $\|(ADA^T)^{-1}ADg\|_2$ but without giving an explicit bound.

At this point, it also deserves mention that the bound provided by Corollary 2.2 is a *uniform bound*, but it can still be arbitrarily large. If $A = (1\ 0)$, then $\sup_{D\in\mathcal{D}} \|(ADA^T)^{-1}AD\| = 1$, but if $A = (1\ \epsilon)$, then $\sup_{D\in\mathcal{D}} \|(ADA^T)^{-1}AD\| = 1/\epsilon$ for $\epsilon \neq 0$. Hence, although the bound is independent of D, it may be impossible to compute in finite precision arithmetic. For a discussion of these matters, see Vavasis [24].

3. Generalization to nondiagonal weight matrices. If W is a general positive definite symmetric matrix, the boundedness property of Corollary 2.2 does not hold. In Stewart [21], a nondiagonal example is given by

$$A = \begin{pmatrix} 0 & 1 \end{pmatrix} \quad \text{and} \quad W(\delta,\epsilon) = \begin{pmatrix} 1+\delta\epsilon^2 & \epsilon(1-\delta) \\ \epsilon(1-\delta) & \epsilon^2+\delta \end{pmatrix}. \tag{3.1}$$

Then $W(\delta,\epsilon)$ is positive definite for any positive δ, and it holds that

$$(AW(\delta,\epsilon)A^T)^{-1}AW(\delta,\epsilon) = \begin{pmatrix} \dfrac{\epsilon(1-\delta)}{\epsilon^2+\delta} & 1 \end{pmatrix}. \tag{3.2}$$

Stewart [21] observes that (3.2) implies that for a *fixed* ϵ ($\epsilon > 0$) we have

$$\lim_{\delta\to 0} (AW(\delta,\epsilon)A^T)^{-1}AW(\delta,\epsilon) = (\epsilon^{-1} \quad 1). \tag{3.3}$$

Hence, (3.3) implies that $\|(AW(\delta,\epsilon)A^T)^{-1}AW(\delta,\epsilon)\|$ can be arbitrarily large when δ and ϵ are close to zero. In this situation $W(\delta,\epsilon)$ is close to diagonal. Note, however, that there is a certain relationship required between δ and ϵ for the norm to become unbounded. This is discussed further in §7.

Theorem 3.1 below gives the characterization of $(AWA^T)^{-1}AW$ for a symmetric matrix W, such that AWA^T is nonsingular when a symmetric diagonal decomposition of W is known; i.e., $W = UDU^T$ for some conformally dimensioned matrices. When U is square and nonsingular and W is positive definite, the generalization is immediate from Theorem 2.1. Typically, this could be the eigenvalue decomposition; see, e.g., Hanke and Neumann [17, §6]. A problem that arises when analyzing nondiagonal matrices using the eigenvalue decomposition is that not only do the eigenvalues change with W but also the eigenvectors. The way to overcome this difficulty, presented in §4, is to define a decomposition where U is rectangular, with more columns than rows, but there are only a finite number of different U-matrices. In this situation D may be indefinite, although W is positive definite.

The proof technique in Theorem 3.1 follows that of Ben-Tal and Teboulle [3, Thm. 2.1] for the unweighted linear least-squares problem. The difference is that we do not require the diagonal matrix to be positive definite, and therefore an "unsymmetric" form of AWA^T is used; see (3.5). For the sake of completeness, we give the total proof.

THEOREM 3.1. *Let A be an $m \times n$ matrix of full row rank, and let W be a symmetric $n \times n$ matrix such that AWA^T is nonsingular. Suppose $W = UDU^T$, where D is diagonal. Then*

$$(AWA^T)^{-1}AW = \sum_{J\in\mathcal{J}(AU)} \left(\frac{\det(D_J)\det(AU_J)^2}{\sum_{K\in\mathcal{J}(AU)} \det(D_K)\det(AU_K)^2} \right) (AU_J)^{-T}U_J^T,$$

where $\mathcal{J}(AU)$ is the set of column indices associated with nonsingular $m \times m$ submatrices of AU.

Proof. Using the nonsingularity of AWA^T and the identity $W = UDU^T$, we obtain

$$\pi = (AWA^T)^{-1}AWg = (AUDU^TA^T)^{-1}AUDU^Tg. \tag{3.4}$$

We first show that AU has full row rank. This is done by contradiction. Suppose $U^TA^T\pi = 0$ for some $\pi \neq 0$. Then $0 = AUDU^TA^T\pi = AWA^T\pi$, contradicting the nonsingularity of AWA^T. Hence, AU has full row rank.

Let $\tilde{A} = AU$, $\tilde{B} = AUD$, and $\tilde{g} = U^Tg$. Then (3.4) gives

$$\pi = (\tilde{B}\tilde{A}^T)^{-1}\tilde{B}\tilde{g}. \tag{3.5}$$

The remainder of the proof consists of rewriting (3.5) using the Cauchy–Binet formula and Cramer's rule (see, e.g., Horn and Johnson [18, pp. 21–22]). Cramer's rule in conjunction with (3.5) gives

$$\pi_i = \frac{\det(\tilde{B}\tilde{A}^T + (\tilde{B}\tilde{g} - \tilde{B}\tilde{A}^Te_i)e_i^T)}{\det(\tilde{B}\tilde{A}^T)}, \quad i = 1, \ldots, m. \tag{3.6}$$

Application of the Cauchy–Binet formula on the denominator of (3.6) now gives

$$\det(\tilde{B}\tilde{A}^T) = \sum_{K \in \mathcal{J}(\tilde{A})} \det(\tilde{B}_K)\det(\tilde{A}_K), \tag{3.7}$$

where $\mathcal{J}(\tilde{A})$ is the set of column indices associated with nonsingular $m \times m$ submatrices of $\tilde{A}$. Note that since D is diagonal, if $J \in \mathcal{J}(\tilde{A})$, then $\tilde{B}_J = D_J\tilde{A}_J$. Hence, $\det(\tilde{B}_J) \neq 0$ only if $\det(\tilde{A}_J) \neq 0$, and application of the Cauchy–Binet formula and Cramer's rule to the numerator of (3.6) gives

$$\begin{aligned}
\det(\tilde{B}\tilde{A}^T + (\tilde{B}\tilde{g} - \tilde{B}\tilde{A}^Te_i)e_i^T) &= \det(\tilde{B}(\tilde{A}^T + (\tilde{g} - \tilde{A}^Te_i)e_i^T)) \\
&= \sum_{J \in \mathcal{J}(\tilde{A})} \det(\tilde{B}_J)\det(\tilde{A}_J^T + (\tilde{g}_J - \tilde{A}_J^Te_i)e_i^T) \\
&= \sum_{J \in \mathcal{J}(\tilde{A})} \det(\tilde{B}_J)\det(\tilde{A}_J)(\tilde{A}_J^{-T}\tilde{g}_J)_i. \qquad (3.8)
\end{aligned}$$

We can now identify $\tilde{A}_J = AU_J$ and $\tilde{g}_J = U_J^Tg$. As was observed above, we obtain $\tilde{B}_J = AU_JD_J$ and $\det(\tilde{B}_J) = \det(AU_J)\det(D_J)$. Substitution of these quantities in (3.7) and (3.8) and insertion into (3.6) yield

$$(AWA^T)^{-1}AWg = \sum_{J \in \mathcal{J}(AU)} \left(\frac{\det(D_J)\det(AU_J)^2}{\sum_{K \in \mathcal{J}(AU)} \det(D_K)\det(AU_K)^2} \right) (AU_J)^{-T}U_J^Tg,$$

where $\mathcal{J}(AU)$ is the set of column indices associated with nonsingular $m \times m$ submatrices of AU. Since g is an arbitrary n-vector, the proof is complete. □

4. The signature decomposition. In this section, we present a symmetric diagonal decomposition of a symmetric $n \times n$ matrix W. This decomposition is referred to throughout as the *signature decomposition.* Loosely speaking, it is a decomposition of W on an elementwise outer-product form. It has the form

$$W = U(s(W))D(W)U(s(W))^T,$$

where we refer to $s(W)$ as the *dense signature vector* associated with W, $U(s(W))$ as the *signature matrix* associated with $s(W)$, and $D(W)$ as the *diagonal-dominance matrix* associated with W. The definition of $s(W)$ is given in (4.1), and the definition

of $D(W)$ is given in (4.6). For a dense signature vector s, with dimension equal to that of $s(W)$, the definition of $U(s)$ is given in (4.4). Finally, Lemma 4.1 shows that with these definitions, the signature decomposition is well defined.

We define the *dense signature vector* of W as the vector of dimension $\frac{n(n-1)}{2}$ with components

$$s_{t(i,j)}(W) = \text{sign}^+(w_{ij}), \quad 1 \le i < j \le n, \tag{4.1}$$

where $t(i,j) = n(i-1) - \frac{i(i+1)}{2} + j$. The notation $t(i,j)$ is used to stress that $t(i,j)$ corresponds to the off-diagonal element w_{ij}. The ordering corresponds to the strict upper-triangular part of W, ordered by rows, i.e., $w_{12}, \ldots, w_{1n}$, $w_{23}, \ldots, w_{2n}, \ldots$, $w_{n-1,n}$. This means that

$$t(i,j) = \sum_{k=1}^{i-1}(n-k) + j - i = n(i-1) - \frac{i(i+1)}{2} + j.$$

The vector $s(W)$ essentially describes the signature of the off-diagonal elements of W; the only difference is that zero elements are given signature one. For example, associated with the matrix W defined as

$$W = \begin{pmatrix} 2 & 0 & -3 \\ 0 & 1 & 1 \\ -3 & 1 & 6 \end{pmatrix} \tag{4.2}$$

we obtain $t(1,2) = 1$, $t(1,3) = 2$, and $t(2,3) = 3$, and the dense signature vector is

$$s(W) = \begin{pmatrix} 1 & -1 & 1 \end{pmatrix}^T. \tag{4.3}$$

Associated with a dense signature vector s of dimension $\frac{n(n-1)}{2}$, we define the associated *signature matrix* $U(s)$ as the matrix of dimension $n \times \frac{n(n+1)}{2}$, with columns $u_i(s)$, $i = 1, \ldots, \frac{n(n+1)}{2}$, defined by

$$u_i(s) = e_i, \qquad i = 1, \ldots, n, \qquad \text{and} \tag{4.4a}$$

$$u_{n+t(i,j)}(s) = e_i + s_{t(i,j)}e_j, \quad 1 \le i < j \le n, \tag{4.4b}$$

where $t(i,j) = n(i-1) - \frac{i(i+1)}{2} + j$. For $s(W)$ of (4.3), the associated $U(s(W))$ is given by

$$U(s(W)) = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & -1 & 1 \end{pmatrix}. \tag{4.5}$$

Associated with W, we define the *diagonal-dominance matrix* $D(W)$ as the diagonal matrix of dimension $\frac{n(n+1)}{2} \times \frac{n(n+1)}{2}$ with diagonal elements $d_i(W)$, $i = 1, \ldots, \frac{n(n+1)}{2}$, defined by

$$d_i(W) = w_{ii} - \sum_{\substack{j=1 \\ j \ne i}}^{n} |w_{ij}|, \qquad i = 1, \ldots, n, \qquad \text{and} \tag{4.6a}$$

$$d_{n+t(i,j)}(W) = |w_{ij}|, \qquad 1 \le i < j \le n, \tag{4.6b}$$

where $t(i,j) = n(i-1) - \frac{i(i+1)}{2} + j$. For W of (4.2), we obtain $D(W)$ as

$$\text{diag}(D(W)) = \begin{pmatrix} -1 & 0 & 2 & 0 & 3 & 1 \end{pmatrix}^T. \tag{4.7}$$

Finally, the following lemma formally defines the signature decomposition.

LEMMA 4.1. *Let W be a symmetric $n \times n$ matrix, and let $s(W)$ be the corresponding dense signature vector defined by* (4.1). *Let $U(s(W))$ be the signature matrix associated with $s(W)$ defined by* (4.4), *and let $D(W)$ be the diagonal-dominance matrix associated with W defined by* (4.6). *Then*

$$W = U(s(W))D(W)U(s(W))^T.$$

Proof. Let W be a symmetric $n \times n$ matrix. We may write W in outer-product form as

$$W = \sum_{i=1}^{n} w_{ii} e_i e_i^T + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} w_{ij} e_i e_j^T.$$

Rearrangement using the identity $w_{ij} = |w_{ij}| \text{sign}^+(w_{ij})$ and the symmetry of W gives

$$\begin{aligned} W &= \sum_{i=1}^{n} \left(w_{ii} - \sum_{\substack{j=1 \\ j \neq i}}^{n} |w_{ij}| \right) e_i e_i^T + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} |w_{ij}| \left(e_i e_i^T + \text{sign}^+(w_{ij})\, e_i e_j^T \right) \\ &= \sum_{i=1}^{n} \left(w_{ii} - \sum_{\substack{j=1 \\ j \neq i}}^{n} |w_{ij}| \right) e_i e_i^T \\ &\quad + \sum_{i=1}^{n} \sum_{j=i+1}^{n} |w_{ij}| \left(e_i + \text{sign}^+(w_{ij}) e_j \right) \left(e_i + \text{sign}^+(w_{ij}) e_j \right)^T. \end{aligned} \tag{4.8}$$

The proof is now complete by comparing (4.8) with the definitions in (4.1), (4.4), and (4.6). ∎

The motivation for considering the signature decomposition of Lemma 4.1 is that there are only a finite number of different signature vectors s and consequently only a finite number of different $U(s)$ matrices. Hence, when studying convergence properties of sequences of W-matrices, all the information that changes continuously is confined to $D(W)$. In particular, this means that Theorem 3.1 in conjunction with the signature decomposition gives the solution of (1.1) as a linear combination of a subset of vectors from a *finite* set. The following corollary makes this precise.

COROLLARY 4.2. *Let A be an $m \times n$ matrix of full row rank, and let W be a symmetric $n \times n$ matrix such that AWA^T is nonsingular. Let $W = U(s(W))D(W)U(s(W))^T$ be the signature decomposition of W given by Lemma* 4.1. *Then*

$$(AWA^T)^{-1}AW = \sum_{J \in \mathcal{J}(AU(s(W)))} \alpha_J(W)(AU_J(s(W)))^{-T}U_J(s(W))^T,$$

with

$$\alpha_J(W) = \frac{\det(D_J(W)) \det(AU_J(s(W)))^2}{\sum_{K \in \mathcal{J}(AU(s(W)))} \det(D_K(W)) \det(AU_K(s(W)))^2},$$

where $\mathcal{J}(AU(s(W)))$ is the set of column indices associated with nonsingular $m \times m$ submatrices of $AU(s(W))$.

Proof. Application of Theorem 3.1 to the signature decomposition defined by Lemma 4.1 gives the result. □

Note that zero diagonals in $D(W)$ occur for two reasons: among the first n columns because the associated row of W has the row sum of absolute values of the off-diagonal elements equal the diagonal element, among the last $\frac{n(n-1)}{2}$ columns because the associated off-diagonal element w_{ij} is zero. In the factors of the example W defined in (4.2) the former occurs in the second diagonal element of $D(W)$ and the latter in diagonal four of $D(W)$; see (4.7). Such diagonal elements of $D(W)$ and associated columns of $U(s(W))$ are redundant and may be removed from the decomposition. In order to keep the notation as simple as possible, we leave them. (If they are removed, the dimensions of $U(s(W))$ and $D(W)$ depend on W.) The following lemma shows that with the above definitions, $D(W)$ is unique if its last $\frac{n(n-1)}{2}$ diagonal elements are nonnegative. It also implies that if $D = D(W)$ and $W = U(s)DU(s)^T$, $U(s)$ differs from $U(s(W))$ only in columns corresponding to zero diagonals of D.

LEMMA 4.3. *Let s be a dense $\frac{n(n-1)}{2}$-dimensional signature vector, let $U(s)$ be the signature matrix associated with s defined by* (4.4)*, and let D be a diagonal matrix of dimension $\frac{n(n+1)}{2} \times \frac{n(n+1)}{2}$ whose last $\frac{n(n-1)}{2}$ diagonal elements are nonnegative. If*

$$W = U(s)DU(s)^T,$$

then $D = D(W)$ and

$$d_j u_j(s) = d_j(W) u_j(s(W)), \quad j = 1, \ldots, \frac{n(n+1)}{2},$$

where $D(W)$ is the diagonal-dominance matrix associated with W, defined by (4.6)*; $s(W)$ is the dense signature vector associated with W, defined by* (4.1)*; and $U(s(W))$ is the signature matrix associated with $s(W)$, defined by* (4.4)*.*

Proof. Let s be a dense $\frac{n(n-1)}{2}$-dimensional signature vector, let $U(s)$ be the associated signature matrix, and let D be a diagonal matrix of dimension $\frac{n(n+1)}{2} \times \frac{n(n+1)}{2}$ whose last $\frac{n(n-1)}{2}$ diagonal elements are nonnegative. If

$$W = U(s)DU(s)^T,$$

then we may write W in outer-product form as

$$W = \sum_{i=1}^{n} d_i e_i e_i^T + \sum_{i=1}^{n} \sum_{j=i+1}^{n} d_{n+t(i,j)} \left(e_i + s_{t(i,j)} e_j\right) \left(e_i + s_{t(i,j)} e_j\right)^T. \tag{4.9}$$

Throughout the proof, let $t(i,j) = n(i-1) - \frac{i(i+1)}{2} + j$. Identification of components in (4.9) gives

$$w_{ii} = d_i + \sum_{j=1}^{i-1} d_{n+t(j,i)} + \sum_{j=i+1}^{n} d_{n+t(i,j)}, \qquad i = 1, \ldots, n, \tag{4.10a}$$

$$w_{ji} = w_{ij} = d_{n+t(i,j)} s_{t(i,j)}, \qquad 1 \le i < j \le n. \tag{4.10b}$$

Since s is a dense signature vector and for $1 \le i < j \le n$ we have $d_{n+t(i,j)} \ge 0$, (4.10b) gives

$$d_{n+t(i,j)} = |w_{ij}| = |w_{ji}| \quad \text{for} \quad 1 \le i < j \le n. \tag{4.11}$$

Insertion of (4.11) into (4.10a) now gives

$$d_i = w_{ii} - \sum_{j=1}^{i-1} |w_{ji}| - \sum_{j=i+1}^{n} |w_{ij}| = w_{ii} - \sum_{\substack{j=1 \\ j \ne i}}^{n} |w_{ij}| \quad \text{for} \quad 1 \le i \le n. \tag{4.12}$$

Comparison of (4.11) and (4.12) with (4.6) shows that $D = D(W)$.

Finally, since $D = D(W)$ and the first n columns of $U(s)$ and $U(s(W))$ are independent of s and $s(W)$, it suffices to show that

$$d_{n+t(i,j)} u_{n+t(i,j)}(s) = d_{n+t(i,j)}(W) u_{n+t(i,j)}(s(W)), \quad 1 \le i < j \le n. \tag{4.13}$$

Since $W = U(s(W))D(W)U(s(W))^T$, (4.10b) gives

$$w_{ji} = w_{ij} = d_{n+t(i,j)}(s(W)) s_{t(i,j)}(W), \quad 1 \le i < j \le n. \tag{4.14}$$

A combination of (4.10b) and (4.14) with the definition of $U(s)$ from (4.4) gives (4.13), completing the proof. □

It may appear a bit counterintuitive that $D(W)$ of (4.7) is indefinite although W of (4.2) is positive definite. However, when looking at the definition of $D(W)$ given by (4.6), it is immediate that a symmetric W is positive semidefinite and *diagonally dominant* if and only if $D(W)$ is positive semidefinite. Hence, the linear combination of Corollary 4.2 is a convex combination if W is positive semidefinite, symmetric, and diagonally dominant. This is essential in the analysis of §5.

5. Results for diagonally dominant weight matrices. In this section, it is shown that the boundedness property of diagonally weighted least-squares solutions can be extended also to weight matrices that are *diagonally dominant.* Using standard terminology, we define a symmetric positive semidefinite $n \times n$ matrix W to be *diagonally dominant* if

$$w_{jj} \ge \sum_{\substack{i=1 \\ i \ne j}}^{n} |w_{ij}| \quad \text{for} \quad j = 1, \ldots, n;$$

see, e.g., Horn and Johnson [18, p. 349]. (Note that the absolute value normally put on w_{jj} is unnecessary, since we require W to be positive semidefinite.) Similar to the diagonal case, we are concerned with two sets of $n \times n$ symmetric weight matrices associated with an $m \times n$ matrix A of full row rank. The first set is $\mathcal{W}_+(A)$ defined by

$$\mathcal{W}_+(A) = \left\{ W \in \mathbb{R}^{n \times n} : \begin{array}{l} W \text{ is symmetric, diagonally dominant, and} \\ \text{positive definite} \end{array} \right\}, \tag{5.1}$$

and the second is $\mathcal{W}_0(A)$ defined by

$$\mathcal{W}_0(A) = \left\{ W \in \mathbb{R}^{n \times n} : \begin{array}{l} W \text{ is symmetric, diagonally dominant, and posi-} \\ \text{tive semidefinite; } AWA^T \text{ is positive definite} \end{array} \right\}. \tag{5.2}$$

Note that $\mathcal{W}_+(A) \subseteq \mathcal{W}_0(A)$, and both $\mathcal{W}_+(A)$ and $\mathcal{W}_0(A)$ are convex sets in which AWA^T is positive definite.

Recall from §4 that the virtue of the signature decomposition is that it transforms the question of positive semidefiniteness and diagonal dominance of a matrix W to positive semidefiniteness of its associated diagonal-dominance matrix $D(W)$ at the same time as there are only a finite number of signature matrices $U(s)$. This means that with W symmetric, positive semidefinite, and diagonally dominant we may rewrite (1.1) as

$$\underset{\pi \in \mathbb{R}^m}{\text{minimize}} \qquad \|D(W)^{1/2}(U(s(W))^T A^T \pi - U(s(W))^T g)\|_2^2.$$

This observation is used below to derive the main results of the paper. Corollary 4.2 and Lemma 4.3 imply that the case of positive semidefinite and diagonally dominant symmetric weight matrices can be reduced to the union of a finite number of cases of positive semidefinite and diagonal weight matrices, one diagonal case for each dense signature vector. The following theorem makes this precise.

THEOREM 5.1. *Let A be an $m \times n$ matrix of full row rank, and let g be a vector of dimension n. Let $\mathcal{S}$ denote the set of dense $\frac{n(n-1)}{2}$-dimensional signature vectors, and for $s \in \mathcal{S}$ let $U(s)$ be the associated signature matrix defined by* (4.4). *Let*

$$\Pi^{\mathcal{W}_0(A)}(A, g) = \{(AWA^T)^{-1}AWg : W \in \mathcal{W}_0(A)\},$$

with $\mathcal{W}_0(A)$ defined by (5.2). *Then*

$$\Pi^{\mathcal{W}_0(A)}(A, g) = \cup_{s \in \mathcal{S}} \Pi^{\mathcal{D}_0(AU(s))}(AU(s), U(s)^T g),$$

where $U(s)$ is defined by (4.4), *and the set $\Pi^{\mathcal{D}_0(AU(s))}(AU(s), U(s)^T g)$ is given by Corollary* 2.5 *with A replaced by $AU(s)$, g replaced by $U(s)^T g$, and n replaced by $\frac{n(n+1)}{2}$.*

Proof. Suppose $\pi \in \Pi^{\mathcal{W}_0(A)}(A, g)$; i.e., $\pi = (AWA^T)^{-1}AWg$ for some $W \in \mathcal{W}_0(A)$. Let $W = U(s(W))D(W)U(s(W))^T$ be the signature decomposition of W from Lemma 4.1. Since W is positive semidefinite and diagonally dominant, (4.6) implies that $D(W)$ is positive semidefinite, and with the notation of Corollary 2.5, it follows that $\pi \in \Pi^{\mathcal{D}_0(AU(s(W)))}(AU(s(W)), U(s(W))^T g)$. Hence, since $s(W) \in \mathcal{S}$, we conclude that $\Pi^{\mathcal{W}_0(A)}(A, g) \subseteq \cup_{s \in \mathcal{S}} \Pi^{\mathcal{D}_0(AU(s))}(AU(s), U(s)^T g)$.

Conversely, suppose $\pi \in \Pi^{\mathcal{D}_0(AU(s))}(AU(s), U(s)^T g)$ for some $s \in \mathcal{S}$. Then

$$\pi = (AU(s)DU(s)^T A^T)^{-1}AU(s)DU(s)^T g$$

for some positive semidefinite and diagonal D such that $AU(s)DU(s)^T A^T$ is positive definite. Let $W = U(s)DU(s)^T$. Lemma 4.3 implies that $D = D(W)$, and hence (4.6) ensures that W is positive semidefinite and diagonally dominant. Consequently, $W \in \mathcal{W}_0(A)$, and hence $\pi \in \Pi^{\mathcal{W}_0(A)}(A, g)$. Since s is an arbitrary element in $\mathcal{S}$, we conclude that $\cup_{s \in \mathcal{S}} \Pi^{\mathcal{D}_0(AU(s))}(AU(s), U(s)^T g) \subseteq \Pi^{\mathcal{W}_0(A)}(A, g)$. ∎

We have preferred to state the result for the case when W is allowed to be positive semidefinite, as long as AWA^T is positive definite. However, as in Corollary 2.5 for the diagonal case, the following corollary shows that the difference between this requirement and requiring W to be positive definite is very small.

COROLLARY 5.2. *Let A be an $m \times n$ matrix of full row rank, let g be a vector of dimension n, and let*

$$\Pi^{\mathcal{W}_+(A)}(A, g) = \{(ADA^T)^{-1}ADg : D \in \mathcal{W}_+(A)\},$$

with $\mathcal{W}_+(A)$ *defined by* (5.1). *Then*

$$\Pi^{\mathcal{W}_+(A)}(A,g) \subseteq \Pi^{\mathcal{W}_0(A)}(A,g) \subseteq \bar{\Pi}^{\mathcal{W}_+(A)}(A,g),$$

where $\Pi^{\mathcal{W}_0(A)}(A,g)$ *is defined by Theorem* 5.1 *and* $\bar{\Pi}^{\mathcal{W}_+(A)}(A,g)$ *denotes the closure of* $\Pi^{\mathcal{W}_+(A)}(A,g)$.

Proof. The proof is analogous to the proof of Corollary 2.5, with $\mathcal{D}_+(A)$ replaced by $\mathcal{W}_+(A)$ and $\mathcal{D}_0(A)$ replaced by $\mathcal{W}_0(A)$. □

Hence, the boundedness properties of the diagonal case carry over to the diagonally dominant case in a straightforward manner. In particular, Corollary 5.2 implies that the same boundedness properties hold for both sets $\mathcal{W}_0(A)$ and $\mathcal{W}_+(A)$.

COROLLARY 5.3. *Let* A *be an* $m \times n$ *matrix of full row rank, and let* $\mathcal{W}_0(A)$ *be defined by* (5.2). *Let* $\mathcal{S}$ *denote the set of dense* $\frac{n(n-1)}{2}$*-dimensional signature vectors, and for* $s \in \mathcal{S}$ *let* $U(s)$ *be the associated signature matrix defined by* (4.4). *Then*

$$\sup_{W \in \mathcal{W}_0(A)} \|(AWA^T)^{-1}AWg\| = \max_{s \in \mathcal{S}} \max_{J \in \mathcal{J}(AU(s))} \|(AU_J(s))^{-T}U_J(s)^T g\| \quad \textit{and}$$

$$\sup_{W \in \mathcal{W}_0(A)} \|(AWA^T)^{-1}AW\| = \max_{s \in \mathcal{S}} \max_{J \in \mathcal{J}(AU(s))} \|(AU_J(s))^{-T}U_J(s)^T\|,$$

where $\mathcal{J}(AU(s))$ *is the set of column indices associated with nonsingular* $m \times m$ *submatrices of* $AU(s)$. *In addition, the same properties hold if* $\mathcal{W}_0(A)$ *is replaced by* $\mathcal{W}_+(A)$, *with* $\mathcal{W}_+(A)$ *defined by* (5.1).

Proof. Theorem 5.1 shows that we may obtain the result for $\|(AWA^T)^{-1}AWg\|$ with $W \in \mathcal{W}_0(A)$ by applying Theorem 2.1 and Corollaries 2.2 and 2.3 for each individual $s \in \mathcal{S}$ and taking the maximum. Since we thus have an identity for each g, the result for $\|(AWA^T)^{-1}AW\|$ with $W \in \mathcal{W}_0(A)$ follows. Corollary 5.2 shows that the same properties hold if $\mathcal{W}_0(A)$ is replaced by $\mathcal{W}_+(A)$. □

Loosely speaking, finding the maximizing $U_J(s)$ in Corollary 5.3 involves finding a particular nonsingular $m \times m$ matrix where each column is either a column of A or the sum or difference of a pair of columns of A. (No pair of columns appears more than once.) As an example of where a simple explicit formula can be given, consider the case of Euclidean norm when A consists of only one row.

COROLLARY 5.4. *Let* a *be a nonzero* n*-vector, and let* $\mathcal{W}_0(a^T)$ *be defined by* (5.2). *Then*

$$\sup_{W \in \mathcal{W}_0(a^T)} \|(a^TWa)^{-1}a^TW\|_2 = \max\left\{ \frac{1}{\min\limits_{i:|a_i|\neq 0} |a_i|}, \frac{\sqrt{2}}{\min\limits_{i,j:|a_i|\neq|a_j|} ||a_i| - |a_j||} \right\}.$$

Proof. The result is straightforward from Corollary 5.3 upon observing that if s is a dense signature vector with associated signature matrix $U(s)$, we have

$$\begin{aligned} a^Tu_i(s) &= a_i \quad \text{and} \quad \|u_i(s)\|_2 = 1, && i = 1,\dots,n, \\ a^Tu_{n+t(i,j)}(s) &= a_i + s_{t(i,j)}a_j \quad \text{and} \quad \|u_{n+t(i,j)}(s)\|_2 = \sqrt{2}, && 1 \le i < j \le n, \end{aligned}$$

where $t(i,j) = n(i-1) - \frac{i(i+1)}{2} + j$. □

6. Extension to equality-constrained linear least-squares problems. The analysis above allows positive diagonal weights of any size, and in particular the

weights may tend to infinity. This implies that the above results in a straightforward manner may be extended to a class of equality-constrained linear least-squares problems of the form

$$\begin{array}{ll} \underset{\pi\in\mathbb{R}^m}{\text{minimize}} & \|W_{11}^{1/2}(A_1^T\pi - g_1)\|_2^2 \\ \text{subject to} & A_2^T\pi = g_2. \end{array} \tag{6.1}$$

Here, we have partitioned A as $A = (A_1\ A_2)$, where A_1 has n_1 columns and A_2 has n_2 columns. Conformally with A, we partition g as $g = (g_1^T\ g_2^T)^T$. The weight matrix W_{11} is symmetric and positive definite of dimension $n_1 \times n_1$. As before, we assume that A has full row rank, and we also assume that A_2 has full column rank. (Note that the assumption that A_2 has full row rank is not significant. If the equations $A_2^T\pi = g_2$ are compatible, linearly dependent columns of A_2 can be removed without changing the problem.) We denote by $\pi(W_{11}, A_1, A_2, g_1, g_2)$ the optimal solution of (6.1).

It is well known that the equality-constrained problem can be viewed as the limiting case of an unconstrained problem when an infinite diagonal weight matrix is associated with A_2; see, e.g., Lawson and Hanson [19, Chap. 22]. This is reviewed in the following lemma.

LEMMA 6.1. *Let A be an $m \times n$ matrix of full row rank, which is partitioned as $A = (A_1\ A_2)$, where A_1 has n_1 columns and A_2 is of full column rank and has n_2 columns. Let g be a vector of dimension n, partitioned conformally with A, as $g = (g_1^T\ g_2^T)^T$. Let W_{11} be a positive definite diagonal matrix of dimension $n_1 \times n_1$, and let $\pi(W_{11}, A_1, A_2, g_1, g_2)$ denote the optimal solution of* (6.1). *Then, with π as the solution of the linear system*

$$\begin{pmatrix} W_{11}^{-1} & 0 & A_1^T \\ 0 & 0 & A_2^T \\ A_1 & A_2 & 0 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ \pi \end{pmatrix} = \begin{pmatrix} g_1 \\ g_2 \\ 0 \end{pmatrix}, \tag{6.2}$$

it holds that $\pi = \pi(W_{11}, A_1, A_2, g_1, g_2)$. Moreover, if $W(\epsilon)$ is defined as

$$W(\epsilon) = \begin{pmatrix} W_{11} & 0 \\ 0 & \dfrac{1}{\epsilon} I \end{pmatrix} \tag{6.3}$$

for $\epsilon > 0$, then $\lim_{\epsilon\to 0+}(AW(\epsilon)A^T)^{-1}AW(\epsilon)g = \pi(W_{11}, A_1, A_2, g_1, g_2)$.

Proof. The characterization and uniqueness of π follows for example from the analysis of Gould [15]. For completeness, we give a direct proof. Since (6.1) is a convex quadratic programming problem, where the objective function is bounded from below, the first-order optimality conditions are necessary and sufficient for optimality; i.e.,

$$A_1W_{11}A_1^T\pi - A_2r_2 = A_1W_{11}g_1, \tag{6.4a}$$

$$A_2^T\pi = g_2. \tag{6.4b}$$

With $r_1 = W_{11}(g_1 - A_1^T\pi)$, (6.4) is equivalent to (6.2). To show that the system (6.2) is nonsingular, assume that there is a solution

$$W_{11}^{-1}u_1 + A_1^Tu_3 = 0, \tag{6.5a}$$

$$A_2^Tu_3 = 0, \tag{6.5b}$$

$$A_1u_1 + A_2u_2 = 0. \tag{6.5c}$$

Premultiplication of (6.5a) by u_1^T, taking into account (6.5c) and (6.5b), gives

$$0 = u_1^T W_{11}^{-1} u_1 + u_1^T A_1^T u_3 = u_1^T W_{11}^{-1} u_1 - u_2^T A_2^T u_3 = u_1^T W_{11}^{-1} u_1.$$

Since W_{11} is positive definite, this means $u_1 = 0$. But if $u_1 = 0$, the full row rank of A in conjunction with (6.5a) and (6.5b) implies $u_3 = 0$. Similarly, (6.5c) and the full column rank of A_2 implies that $u_2 = 0$. Hence, r_1, r_2, and π are the unique solution of (6.2), and $\pi = \pi(W_{11}, A_1, A_2, g_1, g_2)$.

If $\pi(\epsilon) = (AW(\epsilon)A^T)^{-1}AW(\epsilon)g$, with $W(\epsilon)$ defined by (6.3), (1.3) implies that $\pi(\epsilon)$, together with $r_1(\epsilon)$ and $r_2(\epsilon)$, is uniquely defined by

$$\begin{pmatrix} W_{11}^{-1} & 0 & A_1^T \\ 0 & \epsilon I & A_2^T \\ A_1 & A_2 & 0 \end{pmatrix} \begin{pmatrix} r_1(\epsilon) \\ r_2(\epsilon) \\ \pi(\epsilon) \end{pmatrix} = \begin{pmatrix} g_1 \\ g_2 \\ 0 \end{pmatrix}. \tag{6.6}$$

Since (6.2) and (6.6) are nonsingular systems that are identical when $\epsilon = 0$ and the solution of (6.6) is a continuous function of ϵ, we obtain $\lim_{\epsilon \to 0+} \pi(\epsilon) = \pi$. □

Note that for $\epsilon > 0$, $W(\epsilon)$ defined by (6.3) is positive definite and diagonal if and only if W_{11} is positive definite and diagonal, and $W(\epsilon)$ is symmetric, positive definite, and diagonally dominant if and only if W_{11} is symmetric, positive definite, and diagonally dominant. Hence, Lemma 6.1 implies that the boundedness properties of §§2 and 5 apply for the two cases. In addition, the signature vector $s(W(\epsilon))$ is independent of ϵ. Also, the expression for $(AW(\epsilon)A^T)^{-1}AW(\epsilon)g$ provided by Corollary 4.2 involving the signature decomposition of $W(\epsilon)$ is in terms of $m \times m$ matrices where each column is either a column of A or the sum or difference of a pair of columns of A. Hence, it seems plausible that the limiting case can be obtained by letting the infinite diagonal weights on the columns of A_2 mean that A_2 is a submatrix of each $AU_J(s(W(\epsilon)))$ with a nonzero coefficient. This is indeed the case, as the following corollary shows.

COROLLARY 6.2. *Let A be an $m \times n$ matrix of full row rank, which is partitioned as $A = (A_1 \; A_2)$, where A_1 has n_1 columns and A_2 is of full column rank and has n_2 columns. Let g be an n-vector, partitioned conformally with A, as $g = (g_1^T \; g_2^T)^T$. Let W_{11} be a positive definite symmetric matrix of dimension $n_1 \times n_1$. Let W be the $n \times n$ positive definite matrix defined by*

$$W = \begin{pmatrix} W_{11} & 0 \\ 0 & I \end{pmatrix}. \tag{6.7}$$

Suppose $W = U(s(W))D(W)U(s(W))^T$ is the signature decomposition of W given by Lemma 4.1. Then, if $\pi(W_{11}, A_1, A_2, g_1, g_2)$ denotes the optimal solution of (6.1), it holds that

$$\pi(W_{11}, A_1, A_2, g_1, g_2) = \sum_{J \in \mathcal{J}^=(AU(s(W)))} \alpha_J(W)(AU_J(s(W)))^{-T} U_J(s(W))^T g,$$

with

$$\alpha_J(W) = \frac{\det(D_J(W)) \det(AU_J(s(W)))^2}{\sum_{K \in \mathcal{J}^=(AU(s(W)))} \det(D_K(W)) \det(AU_K(s(W)))^2},$$

and $\mathcal{J}^{=}(AU(s(W))) = \{J \in \mathcal{J}(AU(s(W))) : J^{=} \subseteq J \subseteq J^{+}\}$, *where* $\mathcal{J}(AU(s(W)))$ *is the set of column indices associated with nonsingular* $m \times m$ *submatrices of* $AU(s(W))$, $J^{=} = \{n_1 + 1, \ldots, n\}$, *and for* $t(i,j) = n(i-1) - \frac{i(i+1)}{2} + j$

$$J^{+} = \{\{1, \ldots, n\} \cup \{n + t(i,j) : 1 \leq i < j \leq n_1\}\}.$$

Proof. Let W be defined by (6.7), and let $W(\epsilon)$ be defined by (6.3). Since $W(\epsilon)$ and W differ only in some diagonal elements, we have $s(W(\epsilon)) = s(W)$, and hence $U(s(W(\epsilon))) = U(s(W))$. Moreover, (4.6) gives

$$d_j(W(\epsilon)) = d_j(W), \qquad j = 1, \ldots, n_1, \tag{6.8a}$$

$$d_j(W(\epsilon)) = \frac{1}{\epsilon} d_j(W) = \frac{1}{\epsilon}, \qquad j = n_1 + 1, \ldots, n, \tag{6.8b}$$

$$d_{n+t(i,j)}(W(\epsilon)) = d_{n+t(i,j)}(W), \qquad 1 \leq i < j \leq n_1, \tag{6.8c}$$

$$d_{n+t(i,j)}(W(\epsilon)) = d_{n+t(i,j)}(W) = 0, \quad 1 \leq i < j,\ n_1 + 1 \leq j \leq n. \tag{6.8d}$$

For $J \in \mathcal{J}(AU(s(W)))$, let i_2^J denote the number of elements of $J^{=}$ that are included in J. Note that $i_2^J \leq n_2$ for all $J \in \mathcal{J}(AU(s(W)))$. Also, since A_2 has full column rank and A has full row rank and is a submatrix of $AU(s)$, there is at least one $\bar{J} \in \mathcal{J}(AU(s(W)))$ with $i_2^{\bar{J}} = n_2$. Hence, since $\det(D_J(W(\epsilon)))$ occurs both in the numerator and in the denominator of $\alpha_J(W(\epsilon))$ of Corollary 4.2, we may replace $\det(D_J(W(\epsilon)))$ by $\epsilon^{n_2} \det(D_J(W(\epsilon)))$ in the definition of $\alpha_J(W(\epsilon))$ of Corollary 4.2. A combination of (6.11) and the identity $s(W(\epsilon)) = s(W)$ with Corollary 4.2 gives

$$(AW(\epsilon)A^T)^{-1} AW(\epsilon) g = \sum_{J \in \mathcal{J}(AU(s(W)))} \alpha_J(W(\epsilon)) (AU_J(s(W)))^{-T} U_J(s(W))^T g, \tag{6.9}$$

with

$$\alpha_J(W(\epsilon)) = \frac{\epsilon^{n_2} \det(D_J(W(\epsilon))) \det(AU_J(s(W)))^2}{\sum_{K \in \mathcal{J}(AU(s(W)))} \epsilon^{n_2} \det(D_K(W(\epsilon))) \det(AU_K(s(W)))^2}. \tag{6.10}$$

Now (6.8) gives

$$\epsilon^{n_2} \det(D_J(W(\epsilon))) = \epsilon^{n_2 - i_2^J} \det(D_J(W)). \tag{6.11}$$

Thus (6.11) gives

$$\epsilon^{n_2} \det(D_J(W(\epsilon))) = \det(D_J(W)) \quad \text{if} \quad J^{=} \subseteq J, \tag{6.12a}$$

$$\lim_{\epsilon \to 0^+} \epsilon^{n_2} \det(D_J(W(\epsilon))) = 0 \quad \text{if} \quad J^{=} \not\subseteq J. \tag{6.12b}$$

It also follows from (6.8) that $\det(D_J(W(\epsilon))) = \det(D_J(W)) = 0$ unless $J \subseteq J^{+}$. Hence, a combination of (6.10) and (6.12) gives

$$\lim_{\epsilon \to 0^+} \alpha_J(W(\epsilon)) = \frac{\det(D_J(W)) \det(AU_J(s(W)))^2}{\sum_{K \in \mathcal{J}^{=}(AU(s(W)))} \det(D_K(W)) \det(AU_K(s(W)))^2} \quad \text{if} \quad J^{=} \subseteq J, \tag{6.13a}$$

$$\lim_{\epsilon \to 0^+} \alpha_J(W(\epsilon)) = 0 \quad \text{if} \quad J^{=} \not\subseteq J. \tag{6.13b}$$

It follows from (6.13) that the limit of the right-hand side of (6.9) is well defined. Taking the limit of both sides of (6.9) using (6.13) and, as above, the fact that $\det(D_J(W)) = 0$ if $J \not\subseteq J^+$ gives the required result. □

From the result provided by Corollary 6.2, the analysis for the unconstrained case can be extended also to the equality-constrained case. The discussion is analogous to the unconstrained case, and we just summarize it briefly below, first for diagonal weight matrices and then for diagonally dominant weight matrices.

6.1. Diagonal weight matrices and equality constraints. To stress that the weight matrix is diagonal, we replace W_{11} by D_{11}. Similar to the unconstrained case, we associate a set of weight matrices with an $m \times n$ matrix A of full row rank, partitioned as $A = (A_1 \; A_2)$, where A_1 has n_1 columns and A_2 has n_2 columns. We denote by $\mathcal{D}_+^=(A_1)$ the set of $n_1 \times n_1$ positive definite diagonal weight matrices; i.e.,

$$\mathcal{D}_+^=(A_1) = \{D_{11} \in \mathbb{R}^{n_1 \times n_1} : D_{11} \text{ is diagonal and positive definite}\}. \tag{6.14}$$

When the weight matrix of (6.1) is diagonal, Corollary 6.2 takes a simpler form. The following corollary is the counterpart of Theorem 2.1.

COROLLARY 6.3. *Let A be an $m \times n$ matrix of full row rank, which is partitioned as $A = (A_1 \; A_2)$, where A_1 has n_1 columns and A_2 is of full column rank and has n_2 columns. Let g be an n-vector, partitioned conformally with A, as $g = (g_1^T \; g_2^T)^T$. Let D_{11} be a positive definite diagonal matrix of dimension $n_1 \times n_1$. Let D be the $n \times n$ positive definite matrix defined by $D = \operatorname{diag}(D_{11}, I)$. Then if $\pi(D_{11}, A_1, A_2, g_1, g_2)$ denotes the optimal solution of* (6.1) *with weight matrix D_{11}, it holds that*

$$\pi(D_{11}, A_1, A_2, g_1, g_2) = \sum_{J \in \mathcal{J}^=(A)} \left(\frac{\det(D_J) \det(A_J)^2}{\sum_{K \in \mathcal{J}^=(A)} \det(D_K) \det(A_K)^2} \right) A_J^{-T} g_J,$$

with

$$\mathcal{J}^=(A) = \{J \in \mathcal{J}(A) : J^= \subseteq J\},$$

where $\mathcal{J}(A)$ is the set of column indices associated with nonsingular $m \times m$ submatrices of A and $J^= = \{n_1 + 1, \ldots, n\}$.

Proof. If we denote by $D(D)$ the diagonal-dominance matrix associated with the diagonal matrix D defined by (4.6), we have $\det(D_J(D)) = \det(D_J)$ if $J \subseteq \{1, \ldots, n\}$ and $\det(D_J(D)) = 0$ if $J \not\subseteq \{1, \ldots, n\}$. Also, if $J \subseteq \{1, \ldots, n\}$, it follows from (4.4) that $AU_J(s(D)) = A_J$. Hence, the result follows from Corollary 6.2. □

Corollary 6.3 immediately provides the boundedness result corresponding to Corollary 2.2, as the following corollary shows.

COROLLARY 6.4. *Let A be an $m \times n$ matrix of full row rank, which is partitioned as $A = (A_1 \; A_2)$, where A_1 has n_1 columns and A_2 is of full column rank and has n_2 columns. Let g be an n-vector, partitioned conformally with A, as $g = (g_1^T \; g_2^T)^T$. Let $\mathcal{D}_+^=(A_1)$ be defined by* (6.14), *and for $D_{11} \in \mathcal{D}_+^=(A_1)$ let $\pi(D_{11}, A_1, A_2, g_1, g_2)$ denote the optimal solution of* (6.1) *with weight matrix D_{11}. Then*

$$\sup_{D_{11} \in \mathcal{D}_+^=(A_1)} \|\pi(D_{11}, A_1, A_2, g_1, g_2)\| = \max_{J \in \mathcal{J}^=(A)} \|A_J^{-T} g_J\| \quad \textit{and}$$

$$\sup_{D_{11} \in \mathcal{D}_+^=(A_1)} \max_{\|g\|=1} \|\pi(D_{11}, A_1, A_2, g_1, g_2)\| = \max_{J \in \mathcal{J}^=(A)} \|A_J^{-T}\|,$$

with

$$\mathcal{J}^{=}(A) = \{J \in \mathcal{J}(A) : J^{=} \subseteq J\},$$

where $\mathcal{J}(A)$ *is the set of column indices associated with nonsingular* $m \times m$ *submatrices of* A.

Proof. The proof is analogous to the proof of Corollary 2.2 with $\mathcal{J}(A)$ replaced by $\mathcal{J}^{=}(A)$. □

Finally, by considering the signature of the residual vector r_1 of (6.2), Theorem 2.4 is immediately generalized to the constrained case.

COROLLARY 6.5. *Let* A *be an* $m \times n$ *matrix of full row rank, which is partitioned as* $A = (A_1 \; A_2)$, *where* A_1 *has* n_1 *columns and* A_2 *is of full column rank and has* n_2 *columns. Let* g *be an* n*-vector, partitioned conformally with* A, *as* $g = (g_1^T \; g_2^T)^T$. *Let* $\mathcal{D}_+^{=}(A_1)$ *be defined by* (6.14), *and for* $D_{11} \in \mathcal{D}_+^{=}(A_1)$ *let* $\pi(D_{11}, A_1, A_2, g_1, g_2)$ *denote the optimal solution of* (6.1) *with weight matrix* D_{11}. *Let*

$$\Pi^{\mathcal{D}_+^{=}(A_1)}(A_1, A_2, g_1, g_2) = \{\pi(D_{11}, A_1, A_2, g_1, g_2) : D_{11} \in \mathcal{D}_+^{=}(A_1)\}.$$

If $\mathcal{S}_1$ *denotes the set of* n_1*-dimensional signature vectors, then*

$$\Pi^{\mathcal{D}_+^{=}(A_1)}(A_1, A_2, g_1, g_2) = \cup_{s_1 \in \mathcal{S}_1} \Pi^{s_1}(A_1, A_2, g_1, g_2),$$

where

$$\Pi^{s_1}(A_1, A_2, g_1, g_2) = \left\{ \pi : \begin{array}{l} S_1 u_1 + A_1^T \pi = g_1, \;\; A_2^T \pi = g_2, \\ A_1 S_1 v_1 + A_2 v_2 = 0, \;\; u_1 > 0, \;\; v_1 > 0 \end{array} \right\},$$

with $S_1 = \operatorname{diag}(s_1)$.

Proof. Suppose $\pi \in \Pi^{\mathcal{D}_+^{=}(A_1)}(A_1, A_2, g_1, g_2)$. Then Lemma 6.1 implies that there are r_1 and r_2 such that π, r_1, and r_2 solve the augmented system (6.2) given by

$$\begin{pmatrix} D_{11}^{-1} & 0 & A_1^T \\ 0 & 0 & A_2^T \\ A_1 & A_2 & 0 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ \pi \end{pmatrix} = \begin{pmatrix} g_1 \\ g_2 \\ 0 \end{pmatrix}. \tag{6.15}$$

Let $s_1 = \operatorname{sign}(r_1)$ and $S_1 = \operatorname{diag}(s_1)$. Define v_1 with components $(v_1)_i = |(r_1)_i|$ if $(r_1)_i \neq 0$ and $(v_1)_i = 1$ if $(r_1)_i = 0$. Then $S_1 v_1 = r_1$ and $v_1 > 0$. Let $v_2 = r_2$, and let $u_1 = D_{11}^{-1} v_1$. Then $S_1 u_1 = S_1 D_{11}^{-1} v_1 = D_{11}^{-1} S_1 v_1 = D_{11}^{-1} r_1$ and $u_1 > 0$. Hence, we have

$$\begin{pmatrix} S_1 & 0 & 0 & A_1^T \\ 0 & 0 & 0 & A_2^T \\ 0 & A_1 S_1 & A_2 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ v_1 \\ v_2 \\ \pi \end{pmatrix} = \begin{pmatrix} g_1 \\ g_2 \\ 0 \end{pmatrix}, \tag{6.16}$$

with $u_1 > 0$ and $v_1 > 0$. Consequently, $\pi \in \Pi^{s_1}(A_1, A_2, g_1, g_2)$ for $s_1 = \operatorname{sign}(r_1)$, giving $\Pi^{\mathcal{D}_+^{=}(A_1)}(A_1, A_2, g_1, g_2) \subseteq \cup_{s_1 \in \mathcal{S}_1} \Pi^{s_1}(A_1, A_2, g_1, g_2)$.

Conversely, suppose that $\pi \in \Pi^{s_1}(A_1, A_2, g_1, g_2)$ for some signature vector s_1. Then there are u_1, v_1, and v_2, with $u_1 > 0$ and $v_1 > 0$ such that π, u_1, v_1, and v_2 satisfy (6.16) with $S_1 = \operatorname{diag}(s_1)$. Let $r_1 = S_1 v_1$, $r_2 = v_2$, and $D_{11} = V_1 U_1^{-1}$, with $U_1 = \operatorname{diag}(u_1)$ and $V_1 = \operatorname{diag}(v_1)$. Since v_1 and u_1 are positive vectors, D_{11} is well

defined and positive definite, and we have $S_1 u_1 = S_1 U_1 V_1^{-1} v_1 = U_1 V_1^{-1} S_1 v_1 = D_{11}^{-1} r_r$. Hence, this choice of D_{11} gives π, r_1, and r_2 as the solution of (6.15). Since s_1 is an arbitrary element in $\mathcal{S}_1$, we conclude that $\cup_{s_1 \in \mathcal{S}_1} \Pi^{s_1}(A_1, A_2, g_1, g_2) \subseteq \Pi^{\mathcal{D}_+^=(A_1)}(A_1, A_2, g_1, g_2)$. □

Wei [25] generalizes the results of Stewart, discussed in §2, to cover also the constrained linear least-squares problem (6.1) for diagonal weight matrices.

6.2. Diagonally dominant weight matrices and equality constraints. As in the diagonal case, we associate a set of weight matrices with an $m \times n$ matrix A of full row rank, partitioned as $A = (A_1 \; A_2)$, where A_1 has n_1 columns and A_2 has n_2 columns. We denote by $\mathcal{W}_+^=(A_1)$ the set of $n_1 \times n_1$ positive definite symmetric diagonally dominant weight matrices; i.e.,

$$\mathcal{W}_+^=(A_1) = \left\{ W_{11} \in \mathbb{R}^{n_1 \times n_1} : \begin{array}{l} W_{11} \text{ is symmetric, diagonally dominant,} \\ \text{and positive definite} \end{array} \right\}. \tag{6.17}$$

The following corollary gives the result for the constrained case corresponding to the result of Theorem 5.1 for the unconstrained case. For brevity, we sacrifice some precision and only consider the closure of the sets.

COROLLARY 6.6. *Let A be an $m \times n$ matrix of full row rank, which is partitioned as $A = (A_1 \; A_2)$, where A_1 has n_1 columns and A_2 is of full column rank and has n_2 columns. Let g be an n-vector, partitioned conformally with A, as $g = (g_1^T \; g_2^T)^T$. Let $\mathcal{W}_+^=(A_1)$ be defined by* (6.17), *and for $W_{11} \in \mathcal{W}_+^=(A_1)$ let $\pi(W_{11}, A_1, A_2, g_1, g_2)$ denote the optimal solution of* (6.1). *Let $\mathcal{S}_1$ denote the set of dense $\frac{n_1(n_1-1)}{2}$-dimensional signature vectors, and for $s_1 \in \mathcal{S}_1$ let $U_1(s_1)$ be the associated signature matrix defined by* (4.4). *Let*

$$\Pi^{\mathcal{W}_+^=(A_1)}(A_1, A_2, g_1, g_2) = \{\pi(W_{11}, A_1, A_2, g_1, g_2) : W_{11} \in \mathcal{W}_+^=(A_1)\}.$$

Then

$$\bar{\Pi}^{\mathcal{W}_+^=(A_1)}(A_1, A_2, g_1, g_2) = \cup_{s_1 \in \mathcal{S}_1} \bar{\Pi}^{\mathcal{D}_+^=(A_1 U_1(s_1))}(A_1 U_1(s_1), A_2, U_1(s_1)^T g_1, g_2),$$

where the set $\Pi^{\mathcal{D}_+^=(A_1 U_1(s_1))}(A_1 U_1(s_1), A_2, U_1(s_1)^T g_1, g_2)$ is given by Corollary 6.5, *with A_1 replaced by $A_1 U_1(s_1)$, g_1 replaced by $U_1(s_1)^T g_1$, and n_1 replaced by $\frac{n_1(n_1+1)}{2}$. Here $\bar{\Pi}^{\mathcal{W}_+^=(A_1)}(A_1, A_2, g_1, g_2)$ denotes the closure of the set $\Pi^{\mathcal{W}_+^=(A_1)}(A_1, A_2, g_1, g_2)$, and similarly $\bar{\Pi}^{\mathcal{D}_+^=(A_1 U_1(s_1))}(A_1 U_1(s_1), A_2, U_1(s_1)^T g_1, g_2)$ denotes the closure of the set $\Pi^{\mathcal{D}_+^=(A_1 U_1(s_1))}(A_1 U_1(s_1), A_2, U_1(s_1)^T g_1, g_2)$.*

Proof. The proof is analogous to the proof of Theorem 5.1, replacing the optimality conditions (1.2) by (6.4), upon observing that the distinction of the sets considered in Corollary 2.5 is of no significance, since only the closures are considered here. □

Finally, the boundedness properties corresponding to Corollary 5.3 can also be derived.

COROLLARY 6.7. *Let A be an $m \times n$ matrix of full row rank, which is partitioned as $A = (A_1 \; A_2)$, where A_1 has n_1 columns and A_2 is of full column rank and has n_2 columns. Let g be an n-vector, partitioned conformally with A, as $g = (g_1^T \; g_2^T)^T$. Let $\mathcal{W}_+^=(A_1)$ be defined by* (6.17), *and for $W_{11} \in \mathcal{W}_+^=(A_1)$ let $\pi(W_{11}, A_1, A_2, g_1, g_2)$ denote the optimal solution of* (6.1). *Let $\mathcal{S}$ denote the set of dense $\frac{n(n-1)}{2}$-dimensional signature vectors, and for $s \in \mathcal{S}$ let $U(s)$ be the associated signature matrix defined by* (4.4). *Then*

$$\sup_{W_{11} \in \mathcal{W}_+^=(A_1)} \|\pi(W_{11}, A_1, A_2, g_1, g_2)\| = \max_{s \in \mathcal{S}} \max_{J \in \mathcal{J}^=(AU(s))} \|(AU_J(s))^{-T} U_J(s)^T g\|$$

and

$$\sup_{W_{11}\in\mathcal{W}_+^=(A_1)} \max_{\|g\|=1} \|\pi(W_{11},A_1,A_2,g_1,g_2)\| = \max_{s\in\mathcal{S}} \max_{J\in\mathcal{J}^=(AU(s))} \|(AU_J(s))^{-T}U_J(s)^T\|,$$

with $\mathcal{J}^=(AU(s)) = \{J \in \mathcal{J}(AU(s)) : J^= \subseteq J \subseteq J^+\}$, *where* $\mathcal{J}(AU(s))$ *is the set of column indices associated with nonsingular* $m \times m$ *submatrices of* $AU(s)$, $J^= = \{n_1+1,\ldots,n\}$ *and, for* $t(i,j) = n(i-1) - \frac{i(i+1)}{2} + j$,

$$J^+ = \{\{1,\ldots,n\} \cup \{n + t(i,j) : 1 \le i < j \le n_1\}\}.$$

Proof. The proof is analogous to the proof of Corollary 5.3 with the set $\mathcal{J}(AU(s))$ replaced by $\mathcal{J}^=(AU(s))$. □

7. Discussion. Returning to the nondiagonal example (3.1) of Stewart [21], our analysis shows that the unboundedness of the norm predicted by (3.3) can only occur because the absolute value of the off-diagonal element of $W(\delta,\epsilon)$ is at least as large as the smallest diagonal elements. If we let δ and ϵ be such that $W(\delta,\epsilon)$ from (3.1) is diagonally dominant, (3.1) gives $|\epsilon(1-\delta)| \le \min\{|1+\delta\epsilon^2|, |\epsilon^2+\delta|\}$, and (3.2) gives

$$|(AW(\delta,\epsilon)A^T)^{-1}AW(\delta,\epsilon)| \le \begin{pmatrix} 1 & 1 \end{pmatrix}, \tag{7.1}$$

independently of δ. It follows from (3.1) and (3.2) that for the norm of the matrix $(AW(\delta,\epsilon)A^T)^{-1}AW(\delta,\epsilon)$ to remain bounded it suffices if the off-diagonal element of $W(\delta,\epsilon)$ is bounded in comparison with the diagonal elements, and it is not necessary for $W(\delta,\epsilon)$ to be diagonally dominant. Hence, it seems that one could expect to easily find a larger class of matrices for which the norm remains bounded. However, let

$$A = \begin{pmatrix} 0 & 1 \end{pmatrix} \quad \text{and} \quad W(\epsilon) = \begin{pmatrix} 2+\epsilon+\epsilon^2 & \epsilon+\epsilon^2 \\ \epsilon+\epsilon^2 & \epsilon^2 \end{pmatrix}. \tag{7.2}$$

Note that $\det(W(\epsilon)) = \epsilon^2(1-\epsilon)$. It is straightforward to verify that $W(\epsilon)$ has positive diagonal elements if $\epsilon \neq 0$, and hence $W(\epsilon)$ is positive definite for $\epsilon < 0$ and $\epsilon \in (0,1)$. Moreover, as $\epsilon \to 0$, $W(\epsilon)$ converges to a matrix which is diagonal (but it is not diagonally dominant if $\epsilon \in (-\frac{1}{2}, 0)$ or $\epsilon > 0$), and

$$(AW(\epsilon)A^T)^{-1}AW(\epsilon) = \begin{pmatrix} 1+\dfrac{1}{\epsilon} & 1 \end{pmatrix}. \tag{7.3}$$

Hence, we have $\lim_{\epsilon\to 0} \|(AW(\epsilon)A^T)^{-1}AW(\epsilon)\| = \infty$. Consequently, although it is quite possible that there is a larger class of matrices for which this finite-norm property holds, it is not immediate.

This unboundedness of $\lim_{\epsilon\to 0} \|(AW(\epsilon)A^T)^{-1}AW(\epsilon)\|$ is predicted by Corollary 4.2. For $\epsilon \ge 0$ and $\epsilon \le -1$, if the signature decomposition of Lemma 4.1 is denoted by $W(\epsilon) = U^+(\epsilon)D^+(\epsilon)U^+(\epsilon)^T$, then

$$U^+(\epsilon) = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \quad \text{and} \quad D^+(\epsilon) = \begin{pmatrix} 2 & 0 & 0 \\ 0 & -\epsilon & 0 \\ 0 & 0 & \epsilon+\epsilon^2 \end{pmatrix}. \tag{7.4}$$

Note that $D^+(\epsilon)$ is not positive semidefinite if $\epsilon > 0$. This is predicted by (4.6) since $W(\epsilon)$ is positive definite but not diagonally dominant if $\epsilon \in (0,1)$. For $\epsilon > 0$ or $\epsilon \leq -1$, upon observing that $AU^+_{\{1\}}(\epsilon) = 0$, Corollary 4.2 and (7.4) give

$$
\begin{aligned}
(AW(\epsilon)A^T)^{-1}AW(\epsilon) &= \frac{\det(D^+_{\{2\}}(\epsilon))\det(AU^+_{\{2\}}(\epsilon))^2}{\sum_{J\in\{\{2\},\{3\}\}}\det(D^+_J(\epsilon))\det(AU^+_J(\epsilon))^2}(AU^+_{\{2\}}(\epsilon))^{-T}U^+_{\{2\}}(\epsilon)^T \\
&\quad + \frac{\det(D^+_{\{3\}}(\epsilon))\det(AU^+_{\{3\}}(\epsilon))^2}{\sum_{J\in\{\{2\},\{3\}\}}\det(D^+_J(\epsilon))\det(AU^+_J(\epsilon))^2}(AU^+_{\{3\}}(\epsilon))^{-T}U^+_{\{3\}}(\epsilon)^T \\
&= -\frac{1}{\epsilon}\begin{pmatrix} 0 & 1 \end{pmatrix} + \left(1+\frac{1}{\epsilon}\right)\begin{pmatrix} 1 & 1 \end{pmatrix}. \qquad (7.5)
\end{aligned}
$$

For $\epsilon \in (-1,0)$, the signature decomposition of Lemma 4.1 gives $W(\epsilon) = U^-(\epsilon)D^-(\epsilon)U^-(\epsilon)^T$ with

$$
U^-(\epsilon) = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix} \quad \text{and} \quad D^-(\epsilon) = \begin{pmatrix} 2(1+\epsilon+\epsilon^2) & 0 & 0 \\ 0 & \epsilon+2\epsilon^2 & 0 \\ 0 & 0 & -\epsilon-\epsilon^2 \end{pmatrix},
$$

and similarly to (7.5) we obtain

$$
(AW(\epsilon)A^T)^{-1}AW(\epsilon) = \left(2+\frac{1}{\epsilon}\right)\begin{pmatrix} 0 & 1 \end{pmatrix} - \left(1+\frac{1}{\epsilon}\right)\begin{pmatrix} -1 & 1 \end{pmatrix}. \qquad (7.6)
$$

Since $W(\epsilon)$ is positive definite but not diagonally dominant if $\epsilon \in (-\frac{1}{2},0)$ or $\epsilon \in (0,1)$, (4.6) implies that $D^+(\epsilon)$ is not positive semidefinite for $\epsilon \in (0,1)$ and $D^-(\epsilon)$ is not positive semidefinite for $\epsilon \in (-\frac{1}{2},0)$. Hence, the linear combination of Corollary 4.2 is not necessarily a convex combination. All we know is that the coefficients sum up to one. This is manifested in (7.5) and (7.6), where in both cases the two nonzero coefficients have different signs and tend to infinity in magnitude, as ϵ tends to zero from plus and minus, respectively.

8. Summary and further research. It has been shown that the properties of the least-squares solution $(AWA^T)^{-1}AWg$ of (1.1) when W belongs to the set of diagonal and positive definite matrices can be extended also to the larger class of matrices where W is symmetric, diagonally dominant, and positive semidefinite and AWA^T is positive definite. Similar results have been obtained for the matrix operator $(AWA^T)^{-1}AW$.

In essence, the signature decomposition of §4 has provided a tool for deriving the results of this paper for diagonal matrices but in a space where the dimension of the diagonal matrices has been expanded from n to $\frac{n(n+1)}{2}$. The results for the diagonal case also apply for the diagonal-dominant case, taking into account the additional complexity induced by the dense signature vectors associated with the off-diagonal elements of the weight matrices. The decomposition has also been used to generalize the results to the case of infinite diagonal weights, i.e., equality constraints. Finally, it has also given some insight into why the boundedness properties do not hold for general symmetric positive definite weight matrices.

An interesting line of research is to investigate if this insight into the boundedness properties of the solution of the least-squares problem (1.1) can be used to analyze the

stability of solving the normal equations (1.2) by Cholesky factorization for weight matrices that are positive definite and diagonal or diagonally dominant.

Acknowledgments. I thank Philip Gill for helpful comments on the manuscript and Walter Murray for pointing out that Theorem 2.1 was given by Dikin [8], as stated in the paper of Vanderbei and Lagarias [23]. Thanks also to Åke Björck and Lars Eldén for bibliographical assistance and to two referees for their careful reading of the manuscript and helpful comments.

REFERENCES

[1] R. H. Bartels, G. H. Golub, and M. A. Saunders, *Numerical techniques in mathematical programming*, in Nonlinear Programming, J. B. Rosen, O. L. Mangasarian, and K. Ritter, eds., Academic Press, New York, 1970, pp. 123–176.

[2] A. Ben-Israel, *A volume associated with $m \times n$ matrices*, Linear Algebra Appl., 167 (1992), pp. 87–111.

[3] A. Ben-Tal and M. Teboulle, *A geometric property of the least squares solution of linear equations*, Linear Algebra Appl., 139 (1990), pp. 165–170.

[4] L. Berg, *Three results in connection with inverse matrices*, Linear Algebra Appl., 84 (1986), pp. 63–77.

[5] Å. Björck, *Pivoting and stability in the augmented system method*, in Numerical Analysis 1991, Proc. 14th Dundee Conference, Dundee, Scotland, UK, D. F. Griffiths and G. A. Watson, eds., 1991, pp. 1–16.

[6] Å. Björck and C. C. Paige, *Solution of augmented linear systems using orthogonal factorizations*, BIT, 34 (1994), pp. 1–24.

[7] I. I. Dikin, *Iterative solution of problems of linear and quadratic programming*, Soviet Math. Dokl., 8 (1967), pp. 674–675.

[8] ———, *On the speed of an iterative process*, Upravlyaemye Sistemy, 12 (1974), pp. 54–60.

[9] I. S. Duff, N. I. M. Gould, J. K. Reid, J. A. Scott, and K. Turner, *The Factorization of Sparse Symmetric Indefinite Matrices*, Tech. Report CSS 236, Computer Science and Systems Division, AERE Harwell, Oxford, England, 1989.

[10] A. Forsgren, P. E. Gill, and J. R. Shinnerl, *Stability of symmetric ill-conditioned systems arising in interior methods for constrained optimization*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 187–211.

[11] P. E. Gill, W. Murray, and M. H. Wright, *Numerical Linear Algebra and Optimization*, Vol. 1, Addison-Wesley Publishing Company, Reading, MA, 1991.

[12] P. E. Gill, M. A. Saunders, and J. R. Shinnerl, *On the stability of Cholesky factorization for symmetric quasidefinite systems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 35–46.

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

[14] C. C. Gonzaga, *Path-following methods for linear programming*, SIAM Rev., 34 (1992), pp. 167–224.

[15] N. I. M. Gould, *On practical conditions for the existence and uniqueness of solutions to the general equality quadratic programming problem*, Math. Programming, 32 (1985), pp. 90–99.

[16] M. Gullikson and P.-Å. Wedin, *Modifying the QR decomposition to constrained and weighted linear least squares*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1298–1313.

[17] M. Hanke and M. Neumann, *The geometry of the set of scaled projections*, Linear Algebra Appl., 190 (1993), pp. 137–148.

[18] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

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

[20] D. P. O'Leary, *On bounds for scaled projections and pseudoinverses*, Linear Algebra Appl., 132 (1990), pp. 115–117.

[21] G. W. Stewart, *On scaled projections and pseudoinverses*, Linear Algebra Appl., 112 (1989), pp. 189–193.

[22] M. J. Todd, *A Dantzig-Wolfe-like variant of Karmarkar's interior-point linear programming algorithm*, Oper. Res., 38 (1990), pp. 1006–1018.

[23] R. J. VANDERBEI AND J. C. LAGARIAS, *Dikin's convergence result for the affine-scaling algorithm*, Contemp. Math., 114 (1990), pp. 109–119.

[24] S. A. VAVASIS, *Stable numerical algorithms for equilibrium systems*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1108–1131.

[25] M. WEI, *On the boundedness of weighted pseudoinverses and oblique projections*, Tech. Report, Department of Mathematics, East China Normal University, Shanghai, China, 1993.

[26] M. H. WRIGHT, *Interior methods for constrained optimization*, in Acta Numerica 1992, A. Iserles, ed., Cambridge University Press, New York, 1992, pp. 341–407.

[27] S. J. WRIGHT, *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, 1994.

[28] ———, *Stability of linear equations solvers in interior-point methods*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1287–1307.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 4, pp. 789–821, October 1996

006

DEFLATION TECHNIQUES FOR AN IMPLICITLY RESTARTED ARNOLDI ITERATION*

R. B. LEHOUCQ† AND D. C. SORENSEN‡

Abstract. A deflation procedure is introduced that is designed to improve the convergence of an implicitly restarted Arnoldi iteration for computing a few eigenvalues of a large matrix. As the iteration progresses, the Ritz value approximations of the eigenvalues converge at different rates. A numerically stable scheme is introduced that implicitly deflates the converged approximations from the iteration. We present two forms of implicit deflation. The first, a *locking* operation, decouples converged Ritz values and associated vectors from the active part of the iteration. The second, a *purging* operation, removes unwanted but converged Ritz pairs. Convergence of the iteration is improved and a reduction in computational effort is also achieved. The deflation strategies make it possible to compute multiple or clustered eigenvalues with a single vector restart method. A block method is not required. These schemes are analyzed with respect to numerical stability, and computational results are presented.

Key words. Arnoldi method, Lanczos method, eigenvalues, deflation, implicit restarting

AMS subject classifications. 65F15, 65G05

1. Introduction.

The Arnoldi method is an efficient procedure for approximating a subset of the eigensystem of a large sparse $n \times n$ matrix A. The Arnoldi method is a generalization of the Lanczos process and reduces to that method when the matrix A is symmetric. After k steps, the algorithm produces an upper Hessenberg matrix H_k of order k. The eigenvalues of this small matrix H_k are used to approximate a subset of the eigenvalues of the large matrix A. The matrix H_k is an orthogonal projection of A onto a particular *Krylov* subspace, and the eigenvalues of H_k are usually called *Ritz values* or *Ritz approximations*.

There are a number of numerical difficulties with Arnoldi/Lanczos methods. In [34] a variant of this method was developed to overcome these difficulties. This technique, the implicitly restarted Arnoldi iteration (IRA iteration), may be viewed as a truncation of the standard implicitly shifted QR iteration. This connection will be reviewed during the course of the paper. Because of this connection, an IRA iteration shares a number of the QR iteration's desirable properties. These include the well-understood deflation rules of the QR iteration. These deflation techniques are extremely important with respect to the convergence and stability of the QR iteration. Deflation rules have contributed greatly to the emergence of the practical QR algorithm as the method of choice for computing the eigensystem of dense matrices. In particular, the deflation rules allow the QR iteration to compute multiple and clustered eigenvalues.

This paper introduces deflation schemes that may be used within an IRA iteration. This iteration is designed to compute a selected subset of the spectrum of A such as

* Received by the editors February 10, 1995; accepted for publication (in revised form) by A. Greenbaum November 8, 1995. This work was supported in part by ARPA (U.S. Army ORA4466.01), by U.S. Department of Energy contract DE-FG0f-91ER25103, and by National Science Foundation cooperative agreement CCR-9120008.

† Computational and Applied Mathematics Department, Rice University, Houston, TX 77251 (lehoucq@rice.edu). Current address: Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439 (lehoucq@mcs.anl.gov) (http://www.mcs.anl.gov/home/lehoucq/index.html).

‡ Computational and Applied Mathematics Department, Rice University, Houston, TX 77251 (sorensen@rice.edu).

the k eigenvalues of largest real part. We refer to this selected subset as *wanted* and the remainder of the spectrum as *unwanted*. As the iteration progresses some of the Ritz approximations to eigenvalues of A may converge long before the entire set of wanted eigenvalues has been computed. These converged Ritz values may be part of the wanted or the unwanted portion of the spectrum. In either case, it is desirable to *deflate* the converged Ritz values and corresponding Ritz vectors from the unconverged portion of the factorization. If the converged Ritz value is wanted, it is necessary to keep it in the subsequent Arnoldi factorizations. This is called *locking*. If the converged Ritz value is unwanted then it must be removed from the current and subsequent Arnoldi factorizations. This is called *purging*. These notions will be made precise during the course of the paper. For the moment we note that the advantages of a numerically stable deflation strategy include

- reduction of the *working* size of the desired invariant subspace,
- prevention of the effects of the forward instability of the Lanczos and QR algorithms [27, 39],
- the ability to determine clusters of nearby eigenvalues without need for a block Arnoldi method [18, 32, 33].

The fundamentals of the Arnoldi algorithm are introduced in §2 as well as the determination of Ritz value convergence. The IRA iteration is reviewed in §3. Deflating within the IRA iteration is examined in §4. The deflation scheme for converged Ritz values is presented in §5. The practical issues associated with our deflation scheme are examined in §6. These include block generalizations of the ideas examined in §5 for dealing with a number of Ritz values simultaneously and avoiding the use of complex arithmetic when a complex conjugate pair of Ritz values converges. An error analysis of the deflated process is presented in §7. A brief survey of and comparisons with other deflation strategies is given in §8. An interesting connection with the various algorithms used to reorder a Schur form of matrix is presented in §9. Numerical results are presented in §10.

Capital and lowercase letters denote matrices and vectors whereas lowercase Greek letters denote scalars. The jth canonical basis vector is denoted by e_j. The norms used are the Euclidean and Frobenius ones denoted by $\|\cdot\|$ and $\|\cdot\|_F$, respectively. The range of a matrix A is denoted by $\mathcal{R}(A)$.

2. The Arnoldi factorization. Arnoldi's method [1] is an orthogonal projection method for approximating a subset of the eigensystem of a general square matrix. The method builds, step by step, an orthogonal basis for the *Krylov* space

$$\mathcal{K}_k(A, v_1) \equiv \text{span}\{v_1, Av_1, \ldots, A^{k-1}v_1\}$$

for A generated by the vector v_1. The original algorithm in [1] was designed to reduce a dense matrix to upper Hessenberg form. However, the method only requires knowledge of A through matrix–vector products, and its ultimate value as a technique for approximating a few eigenvalues of a large sparse matrix was soon realized. When the matrix A is symmetric, the procedure reduces to the Lanczos method [22].

Over a decade of research was devoted to understanding and overcoming the numerical difficulties of the Lanczos method [26]. Development of the Arnoldi method lagged behind due to the inordinate computational and storage requirements associated with the original method when a large number of steps are required for convergence. Not only is more storage required for V_k and H_k when A is nonsymmetric, but in general more steps are required to compute the desired Ritz value approximations. An explicitly restarted Arnoldi iteration (ERA iteration) was introduced

by Saad [30] to overcome these difficulties. The idea is based on similar ones developed for the Lanczos process by Paige [25], Cullum and Donath [10], and Golub and Underwood [17]. Karush proposed the first example of a restarted iteration in [21].

After k steps, the Arnoldi algorithm computes a truncated factorization,

$$AV_k = V_k H_k + f_k e_k^T, \tag{2.1}$$

of $A \in \mathbf{R}^{n\times n}$ into upper Hessenberg form where $V_k^T V_k = I_k$. The vector f_k is the residual and is orthogonal to the columns of V_k. The matrix $H_k \in \mathbf{R}^{k\times k}$ is an upper Hessenberg matrix that is the orthogonal projection of A onto $\mathcal{R}(V_k) \equiv \mathcal{K}_k(A, v_1)$.

The following procedure shows how the factorization is extended from length k to $k+p$.

ALGORITHM 2.1.

function $[V_{k+p}, H_{k+p}, f_{k+p}]$ = `Arnoldi` (A, V_k, H_k, f_k, k, p)

Input: $AV_k - V_k H_k = f_k e_k^T$ *with* $V_k^T V_k = I_k$, *&* $V_k^T f_k = 0$.

Output: $AV_{k+p} - V_{k+p} H_{k+p} = f_{k+p} e_{k+p}^T$ *with* $V_{k+p}^T V_{k+p} = I_{k+p}$, *&* $V_{k+p}^T f_{k+p} = 0$.

1. *For* $j = 1, 2 \ldots p$
2. $\beta_{k+j} \leftarrow \|f_{k+j-1}\|$; *if* $\beta_{k+j} = 0$ *then stop*;
3. $v_{k+j} \leftarrow f_{k+j-1}\beta_{k+j}^{-1}$; $V_{k+j} \leftarrow \begin{bmatrix} V_{k+j-1} & v_{k+j} \end{bmatrix}$;
4. $w \leftarrow Av_{k+j}$;
5. $h_{k+j} \leftarrow V_{k+j-1}^T w$; $\alpha_{k+j} \leftarrow v_{k+j}^T w$;
6. $H_{k+j} \leftarrow \begin{bmatrix} H_{k+j-1} & h_{k+j} \\ \beta_{k+j} e_{k+j-1}^T & \alpha_{k+j} \end{bmatrix}$
7. $f_{k+j} \leftarrow w - V_{k+j-1} h_{k+j} - v_{k+j}\alpha_{k+j}$;

If $k = 0$ then $V_1 = v_1$ represents the initial vector. In order to ensure that $V_k^T f_k \approx 0$ in finite precision arithmetic, the above algorithm requires some form of reorthogonalization at step 7; see Chapter 7 of [23].

In exact arithmetic, the algorithm continues until $f_k = 0$ for some $k \leq n$. All of the intermediate Hessenberg matrices H_j are *unreduced* for $j \leq k$. A Hessenberg matrix is said to be unreduced if all of its main subdiagonal elements are nonzero. The residual vanishes at the first step k such that $\dim\mathcal{K}_{k+1}(A, v_1) = k$ and hence is guaranteed to vanish for some $k \leq n$. The following result indicates when an exact truncated factorization occurs. This is desirable since the columns of V_k form a basis for an invariant subspace and the eigenvalues of H_k are a subset of those of A.

THEOREM 2.2. *Let equation* (2.1) *define a k step Arnoldi factorization of A, with* H_k *unreduced. Then* $f_k = 0$ *if and only if* $v_1 = Q_k y$ *where* $AQ_k = Q_k R_k$ *with* $Q_k^T Q_k = I_k$, *and* R_k *is an upper quasi-triangular matrix of order k.*

Proof. See Chapter 2 of [23] or [34] for a proof based on the Jordan canonical form. □

In Theorem 2.2, the span of the k columns of Q_k represents an invariant subspace for A. The matrix equation $AQ_k = Q_k R_k$ is a partial real Schur decomposition of order k for A. The diagonal blocks of R_k contain the eigenvalues of A. The complex conjugate pairs are in blocks of order-2 and the real eigenvalues are on the diagonal of R_k, respectively. In particular, the theorem gives that if the initial vector is a linear combination of k linearly independent eigenvectors then the kth residual vector vanishes. It is therefore desirable to devise a method that forces the starting vector v_1 to lie in the invariant subspace associated with the wanted eigenvalues.

The algorithms of this paper are appropriate when the order of A is so large that storage and computational requirements prohibit completion of the algorithm

that produces V_n and H_n. We also remark that working in finite precision arithmetic generally removes the possibility of the computed residual ever vanishing exactly.

As the norm of f_k decreases, the eigenvalues of H_k become better approximations to those of A. Experience indicates that $\|f_k\|$ rarely becomes small, let alone zero. However, as the order of H_k increases, certain eigenvalues of H_k may emerge as excellent estimates to eigenvalues of A. When an eigenvalue H_k is sufficiently near one of A, we will say that convergence occurred. Since the interest is in a small subset of the eigensystem of A, alternate criteria that allow termination for $k \ll n$ are needed. Let $H_k y = y\theta$ where $\|y\| = 1$. Define the vector $x = V_k y$ to be a *Ritz vector* and θ to be *Ritz value*. Then

$$\begin{aligned} \|AV_k y - V_k H_k y\| &= \|Ax - x\theta\| \\ &= \|f_k\| \, |e_k^T y| \end{aligned} \tag{2.2}$$

indicates that if the last component of an eigenvector for H_k is small the Ritz pair (x, θ) is an approximation to an eigenpair of A. This pair is exact for a nearby problem: it is easily shown that $(A + E)x = x\theta$ with $E = -(e_k^T y) f_k x^H$. The advantage of using the *Ritz estimate* (2.2) is to avoid explicit formation of the quantity $AV_k y - V_k y\theta$ when accessing the numerical accuracy of an approximate eigenpair. Recent work by Chatelin [8], Chatelin and Fraysée [9], and Godet-Thobie [14] suggests that when A is highly non-normal, the size of $e_k^T y$ is not an appropriate guide for detecting convergence. If the relative *departure from normality* defined by the Henrici number $\|AA^T - A^T A\|_F / \|A^2\|_F$ is large, the matrix A is considered highly non-normal. Assuming that A is diagonalizable, a large Henrici number implies that the basis of eigenvectors is ill conditioned [8]. Bennani and Braconnier compare the use of the Ritz estimate and direct residual $\|Ax - x\theta\|$ in Arnoldi algorithms [4]. They suggest normalizing the Ritz estimate by the norm of A, resulting in a stopping criteria based on the *backward* error. The backward error is defined as the smallest, in norm, perturbation ΔA such that the Ritz pair is an eigenpair for $A + \Delta A$. Scott [33] presents a lucid account of the many issues involved in determining stopping criteria for the unsymmetric problem.

3. The implicitly restarted Arnoldi iteration. Theorem 2.2 motivates the selection of a starting vector that will lead to the construction of an approximate basis for the desired invariant subspace of A. The best possible starting vector would be a linear combination of a Schur basis for the desired invariant subspace. The IRA iteration iteratively restarts the Arnoldi factorization with the goal of forcing the starting vector closer and closer to the desired invariant subspace. The scheme is called *implicit* because the updating of the starting vector is accomplished with an implicitly shifted QR mechanism on H_k. This will allow us to update the starting vector by working with orthogonal matrices that live in $\mathbf{R}^{k \times k}$ rather than in $\mathbf{R}^{n \times n}$.

The iteration starts by extending a length k Arnoldi factorization by p steps. Next, p shifted QR steps are performed on H_{k+p}. The last p columns of the factorization are discarded resulting in a length k factorization. The iteration is defined by repeating the above process until convergence.

As an example, suppose that $p = 1$ and that k represents the dimension of the desired invariant subspace. Let μ be a real shift and let $H_{k+1} - \mu I = QR$ with Q orthogonal and R upper triangular matrices, respectively. Then from (2.1)

$$\begin{aligned} (A - \mu I)V_{k+1} - V_{k+1}(H_{k+1} - \mu I) &= f_{k+1} e_{k+1}^T, \\ (A - \mu I)V_{k+1} - V_{k+1} QR &= f_{k+1} e_{k+1}^T, \end{aligned} \tag{3.1}$$

$$(A - \mu I)(V_{k+1}Q) - (V_{k+1}Q)(RQ) = f_{k+1}e_{k+1}^T Q,$$
$$A(V_{k+1}Q) - (V_{k+1}Q)(RQ + \mu I) = f_{k+1}e_{k+1}^T Q. \tag{3.2}$$

The matrices are updated via $V_{k+1}^+ \leftarrow V_{k+1}Q$ and $H_{k+1}^+ \leftarrow RQ + \mu I$ and the latter matrix remains upper Hessenberg since R is upper triangular and Q is upper Hessenberg. However, equation (3.2) is not quite a legitimate Arnoldi factorization. The relation of equation (3.2) fails to be an Arnoldi factorization since the matrix $f_{k+1}e_{k+1}^T Q$ has a nonzero kth column. Partitioning the matrices in the updated equation results in

$$A \begin{bmatrix} V_k^+ & v_{k+1}^+ \end{bmatrix} = \begin{bmatrix} V_k^+ & v_{k+1}^+ \end{bmatrix} \begin{bmatrix} H_k^+ & h_{k+1}^+ \\ \beta_{k+1}^+ e_k^T & \alpha_{k+1}^+ \end{bmatrix} + f_{k+1} \begin{bmatrix} \sigma_k e_k^T & \gamma_k \end{bmatrix}, \tag{3.3}$$

where $\sigma_k = e_{k+1}^T Q e_k$ and $\gamma_k = e_{k+1}^T Q e_{k+1}$. Equating the first k columns of (3.3) gives

$$AV_k^+ = V_k^+ H_k^+ + (\beta_{k+1}^+ v_{k+1}^+ + \sigma_k f_{k+1})e_k^T. \tag{3.4}$$

Performing the update $f_k^+ \leftarrow \beta_{k+1}^+ v_{k+1}^+ + \sigma_k f_{k+1}$ and noting that $(V_k^+)^T f_k^+ = 0$, it follows that equation (3.4) is a length k Arnoldi factorization.

We now show that the IRA iteration is equivalent to forming the leading portion of an implicitly shifted QR iteration. Note that equations (3.1) and (3.2) are valid for $1 \le k \le n$. In particular, extending the factorization of equation (3.1) by $n - k$ steps gives $f_n = 0$, and $AV_n - V_n H_n = 0$ defines a decomposition of A into upper Hessenberg form. Let $Q_n R_n = H_n - \mu I$ where Q_n and R_n are orthogonal and upper triangular matrices of order n, respectively. Since Q and R are the leading principal submatrices of order $k+1$ for Q_n and R_n, respectively, $V_n Q_n R_n e_1 = V_{k+1} Q R e_1$ and $e_1^T R_n e_1 = e_1^T R e_1$ follow. Postmultiplication of equation (3.2) with e_1 exposes the relationship

$$(A - \mu I)v_1 = V_{k+1}Q e_1 \rho_{11} = V_n Q_n e_1 \rho_{11} = v_1^+,$$

where $\rho_{11} = e_1^T R e_1$, $v_1 = V_{k+1}e_1$, and $V_{k+1}^+ e_1 = v_1^+$. In other words, the first column of the updated k step factorization matrix is the *same* as the first column of the orthogonal matrix obtained after a complete QR step on A with shift μ. Thus, the IRA iteration may be viewed as a truncated version of the standard implicitly shifted QR iteration. This idea may be extended for up to $p > 1$ shifts [34]. One cycle of the iteration is pictured in Figures 3.1–3.3. Application of the shifts may be performed implicitly as in the QR algorithm. If the shifts are in complex conjugate pairs then the implicit double shift can be used to avoid complex arithmetic.

Numerous choices are possible for the selection of the p shifts. One immediate choice is to use the p unwanted eigenvalues of H_{k+p}. In exact arithmetic, the last p subdiagonal elements of H_{k+p} are zero and the Arnoldi factorization decouples. For example, in equation (3.4), $\beta_{k+1}^+ = 0$ when μ is an eigenvalue of H_k. The reader is referred to [7, 23, 34] for further information.

The number of shifts to apply at each cycle of the above iteration is problem dependent. At present there is no a priori analysis to guide the selection of p relative to k. The only formal requirement is that $1 \le p \le n - k$. However, computational experience suggests that $p \ge k$ is preferable. If many problems of the same type are to be solved, experimentation with p for a fixed k should be undertaken. This usually decreases the required number of matrix–vector operations but increases the work and

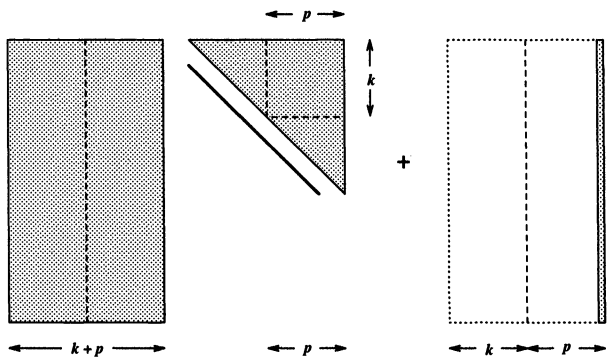


FIG. 3.1. *The set of rectangles represents the matrix equation* $V_{k+p}H_{k+p} + f_{k+p}e_{k+p}^T$ *of an Arnoldi factorization. The unshaded region on the right is a zero matrix of* $k+p-1$ *columns.*

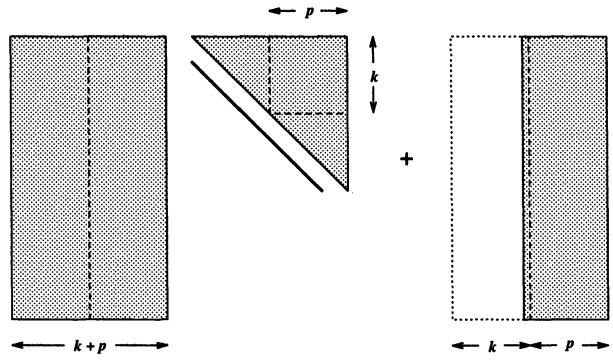


FIG. 3.2. *After performing* p *implicitly shifted QR steps on* H_{k+p}*, the middle set of pictures illustrates* $V_{k+p}QQ^TH_{k+p}Q + f_{k+p}e_{k+p}^TQ$*. The last* $p+1$ *columns of* $f_{k+p}e_{k+p}^TQ$ *are nonzero because of the QR iteration.*

storage required to maintain the orthogonal basis vectors. The optimal *crossover* with respect to CPU time varies and must be determined empirically. Lehoucq makes a connection with subspace iteration, in Chapter 8 of [23]. There has been considerable experience with subspace iteration, and this connection may eventually shed light on how to select p relative to k. For example, it is well known that performing subspace iteration on a subspace of dimension larger than the number of eigenvalues required typically leads to improved convergence rates; see the paper of Duff and Scott [12] for a discussion and further references.

Among the several advantages an implicit updating scheme possesses are

- fixed storage requirements,
- the ability to maintain a prescribed level of orthogonality for the columns of V since k is of modest size,
- application of the matrix polynomial $v_1^+ \leftarrow \Psi(A)v_1$ without needing to apply matrix–vector products with A,
- the incorporation of the well-understood numerical and theoretical behavior of the QR algorithm.

These last two points warrant further discussion. Quite often, the dominant cost during Arnoldi iterations is the matrix–vector products with A. Thus, the IRA iteration

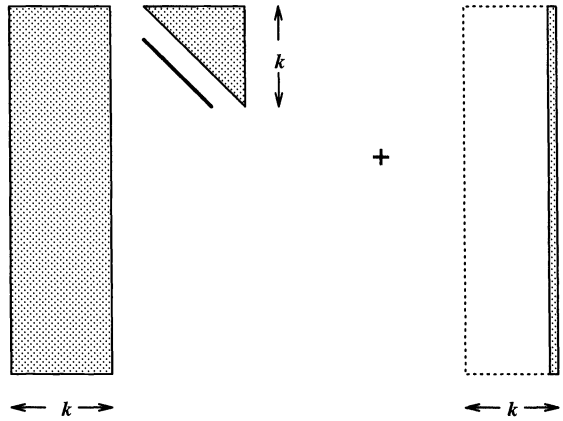


FIG. 3.3. *After discarding the last p columns, the final set represents* $V_k H_k + f_k e_k^T$ *of a length k Arnoldi factorization.*

may result in a substantial reduction in time when building a length $k+p$ Arnoldi factorization. The last point is important since it allows the possibility of constructing general purpose and reliable software for the large-scale eigenvalue problem.

4. Deflation within an IRA iteration. As the iteration progresses, the Ritz estimates (2.2) decrease at different rates. When a Ritz estimate is small enough, the corresponding Ritz value is said to have converged. The converged Ritz value may be wanted or unwanted. In either case, a mechanism to deflate the converged Ritz value from the current factorization is desired. Depending on whether the converged Ritz value is wanted or not, it is useful to define two types of deflation. Before we do this, it will prove helpful to illustrate how deflation is achieved. Suppose that after m steps of the Arnoldi algorithm we have

$$A \begin{bmatrix} V_1 & V_2 \end{bmatrix} = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} H_1 & G \\ \epsilon e_1 e_j^T & H_2 \end{bmatrix} + f e_m^T, \tag{4.1}$$

where $V_1 \in \mathbf{R}^{n \times j}$, $H_1 \in \mathbf{R}^{j \times j}$ for $1 \leq j < m$. If ϵ is suitably small then the factorization *decouples* in the sense that a Ritz pair (y, θ) for H_1 provides an approximate eigenpair $(x = V_1 y, \theta)$ with a Ritz estimate of $|\epsilon e_j^T y|$. Setting ϵ to zero splits a nearby problem exactly and setting $\epsilon = 0$ is called *deflation*. If ϵ is suitably small, then all the eigenvalues of H_1 may be regarded as converged Ritz values.

4.1. Locking. If deflation has taken place, the column vectors in V_1 are considered *locked*. This means that subsequent implicit restarting is done on the basis V_2. The submatrices affected during implicit restarting are G, H_2, and V_2. However, during the phase of the iteration that extends the Arnoldi factorization from k to $k+p$ steps, all of the columns of $\begin{bmatrix} V_1 & V_2 \end{bmatrix}$ participate just as if no deflation had occurred. This ensures that all of the new Arnoldi basis vectors are orthogonalized against converged Ritz vectors and prevents the introduction of spurious eigenvalues.

After deflation, equating the last $m-j$ columns of (4.1) results in $(I - V_1 V_1^T) A V_2 = V_2 H_2 + f e_{m-j}^T$. Thus, deflating V_1 and H_1 from the factorization defines a new Arnoldi factorization with the matrix $(I - V_1 V_1^T) A$ and starting vector $V_2 e_1$. This equivalence was noted by Saad [31, p. 182]. Moreover, this provides a means to safely compute multiple eigenvalues when they are present. A block method is not required if deflation

and locking are used. The concept of locking was introduced by Stewart and Jennings [37] as a deflation technique for simultaneous iteration.

4.2. Purging. If deflation has occurred but some of the deflated Ritz values are unwanted then a further mechanism, purging, must be introduced to remove the unwanted Ritz values and corresponding vectors from the factorization. The basic idea of purging is perhaps best explained with the case of a single deflated Ritz value.

Let $j = 1$ in (4.1) and equate the first columns of both sides to obtain

$$Av_1 = v_1\alpha_1 + \epsilon V_2 e_1, \tag{4.2}$$

where $v_1 = V_1 e_1$ and $H_1 = \alpha_1$. Equation (4.2) is an Arnoldi factorization of length one. The Ritz value α_1 has Ritz estimate $|\epsilon|$.

Equating the last $m-1$ columns of (4.1) results in

$$AV_2 = V_1 G + V_2 H_2 + f e_{m-1}^T. \tag{4.3}$$

Suppose that α_1 represents an unwanted Ritz value. If A were symmetric then $G = \epsilon e_1^T$ and equation (4.3) would become

$$(A + E)V_2 = V_2 H_2 + f e_{m-1}^T,$$

where $E = -\epsilon v_1 (V_2 e_1)^T - \epsilon (V_2 e_1) v_1^T$. Since $\|E\| = \epsilon$ equation (4.3) defines a length $m-1$ Arnoldi factorization for a nearby problem. The unwanted Ritz pair (v_1, α_1) may be *purged* from the factorization simply by taking $V = V_2$ and $H = H_2$ and setting $G = 0$ in (4.3). If A is not symmetric, the $1 \times (m-1)$ matrix G couples v_1 to the rest of the basis vectors V_2. This vector may be decoupled using the standard Sylvester equation approach [15, pp. 386–387]. Purging then takes place as in the symmetric case. However, the new set of basis vectors must be reorthogonalized in order to return to an Arnoldi factorization. This procedure is developed in §§5 and 6 including the case of purging several vectors.

4.3. Complications. An immediate question follows: do any subdiagonal elements in the Hessenberg matrix of the factorization (4.1) become negligible as an IRA iteration progresses? Since a cycle of the Arnoldi iteration involves performing a sequence of QR steps, the question is answered by considering the behavior of the QR iteration upon upper Hessenberg matrices. In exact arithmetic, under the assumption that the Hessenberg matrix is unreduced, only the last subdiagonal element may become zero when shifting. But the other subdiagonal elements may become arbitrarily small.

In addition, in exact arithmetic, the purging technique would not be necessary as the implicit shift technique would accomplish the removal of the unwanted Ritz pairs from the leading portion of the iteration. For example, using the unwanted Ritz values as shifts accomplishes this removal.

Computing in finite precision arithmetic complicates the situation. A robust implementation of the QR algorithm sets a subdiagonal element to zero if it is in magnitude less than some prescribed threshold and this technique is also adopted for deflation. This deflation overcomes the technical difficulty associated with tiny subdiagonals and improves the convergence of the IRA iteration. In addition, it may be impossible to accomplish the removal of the unwanted Ritz values from the leading portion of the iteration due to the forward instability [27, 39] of the QR algorithm.

The phenomena of the forward instability of the tridiagonal QR iteration [27] was initially explored by Parlett and Le. They observe that while the implicitly shifted QR

iteration is always backward stable, there are cases where severe forward instability can occur. It is possible for a QR iteration to result in a computed Hessenberg matrix with entries that have no significant digits in common with the corresponding entries of the Hessenberg matrix that would have been determined in exact arithmetic. The implication is that the computed subdiagonal entries may not be reliable indicators for decoupling the Arnoldi factorization. Parlett and Le's analysis formally implies that the computed Hessenberg matrix may lose significant digits when the shift used is nearly an eigenvalue of H and the last component of the normalized eigenvector is small. We also mention the work of Watkins [39], which investigates the transmission of the shift during a QR step through H.

Since convergence of a Ritz value is predicated upon the associated Ritz estimate being small, using shifts that are near these converged values may force the IRA iteration to undergo forward instability. This indicates that it may be impossible to filter out unwanted eigenvalues with the implicit restarting technique, and this is the motivation for developing both the locking and the purging techniques. Further details may be found in Chapter 5 of [23].

5. Deflating converged Ritz values. During an Arnoldi iteration, a Ritz value may be near an eigenvalue of A with no small elements appearing on the subdiagonal of H_k. However, when a Ritz value converges, it is always possible to make an orthogonal change of basis in which the appropriate subdiagonal of H_k is zero. The following result indicates how to exploit the convergence information available in the last row of the eigenvector matrix for H_k. For notational convenience, all subscripts are dropped on the Arnoldi matrices V, H, and f for the remainder of this section.

LEMMA 5.1. *Let $Hy = y\theta$ where $H \in \mathbf{R}^{k\times k}$ is an unreduced upper Hessenberg matrix and $\theta \in \mathbf{R}$ with $\|y\| = 1$. Let W be a Householder matrix such that $Wy = e_1\tau$ where $\tau = -\text{sign}(e_1^T y)$. Then*

$$e_k^T W = e_k^T + w^T, \tag{5.1}$$

where $\|w\| \le \sqrt{2}|e_k^T y|$ and

$$W^T H W e_1 = e_1\theta. \tag{5.2}$$

Proof. The required Householder matrix has the form

$$W = I - \gamma(y - \tau e_1)(y - \tau e_1)^T,$$

where $\gamma = (1 + |e_1^T y|)^{-1}$ and $\tau = -\text{sign}(e_1^T y)$. A direct computation reveals that

$$e_k^T W = e_k^T + w^T, \tag{5.3}$$

where $w^T = \gamma e_k^T y(\tau e_1^T - y^T)$. Estimating

$$\begin{aligned}\|w\| &= \frac{|e_k^T y|}{1 + |e_1^T y|}\|y - \tau e_1\| \\ &= \frac{|e_k^T y|}{1 + |e_1^T y|}\sqrt{2(1 + |e_1^T y|)} \\ &\le \sqrt{2}|e_k^T y|\end{aligned}$$

establishes the bound on $\|w\|$. The final assertion (5.2) follows from

$$\begin{aligned} W^T H W e_1 &= \tau^{-1} W^T H y \\ &= \tau^{-1} \theta W^T y \\ &= \tau^{-1} \theta W y \quad (W^T = W) \\ &= \theta e_1. \qquad \square \end{aligned}$$

Lemma 5.1 indicates that the last row and column of W differ from the last row and column of I_k by terms of order $|e_k^T y|$. The Ritz estimate (2.2) will indicate when it is safe to deflate the corresponding Ritz value θ. Rewriting (2.1) as

$$AVW = VWW^T HW + f e_k^T W$$

and using both (5.1) and (5.2) and partitioning we obtain

$$AVW = VW \begin{bmatrix} \theta & \bar{h}^T \\ 0 & \bar{H} \end{bmatrix} + f e_k^T + f w^T. \tag{5.4}$$

Equation (5.4) is not an Arnoldi factorization. In order to return to an Arnoldi factorization, the matrix $\bar{H}$ of order $k-1$ needs to be returned to upper Hessenberg form and the term fw^T dropped. Care must be taken not to disturb the matrix fe_k^T and the first column of $W^T HW$. To start the process we compute a Householder matrix Y_1 such that

$$Y_1^T \bar{H} Y_1 = \begin{bmatrix} \bar{G} & \bar{g} \\ \bar{\beta}_k e_{k-2}^T & \gamma \end{bmatrix},$$

with $e_{k-1}^T Y_1 = e_{k-1}^T$. The above idea is repeated resulting in Householder matrices $Y_1, Y_2, \ldots, Y_{k-3}$ that return $\bar{H}$ to upper Hessenberg form. Defining

$$Y = \begin{bmatrix} 1 & 0 \\ 0 & Y_1 Y_2 \cdots Y_{k-3} \end{bmatrix},$$

it follows by the construction of the Y_j that $e_k^T Y = e_k^T$ and

$$Y^T W^T H W Y e_1 = \theta e_1. \tag{5.5}$$

The process of computing a similarity transformation as in equation (5.5) is not new. Wilkinson discusses similar techniques in [40, pp. 587–596]. Wilkinson references the work of Feller and Forsythe [13], who appear to be the first to use elementary Householder transformations for deflation. Problem 7.4.8 of [15, p. 371] addresses the case when working with upper Hessenberg matrices. What appears to be new is the application to the Arnoldi factorization for converged Ritz values.

Since $\|fw^T Y\| = \|f\| \, \|Y^T w\| = \|f\| \, \|w\|$, the size of $\|fw^T\|$ remains unchanged. Making the updates

$$V \leftarrow VWY, \quad H \leftarrow Y^T W^T H W Y, \quad w^T \leftarrow w^T Y,$$

we obtain the relation

$$AV = VH + f e_k^T + f w^T. \tag{5.6}$$

A deflated Arnoldi factorization is obtained from equation (5.6) by discarding the term fw^T.

The following theorem shows that the deflated Arnoldi factorization resulting from this scheme is an exact k step factorization of a nearby matrix.

THEOREM 5.2. *Let an Arnoldi factorization of length k be given by* (5.6), *where* $Hy = y\theta$ *and* $\sqrt{2}|e_k^T y|\,\|f\| \leq \epsilon\|A\|$ *for some* $\epsilon > 0$. *Then there exists a matrix* $E \in \mathbf{R}^{n\times n}$ *such that*

$$(A+E)V = VH + fe_k^T, \tag{5.7}$$

where

$$\|E\| \leq \epsilon\|A\|.$$

Proof. Subtract fw^T from both sides of equation (5.6). Set $E = -f(Vw)^T$ and then

$$EV = -f(Vw)^T V = -fw^T$$

and equation (5.7) follows. Using Lemma 5.1 gives

$$\|E\| = \|f\|\,\|w\| = \sqrt{2}|e_k^T y|\,\|f\| \leq \epsilon\|A\|. \qquad \square$$

If A is symmetric then the choice $E = -f(Vw)^T - (Vw)f^T$ results in a symmetric perturbation. If ϵ is on the order of unit round-off then the deflation scheme introduces a perturbation of the same order to those already present from computing the Arnoldi factorization in floating point arithmetic.

Once a converged Ritz value θ is deflated, the Arnoldi vector corresponding to θ is locked or purged as described in the previous section. The only difficulty that remains is purging when A is nonsymmetric.

If A is not symmetric, then the Ritz pair may not be purged immediately because of the presence of $\bar{h}$. A standard reduction of H to block diagonal form is used. If θ is not an eigenvalue of $\bar{H}$, then we may construct a vector $z \in \mathbf{R}^{k-1}$ so that

$$\begin{bmatrix} \theta & \bar{h}^T \\ & \bar{H} \end{bmatrix} \begin{bmatrix} 1 & z^T \\ & I_{k-1} \end{bmatrix} = \begin{bmatrix} 1 & z^T \\ & I_{k-1} \end{bmatrix} \begin{bmatrix} \theta & \\ & \bar{H} \end{bmatrix}. \tag{5.8}$$

Solving the linear system

$$(\bar{H}^T - \theta I_{k-1})z = \bar{h} \tag{5.9}$$

determines z. Define

$$Z \equiv \begin{bmatrix} 1 & z^T \\ & I_{k-1} \end{bmatrix}.$$

Postmultiplication of equation (5.6) by Z results in

$$AVZ = VZ \begin{bmatrix} \theta & \\ & \bar{H} \end{bmatrix} + fe_k^T + fw^T Z$$

since $e_k^T Z = e_k^T$. Equating the last $k-1$ columns of the previous expression results in

$$AV \begin{bmatrix} z^T \\ I_{k-1} \end{bmatrix} = V \begin{bmatrix} z^T \\ I_{k-1} \end{bmatrix} \bar{H} + fe_{k-1}^T + fw^T \begin{bmatrix} z^T \\ I_{k-1} \end{bmatrix}. \tag{5.10}$$

Compute the factorization (using $k-1$ Givens rotations)

$$QR = \begin{bmatrix} z^T \\ I_{k-1} \end{bmatrix}, \tag{5.11}$$

where $Q \in \mathbf{R}^{k \times k-1}$ with $Q^TQ = I_{k-1}$ and R is an upper triangular matrix of order $k-1$. Since the last $k-1$ columns of Z are linearly independent, R is nonsingular. Postmultiplying equation (5.10) by R^{-1} gives

$$AVQ = VQR\bar{H}R^{-1} + \rho_{k-1}^{-1} f e_{k-1}^T + f w^T Q, \tag{5.12}$$

where $\rho_{k-1} = e_{k-1}^T R e_{k-1}$. The last term fw^TQ in (5.12) is discarded by the deflation scheme, and this relation shows that the discarded term is not magnified in norm by the purging procedure. The matrix $R\bar{H}R^{-1}$ remains upper Hessenberg since R is upper triangular.

Partitioning Q conformally with the right side of equation (5.11) results in

$$\begin{bmatrix} q_{11}^T \\ Q_{21} \end{bmatrix} R = \begin{bmatrix} z^T \\ I_{k-1} \end{bmatrix},$$

and it follows that $R^{-1} = Q_{21}$. Using the Cauchy–Schwarz inequality, it follows that $|\rho_{k-1}^{-1}| = |e_{k-1}^T Q_{21} e_{k-1}| \leq 1$ and hence the Arnoldi residual is not amplified by the purging. The final purged Arnoldi factorization is

$$AVQ = VQR\bar{H}Q_{21} + \rho_{k-1}^{-1} f e_{k-1}^T. \tag{5.13}$$

Performing the set of updates

$$V \leftarrow VQ, \quad H \leftarrow R\bar{H}Q_{21}, \quad f \leftarrow \rho_{k-1}^{-1} f$$

defines equation (5.13) as an Arnoldi factorization of length $k-1$. Theorem 5.2 implies this is an Arnoldi factorization for a nearby matrix. It is easily verified that $V^T f(e_{k-1}^T + w^T) = 0$ and that H is an upper Hessenberg matrix of order $k-1$. Since the term fw^T is discarded, the Ritz estimates given by the updated Arnoldi factorization for the remaining Ritz values will be slightly inaccurate. Lemma 5.1 and the fact that $\|R^{-1}\| \leq 1$ may be used to show that the errors in these estimates are bounded above by $\|f\|(\sqrt{2}|e_k^T y|)$. If $w = 0$ then the Ritz estimates for the updated factorization would be exactly the same as the Ritz residuals and estimates for the original one.

6. A practical deflating procedure for the Arnoldi factorization. The practical issues associated with a numerically stable deflating procedure are addressed in this section. These include

1. performing the deflation in real arithmetic when a converged Ritz value has a nonzero imaginary component,
2. deflation with more than one converged Ritz value,
3. error analysis.

Section 6.2 presents two algorithms that implement the deflation schemes. The error analysis of the two deflation schemes is presented in the next section.

6.1. Deflation with real arithmetic. Suppose $H(y + iz) = (\theta + i\mu)(y + iz)$ where y and z are unit vectors in $\mathbf{R}^k$, $H \in \mathbf{R}^{k\times k}$, and $\mu \neq 0$. It then follows that

$$H \begin{bmatrix} y & z \end{bmatrix} = \begin{bmatrix} y & z \end{bmatrix} \begin{bmatrix} \theta & \mu \\ -\mu & \theta \end{bmatrix} \equiv \begin{bmatrix} y & z \end{bmatrix} C.$$

Thus, we may deflate a complex Ritz value in real arithmetic if $|e_k^T y|$ and $|e_k^T z|$ are small enough.

Suppose that H corresponds to an Arnoldi factorization of length k and that $|e_k^T y| = O(\epsilon) = |e_k^T z|$. Factor

$$\begin{bmatrix} y & z \end{bmatrix} = U \begin{bmatrix} T \\ 0 \end{bmatrix}, \tag{6.1}$$

where $U^T U = I_k$ and T is an upper triangular matrix. It is easily shown that y and z are linearly independent as vectors in $\mathbf{R}^k$ since $\mu \neq 0$ and the nonsingularity of T follows. Performing a similarity transformation on H with U gives

$$U^T H U \begin{bmatrix} e_1 & e_2 \end{bmatrix} = \begin{bmatrix} TCT^{-1} \\ 0 \end{bmatrix}.$$

In order to deflate the complex conjugate pair of eigenvalues from the factorization in an implicit manner, we require that $e_k^T U = e_k^T + u^T$ where $\|u\| = O(\epsilon)$.

We now show that the magnitudes of the last components of y and z are not sufficient to guarantee the required form for U. Suppose that $z = y\cos\phi + r\sin\phi$ where r is a unit vector orthogonal to y and ϕ measures the positive angle between y and z. Lemma 5.1 implies that a Householder W matrix may be constructed such that

$$W^T \begin{bmatrix} y & z \end{bmatrix} = \begin{bmatrix} \tau_1 e_1 & \tau_1 e_1 \cos\phi + W^T r \sin\phi \end{bmatrix} \equiv \begin{bmatrix} \tau_1 & \zeta \\ 0 & \hat{z} \end{bmatrix},$$

where $\tau_1 = \pm 1$ and the last column and row of W and I_k are the same up to order $e_k^T y$. To compute the required orthogonal factorization in equation (6.1) another Householder matrix $Q = \left[\begin{smallmatrix} 1 & 0 \\ 0 & \hat{Q} \end{smallmatrix}\right]$ is needed so that $\hat{Q}^T \hat{z} = \pm\|\hat{z}\| e_1$. But Lemma 5.1 only results in $e_{k-1}^T \hat{Q} = e_{k-1}^T + \hat{q}^T$ with $\|\hat{q}\| = O(\epsilon)$ if $e_{k-1}^T \hat{z}$ is small relative to $\|\hat{z}\|$. Unfortunately, if ϕ is small, $W^T z \approx \tau_1 e_1$ and $\|\hat{z}\| \approx \phi$. Hence we cannot obtain the required form for $U = WQ$.

Fortunately, when y and z are nearly aligned, μ may be neglected, as the following result demonstrates.

LEMMA 6.1. *Let $H(y + iz) = (\theta + i\mu)(y + iz)$ where y and z are unit vectors in $\mathbf{R}^k$, $H \in \mathbf{R}^{k\times k}$, and $\mu \neq 0$. Suppose that ϕ measures the positive angle between y and z. Then*

$$|\mu| \leq \sin\phi \|H\|. \tag{6.2}$$

Proof. Let $z = y\cos\phi + r\sin\phi$ where r is a unit vector orthogonal to y and ϕ measures the positive angle between y and z. Equating real and imaginary parts of $H(y + iz) = (\theta + i\mu)(y + iz)$ results in $Hy = y\theta - z\mu$ and $Hz = y\mu + z\theta$. The desired estimate follows since

$$2\mu = y^T H z - z^T H y = \sin\phi(y^T H r - r^T H y)$$

results in $|\mu| \leq \sin\phi \|H\|$. □

For small ϕ, y and z are almost parallel eigenvectors of H corresponding to a nearly multiple eigenvalue. Numerically, we set μ to zero and deflate one copy of θ from the Arnoldi factorization.

A computable bound on the size of the angle ϕ is now determined using only the real and imaginary parts of the eigenvector. The second Householder matrix Q should not be computed if

$$|e_{k-1}^T \hat{z}| > \|\hat{z}\| |e_k^T z|. \tag{6.3}$$

Recall that Lemma 5.1 gives $e_k^T W = e_k^T + w^T$ where $w^T = \gamma e_k^T y(\tau_1 e_1^T - y^T)$ and $\gamma = (1 + |e_1^T y|)^{-1}$. Thus

$$e_{k-1}^T \hat{z} = e_k^T W^T z = e_k^T W z = e_k^T z + w^T z$$

where the symmetry of W is used. The estimate

$$\|\hat{z}\| = \| \begin{bmatrix} 0 & \hat{z}^T \end{bmatrix}^T \| = \|W^T r\| \sin\phi = \sin\phi$$

follows since W is orthogonal and r is a unit vector. Rewriting equation (6.3), we obtain

$$\begin{aligned} \sin\phi &< \left| \frac{e_k^T z + w^T z}{e_k^T z} \right| \\ &= \left| 1 + \frac{w^T z}{e_k^T z} \right| \\ &= \left| 1 + \gamma(\tau_1 e_1^T z - y^T z) \frac{e_k^T y}{e_k^T z} \right| \end{aligned} \tag{6.4}$$

as our computable bound.

Suppose that $HX = XD$ where $X \in \mathbf{R}^{k \times j}$ and D is a quasi-diagonal matrix. The eigenvalues of H are on the diagonal of D if they have zero imaginary component and in blocks of two for the complex conjugate pairs. The columns of X span the eigenspace corresponding to diagonal values of D. For the blocks of order-2 on the diagonal the corresponding complex eigenvector is stored in two consecutive columns of X, the first holding the real part, and the second the imaginary part. If we want to block deflate X, where the last row is small, from H we could proceed as follows. Compute the orthogonal factorization $X = Q\begin{bmatrix} R \\ 0 \end{bmatrix}$ via Householder reflectors where $Q^T Q = I_k$ and $R \in \mathbf{R}^{k \times k}$ is upper triangular. Then the last row and column of Q differ from that of I_k with terms on the same order of the entries in the last row of X if the condition number of R is modest. Thus, if the columns of X are not almost linearly dependent, an appropriate Q may be determined. Finally, we note that when H is a symmetric tridiagonal matrix, an appropriate Q may always be determined.

6.2. Algorithms for deflating converged Ritz values. The two procedures presented in this section extend the ideas of §4 to provide deflation of more than one converged Ritz value at a time. The first purges the factorization of the unwanted converged Ritz values. The second locks the Arnoldi vectors corresponding to the desired converged Ritz values. When both deflation algorithms are incorporated within an IRA iteration, the locked vectors form a basis for an approximate invariant subspace of A. This truncated factorization is an approximate partial Schur decomposition.

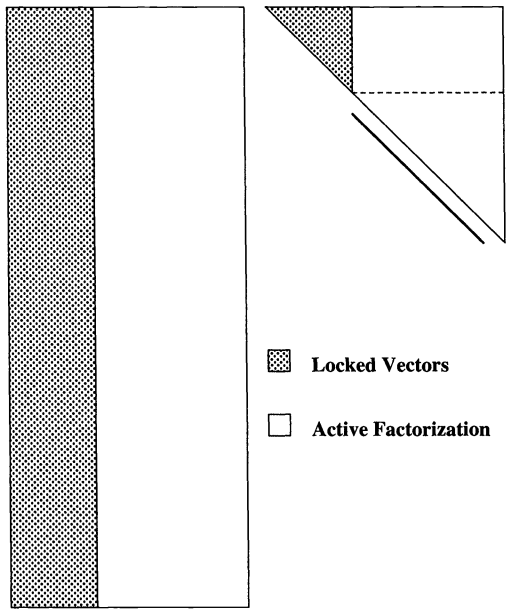


FIG. 6.1. *The matrix product $V_m H_m$ of the factorization upon entering Algorithm* 6.2 *or* 6.3. *The shaded region corresponds to the converged portion of the factorization.*

When A is symmetric, the approximate Schur vectors are Ritz vectors and the upper quasi-triangular matrix is the diagonal matrix of Ritz values.

Partition a length m Arnoldi factorization as

$$A \begin{bmatrix} V_j & \bar{V}_{m-j} \end{bmatrix} = \begin{bmatrix} V_j & \bar{V}_{m-j} \end{bmatrix} \begin{bmatrix} H_j & G_j \\ 0 & \bar{H}_{m-j} \end{bmatrix} + f_m e_m^T + f w^T, \tag{6.5}$$

where H_j and $\bar{H}_{m-j}$ are upper quasi-triangular and unreduced upper Hessenberg matrices, respectively. The matrix $H_j \in \mathbf{R}^{j \times j}$ contains the wanted converged Ritz values of the matrix H_m. The columns of $V_j \in \mathbf{R}^{n \times j}$ are the locked Arnoldi vectors that represent an approximate Schur basis for the invariant subspace of interest. The matrix $\bar{H}_{m-j}$ designates the trailing submatrix of order $m - j$. Analogously, the last $m - j$ columns of V_m are denoted by $\bar{V}_{m-j}$. We shall refer to the last $m - j$ columns of (6.5) as the *active* part of the factorization. Finally, $G_j \in \mathbf{R}^{j \times (m-j)}$ denotes the submatrix in the northeast corner of H_m. Figure 6.1 illustrates the matrix product $V_m H_m$ of equation (6.5).

If A is symmetric, the two deflation procedures simplify considerably. In fact, purging is only used when A is nonsymmetric for otherwise $G_j = 0_{j \times (m-j)}$ and both H_j and $\bar{H}_{m-j}$ are symmetric tridiagonal matrices. Both algorithms are followed by remarks concerning some of the specific details.

ALGORITHM 6.2.

function $[V_m, H_m, f_m] =$ `Lock` (V_m, H_m, f_m, X_i, j)

INPUT: A length m Arnoldi factorization $AV_m = V_m H_m + f_m e_m^T$. The first j columns of V_m represent an approximate invariant subspace for A. The leading principal submatrix H_j of order j of H_m is upper quasi-triangular and contains the converged Ritz values of interest. The columns of $X_i \in \mathbf{R}^{(m-j) \times i}$ are the eigenvectors corresponding to the eigenvalues that are to be locked.

OUTPUT: A length m Arnoldi factorization defined by V_m, H_m, and f_m where the first $j+i$ columns of V_m are an approximate invariant subspace for A.

1. *Compute the orthogonal factorization*

$$Q \begin{bmatrix} R_i \\ 0_{m-j-i} \end{bmatrix} = X_i,$$

where $Q \in \mathbf{R}^{(m-j)\times(m-j)}$ using Householder matrices;

2. *Update the factorization*
$\bar{H}_{m-j} \leftarrow Q^T \bar{H}_{m-j} Q$; $\bar{V}_{m-j} \leftarrow \bar{V}_{m-j} Q$; $G_j \leftarrow G_j Q$;
3. *Compute an orthogonal matrix $P \in \mathbf{R}^{(m-j-i)\times(m-j-i)}$ using Householder matrices that restore $\bar{H}_{m-j-i}$ to upper Hessenberg form;*
4. *Update the factorization*
$\bar{H}_{m-j-i} \leftarrow P^T \bar{H}_{m-j-i} P$; $\bar{V}_{m-j-i} \leftarrow \bar{V}_{m-j-i} P$; $G_{j+i} \leftarrow G_{j+i} P$;

Line 1 computes an orthogonal basis for the eigenvectors of $\bar{H}_{m-j}$ that correspond to the Ritz estimates that are converged. The matrix of eigenvectors in line 1 satisfies the equation $\bar{H}_{m-j} X_i = X_i D_i$ where D_i is a quasi-diagonal matrix containing the eigenvalues to be locked. From §6.1, we see that the leading submatrix of $Q^T \bar{H}_{m-j} Q$ of order i is upper quasi-triangular. The required relation $e_m^T Q = e_m^T + q^T$, with $\|q\|$ small, is guaranteed if the condition number of R_i is modest. Since i is typically a small number, we compute the condition number of R_i. The number of vectors to be locked is assumed to be such that the condition number of R_i is small. In particular, if H_m is a symmetric tridiagonal matrix, Q always has the required form. Lines 3–4 return the updated $\bar{H}_{m-j}$ to upper Hessenberg form.

Before entering `Purge`, the unwanted converged Ritz pairs are placed at the front of the factorization. A prior call to `Lock` places the unwanted values and vectors to the beginning of the factorization. Unlike `Lock`, the procedure `Purge` requires accessing and updating the entire factorization when A is nonsymmetric. Thus, for large-scale nonsymmetric eigenvalue computations, the amount of purging performed should be kept to a minimum.

ALGORITHM 6.3.

function $[V_{m-i}, H_{m-i}, f_{m-i}] =$ `Purge` (V_m, H_m, f_m, j, i)

INPUT: A length m Arnoldi factorization $AV_m = V_m H_m + f_m e_m^T$. The first $i+j$ columns of V_m represent an approximate invariant subspace for A. The leading principal submatrix H_{i+j} of order $i+j$ of H_m is upper quasi-triangular and contains the converged Ritz values. The i unwanted converged eigenvalues are in the leading portion of H_{i+j}. The converged complex conjugate Ritz pairs are stored in 2×2 blocks on the diagonal of H_{i+j}.

OUTPUT: A length $m-i$ Arnoldi factorization defined by V_{m-i}, H_{m-i}, and f_{m-i} purged of the unwanted converged Ritz values and corresponding Schur vectors.

Lines 1–3 *purge the factorization of the unwanted converged Ritz values contained in the leading portion of H_m;*

1. *Solve the Sylvester set of equations,*

$$Z \bar{H}_{m-i} - H_i Z = G_i,$$

for $Z \in \mathbf{R}^{i\times(m-i)}$ that arise from block diagonalizing H_m

$$H_m \begin{bmatrix} I_i & Z \\ & I_{m-i} \end{bmatrix} = \begin{bmatrix} I_i & Z \\ & I_{m-i} \end{bmatrix} \begin{bmatrix} H_i & \\ & \bar{H}_{m-i} \end{bmatrix};$$

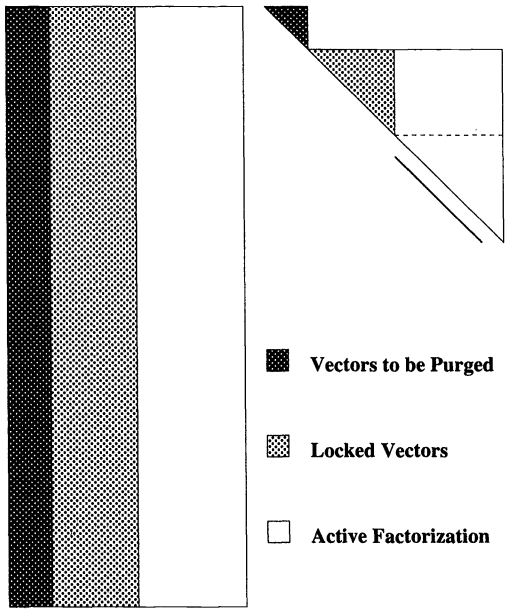


FIG. 6.2. *The matrix product $V_m H_m$ of the factorization just prior to discarding in Algorithm 6.3. The darkly shaded regions may now be dropped from the factorization.*

2. *Compute the orthogonal factorization*

$$QR_{m-i} = \begin{bmatrix} Q_i \\ Q_{m-i} \end{bmatrix} R_{m-i} = \begin{bmatrix} Z \\ I_{m-i} \end{bmatrix},$$

where $Q \in \mathbf{R}^{m\times(m-i)}$ using Householder matrices;

3. *Update the factorization and obtain a length $m-i$ factorization*
$H_{m-i} \leftarrow R_{m-i}\bar{H}_{m-i}Q_{m-i}$; $V_{m-i} \leftarrow V_m Q$; $f_{m-i} \leftarrow \rho^{-1}_{m-i,m-i} f_m$;
where $\rho_{m-i,m-i} = e^T_{m-i} R_{m-i} e_{m-i}$;

At the completion of Algorithm 6.3 the factorization is of length $m-i$ and the leading submatrix of order j will be upper quasi-triangular. The wanted converged Ritz values will be either on the diagonal if real or in blocks of two for the complex conjugate pairs. Figure 6.2 shows the structure of the updated $V_m H_m$ just prior to discarding the unwanted portions.

The solution of the Sylvester equation at line 1 determines the matrix Z that block diagonalizes the spectrum of H_m into two submatrices. The unwanted portion is in the leading corner and the remaining eigenvalues of H_m are in the other block. A solution Z exists when the H_i and $\bar{H}_{m-i}$ do not have a common eigenvalue. If there is an eigenvalue that is shared by H_i and $\bar{H}_{m-i}$, then H_m has an eigenvalue of multiplicity greater than one. The remedy is a criterion that determines whether to increase or decrease i, the number of Ritz values that require purging. Analysis similar to that in §5 demonstrates that after line 3 the Ritz estimates for the eigenvalues of H_{m-i} are not altered. We also remark that R_{m-i} is nonsingular since the matrix

$$\begin{bmatrix} Z \\ I_{m-i} \end{bmatrix}$$

is of full column rank and $|\rho^{-1}_{m-i,m-i}| \leq 1$.

7. Error analysis. This section examines the numerical stability of the two deflation algorithms when computing in finite precision arithmetic. A stable algorithm computes the exact solution of a nearby problem. It will be shown that Algorithms 6.3 and 6.2 deflate slightly perturbed matrices.

For ease of notation

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}$$

replaces $H_m \in \mathbf{R}^{m\times m}$ used by procedures `Lock` and `Purge` of §6.2. The submatrix H_{11} is of order i and H_{21} is zero except for the subdiagonal entry of H located in the northeast corner. Analogously, $\hat{H}$ represents H after the similarity transformation performed by `Lock` or `Purge` partitioned conformally.

7.1. Locking. The locking scheme is considered successful if the desired eigenvalues end up in $\hat{H}_{11}$ and $\hat{H}_{21}$ is small in norm. The largest source of error is from computing an orthogonal factorization from the approximate eigenvector matrix containing the vectors to be locked.

The matrix pair (X, D) represents an approximate quasi-diagonal form for H. The computed eigenvalues of H are on the diagonal of D if they have zero imaginary component and in blocks of two for the complex conjugate pairs. The computed columns of X span the right eigenspace corresponding to diagonal values of D. For the blocks of order-2 on the diagonal, the corresponding complex eigenvector is stored in two consecutive columns of X, the first holding the real part and the second the imaginary part. We assume that X is a nonsingular matrix and that each column is a unit vector.

Standard results give $\|XD - HX\| \leq \epsilon_1\|H\|$ where ϵ_1 is a small multiple of machine precision for a stable algorithm. Defining the matrix $E = (XD-HX)Y^T$ where $X^{-1} = Y^T$ it follows that $(H + E)X = XD$. If $\sigma_m^{-1}(X)$ is the smallest singular value of X then $\|X^{-1}\| = \sigma_m^{-1}(X)$. Since each column of X is a unit vector, $\|X\| \leq \sqrt{m}$. If $\kappa(X) = \|X\|\|X^{-1}\|$ is the condition number for the matrix of approximate eigenvectors, $\|E\| \leq \epsilon_1\kappa(X)\|H\|$. If X is a well-conditioned matrix, then the approximate quasi-diagonal form for H is exact for a nearby matrix. In particular, if H is symmetric then E is always a small perturbation. As the columns of X become linearly dependent, $\sigma_m(X)$ decreases and E may represent a large perturbation.

The following result informs us that locking is a conditionally stable process.

THEOREM 7.1. *Let $H \in \mathbf{R}^{m\times m}$ be an unreduced upper Hessenberg matrix with distinct eigenvalues. Suppose that*

$$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \text{ and } D = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}$$

is an approximate quasi-diagonal form for H that satisfies $(H + E)X = XD$ where $\|E\| \leq \epsilon_1\kappa(X)\|H\|$. Let $Q_1R_1 = X_1 \in \mathbf{R}^{m\times j}$ where $Q_1^TQ_1 = I_j$. Suppose a QR factorization of X_1 is computed so that $\hat{Q}\hat{R} = X_1 + \hat{E}$ where $\hat{Q}^T\hat{Q} = I_m$ and $\|\hat{E}\| \leq \epsilon_2\|X_1\|$. Both ϵ_1 and ϵ_2 are small multiples of the machine precision ϵ_M. Let $\epsilon = \max(\epsilon_1, 2\epsilon_2)$ and let $\kappa(R_1) = \|R_1\|\|R_1^{-1}\|$ be the condition number for R_1 where

$$\mu \equiv \frac{\kappa(R_1)}{1 - \epsilon_2\kappa(R_1)}.$$

If $\theta \equiv \epsilon(\kappa(X)+\epsilon\mu(1+\epsilon\mu\kappa(R_1))) < 1$, *then there exists a matrix* $C \in \mathbf{R}^{m\times m}$ *such that*

$$\hat{Q}^T(H-C)\hat{Q} = \hat{H} = \begin{bmatrix} \hat{H}_{11} & \hat{H}_{12} \\ 0 & \hat{H}_{22} \end{bmatrix},$$

where $\hat{H}_{11}$ *is an upper quasi-triangular matrix similar to* D_1 *and*

$$\|C\| \leq \epsilon(\kappa(X)+\mu)\|H\| + O(\epsilon^2). \tag{7.1}$$

A few remarks are in order.

1. If H is symmetric, $\hat{H}_{12} = 0$ and $\hat{H}_{11}$ is diagonal. Procedure `Lock` is stable, since noted previously $\kappa(X) = 1$ and $\mu \approx 1$. Parlett [26, pp. 85–86] proves Theorem 7.1 for symmetric matrices when locking one approximate eigenvector.
2. If only one column is locked, then $\mu = 1 + O(\epsilon)$ and $\|C\|$ is small relative to $\kappa(X)\|H\|$.
3. If $\kappa(R_1)$ is large, the columns of X_1 are nearly dependent. In this case, $\kappa(X)$ will also be large and locking will likely introduce no more error into the computation than already present from computing the quasi-diagonal pair (X, D). The factor of μ may be minimized by decreasing j, the number of columns locked.
4. A conservative strategy locks only one vector at a time. The only real concern is when locking two vectors corresponding to a complex conjugate pair. If the real and imaginary parts of the complex eigenvector are nearly aligned, μ will be large and locking may be unstable. But as §6.1 explains, the complex conjugate pair may be numerically regarded as a double eigenvalue with zero imaginary part. Only one copy is deflated and $\mu \approx 1$.

Proof. Partition

$$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \text{ and } D = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}.$$

The i columns of X_1 are a basis for the right eigenspace to be locked, and D_1 contains the corresponding eigenvalues. We assume that the eigenvalues of D_1 and D_2 are distinct and that X is nonsingular. Let

$$Y^T = \begin{bmatrix} Y_1^T \\ Y_2^T \end{bmatrix}$$

denote the inverse of X. The rows of Y_1^T span the left eigenspace associated with the computed eigenvalues of D_1.

Let the product $\hat{Q}\hat{R}$ be an exact QR factorization of a matrix near X_1:

$$\hat{Q}\hat{R} = \begin{bmatrix} \hat{Q}_1 & \hat{Q}_2 \end{bmatrix} \begin{bmatrix} \hat{R}_1 \\ 0 \end{bmatrix} = X_1 + \hat{E}$$

where $\|\hat{E}\| \leq \epsilon_2\|X_1\|$. Using Theorem 1.1 of Stewart [36], since $\|R_1^{-1}\|\|\hat{E}\| < \theta < 1$ there exist matrices $W_1 \in \mathbf{R}^{m\times j}$ and $F_1 \in \mathbf{R}^{j\times j}$ such that $(Q_1+W_1)(R_1+F_1) = \hat{Q}_1\hat{R}_1$ where

$$QR = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = X_1$$

and $(Q_1 + W_1)^T(Q_1 + W_1) = I_j$. Define

$$F = \begin{bmatrix} F_1 \\ 0 \end{bmatrix} \text{ and } W = \begin{bmatrix} W_1 & 0 \end{bmatrix}.$$

The matrices W and F are the perturbations that account for the backward error $\hat{E}$ produced by computation.

Partitioning W conformally with Q gives

$$\begin{aligned} \hat{Q}^T H \hat{Q} &= \hat{Q}^T X D Y^T \hat{Q} - \hat{Q}^T E \hat{Q} \\ &= \hat{Q}^T (X_1 D_1 Y_1^T + X_2 D_2 Y_2^T)\hat{Q} - \hat{Q}^T E \hat{Q} \\ &\approx \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} (X_1 D_1 Y_1^T + X_2 D_2 Y_2^T) \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \\ &\quad + W^T (X_1 D_1 Y_1^T + X_2 D_2 Y_2^T) \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \\ &\quad + \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} (X_1 D_1 Y_1^T + X_2 D_2 Y_2^T) W - \hat{Q}^T E \hat{Q}, \end{aligned} \tag{7.2}$$

where the second-order terms involving W are ignored. From the decomposition $X_1 = Q_1 R_1$ it follows that $Q_1 = X_1 R_1^{-1}$ which gives $Q_2^T X_1 = 0$. The equality $Y^T = X^{-1}$ implies that $Y_l^T X_l = I$ for $l = 1, 2$ and $Y_2^T X_1 = 0 = Y_1^T X_2$ and hence $Y_2^T Q_1 = 0$.

Using these relationships, equation (7.2) becomes

$$\hat{Q}^T H \hat{Q} = \begin{bmatrix} R_1 D_1 R_1^{-1} & Q_1^T X D Y^T Q_2 \\ 0 & Q_2^T X_2 D_2 Y_2^T Q_2 \end{bmatrix} + \hat{C} \tag{7.3}$$

$$\equiv \hat{H} + \hat{C}, \tag{7.4}$$

where the matrix $\hat{C}$ absorbs the three matrix products involving W or E on the right-hand side of equation (7.2). We note that if H is symmetric, $Q_1^T X_2 = 0 = Y_1^T Q_2$, R_1 is a diagonal matrix, and hence $R_1 D_1 R_1^{-1} = D_1$. Thus $\hat{H}$ is also a symmetric matrix. Defining $C = \hat{Q}\hat{C}\hat{Q}^T$ equation (7.4) is rewritten as $\hat{Q}^T(H - C)\hat{Q} = \hat{H}$. Since $Q\hat{H} = (X_1 D_1 Y_1^T + X_2 D_2 Y_2^T)Q$ and using the definition of $\hat{C}$ from equation (7.2),

$$\hat{C} = W^T Q \hat{H} + Q^T W \hat{H} - \hat{Q}^T E \hat{Q}, \tag{7.5}$$

it follows that $\|C\| \leq 2\|W^T Q\|\|\hat{H}\| + \|E\|$. The result of Theorem 1.1 of Stewart [36] also provides the estimate

$$\|W^T Q\| \leq \|W\| \leq \epsilon_2 \mu (1 + \epsilon_2 \mu \kappa(R_1)),$$

where $O(\epsilon^3)$ terms are ignored. For modest values of μ, W is numerically orthogonal to Q. From equation (7.5)

$$\begin{aligned} \|C\| &= \|\hat{C}\| \\ &\leq 2\epsilon_2 \mu (1 + \epsilon_2 \mu \kappa(R_1))\|\hat{H}\| + \epsilon_1 \kappa(X)\|H\| \\ &\leq 2\epsilon_2 \mu (1 + \epsilon_2 \mu \kappa(R_1))(\|H\| + \|C\|) + \epsilon_1 \kappa(X)\|H\| \\ &\leq \epsilon(\kappa(X) + \mu(1 + \epsilon\mu\kappa(R_1)))\|H\| + \epsilon\mu(1 + \epsilon\mu\kappa(R_1))\|C\| \\ &\equiv \theta\|H\| + \hat{\theta}\|C\|, \end{aligned}$$

where the second inequality uses equation (7.4). Since $\hat{\theta} < \theta$, rearranging the last inequality gives $\|C\|(1-\theta) \leq \theta\|H\|$. Ignoring $O(\theta^2)$ terms $\|C\| \leq \theta\|H\|$. The estimate on the size of C in equation (7.1) now follows since $\theta = \epsilon(\kappa(X) + \mu(1 + \epsilon\mu\kappa(X))) \leq \epsilon(\kappa(X) + \mu) + O(\epsilon^2)$. □

7.2. Purging. The success of the purging scheme depends upon the solution of the Sylvester set of equations required by Algorithm 6.3. We rewrite the Sylvester set of equations in Algorithm 6.3 as $ZH_{22} - H_{11}Z = H_{12}$. The job is to examine the effect of performing the similarity transformation $RH_{22}R^{-1}$, where

$$QR \equiv \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} R = \begin{bmatrix} Z \\ I \end{bmatrix} \equiv S.$$

The last relation implies that $R^{-1} = Q_2^T$. In actual computation, this equality obviates the need to solve linear systems with R necessary for the similarity transformation. For the error analysis that follows, R^{-1} is used in a formal sense.

Let $\hat{Z}$ be the computed solution to the Sylvester set of equations. In a similar analysis, Bai and Demmel [2] assume that the QR factorization of S is performed exactly and we do also. The major source of error is that arising from computing $\hat{Z}$.

Suppose that $\hat{Q}\hat{R} = \left[\begin{smallmatrix} \hat{Z} \\ I \end{smallmatrix}\right] \equiv \hat{S}$. Write $\hat{Z} = Z + E$ where E is the error in $\hat{Z}$. If $QR = S$ and $\|R^{-1}\|\|E\| < 1$, then Theorem 1.1 of Stewart [36] gives matrices W and F such that $(Q+W)(R+F) = \hat{Q}\hat{R}$ where $(Q+W)^T(Q+W) = I_m$. The result gives the bound $\|F\| \leq \|R\|\|E\| + O(\|E\|^2)$. Up to first-order perturbation terms,

$$\hat{R}H_{22}\hat{R}^{-1} = (R+F)H_{22}(R+F)^{-1} = RH_{22}R^{-1} + RH_{22}R^{-1}FR^{-1} + FH_{22}R^{-1}.$$

Defining the error matrix $C = H_{22}R^{-1}F + R^{-1}FH_{22}$ it follows that

$$\hat{R}H_{22}\hat{R}^{-1} = R(H_{22} + C)R^{-1}.$$

Ignoring second-order terms, we obtain the estimate

$$\|C\| \leq 2\|R^{-1}\|\|F\|\|H_{22}\| \leq 2\kappa(S)\|E\|\|H_{22}\|.$$

The invariance of $\|\cdot\|$ under orthogonal transformations gives $\kappa(S) = \|R^{-1}\|\|R\|$. Since the singular values of S are the square roots of the eigenvalues of S^TS it follows that

$$\kappa(S) = \sqrt{\frac{1+\sigma_{\max}^2(Z)}{1+\sigma_{\min}^2(Z)}},$$

where $\sigma_{\max}(Z)$ and $\sigma_{\min}(Z)$ are the largest and smallest singular values of Z. Since Z^TZ is a symmetric positive semidefinite matrix, $\lambda_{\max}(Z^TZ) = \|Z\|^2$, and then $\kappa(S) \leq \sqrt{1+\|Z\|^2}$, with equality if zero is an eigenvalue of Z^TZ.

The previous discussion is summarized in the following result.

THEOREM 7.2. *Let $\hat{Z}$ be the computed solution to the Sylvester set of equations, $ZH_{22} - H_{11}Z = H_{12}$, where the eigenvalues of H_{11} and H_{22} are distinct. Let $\hat{Z} = Z+E$ where E is the error in $\hat{Z}$ and suppose that $\|R^{-1}\|\|E\| < 1$ where $QR = \left[\begin{smallmatrix} Z \\ I \end{smallmatrix}\right]$.*

Then there exists a matrix C such that

$$\hat{R}H_{22}\hat{R}^{-1} = R(H_{22} + C)R^{-1},$$

where

$$\|C\| \leq 2\sqrt{1+\|Z\|^2}\,\|E\|\,\|H\|. \tag{7.6}$$

If $\|E\|$ is a modest multiple of machine precision and the solution of Sylvester's equations is not large in norm, then purging is backward stable since $\|C\|$ is small relative to $\|H\|$.

The two standard approaches [3, 16] for solving Sylvester's equation show that $\|\hat{F}\|_F \leq \epsilon_3(\|H_{11}\|_F + \|H_{22}\|_F)\|\hat{Z}\|_F$ where $\hat{F} \equiv H_{12} - \hat{Z}H_{22} + H_{11}\hat{Z}$ and ϵ_3 is a modest multiple of machine precision. Standard bounds [8, 15] also give $\|Z\|_F \leq \text{sep}^{-1}(H_{11}, H_{22})\|H_{12}\|_F$ where

$$\text{sep}(H_{11}, H_{22}) \equiv \min_{X \neq 0} \frac{\|XH_{22} - H_{11}X\|_F}{\|X\|_F}$$

is the *separation* between H_{11} and H_{22}. Although

$$\text{sep}(H_{11}, H_{22}) \leq \min_{k,l} |\lambda_k(H_{11}) - \lambda_l(H_{22})|,$$

Varah [38] indicates that if the matrices involved are highly non-normal, the smallest difference between the spectrums of H_{11} and H_{22} may be an overestimate of the actual separation. Recently, Higham [19] gave a detailed error analysis for the solution of Sylvester's equation. The analysis takes into account the special structure of the equations involved. For example, Higham shows that $\|E\|_F \leq \text{sep}^{-1}(H_{11}, H_{22})\|\hat{F}\|_F$, but this may lead to an arbitrarily large estimate of the true forward error. For use in practical error estimation, LAPACK-style software is available.

A robust implementation of procedure `Lock` determines the backward stability by estimating both $\|Z\|$ and $\|E\|$.

8. Other deflation techniques. Wilkinson [40, pp. 584–602] has given a comprehensive treatment of various deflation schemes associated with iterative methods. Recently, Saad [31, pp. 117–125, 180–182] discussed several deflation strategies used with both simultaneous iteration and Arnoldi's method. Algorithm 6.2 is an in-place version of one of these schemes [31, p. 181]. Saad's version explicitly orthonormalizes the newly converged Ritz vectors against the already computed approximate j Schur vectors. This is the form of locking used by Scott [33]. Instead, procedure `Lock` achieves the same task implicitly through the use of Householder matrices in $\mathbf{R}^{m \times m}$. Thus, we are able to orthogonalize vectors in $\mathbf{R}^n$ at a reduced expense since $m \ll n$.

Other deflation strategies include the various Wielandt deflation techniques [31, 40]. We briefly review those that do not require the approximate left eigenvectors of A or complex arithmetic. Denote by $\lambda_1, \ldots, \lambda_j$ the wanted eigenvalues of A. The Wielandt and Schur–Wielandt forms of deflation determine a rank j modification of A,

$$A_j = A - U_j S_j U_j^T, \tag{8.1}$$

where $S_j \in \mathbf{R}^{j \times j}$ and j represents the dimension of the approximate invariant subspace already computed. The idea is to choose S_j so that A_j will converge to the remainder of the invariant subspace desired. For example, S_j is selected to be a diagonal matrix of shifts $\sigma_1, \ldots, \sigma_j$ so that A_j has eigenvalues $\{\lambda_1 - \sigma_1, \ldots, \lambda_j - \sigma_j, \lambda_{j+1}, \ldots, \lambda_n\}$.

Both forms of deflation differ in the choice of U_j. The Wielandt variant uses converged Ritz vectors while the Schur–Wielandt uses approximate Schur vectors. With either form of deflation, the eigenvalues of A_j are $\lambda_i - \sigma_i$ for $i \leq j$ and λ_i otherwise, and both forms leave the Schur vectors unchanged. This motivates Saad to suggest that an approximate Schur basis should be incrementally built as Ritz vectors of A_j converge. Braconnier [6] employs the Wielandt variant and discusses the details of deflating a converged Ritz value that has a nonzero imaginary part in real arithmetic.

We now compare our locking scheme to the Schur–Wielandt deflation technique. We shall assume that $AU_j = U_j R_j$ is a real partial Schur form of order j for A, and we will put $S_j = R_j$ in the Schur–Wielandt deflation scheme. Suppose that

$$A \begin{bmatrix} U_j & V_m \end{bmatrix} = \begin{bmatrix} U_j & V_m \end{bmatrix} \begin{bmatrix} R_j & M_j \\ 0 & H_m \end{bmatrix} + f_{m+j} e_{m+j}^T \tag{8.2}$$

is a length $m + j$ Arnoldi factorization obtained after locking. Consider any associated round-off errors as being absorbed in A here. Equate the last m columns of equation (8.2) to obtain

$$AV_m = U_j M_j + V_m H_m + f_{m+j} e_m^T. \tag{8.3}$$

Since U_j is orthogonal to V_m, it follows that $(I - U_j U_j^T)A(I - U_j U_j^T)V_m = V_m H_m + f_{m+j} e_m^T$. This implies that the Arnoldi factorization (8.2) is equivalent to applying Arnoldi's method to the projected matrix $(I - U_j U_j^T)A(I - U_j U_j^T)$ with the first column of V_m as the starting vector. Keeping the locked vectors active in the construction and the IRA update of this Arnoldi factorization ensures that the Krylov space generated by V_m remains free of components corresponding to locked Ritz values. The appearance of spurious Ritz values in the subsequent factorization is automatically avoided. Note that when A is symmetric, this is equivalent to the selective orthogonalization scheme proposed by Parlett [26, pp. 275–284] and Scott.

In contrast to locking, consider the consequences of applying the Schur–Wielandt deflation scheme to construct a new Arnoldi factorization using $V_m e_1$ as a starting vector. In the symmetric case with exact arithmetic, the two schemes would be mathematically equivalent. Without these assumptions, there may be considerable differences. From equation (8.3), it follows that (with A replaced with A_j of equation (8.1))

$$(A - U_j R_j U_j^T)V_m = A(I - U_j U_j^T)V_m = U_j M_j + V_m H_m + f_{m+j} e_m^T. \tag{8.4}$$

From equation (8.4) we can use an easy induction to derive the relations

$$(A - U_j R_j U_j^T)^i V_m e_1 = (U_j M_j + V_m H_m) H_m^{i-1} e_1, \quad i \geq 1.$$

Thus, the Krylov subspace $\mathcal{K}_k(A - U_j R_j U_j^T, V_m e_1)$ and hence the corresponding Arnoldi factorization of $A - U_j R_j U_j^T$ *must* be corrupted with components in $\mathcal{R}(U_j)$ when the starting vector is orthogonal to $\mathcal{R}(U_j)$. Within the context of Arnoldi iterations, the Schur–Wielandt technique does not deflate the invariant subspace information contained in the $\mathcal{R}(U_j)$ from the remainder of the iteration. In other words, Schur–Wielandt deflation is unstable.

This helps to explain why Saad suggests that Wielandt and Schur–Wielandt deflation techniques should not be used "to compute more than a few eigenvalues and eigenvectors" [31, p. 125]. We note that if $M_j \approx 0$, then the Wielandt forms of deflation may safely be used within an Arnoldi iteration. This will always be true when A is symmetric.

The cost of matrix–vector products with A_j increases due to the rank j modifications of A required. Moreover, every time an approximate Schur vector or a Ritz vector converges, the iteration needs to be explicitly restarted with A_j. The two deflation techniques introduced in this paper allow the iteration to be implicitly restarted—avoiding the need to build a new factorization from scratch.

Finally, we mention that the idea of deflating a converged Ritz value from a Lanczos iteration is also discussed by Parlett and Nour-Omid [28]. They present an explicit deflation technique by using the QR algorithm with converged Ritz values as shifts. Parlett indicates that this was a primary reason for undertaking the study concerning the forward instability of the QR algorithm [27].

9. Reordering the Schur form of a matrix. We now establish a connection between the IRA iteration with locking and the algorithms used to reorder the Schur form of a matrix. Suppose a matrix A is reduced to upper quasi-triangular form by the QR algorithm

$$Q^T A Q = T \equiv \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}, \tag{9.1}$$

where Q is the orthogonal matrix computed by the algorithm. Equation (9.1) is a Schur form for A of order $p+q$ where the submatrices T_{11} and T_{22} are of order p and q, respectively. Assume that the spectrums of T_{11} and T_{22} are distinct. In practice, the order in which the computed eigenvalues of A appear on the diagonal of T is somewhat random. The first p columns of Q are an orthogonal basis for the unique invariant subspace associated with the eigenvalues of T_{11}. If the eigenvalues of interest are located in T_{22} and an orthonormal basis for them is wanted, we must either increase the number of columns of Q used or somehow place them at the top of T. Algorithms for reordering a Schur form accomplish this task by using orthogonal matrices that move the wanted eigenvalues to the top of T. The recent work of Bai and Demmel [2] attempts to correct the occasional numerical problems encountered by Stewart's algorithm [35] EXCHNG. Their work was motivated by that of Ruhe [29] and that of Dongarra, Hammarling, and Wilkinson [11]. Both algorithms swap consecutive 1×1 and 2×2 blocks of a quasi-triangular matrix to attain the desired ordering.

Let both T_{11} and T_{22} of equation (9.1) be matrices of at most order-2. When swapping adjacent blocks of order-1, $p = 1 = q$, EXCHNG constructs a plane rotation that zeros the second component of the eigenvector corresponding to the eigenvalue $\lambda_2 = T_{22}$. A similarity transformation is performed on T with the plane rotation and the diagonal blocks are interchanged. We refer to a strategy that constructs an orthogonal matrix and performs a similarity transformation to interchange the eigenvalues as a *direct* swapping algorithm. Consider the following alternate *iterative* swapping algorithm: Perform a similarity transformation on T with an arbitrary orthogonal matrix followed by one step of the QR iteration with shift equal to λ_2. The arbitrary orthogonal similarity transformation introduces a nonzero off-diagonal element in the $2,1$ entry so that the transformed T is an unreduced upper Hessenberg matrix with the diagonal blocks now coupled. The standard convergence theory of the QR algorithm dictates that λ_1 and λ_2 are switched and the $2,1$ entry is zero. If the order of T_{22} is equal to two, EXCHNG uses the iterative swapping strategy using a standard double shift to reorder the diagonal blocks. The direct swapping algorithm instead computes an appropriate orthogonal matrix by computing the QR factorization of a basis of two vectors that span the desired invariant subspace. For example, the factorization used in equation (6.1) in §6.1 may be used. The reader is referred to [2, 11] for further details.

The iterative swapping algorithm is equivalent to the implicit restarting technique used by the IRA iteration, since both depend upon an implicitly shifted QR step applied to an unreduced upper Hessenberg matrix to interchange T_{11} and T_{22}. The direct swapping algorithm is equivalent to the locking technique. An orthogonal

matrix is constructed from a basis for the invariant subspace corresponding to T_{22}. When this is applied as a similarity transformation, the diagonal blocks of T are swapped. In exact arithmetic, both swapping variants result in a matrix that is upper quasi-triangular with the blocks interchanged. Unfortunately, these existing reordering techniques do not preserve the leading portion of the Arnoldi factorization and thus explicit restarting would have to be used.

The following example demonstrates that the two variants may produce drastically different output matrices when computed in floating point arithmetic. The following experiment was carried out in MATLAB, Version 4.2a, on a SUN SPARC station IPX. The floating point arithmetic is IEEE standard double precision with machine precision of $\epsilon_M \equiv 2^{-52} \approx 2.2204 \cdot 10^{-16}$. Let

$$T = \begin{bmatrix} 1 + 10\epsilon_M & 1 \\ 0 & 1 \end{bmatrix}.$$

An eigenvector corresponding to $\lambda_2 = 1$ is

$$\begin{bmatrix} -1 \\ 10\epsilon_M \end{bmatrix}.$$

Denote by Z the plane rotation that transforms this eigenvector to a multiple of the first column of the identity matrix in $\mathbf{R}^{2\times 2}$. Let

$$U = \begin{bmatrix} 1 & -5\epsilon_M \\ 10\epsilon_M & 1 \end{bmatrix},$$

so that U is orthogonal up to a small multiple of machine precision. The matrix U acts as the arbitrary orthogonal transformation required by the iterative algorithm. Let $\hat{T}$ denote the matrix computed by performing one step of the QR iteration to the matrix $U^T T U$ with shift equal to $\lambda_1 = 1 + 10\epsilon_M$. We remark that for matrices of order-2, the explicit and implicit formulations of the QR iteration are equivalent. The two computed matrices are

$$Z^T T Z = \begin{bmatrix} 1 & -1 \\ 0 & 1 + 10\epsilon_M \end{bmatrix},$$

$$\hat{T} = \begin{bmatrix} 1.400000000000003 & -7.999999999999996 \cdot 10^{-1} \\ 2.000000000000002 \cdot 10^{-1} & 6.000000000000001 \cdot 10^{-1} \end{bmatrix}.$$

The computed eigenvalues of $\hat{T}$ are 1.000000033320011 and $9.999999666799921 \cdot 10^{-1}$ which both lost eight digits of accuracy. If we perform another QR step on the matrix $\hat{T}$ with the same shift,

$$\begin{bmatrix} 1.000000000000003 & 1.000000000000001 \\ \approx 1.09 \cdot 10^{-15} & 1 \end{bmatrix}$$

is computed. Note that the off-diagonal element is slightly larger than machine precision so that a standard QR algorithm does not set it to zero. Moreover, even if the off-diagonal element is set to zero, the iterative swapping algorithm fails to interchange the eigenvalues. Continuing to apply QR steps with the shift equal to λ_1 does not result in a properly interchanged matrix.

The explanation of why the iterative algorithm fails to work is simple enough. The matrix T constructed is poorly conditioned with respect to the eigenvalue problem since the eigenvectors are nearly aligned. The eigenvalues of $U^T T U$ are

$$1.000000033320011 \quad \text{and} \quad 9.999999666799921 \cdot 10^{-1}.$$

Thus, the small relative errors on the order of machine precision that occur when computing U^TTU produce a nearby matrix in which both eigenvalues differ by eight digits of accuracy. Performing a shifted QR step with λ_1 incurs forward instability since the last components of the eigenvectors for U^TTU are on the order of $\sqrt{\epsilon_M}$. This is the necessary and sufficient condition of Parlett and Le [27]. Another QR step with the same shift on $\hat{T}$ almost zeros out the subdiagonal element since the last components of the eigenvectors for $\hat{T}$ are of order 10^{-1} and the shift is almost the average of the eigenvalues of $\hat{T}$ and quite close to both. We emphasize that the loss of accuracy of the computed eigenvalues is one of the deleterious effects of forward instability.

Bai and Demmel [2] present an example which compares their direct swapping approach with Stewart's algorithm EXCHNG. The matrix considered is

$$A(\tau) = \begin{bmatrix} 7.001 & -87 & 39.4\tau & 22.2\tau \\ 5 & 7.001 & -12.2\tau & 36.0\tau \\ 0 & 0 & 7.01 & -11.7567 \\ 0 & 0 & 37 & 7.01 \end{bmatrix}.$$

When $\tau = 10$, ten QR iterations are required to interchange the two blocks. As before, the eigenvalues undergo a loss of accuracy. The iterative swapping algorithm fails for the matrix $A(100)$. No explanation is given for the failure of Stewart's algorithm. The explanation for the failure is the same as for the previous example. Using a direct algorithm, the eigenvalues of $A(10)$ and $A(100)$ are correctly swapped and the eigenvalues lose only a tiny amount of accuracy.

Bai and Demmel present a rigorous analysis of their direct swapping algorithm. Although backward stability is not guaranteed, it appears that only when both T_{11} and T_{22} are of order-2 and have almost indistinguishable eigenvalues [5] is stability lost. In this case, the interchange is not performed. Bojanczyk and Van Dooren [5] present an alternate swapping algorithm that appears to be backward stable.

10. Numerical results. An IRA iteration using the two deflation procedures of §6.2 was written in MATLAB, Version 4.2a. An informal description given parameters k and p is given in Table 10.1. The codes are available from the first author upon request. A high-quality and robust implementation of the deflation procedures is planned for the Fortran software package ARPACK [24].

In the examples that follow, Q_k and R_k denote the approximate Schur factors for an invariant subspace of order-k computed by an IRA iteration. All the experiments used the starting vector equal to `randn(n, 1)`, where the seed is set with `randn('seed', 0)` and `n` is the order of the matrix. The shifting strategy uses the unwanted eigenvalues of H_{k+p} that have not converged. An eigenpair (θ, y) of H_{k+p} is accepted if its Ritz estimate (2.2) satisfies

$$|e_{k+p}^T y| \, \|f_{k+p}\| \leq \eta |\theta|. \tag{10.1}$$

The value of η is chosen according to the relative accuracy of the Ritz value desired.

10.1. Example 1. The first example illustrates the use of the deflation techniques when the underlying matrix has several complex repeated eigenvalues. The example also demonstrates how the iteration locks and purges blocks of Ritz values in real arithmetic. A block diagonal matrix C was generated having n blocks of order-2. Each block was of the form

$$\begin{bmatrix} \xi_l & \eta_l \\ -\eta_l & \xi_l \end{bmatrix},$$

TABLE 10.1
Formal description of an IRA iteration.

```
1. Initialize an Arnoldi factorization of length k
2. Main Loop
   3. Extend an Arnoldi factorization to length k + p
   4. Check for convergence
      Exit if k wanted Ritz values converge
      Let i and j denote the wanted and unwanted converged
      Ritz values, respectively
   5. Lock the i + j converged Ritz values
   6. Implicit application of shifts resulting in an
      Arnoldi factorization of length k + j
   7. Purge the j unwanted converged Ritz values.
```

where

$$\xi_{l=i+j-1} \equiv 4\sin^2\left(\frac{i\pi}{2(n+1)}\right) + 4\sin^2\left(\frac{j\pi}{2(n+1)}\right)$$

for $1 \le i, j \le n$, and $\eta_l \equiv \sqrt{\xi_l}$. The eigenvalues of C are $\xi_l \pm \eta_l i$ where $i = \sqrt{-1}$. Since the eigenvalues of a quasi-diagonal matrix are invariant under orthogonal similarity transformations, using an IRA iteration on C with a randomly generated starting vector is general. An IRA iteration was used to compute the $k = 12$ eigenvalues of C_{450} with smallest real part. The number of shifts used was $p = 16$ and the convergence tolerance η was set equal to 10^{-10}. With these choices of k and p, the iteration stores at most 28 Arnoldi vectors. There are four eigenvalues with multiplicity two. Table 10.2 shows the results attained. Let the diagonal matrix D_{12} denote the eigenvalues of the upper triangular matrix R_{12} computed by the iteration. The diagonal matrix Λ_{12} contains the wanted eigenvalues. After 24 iterations, 12 Ritz values converged. But the pair of Ritz values purged at iteration 21 was a previously locked value which the iteration discarded. This behavior is typical when there are clusters of eigenvalues.

10.2. Example 2. Consider the eigenvalue problem for the convection–diffusion operator

$$-\Delta u(x,y) + \rho(u_x(x,y) + u_y(x,y)) = \lambda u(x,y)$$

on the unit square $[0,1] \times [0,1]$ with zero boundary data. Using a standard five-point scheme with centered finite differences, the matrix L_{n^2} that arises from the discretization is of order n^2 where $h = 1/(n+1)$ is the cell size. The eigenvalues of L_{n^2} are

$$\lambda_{ij} = 2\sqrt{1-\gamma}\cos\left(\frac{i\pi}{n+1}\right) + 2\sqrt{1-\gamma}\cos\left(\frac{j\pi}{n+1}\right)$$

for $1 \le i, j \le n$ where $\gamma = \rho h/2$. An IRA iteration was used to compute the $k = 6$ smallest eigenvalues of L_{625} where $\rho = 25$. The number of shifts used was $p = 10$ and the convergence tolerance η was set equal to 10^{-8}. With these choices of k and p, the iteration stores at most 16 Lanczos vectors. Let the diagonal matrix D_6 denote the eigenvalues of the upper triangular matrix R_6 computed by the iteration. The diagonal matrix $\Lambda_6 \in \mathbf{R}^{6\times 6}$ contains the six smallest eigenvalues. We note that there are two eigenvalues with multiplicity two. Table 10.3 shows the results attained. The diagonal matrix D_6 approximates Λ_6. After 30 iterations six Ritz values converged.

TABLE 10.2
Convergence history for Example 1.

IRA iteration for C_{450}		
$k = 12$ and $p = 16$ with convergence tolerance $\eta = 10^{-10}$		
Iteration	Ritz values Locked	Ritz values Purged
9	2	0
10	2	0
12	2	0
13	2	0
17	2	0
21	0	2
24	2	0
28	0	2
31	2	0
Totals	14	4
Number of matrix–vector products		436
$\|C_{450}Q_{12} - Q_{12}R_{12}\| \approx 10^{-12}$		
$\|Q_{12}^T C_{450}Q_{12} - R_{12}\| \approx 10^{-11}$		
$\|Q_{12}^T Q_{12} - I_{12}\| \approx 10^{-14}$		
$\|D_{12} - \Lambda_{12}\|_\infty \approx 10^{-15}$		

But the Ritz value purged at iteration 24 was a previously locked value. The other purged Ritz values are approximations to the eigenvalues of L_{625} larger than λ_6.

Figure 10.1 gives a graphical interpretation of the expense of an IRA iteration in terms of matrix–vector products when the value of p is increased. For all values of p shown, the results of the iteration were similar to those of Table 10.3. The results presented in Table 10.3 correspond to the value of p that gave the minimum number of matrix–vector products. For the value of $p = 1$, the iteration converged to the five smallest eigenvalues after 999 matrix–vector products. But the iteration was not able to converge to the second copy of λ_5. For $p = 2$, the only form of deflation employed was locking. All other values of p shown demonstrated similar behavior to that of Table 10.3.

In order to determine the benefit of the two deflation techniques, experiments were repeated without the use of locking or purging. In addition, all the unwanted Ritz values were used as shifts, converged or not. The first run used the same parameters as given in Table 10.3. After 210 matrix–vector products, the iteration converged to six Ritz values. But the second copy of the fifth smallest eigenvalue was not among the final six. The value of p was increased to 23 with the same results.

10.3. Example 3. The following example shows the behavior of the iteration on a matrix with a very ill-conditioned basis of eigenvectors. Define the Clement tridiagonal matrix [20] of order $n + 1$:

$$B_{n+1} = \begin{bmatrix} 0 & n & \cdots & 0 \\ 1 & 0 & n-1 & \\ \vdots & \ddots & \ddots & \\ 0 & & n & 0 \end{bmatrix}.$$

TABLE 10.3
Convergence history for Example 2.

IRA iteration on L_{625}		
$k = 6$ and $p = 10$ with convergence tolerance $\eta = 10^{-8}$		
Iteration	Ritz values Locked	Ritz values Purged
14	1	0
16	1	0
19	1	0
21	1	0
23	1	1
24	0	1
30	1	0
35	0	1
38	1	1
Totals	7	4
Number of matrix–vector products		325
$\|L_{625}Q_6 - Q_6R_6\| \approx 10^{-9}$		
$\|Q_6^T L_{625}Q_6 - R_6\| \approx 10^{-9}$		
$\|Q_6^T Q_6 - I_6\| \approx 10^{-14}$		
$\|D_6 - \Lambda_6\|_\infty \approx 10^{-7}$		

The eigenvalues are $\pm n, \pm n-2, \ldots, \pm 1$, and zero if n is even. We note that $B_{n+1} = S_{n+1}A_{n+1}S_{n+1}^{-1}$ where $S_{n+1}^2 = \text{diag}(1, \frac{n}{1}, \frac{n}{1}\frac{n-1}{2}, \ldots, \frac{n!}{n!})$ is a diagonal matrix. Thus the condition number of the basis of eigenvectors for B_{n+1} is $\|S_{n+1}\|\,\|S_{n+1}^{-1}\|$ which implies that the eigenvalue problem for B_{n+1} is quite ill conditioned. An IRA iteration was used to compute the $k = 4$ largest in magnitude eigenvalues of B_{1000}. The number of shifts used was $p = 16$, and the convergence tolerance η was set equal to 10^{-6}. With these choices of k and p, the iteration stores at most 20 Arnoldi vectors. Let the diagonal matrix D_4 denote the eigenvalues of the upper triangular matrix R_4 computed by the iteration. The diagonal matrix $\Lambda_4 \in \mathbf{R}^{4\times 4}$ contains the four largest in magnitude eigenvalues. Table 10.4 shows the results attained. Although the iteration needed a large number of matrix–vector products, the iteration was able to extract accurate Ritz values given the convergence tolerance.

10.4. Example 4. Finally, we present a dramatic example of how the convergence of an IRA iteration benefits from the two deflation procedures. A matrix T of order-10 had the values

$$\tau_1 = 10^{-6}, \tau_{i=2:8} = i \cdot 10^{-3}, \tau_{9:10} = 1$$

on the diagonal. Since the eigenvalues of a matrix are invariant under orthogonal similarity transformations, using an IRA iteration on T with a randomly generated starting vector is general. An IRA iteration was used to compute an approximation to the smallest eigenvalue. The number of shifts used was $p = 3$ and the convergence tolerance η was set equal to 10^{-3}. Table 10.5 shows the results attained. Another experiment was run with the locking and purging mechanisms turned off. Additionally, all unwanted Ritz values were used as shifts. The same parameters were used as in Table 10.5 but the iteration now consumed 41 matrix–vector products. As in

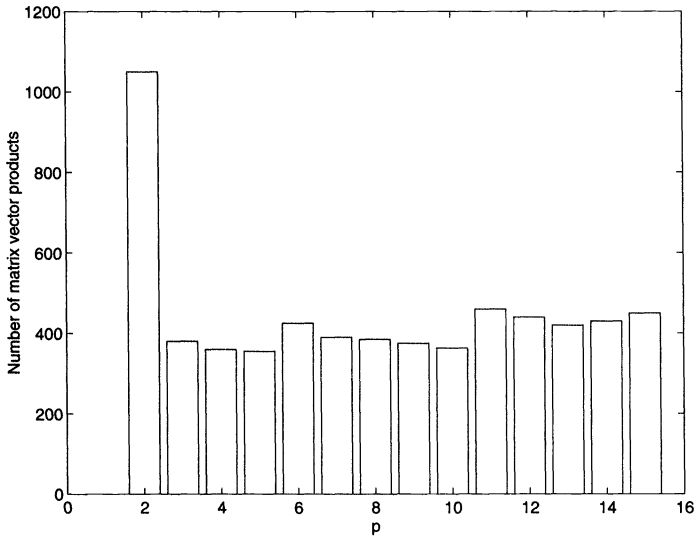


FIG. 10.1. *Bar graph of the number of matrix–vector products used by an IRA iteration for Example* 2 *as a function of* p.

TABLE 10.4
Convergence history for Example 3.

IRA iteration on B_{1000}		
$k = 4$ and $p = 16$ with convergence tolerance $\eta = 10^{-6}$		
Iteration	Ritz values Locked	Ritz values Purged
76	1	0
85	1	0
91	2	0
Totals	4	0
Number of matrix–vector products		1423
$\|B_{1000}Q_4 - Q_4R_4\|/\|B_{1000}\| \approx 10^{-6}$		
$\|Q_4^T B_{1000}Q_4 - R_4\| \approx 10^{-6}$		
$\|Q_4^T Q_4 - I_4\| \approx 10^{-14}$		
$\|D_4 - \Lambda_4\|_\infty/\|B_{1000}\|_\infty \approx 10^{-6}$		

the results for Table 10.5, the modified iteration converged to one of the dominant eigenvalues after one iteration. After six iterations, the leading block of H_4 split off, having converged to the invariant subspace corresponding to $\tau_{9:10}$. But since purging was turned off, the modified iteration had to continue attempting to converge to τ_1 using only the lower block of order-2 in H_4. Incidentally, if the iteration instead simply discarded the leading portion of the factorization corresponding to $\tau_{9:10}$ after the sixth iteration, convergence to τ_1 never occurred. Crucial to the success of an IRA iteration is the ability to deflate converged Ritz values in a stable manner. Both purging and locking allow faster convergence.

11. Conclusions. In the paper, we developed deflation techniques for an implicitly restarted Arnoldi iteration. The first technique, locking, allows an orthogonal

TABLE 10.5
Convergence history for Example 4.

IRA iteration on T		
$k = 1$ and $p = 3$ with convergence tolerance $\eta = 10^{-3}$		
Iteration	Ritz values Locked	Ritz values Purged
1	0	1
15	1	1
Totals	1	2
Number of matrix–vector products		32
$\|TQ_1 - Q_1R_1\|/\tau_1 \approx 10^{-3}$		
$\|Q_1^TTQ_1 - R_1\|/\tau_1 \approx 10^{-3}$		
$\|Q_1^TQ_1 - I_1\| \approx 10^{-15}$		
$\|R_1 - \tau_1\|_\infty/\tau_1 \approx 10^{-3}$		

change of basis for an Arnoldi factorization which results in a partial Schur decomposition containing the converged Ritz values. The corresponding Ritz value is deflated in an implicit but direct manner. The second technique, purging, allows implicit removal of unwanted converged Ritz values from the Arnoldi iteration. Both deflation techniques are accomplished by working with matrices in the projected Krylov space which for large eigenvalue problems is a fraction of the order of the matrix from which estimates are sought. Since both deflation techniques are implicitly applied to the Arnoldi factorization the need for explicit restarting associated with all other deflation strategies is avoided. Both techniques were carefully examined with respect to numerical stability and computational results were presented. Convergence of the Arnoldi iteration is improved and a reduction in computational effort is realized. Although a direct comparison with block Arnoldi/Lanczos methods was not given, computational experience shows that if an IRA iteration builds the same size factorization used by the block methods and the convergence tolerance is small enough, multiple or clustered eigenvalues are correctly computed. The connection between an IRA and QR iteration explains the reason for the size of the convergence tolerance used.

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] Z. BAI AND J. W. DEMMEL, *On swapping diagonal blocks in real Schur form*, Linear Algebra Appl., 186 (1993), pp. 73–95.

[3] R. H. BARTELS AND G. W. STEWART, *Algorithm* 432: *Solution of the matrix equation* $AX + XB = C$, Comm. ACM, 15 (1972), pp. 820–826.

[4] M. BENNANI AND T. BRACONNIER, *Stopping Criteria for Eigensolvers*, Tech. report TR/PA/93/25, CERFACS, Toulouse, France, 1993.

[5] A. BOJANCZYK AND P. VAN DOOREN, *Reordering diagonal blocks in the Schur form*, in Linear Algebra for Large Scale and Real Time Applications, NATO ASI Series, Kluwer Academic Publishers, Norwell, MA, 1993, pp. 351–352.

[6] T. BRACONNIER, *The Arnoldi–Tchebycheff Algorithm for Solving Large Nonsymmetric Eigenproblems*, Tech. report TR/PA/93/25, CERFACS, Toulouse, France, 1993.

[7] D. CALVETTI, L. REICHEL, AND D. C. SORENSEN, *An implicitly restarted Lanczos method for large symmetric eigenvalue problems*, Electron. Trans. Numer. Anal., 2 (1994), pp. 1–21.

[8] F. CHATELIN, *Eigenvalues of Matrices*, John Wiley, New York, 1993.
[9] F. CHATELIN AND V. FRAYSÉE, *Qualitative Computing: Elements of a Theory for Finite-Precision Computation*, Tech. report, CERFACS and THOMSON–CSF, June 1993. Lecture Notes for the Commett European Course, June 8–10, Orsay, France.
[10] J. CULLUM AND W. E. DONATH, *A block Lanczos algorithm for computing the q algebraically largest eigenvalues and a corresponding eigenspace for large, sparse symmetric matrices*, in Proc. 1974 IEEE Conference on Decision and Control, New York, 1974, pp. 505–509.
[11] J. DONGARRA, S. HAMMARLING, AND J. WILKINSON, *Numerical considerations in computing invariant subspaces*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 145–161.
[12] I. S. DUFF AND J. A. SCOTT, *Computing selected eigenvalues of large sparse unsymmetric matrices using subspace iteration*, ACM Trans. Math. Software, 19 (1993), pp. 137–159.
[13] W. FELLER AND G. FORSYTHE, *New matrix transformations for obtaining characteristic vectors*, Quart. Appl. Math., 8 (1951), pp. 325–331.
[14] S. GODET-THOBIE, *Eigenvalues of Large Highly Nonnormal Matrices*, Ph.D. thesis, University Paris IX, Dauphine, Paris, France, 1993; Tech. report TH/PA/93/06, CERFACS, Toulouse, France, 1993.
[15] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
[16] G. H. GOLUB, S. NASH, AND C. F. VAN LOAN, *A Hessenberg–Schur method for the problem* $AX + XB = C$, IEEE Trans. Automat. Control, AC-24 (1979), pp. 909–913.
[17] G. H. GOLUB AND R. UNDERWOOD, *The block Lanczos method for computing eigenvalues*, in Mathematical Software III, J. R. Rice, ed., Academic Press, New York, 1977, pp. 361–377.
[18] R. G. GRIMES, J. G. LEWIS, AND H. D. SIMON, *A shifted block Lanczos algorithm for solving sparse symmetric generalized eigenproblems*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 228–272.
[19] N. J. HIGHAM, *Perturbation theory and backward error for* $AX - XB = C$, BIT, 33 (1993), pp. 124–136.
[20] ———, *The Test Matrix Toolbox for Matlab*, Numerical Analysis Report 237, University of Manchester, England, 1993.
[21] W. KARUSH, *An iterative method for finding characteristics vectors of a symmetric matrix*, Pacific J. Math., 1 (1951), pp. 233–248.
[22] C. LANCZOS, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Research Nat. Bur. Standards, 45 (1950), pp. 255–282.
[23] R. B. LEHOUCQ, *Analysis and Implementation of an Implicitly Restarted Iteration*, Ph.D. thesis, Rice University, Houston, TX, May 1995; Tech. report TR95-13, Dept. of Computational and Applied Mathematics, Rice University, Houston, TX, 1993.
[24] R. B. LEHOUCQ, D. C. SORENSEN, AND P. VU, *ARPACK:* An Implementation of the Implicitly Re-started Arnoldi Iteration That Computes Some of the Eigenvalues and Eigenvectors of a Large Sparse Matrix, 1995. Available from netlib@ornl.gov under the directory scalapack.
[25] C. C. PAIGE, *The Computation of Eigenvalues and Eigenvectors of Very Large Sparse Matrices*, Ph.D. thesis, University of London, London, England, 1971.
[26] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
[27] B. N. PARLETT AND J. LE, *Forward instability of tridiagonal QR*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 279–316.
[28] B. N. PARLETT AND B. NOUR-OMID, *The use of a refined error bound when updating eigenvalues of tridiagonals*, Linear Algebra Appl., 68 (1984), pp. 179–219.
[29] A. RUHE, *An algorithm for numerical determination of the structure of a general matrix*, BIT, 10 (1970), pp. 196–216.
[30] Y. SAAD, *Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices*, Linear Algebra Appl., 34 (1980), pp. 269–295.
[31] ———, *Numerical Methods for Large Eigenvalue Problems*, Halsted Press, New York, 1992.
[32] M. SADKANE, *A block Arnoldi–Chebyshev method for computing the leading eigenpairs of large sparse unsymmetric matrices*, Numer. Math., 64 (1993), pp. 181–193.
[33] J. A. SCOTT, *An Arnoldi code for computing selected eigenvalues of sparse real unsymmetric matrices*, ACM Trans. Math. Software, 21 (1995), pp. 432–475.
[34] D. C. SORENSEN, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 357–385.
[35] G. W. STEWART, *ALGORITHM* 506: *HQR3 and EXCHANG: Fortran subroutines for calculating and ordering the eigenvalues of a real upper Hessenberg matrix* [F2], ACM Trans. Math. Software, 2 (1976), pp. 275–280.
[36] ———, *Perturbation bounds for the QR factorization of a matrix*, SIAM J. Numer. Anal., 14

(1977), pp. 509–518.
[37] W. STEWART AND A. JENNINGS, *A simultaneous iteration algorithm for real matrices*, ACM Trans. Math. Software, 7 (1981), pp. 184–198.
[38] J. M. VARAH, *On the separation of two matrices*, SIAM J. Numer. Anal., 16 (1979), pp. 216–222.
[39] D. S. WATKINS, *Forward stability and transmission of shifts in the QR algorithm*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 469–487.
[40] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, UK, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 4, pp. 822–833, October 1996

007

ON THE DYNAMICS OF THE LINEAR PROCESS $Y(k) = A(k)Y(k-1)$ WITH IRREDUCIBLE MATRICES $A(k)$*

MARC ARTZROUNI†

Abstract. An upper bound is given for the projective distance $d(Y(k), V(k))$, where $Y(k) = A(k)Y(k-1)$ is a nonautonomous linear process in the positive quadrant of $\mathbb{R}^n$ ($A(i)$ irreducible for all i) and the $V(i)$'s are the Perron vectors of the $A(i)$'s. This bound shows that the larger k is, and the more slowly the Perron vectors have varied in the recent past preceding k, the smaller the distance $d(Y(k), V(k))$ will be. Corollaries are given concerning the growth of the $Y(k)$'s and the behavior of the backward product of matrices $A(1)A(2)\dots A(k)$. The results are illustrated with a numerical simulation. The cases of stochastic matrices and cyclic matrices are investigated.

Key words. nonnegative matrices, projective distance, products, ergodicity, Perron root, Perron vector, cyclic matrix, stochastic matrix

AMS subject classifications. 15A04, 15A42, 15A48, 15A51

1. Introduction.

Infinite products of matrices are used in the study of linear inhomogeneous processes. In many applications the matrices are nonnegative, and this case has received a fair amount of attention ([3–10], among others).

One approach to this problem is the consideration of the linear nonautonomous process $Y(k) = A(k)Y(k-1)$, where the $Y(k)$'s are vectors in the positive quadrant of $\mathbb{R}^n$ and the $A(k)$'s are nonnegative square matrices of order n. Indeed, when $Y(0)$ is the vector with rth component "1" and zeros elsewhere, $Y(k)$ is the rth column of the backward product $U(0,k) = A(k)A(k-1)\dots A(1)$. We let $U(0,k)_{rs}$ denote the entry in the rth row, sth column of $U(0,k)$. Results on the growth and structure of the product $U(0,k)$ were then derived when the $A(k)$'s are slowly varying primitive matrices that satisfy a number of technical conditions (for example, a condition of "uniform primitivity") [1]. In particular, it was shown that the slower the Perron vectors[1] $V(k)$ of the $A(k)$'s vary in the recent past preceding any large k_0, the closer the ratios $U(0,k_0+1)_{rs}/U(0,k_0)_{rs}$ are to $\lambda(k_0+1)$ (where $\lambda(k)$ denotes the Perron root of $A(k)$). This result (which extends to inhomogeneous products a well-known result on the powers of a primitive matrix) hinged crucially on the slow variation of the matrices $A(k)$. This means that there exists an ε such that the distances $\|A(k+1) - A(k)\|$ between all consecutive matrices must be less than ε in order for the results to hold; ε exists but its value is not known, and therefore there is no way of knowing how slowly the matrices must vary in order for the results to hold.

Numerical simulations, however, suggested that the results concerning the proximity between $U(0,k_0+1)_{rs}/U(0,k_0)_{rs}$ and $\lambda(k_0+1)$ for slowly varying Perron vectors $V(k)$ in the recent past may in fact be true even without the matrices $A(k)$ themselves varying slowly. This possibility, clearly hinted at in [1, p. 55], is the subject of the present paper.

* Received by the editors November 30, 1994; accepted for publication (in revised form) by G. P. Styan November 12, 1995.

† Department of Applied Mathematics, University of Pau, 64000 Pau, France (artzrouni@crisv1.univ-pau.fr).

[1] The term "Perron vector" will refer to the right Perron vector. We will later add "left" to refer to the left Perron vector.

The main theorem proved below assumes only that the matrices are irreducible and that the product $U(0,k)$ is weakly ergodic at a geometric rate. No assumption is made concerning the rate of change of the $A(k)$'s and the $V(k)$'s. For any k_0 the theorem provides an upper bound for the projective distance $d(Y(k_0), V(k_0))$. This bound shows that the slower the Perron vectors $V(k)$ have changed in the recent past preceding some large k_0 (i.e., the smaller the quantities $d(V(k), V(k-1))$ are for k less than k_0) the smaller $d(Y(k_0), V(k_0))$ will be. Thus $Y(k_0)$ is close in direction to $V(k_0)$. Corollaries that extend previous results are given concerning the growth and structure of the vectors $Y(k)$ and the product of matrices $U(0,k)$.

2. Results on the structure of each $Y(k)$. The projective distance between two vectors $X = (x_i)$ and $Y = (y_i)$ in the positive quadrant of $\mathbb{R}^n$ is defined as

$$d(X^T, Y^T) = \max_{i,j} \ln \left(\frac{x_i/y_i}{x_j/y_j} \right), \tag{1}$$

where the superscript "T" denotes the transpose of a vector or matrix.

Given a column-allowable nonnegative matrix A, its coefficient of ergodicity $\tau(A)$ is defined as

$$\tau(A) = \sup_{X,Y>0; X \neq \lambda Y} \frac{d(X^T A, Y^T A)}{d(X^T, Y^T)}. \tag{2}$$

This coefficient satisfies $0 \leq \tau(A) \leq 1$, and for any $X, Y > 0$ we have

$$d(X^T A, Y^T A) \leq d(X^T, Y^T)\tau(A) \tag{3}$$

and $\tau(A_1 A_2) \leq \tau(A_1)\tau(A_2)$ when A_1 and A_2 are column-allowable.

When $A = (a_{ij})$ is column-allowable, its coefficient of ergodicity is [9, 2]

$$\tau(A) = \frac{1 - \phi(A)^{0.5}}{1 + \phi(A)^{0.5}}, \tag{4}$$

where

$$\phi(A) = \min_{a_{jk}a_{il} \neq 0} \frac{a_{ik}a_{jl}}{a_{jk}a_{il}}. \tag{5}$$

This explicit expression for $\tau(A)$ shows that $\tau(A) = \tau(A^T)$—a matrix and its transpose have the same coefficient of ergodicity. Therefore, when putting A^T instead of A in (3), we get

$$d(X^T A^T, Y^T A^T) \leq d(X^T, Y^T)\tau(A^T) = d(X^T, Y^T)\tau(A). \tag{6}$$

If, as a notational matter, the vectors appearing in (6) are written as columns instead of rows, (6) is equivalent to

$$d(AX, AY) \leq d(X, Y)\tau(A). \tag{7}$$

Our starting point is a linear nonautonomous process in the positive quadrant of $\mathbb{R}^n$: $Y(k) = A(k)Y(k-1)\,(k = 1, 2, \ldots)$ with $Y(0) \geq 0$ and $A(k)\,(k = 1, 2, \ldots)$ a

sequence of nonnegative $n \times n$ matrices. We now define for $p \geq 0, r \geq 1$ the backward product $U(p,r)$ of matrices $A(k)$ as

$$U(p,r) = A(p+r)A(p+r-1)\dots A(p+1), \tag{8}$$

and therefore $Y(k) = A(k)Y(k-1) = U(0,k)Y(0)$.

We will assume that the products $U(p,r)$ satisfy a set of conditions that are known to be sufficient to ensure their weak ergodicity at a geometric rate (i.e., $\tau(U(p,r))$ approaches 0 geometrically for $r \to \infty$ [9]). These conditions are as follows.

Assumption A1. $\exists m, M > 0$ such that $\forall k\ 0 < m \leq \min^+_{i,j} A(k)_{i,j}$ and $\max_{i,j} A(k)_{i,j} \leq M$, where $A(k)_{i,j}$ is the entry in the ith row, jth column of $A(k)$, and $\min^+_{i,j} A(k)_{i,j}$ denotes the smallest among the positive elements of $A(k)$. If $\gamma \stackrel{\text{def}}{=} m/M$ then we have $\min^+_{i,j} A(k)_{i,j} / \max_{i,j} A(k)_{i,j} \geq \gamma > 0$ for all k.

Assumption A2. $\exists r^* \in N$ such that $U(k, r^*) > 0\ \forall k$.

Under these assumptions it can be seen that $\phi(U(p,r^*))$ (defined in Eq. (5)) satisfies, for every p,

$$\phi(U(p,r^*)) \geq \left(\frac{m^{r^*}}{n^{r^*-1} M^{r^*}} \right)^2. \tag{9}$$

If we define

$$C = \frac{1 - \frac{m^{r^*}}{n^{r^*-1}M^{r^*}}}{1 + \frac{m^{r^*}}{n^{r^*-1}M^{r^*}}}, \tag{10}$$

then (4) and (9) show that $\tau(U(p,r^*)) \leq C$. Therefore,

$$\tau(U(p,r)) \leq C^{[r/r^*]}, \tag{11}$$

where $[\ ]$ denotes the integral part function. We note in particular the following.

(a) For $r < r^*$ we know only that $\tau(U(p,r)) \leq 1$.

(b) For $r \geq r^* > 1$ we have

$$\tau(U(p,r)) \leq C^{[r/r^*]} \leq C^{r/r^*-1}. \tag{12}$$

(c) For $r^* = 1$ we have $\tau(U(p,r)) \leq C^{[r/r^*]} = C^r$.

We will assume that the matrices $A(k)$ are irreducible, each with Perron root $\lambda(k) > 0$ and probability-normed Perron vector $V(k) > 0$. (We note that an irreducible matrix is necessarily allowable.) In the main theorem below we give a bound for the projective distance $d(Y(k), V(k))$ between $Y(k)$ and the Perron vector $V(k)$ of $A(k)$. In essence, this bound will show that the slower the Perron vectors vary prior to the index k, the closer $Y(k)$ will be to $V(k)$ (for the projective distance).

THEOREM 2.1. *Let $\{A(i)\}, i = 1, 2, \dots$ be a sequence of irreducible matrices, each with probability-normed Perron vector $V(i)$ and Perron root $\lambda(i)$. Consider the linear process $Y(k) = A(k)Y(k-1)$ $(0 \leq Y(0) \neq 0)$ and assume that the backward products $U(p,r)$ satisfy Assumptions* A1 *and* A2. *Given C of Eq.* (10) *we define $C_1 = C^{1/r^*}$. We then have*

$$\begin{aligned} d(Y(k), V(k)) \leq\ & W(r^*) C^{-1} C_1^k d(U(0,r^*)Y(0), U(0,r^*)V(1)) \\ & + \sum_{j=1}^{r^*-1} d(V(k-j), V(k-j+1)) \\ & + W(r^*) \sum_{j=r^*}^{k-1} C_1^j d(V(k-j), V(k-j+1)), \\ & k = 2r^*, 2r^*+1, \dots, \end{aligned} \tag{13}$$

where $W(1) = 1$ *and* $W(r^*) = C^{-1}$ *for* $r^* \geq 2$; *the sum to* $r^* - 1$ *is* 0 *if* $r^* = 1$.

Proof. Under Assumption A2 the vector $Y(k)$ will be positive for any $k \geq r^*$ since $Y(k) = U(0,k)Y(0)$ and $Y(0) \neq 0, U(0,k) > 0$ for $k \geq r^*$. This implies that the projective distances $d(Y(k), V(k))$ will be defined for $k \geq r^*$. We now recall that $d(X, \lambda Y) = d(X, Y)$ for any $\lambda > 0$ and observe that

$$U(i, k-i)V(i+1) = \lambda(i+1)U(i+1, k-i-1)V(i+1) \quad \forall i, \tag{14}$$

$$A(k)V(k) = U(k-1,1)V(k) = \lambda(k)V(k). \tag{15}$$

The triangle inequality, combined with Eqs. (7), (14), and (15), yields

$$\begin{aligned} d(Y(k), V(k)) &\leq d(U(0,k)Y(0), U(0,k)V(1)) + d(U(0,k)V(1), U(1,k-1)V(2)) \\ &\quad + d(U(1,k-1)V(2), U(2,k-2)V(3)) + \cdots + d(U(k-3,3)V(k-2), U(k-2,2)V(k-1)) \\ &\quad + d(U(k-2,2)V(k-1), V(k)) \qquad (16) \\ &= d(U(0,k)Y(0), U(0,k)V(1)) + d(U(1,k-1)V(1), U(1,k-1)V(2)) \\ &\quad + d(U(2,k-2)V(2), U(2,k-2)V(3)) + \cdots + d(U(k-2,2)V(k-2), U(k-2,2)V(k-1)) \\ &\quad + d(U(k-1,1)V(k-1), U(k-1,1)V(k)) \\ &\leq \tau(U(r^*, k-r^*))d(U(0,r^*)Y(0), U(0,r^*)V(1)) + \tau(U(1,k-1))d(V(1), V(2)) \\ &\quad + \tau(U(2,k-2))d(V(2), V(3)) + \cdots + \tau(U(k-1,1))d(V(k-1), V(k)). \qquad (17) \end{aligned}$$

By virtue of (11), $\tau(U(r^*, k-r^*)) \leq C^{[(k-r^*)/r^*]}$ and $\tau(U(k-j, j)) \leq C^{[j/r^*]}$, and therefore (17) yields

$$\begin{aligned} d(Y(k), V(k)) &\leq C^{[(k-r^*)/r^*]}d(U(0,r^*)Y(0), U(0,r^*)V(1)) \\ &\quad + \sum_{j=1}^{k-1} C^{[j/r^*]}d(V(k-j), V(k-j+1)). \end{aligned} \tag{18}$$

The result of (13) follows directly by a consideration of observations (a), (b) (Eq. (12)), and (c) made earlier. This completes the proof.

The inequality of (13) shows that the projective distance between $Y(k)$ and $V(k)$ is bounded by a sum of three terms.

1. The first term on the right-hand side of (13) approaches 0 geometrically as $k \to \infty$.

2. The next two terms represent a weighted sum of the distances between all consecutive Perron vectors prior to the index k with the $r^* - 1$ most recent distances having weight 1 and the "older" distances having weights decreasing as powers of C_1.

The result of (13), which is not based on any assumption concerning the $V(i)$'s, is of particular interest when k is large and the $V(i)$'s have varied slowly (for the projective distance) in the recent past preceding k. Indeed, in such a case (13) shows that the projective distance $d(Y(k), V(k))$ is small, which means that $Y(k)$ is close in structure to $V(k)$.

We note that the quantities C, C_1, and r^* appearing in (13) are known as soon as the matrices $A(k)$ are known, and it is therefore not difficult to assess the upper bound obtained for $d(Y(k), V(k))$.

We have the following results under stronger assumptions concerning the distances $\delta(k) \stackrel{\text{def}}{=} d(V(k), V(k+1))$ between consecutive Perron vectors.

COROLLARY 2.1. *If the projective distances* $\delta(k) = d(V(k), V(k+1))$ *are bounded by some* $d^* > 0$, *then*

$$
\begin{aligned}
d(Y(k), V(k)) &\leq W(r^*) C^{-1} C_1^k d(U(0, r^*) Y(0), U(0, r^*) V(1)) \\
&\quad + d^* \left((r^* - 1) + \frac{C_1^{r^*} W(r^*)}{1 - C_1} \right).
\end{aligned} \tag{19}
$$

If, in addition, $\delta(k) \to 0$ *for* $k \to \infty$, *then*

$$
d(Y(k), V(k)) \to 0 \quad \text{for } k \to \infty. \tag{20}
$$

In particular, if the Perron vectors $V(k)$ *converge to a limit* $V > 0$, *then* $d(Y(k), V) \to 0$.

Proof. These results follow directly from (13); (20) results from the fact that $\sum_{j=r^*}^{k-1} C_1^j \delta(k-j) \to 0$ when $k \to \infty$ because $\delta(k) \to 0$.

The result of (19) shows that the smaller d^* is, the smaller $d(Y(k), V(k))$ will be, at least asymptotically as $k \to \infty$. Also, (20) shows that $Y(k)$ will approach $V(k)$ in direction if the distances $\delta(k)$ tend to 0.

We note that in (16) we could not write

$$
d(U(0,k)Y(0), U(0,k)V(1)) \leq \tau(U(0,k)) d((Y(0), V(1)) \tag{21}
$$

because $Y(0)$ is not necessarily positive, and therefore $d(Y(0), V(1))$ may not be defined. This situation occurs when $Y(0)$ is taken equal to the vector Y_r having "1" in its rth position and zeros elsewhere. The corresponding vector $Y(k)$ is then the rth column of the product $U(0,k)$. Therefore, the slower the Perron vectors have varied for recent indices preceding k, and the larger k is, the closer (for the projective distance) each column of $U(0,k)$ will be to the Perron vector $V(k)$ of $A(k)$. This extends an earlier result that hinged on stringent conditions on the matrices $A(i)$, which were assumed to be in some sense uniformly primitive and to vary slowly [1].

Finally, when $r^* = 1$ (i.e., the matrices $A(k)$ are positive) we let $H(k)$ denote the right- hand side of (13), i.e., the upper bound for $d(Y(k), V(k))$. A straightforward calculation shows that

$$
H(k+1) = CH(k) + \delta(k), \qquad k = 1, 2, \ldots; H(1) = d[Y(1), V(1)]. \tag{22}
$$

Equation (22) shows that when $r^* = 1$, the bounds $H(k+1)$ can be generated as a particularly simple nonautonomous iterative process in $\mathbb{R}$.

3. Results on the growth of each $Y(k)$. The previous section provides information on the *structure* of the $Y(k)$'s but not on their *growth*. Now in the context of Theorem 2.1 $Y(k)$ is close to $V(k)$ (for the projective distance). If $V(k)$ is also close to $V(k+1)$ for the projective distance, then $Y(k)$ is close to $V(k+1)$ (triangle inequality), and therefore the vector $Y(k+1) = A(k+1)Y(k)$ will be close to $\lambda(k+1)Y(k)$ for the Euclidean norm. This observation will be made more precise in

a result on the growth of the vectors $Y(k)$. We begin with a lemma on the projective distance between two vectors $X, Y > 0$.

LEMMA 3.1. *For* $X = (x_i) > 0, Y = (y_i) > 0$,

$$d(X,Y) \le \varepsilon \Leftrightarrow \exists c > 0, \text{ and there is a diagonal matrix } M = [m(i)] \ge 0 \text{ with } m(i) \le e^{\varepsilon} - 1 \text{ such that } X = c(Y + MY).$$

Proof. We define $d(i) = x_i/y_i\ (i = 1, 2, \ldots, n)$. If $d(X,Y) \le \varepsilon$, then $\forall i, j\ d(i)/d(j) \le e^{\varepsilon}$. We let $c = \min_i d(i)$. Then if we define $m(i) = (d(i) - c)/c$ we have $m(i) \le e^{\varepsilon} - 1$ and $d(i) = c(1 + m(i))$, which is the desired result. Conversely, if $X = c(Y + MY)$ (with $m(i) \le e^{\varepsilon} - 1$), then $d(i) = c(1 + m(i))$, and $d(i)/d(j) = (1 + m(i))/(1 + m(j))$; therefore, $\ln(d(i)/d(j)) \le \ln(1 + m(i)) \le \varepsilon$ for any i, j, which proves that $d(X,Y) \le \varepsilon$.

Under the conditions of Theorem 2.1 the triangle inequality and (13) yield, for $k \ge 2r^*$,

$$\begin{aligned} d(Y(k), V(k+1)) &\le d(Y(k), V(k)) + d(V(k), V(k+1)) \\ &\le W(r^*)C^{-1}C_1^k d(U(0,r^*)Y(0), U(0,r^*)V(1)) \\ &\quad + \sum_{j=0}^{r^*-1} d(V(k-j), V(k-j+1)) \\ &\quad + W(r^*) \sum_{j=r^*}^{k-1} C_1^j d(V(k-j), V(k-j+1)) \stackrel{\text{def}}{=} B(k+1). \end{aligned} \tag{23}$$

This right-hand side $B(k+1)$ of (23) is the right-hand side of (13), to which is added $d(V(k), V(k+1))$ (appearing in the first sum for $j = 0$).

As in (19), if $\delta(k) \stackrel{\text{def}}{=} d(V(k), V(k+1)) \le d^*$ for every k we have

$$B(k+1) \le W(r^*)C^{-1}C_1^k d(U(0,r^*)Y(0), U(0,r^*)V(1)) + d^*\left(r^* + \frac{C_1^{r^*}W(r^*)}{1 - C_1}\right). \tag{24}$$

We now define B^* as the right-hand side of (24) for $k = 2r^*$; i.e.,

$$B^* = W(r^*)C^{-1}C_1^{2r^*} d(U(0,r^*)Y(0), U(0,r^*)V(1)) + d^*\left(r^* + \frac{C_1^{r^*}W(r^*)}{1 - C_1}\right). \tag{25}$$

Because the right-hand side of (24) is a decreasing function of k we have

$$B(k+1) \le B^*, \qquad k = 2r^*, 2r^* + 1, \ldots. \tag{26}$$

The results of (25)–(26) will be used in what follows. We will also make use of the row–sum norm $\|A\| = \max_i \sum_{j=1}^n |a_{ij}|$ of a matrix $A = (a_{ij})$ and of the max norm $|V| = \max_i |v_i|$ of a vector $V = (v_i)$. These norms satisfy $\|AB\| \le \|A\|\|B\|$ and $|AV| \le \|A\||V|$. These remarks set the stage for the following result on the values of $Y(k+1)_i/Y(k)_i$, where a subscript i denotes the ith component of a vector.

THEOREM 3.1. *In the context of Theorem* 2.1 *and with the notation defined above we assume the following.*

(i) *The matrices* $A(i)$ *are bounded:* $K_1 = \sup_i \|A(i)\|$ *and* $\Lambda = \sup_i \lambda(i)$.

(ii) *The components* $V(k)_i$ $(i = 1, 2, \ldots, n)$ *of the probability-normed Perron vectors* $V(k)$ *are uniformly bounded from below by some* $\delta > 0$; *i.e.,* $\forall k\ V(k)_i > \delta$. *We let* $d^* = \sup_i d(V(i), V(i+1))$ *and* $A_1 = (K_1/\delta + \Lambda)(e^{B^*} - 1)/e^{B^*}$. *We then have*

$$
\begin{aligned}
&\left| \frac{Y(k+1)_i}{Y(k)_i} - \lambda(k+1) \right| \\
&\quad \leq A_1 B(k+1) \\
&\quad = A_1 \Bigg[W(r^*) C^{-1} C_1^k d[U(0, r^*) Y(0), U(0, r^*) V(1)] \\
&\qquad + \sum_{j=0}^{r^*-1} d(V(k-j), V(k-j+1)) \\
&\qquad + W(r^*) \sum_{j=r}^{k-1} C_1^j d(V(k-j), V(k-j+1)) \Bigg], \qquad k = 2r^*, 2r^*+1, \ldots .
\end{aligned} \tag{27}
$$

Proof. We have $d(Y(k), V(k+1)) \leq B(k+1)$. By virtue of Lemma 3.1 there is $c > 0$ and a diagonal matrix M (with diagonal entries $m(i)$ satisfying $0 < m(i) \leq e^{B(k+1)} - 1$) such that

$$
Y(k) = c(V(k+1) + MV(k+1)). \tag{28}
$$

Therefore,

$$
Y(k+1) = A(k+1)Y(k) = c(\lambda(k+1)V(k+1) + A(k+1)MV(k+1)). \tag{29}
$$

We note that $|V(k+1)| \leq 1$. If $\{X\}_i$ denotes the ith component of a vector X and I is the identity matrix, then $\{(I+M)V(k+1)\}_i \geq \{V(k+1)\}_i$. Therefore,

$$
\begin{aligned}
\left| \frac{Y(k+1)_i}{Y(k)_i} - \lambda(k+1) \right| &= \left| \frac{\{A(k+1)MV(k+1)\}_i - \lambda(k+1)\{MV(k+1)\}_i}{\{(I+M)V(k+1)\}_i} \right| \\
&\leq \left| \frac{\{A(k+1)MV(k+1)\}_i}{\{V(k+1)\}_i} - \lambda(k+1)m(i) \right| \\
&\leq \frac{\|A(k+1)\| \times \|M\|}{\delta} + \Lambda \|M\| \\
&\leq \left(\frac{K_1}{\delta} + \Lambda \right) (e^{B(k+1)} - 1).
\end{aligned} \tag{30}
$$

We now observe that if a positive number u is less than B^*, then $e^u - 1 \leq u(e^{B^*} - 1)/e^{B^*}$, and therefore

$$
e^{B(k+1)} - 1 \leq \frac{e^{B^*} - 1}{e^{B^*}} B(k+1), \tag{31}
$$

which completes the proof.

The result of (27) shows that the ratio $Y(k+1)_i/Y(k)_i$ is close to $\lambda(k+1)$ when k is large and the Perron vectors $V(k)$ have varied slowly for recent indices preceding k. If we take $Y(0) = Y_r$, say (Y_r is the vector having rth component "1" and zeros elsewhere), then the ratio $Y(k+1)_j/Y(k)_j$ is $U(0,k+1)_{jr}/U(0,k)_{jr}$, which is the entry in the jth row, rth column of $U(0,k+1)$ divided by the corresponding entry in $U(0,k)$.

We will now examine what information these results add concerning the structure of the product $U(0,k)$, beyond that already implied by weak ergodicity. Weak ergodicity of the backward product means that $U(0,k)$ may be written as

(32)
$$U(0,k) = \begin{pmatrix} u(k)_1[z_1+\nu(k)_{11}] & u(k)_1[z_2+\nu(k)_{12}] & \dots & u(k)_1[z_n+\nu(k)_{1n}] \\ u(k)_2[z_1+\nu(k)_{21}] & u(k)_2[z_2+\nu(k)_{22}] & \dots & u(k)_2[z_n+\nu(k)_{2n}] \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ u(k)_n[z_1+\nu(k)_{n1}] & \cdot & \dots & u(k)_n[z_n+\nu(k)_{nn}] \end{pmatrix},$$

where

(i) the $\nu(k)_{ij}$'s tend to 0 for $k \to \infty$,

(ii) $Z = (z_1, z_2, \dots, z_n)$ is a probability-normed vector,

(iii) the vectors $U(k) = (u(k)_1, u(k)_2, \dots, u(k)_n)$ are positive and change with each k.

In the present context, when the Perron vectors $V(t)$ vary slowly for the projective distance, each $U(k)$ will be for large k close in structure to $V(k)$. Furthermore, for each i the ratio $u(k+1)_i/u(k)_i$ will be close to $\lambda(k+1)$.

4. Numerical illustrations. In order to illustrate the results we constructed a sequence of 2×2 positive matrices $A(k)$ with Perron vectors $V(k)$ varying slowly for the projective distance (i.e., the projective distance $d(V(k+1), V(k))$ remains small for all k). The goal is to see how close $Y(k+1)_i/Y(k)_i$ remains to $\lambda(k+1)$.

The first component $V(k)_1$ of the Perron vectors $V(k)$ is chosen equal to 0.5 plus a cosine function of k:

$$V(k)_1 = 0.5 + 0.1\cos(k\pi/10)$$

and $V(k)_2 = 1 - V(k)_1$. The idea is that the projective distance between consecutive vectors $V(k)$ will be small when the $V(k)$'s vary smoothly for the Euclidean distance (and the $V(k)$'s are bounded away from 0).

Next we will generate the Perron roots $\lambda(k)$ as arbitrary numbers between 0.8 and 1.2: each $\lambda(k)$ is a uniformly distributed random number between 0.8 and 1.2. This range is chosen so that the product of matrices neither grows too fast nor goes to zero. We will now determine the four entries of each matrix $A(k)$ in such a way that each $V(k)$ and $\lambda(k)$ chosen above is the Perron vector and the Perron root of each $A(k)$. The entries $A(k)_{11}$ and $A(k)_{22}$ are chosen as arbitrary numbers in the range $0.5\lambda(k)$–$0.8\lambda(k)$ for $A(k)_{11}$ and $0.4\lambda(k)$–$0.8\lambda(k)$ for $A(k)_{22}$ (i.e., once $\lambda(k)$ is chosen, these entries are uniformly distributed random numbers in these intervals). The numbers 0.4, 0.5, and 0.8 are arbitrary. Their only purpose is to ensure that $A(k)_{11}, \lambda(k) - A(k)_{11}, A(k)_{22}$, and $\lambda(k) - A(k)_{22}$ remain bounded away from 0; see (33)–(34) below. Once these values are obtained, the remaining entries are

$$A(k)_{12} = [\lambda(k) - A(k)_{11}]V(k)_1/V(k)_2, \tag{33}$$

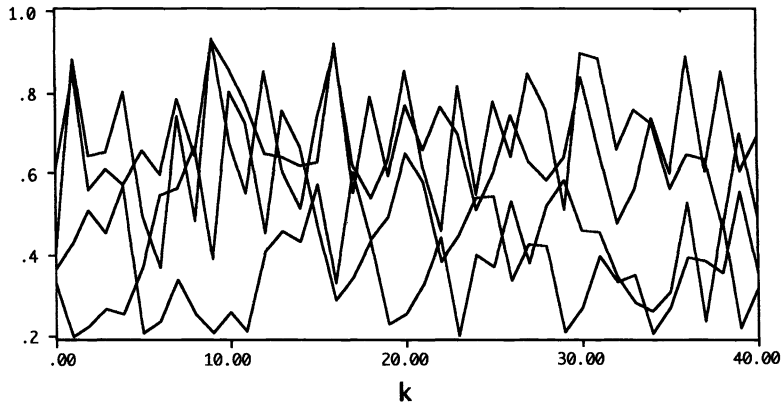


FIG. 1. *Four entries of the matrices* $A(k); k = 0, 1, \ldots, 40$.

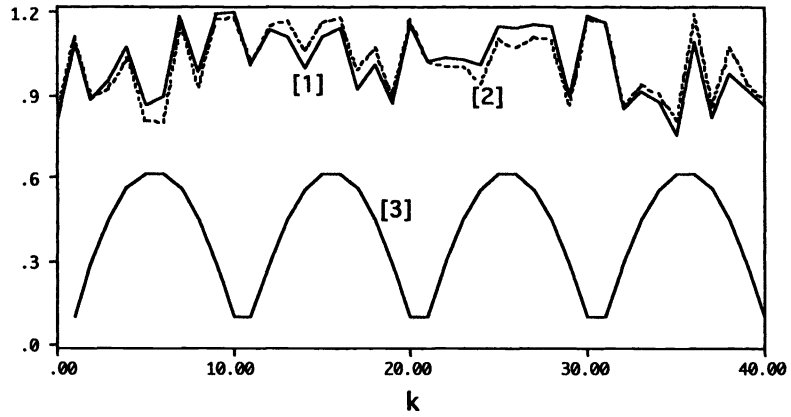


FIG. 2. *Ratio* $Y(k+1)_2/Y(k)_2$ [1]; $\lambda(k+1)$ [2]; $5d(V(k+1), V(k))$ [3].

$$A(k)_{21} = [\lambda(k) - A(k)_{22}]V(k)_2/V(k)_1. \tag{34}$$

The ranges of possible values for $V(k)_1$, $A(k)_{11}$, and $A(k)_{22}$ and the specifications of (33)–(34) ensure that the four entries of $A(k)$ are bounded and bounded away from 0. The product of matrices $A(k)$ is then weakly ergodic. Of course (33)–(34) also guarantee that for every k $V(k)$ and $\lambda(k)$ are the Perron vector and Perron root of $A(k)$.

Figure 1 depicts the four entries of the $A(k)$'s as defined above.

The elaborate construction of the matrices $A(k)$ is to emphasize the only constraint we are imposing— the slowly varying Perron vectors $V(k)$. Aside from this constraint, the matrices may vary erratically, as illustrated in Figure 1.

In Figure 2 the values of $Y(k+1)_2/Y(k)_2$ and $\lambda(k+1)$ were plotted in order to assess how close these two quantities are to each other (see (27)). (The initial vector $Y(0)$ was taken arbitrarily as $(0.4\ 0.6)^T$, but there is nothing critical about this specification.) The distances $5d(V(k+1), V(k))$ were also plotted in order to see how these values impacted the "closeness" between $Y(k+1)_2/Y(k)_2$ and $\lambda(k+1)$. (The factor "5" is to make the distances $d(V(k+1), V(k))$ more visible on the same graph.)

First, it is apparent that $Y(k+1)_2/Y(k)_2$ is close to $\lambda(k+1)$, even for small k's. The same holds true for $Y(k+1)_1/Y(k)_1$. This is because $r^* = 1$ and the factor C of

Eq. (10) is relatively small, i.e., well under 1 ($C = 0.87$), and therefore $C_1^k = C^k$ and C_1^k approaches 0 quite rapidly.

When C is closer to 1 and r^* is larger than 1, then $C_1^k = C^{k/r^*}$ and the convergence of C_1^k to 0 will be slower; the right-hand side of (27) will be altogether larger, and each $Y(k+1)_2/Y(k)_2$ may not be as close to $\lambda(k+1)$. It remains that here the $\lambda(k+1)$'s vary erratically, and the $Y(k+1)_2/Y(k)_2$'s follow closely.

Second, we clearly see how the agreement between curves [1] and [2] in Figure 2 improves when the distances $d(V(k+1), V(k))$ become closer to 0 (as predicted by (27)). Indeed, the figure shows that when $d(V(k+1), V(k))$ becomes smaller (for $k = 10, 20$, and 30), the Perron root $\lambda(k+1)$ becomes closer to $Y(k+1)_2/Y(k)_2$.

Many other simulations illustrate the findings. For example, if the $V(k)$'s vary more slowly, the agreement becomes better between each $\lambda(k+1)$ and $Y(k+1)_2/Y(k)_2$ (and vice versa).

5. Discussion.

5.1. Forward products. The results proven for a backward product carry over to a forward product since one is a transposition of the other. Given the forward product $T(0,k) = A(1)A(2)\dots A(k)$ we may consider the linear process

$$Y(k)^T = Y(0)^T A(1)A(2)\dots A(k) = Y(0)^T T(0,k). \tag{35}$$

Transposing the results requires the weak ergodicity of the forward product $T(0,k)$ at a geometric rate, which is then a transposition of Eq. (32):

(36)

$$T(0,k) = \begin{pmatrix} u(k)_1[z_1 + \nu(k)_{11}] & u(k)_2[z_1 + \nu(k)_{12}] & \dots & u(k)_n[z_1 + \nu(k)_{1n}] \\ u(k)_1[z_2 + \nu(k)_{21}] & u(k)_2[z_2 + \nu(k)_{22}] & \dots & u(k)_n[z_2 + \nu(k)_{2n}] \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ u(k)_1[z_n + \nu(k)_{n1}] & \cdot & \dots & u(k)_n[z_n + \nu(k)_{nn}] \end{pmatrix}.$$

For every k we now define $W(k)$ as the *left* probability-normed Perron vector of $A(k)$. A straightforward transposition of the results proven here shows that when the $W(k)$'s vary slowly, the structure of $Y(k)$ and of each *row* of $T(0,k)$ is close to that of $W(k)$ (i.e., each vector $U(k) = (u(k)_1, u(k)_2, \dots, u(k)_n)$ will be close in structure to $W(k)$). Also, for each i the ratio $u(k+1)_i/u(k)_i$ will be close to $\lambda(k+1)$, the Perron root of $A(k+1)$.

5.2. Stochastic matrices. If the matrices $A(k)$ are stochastic (i.e., their rows sum to 1) then their product is stochastic. For every k the Perron root $\lambda(k)$ is 1 and the probability-normed Perron vector $V(k)$ is $V^* = (1/n, 1/n, \dots, 1/n)$, where n is the order of the matrices. As in §5.1 we consider the *left* probability-normed Perron vector $W(k)$ of each $A(k)$. We now distinguish between forward and backward products.

(a) *Forward products of stochastic matrices.* The process

$$Y(k)^T = Y(0)^T A(1)A(2)\dots A(k) = Y(0)^T T(0,k) \tag{37}$$

gives at each time k the row vector $Y(k)^T$ of probabilities of being in the n different states of an inhomogeneous Markov chain with matrices $A(k)$. In order to apply the results we assume that the product is weakly ergodic at a geometric rate. Given that $T(0,k)$ is stochastic, Eq. (36) implies that the z_i's are all equal to $1/n$ and that

$\sum_{i=1}^{n} u(k)_i \to n$ for $k \to \infty$. If the $W(k)$'s now vary slowly, then the structure of each $Y(k)$ is close to that of $W(k)$. The statistical interpretation of this is that for slowly varying $W(k)$'s the system somewhat forgets its distant past (as if the $A(k)$'s were constant), and as an approximation the probabilities of being in the different states at time k are given by $W(k)$. We note that $W(k)$ is also the stationary distribution of a Markov chain with a constant unchanging matrix of transition probabilities equal to $A(k)$, at least if $A(k)$ is primitive. In the present inhomogeneous situation no primitivity assumption is made—one could say that in the inhomogeneous context primitivity is replaced by the weak ergodicity assumption.

(b) *Backward products of stochastic matrices.* When the backward product is weakly ergodic at a geometric rate then Corollary 2.1 ensures that the structure of $Y(k) = A(k)Y(k-1)$ and of each column of the stochastic matrix $U(0,k) = A(k)A(k-1)\dots A(1)$ converges to $V^* = (1/n, 1/n, \dots, 1/n)$ (since the Perron vectors $V(k)$ are all equal to V^*). A direct inspection of (32) shows that the stochasticity of $U(0,k)$ in fact implies that for each i the $u(k)_i$'s converge to 1 for $k \to \infty$ and each row converges to the same probability-normed vector $(z_1, z_2, \dots, z_n)$. This is a known result: for a backward product of stochastic matrices strong and weak ergodicity are equivalent with convergence of $Y(k)$ to a vector having equal components [9, p. 154].

5.3. Periodic irreducible matrices. No primitivity assumption is made in the theorems proven above, and one may wonder about the application of these theorems to periodic irreducible matrices (i.e., matrices A that are irreducible but imprimitive, which means that A^k has at least one zero entry for all k).

Although weak ergodicity at a geometric rate (obtained via Assumptions A1 and A2) is in general a rather weak condition, it does not hold for powers of periodic irreducible matrices. In fact, it is readily seen that the powers A^k of an irreducible matrix are weakly (and strongly) ergodic at a geometric rate if and only if A is primitive. Indeed, primitivity implies weak (and strong) ergodicity at a geometric rate, and imprimitivity implies that $\tau(A^k) = 1$ for all k (since each A^k has at least one zero entry); $\tau(A^k) = 1$ for all k precludes both weak and strong ergodicity (at any rate). This shows that the results proven here will not be applicable to the powers of a periodic irreducible matrix simply because such powers are not weakly ergodic.

However, an inhomogeneous backward product that includes periodic matrices can be weakly ergodic at a geometric rate. For example, consider two sequences of positive numbers $v(k)$ and $w(k)$ (with $1/2 \le v(k) \le 5, 1/2 \le w(k) \le 5, k = 1, 2, \dots$) and the backward product

$$\begin{pmatrix} 0 & v(k) \\ w(k) & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & v(k-1) \\ w(k-1) & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \tag{38}$$
$$\dots \begin{pmatrix} 0 & v(1) \\ w(1) & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & v(0) \\ w(0) & 0 \end{pmatrix}$$

of matrices that alternate between a matrix of ones and the periodic matrix having $v(k)$ and $w(k)$ on the second diagonal and zeros on the main diagonal. Assumptions A1 and A2 are clearly satisfied with $m = 1/2, M = 5$, and $r^* = 2$, and the results proven here become applicable. These results are of interest when the probability-normed Perron vectors vary slowly for the projective distance, i.e., when $v(k)$ is close to $w(k)$ for all k. In such a case the probability-normed Perron vector of each matrix remains close to $V = (1/2, 1/2)$ (and equal to V for the matrix of ones). Hence for large k the structure of the columns of the product remains close to that of V.

6. Conclusion. The results proven here show that when the Perron vectors vary slowly for the projective distance, these Perron vectors and the Perron roots drive the structure and growth of the vectors $Y(k)$ and of the columns of the product $U(0,k)$. No assumption is made concerning the rate of change of the matrices $A(k)$. It can be useful to know that in this context the past history of the process is not needed in order to assess the structure and growth of a given vector $Y(k)$: its structure is close to that of the Perron vector $V(k)$ and the ratios $Y(k+1)_i/Y(k)_i$ can be approximated by the Perron roots $\lambda(k+1)$. This generalizes to inhomogeneous products of matrices the situation in which the matrices $A(k)$ are equal to the same primitive matrix A with Perron root λ, since in such a case the ratios $Y(k+1)_i/Y(k)_i$ tend to λ.

Acknowledgment. The author wishes to thank an anonymous referee whose comments greatly improved this paper.

REFERENCES

[1] M. ARTZROUNI, *On the growth of infinite products of slowly varying primitive matrices*, Linear Algebra Appl., 145 (1991), pp. 33–57.

[2] M. ARTZROUNI AND X. LI, *A note on the coefficient of ergodicity of a column-allowable matrix*, Linear Algebra Appl., 214 (1995), pp. 93–101.

[3] J. E. COHEN, *Finite non-homogeneous Markov chains: Asymptotic behavior*, Adv. in Appl. Probab., 8 (1976), pp. 502–516.

[4] ———, *Contractive inhomogeneous products of non-negative matrices*, Math. Proc. Cambridge Philos. Soc., 86 (1979), pp. 351–364.

[5] H. COHN AND O. NERMANN, *On products of nonnegative matrices*, Ann. Probab., 18 (1990), pp. 1806–1815.

[6] J. HAJNAL, *On products of non-negative matrices*, Math. Proc. Cambridge Philos. Soc., 79 (1976), pp. 521–530.

[7] D. J. HARTFIEL, *On infinite products of nonnegative matrices*, SIAM J. Appl. Math., 26 (1974), pp. 297–301.

[8] A. LEIZAROWITZ, *On infinite products of stochastic matrices*, Linear Algebra Appl., 168 (1992), pp. 189–219.

[9] E. SENETA, *Non-negative Matrices and Markov Chains*, 2nd ed., Springer-Verlag, Berlin, 1981.

[10] J. WOLFOVITZ, *Products of indecomposable, aperiodic, stochastic matrices*, Proc. Amer. Math. Soc., 14 (1963), pp. 733–737.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 4, pp. 834–850, October 1996

008

EXTENSIONS OF $\mathcal{G}$-BASED MATRIX PARTIAL ORDERS*

S. K. JAIN†, S. K. MITRA‡, AND H. J. WERNER§

Abstract. We prove that a partial order $\preceq^{\mathcal{G}}$ on $\mathbf{R}^{m\times n}$ can always be extended to a $\mathcal{G}$-based matrix partial order $\preceq^{\mathcal{G}^*}$ such that $\mathcal{G}^*(A) \neq \emptyset$ for all $A \in \mathbf{R}^{m\times n}$, thus answering an open question [Mitra, *Linear Algebra Appl.*, 148 (1991), pp. 237–263]. It is further shown that this result does not in general remain true if besides $\mathcal{G}$, we also insist that $\mathcal{G}^*$ be semicomplete. And even if in a special situation this is possible and if card $\mathcal{G}(A) \leq 1$ for each A, this does not mean that there also need be a semicomplete extension such that $\mathcal{G}^*(A)$ is a singleton for all A. In addition, some other interesting results on matrix partial orders are given. For instance, a useful characterization for a semicomplete map to induce a partial order on the set of square matrices is derived.

Key words. $\mathcal{G}$-based matrix partial order, star order, minus order, sharp order, semicomplete map, property-p map

AMS subject classifications. 15A30, 15A09

1. Introduction.

This paper continues a series of recent articles investigating different types of matrix orders and discussing their properties and relations; see [6], [7], [8], [18]. To facilitate reading, we first present the order concept, which is of interest to us in this paper.

Let $\mathbf{R}^{m\times n}$ denote the set of real $m \times n$ matrices, and let $\mathcal{P}(S)$ denote the power set of a set S. Moreover, let

$$\mathcal{G}: \quad \mathbf{R}^{m\times n} \to \mathcal{P}(\mathbf{R}^{n\times m})$$

be a map such that

$$\mathcal{G}(A) \subseteq A\{1\}, \tag{1.1}$$

where $A\{1\}$ denotes the set of all g-inverses of A; see §2. The map $\mathcal{G}$ is called *semicomplete* if for every matrix $A \in \mathbf{R}^{m\times n}$ one has $GAG \in \mathcal{G}(A)$ whenever $G \in \mathcal{G}(A)$. Define the relation $\preceq^{\mathcal{G}}$ on $\mathbf{R}^{m\times n}$ by saying

$$A \preceq^{\mathcal{G}} B \quad \text{if} \quad (B-A)X = 0,\ X(B-A) = 0 \quad \text{for some} \quad X \in \mathcal{G}(A). \tag{1.2}$$

This relation is said to be a *$\mathcal{G}$-based relation*, and if $\mathcal{G}$ is semicomplete, then $\preceq^{\mathcal{G}}$ is also said to be *semicomplete*. Call the set

$$\Omega_{\mathcal{G}} := \{A \in \mathbf{R}^{m\times n} \mid \mathcal{G}(A) \neq \emptyset\}$$

*Received by the editors October 28, 1991; accepted for publication (in revised form) by G. P. Styan November 12, 1995.

†Department of Mathematics, Ohio University, Athens, OH 45701 (jain@oucsace.cs.ohio.edu). This author was supported by a U.S. Fulbright Award (1991–92), NSF grant INT 9210491 (1992–93), and Baker Fund Award of Ohio University (1993–94).

‡Indian Statistical Institute, New Delhi 110016, India (mitra@isid.ernet.in). This author's research was partially supported by the Council of Scientific and Industrial Research under its Emeritus Scientists Scheme.

§Institute for Econometrics and Operations Research, Econometrics Unit, University of Bonn, D-53113 Bonn, Germany (or470@unitas.or.uni-bonn.de, na.werner@na-net.ornl.gov). Financial support by Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 303 at the University of Bonn, is gratefully acknowledged.

the *support* of $\mathcal{G}$ or $\preceq^{\mathcal{G}}$. The relation $\preceq^{\mathcal{G}}$ is automatically *antisymmetric*. Furthermore, it is trivially *reflexive on its support*. Throughout this paper, a $\mathcal{G}$-based relation is therefore said to be a *partial order* (on $\mathbf{R}^{m\times n}$) if it is *transitive*. Since in the literature the notion of a partial order is sometimes defined in a stronger manner, it is pertinent to emphasize here that in our definition the condition of reflexivity is not required to hold on the whole of $\mathbf{R}^{m\times n}$. That $\preceq^{\mathcal{G}}$ as defined by (1.2) need not always correspond to a partial order, that is, that $\preceq^{\mathcal{G}}$ need not necessarily be transitive, is illustrated by Example 1 in [6, p. 242]. Sufficient conditions that make $\preceq^{\mathcal{G}}$ a partial order are studied in [6]. Observe that the above $\mathcal{G}$-based relation concept, which is due to Mitra [6], covers as special cases the various known matrix orders such as the *star order* $\preceq^{*}$ (set $\mathcal{G}(A) = \{A^{\dagger}\}$), the *minus order* $\preceq^{-}$ (set $\mathcal{G}(A) = A\{1\}$), and the *sharp order* $\preceq^{\#}$ (set $\mathcal{G}(A) = \{A^{\#}\}$ if A has index 1, and set $\mathcal{G}(A) = \emptyset$ otherwise), to mention only a few. Next, let $\preceq^{\mathcal{G}}$ correspond to a partial order. The sharp order provides us with a typical $\mathcal{G}$-based partial order that does not support the whole set of square matrices. For such a case the following question was an interesting open problem (see [6, p. 252]): Is it possible to modify $\mathcal{G}$ only on the complement of $\Omega_{\mathcal{G}}$ so that the modified map, say $\mathcal{G}^{*}$, continues to induce a partial order? Let us call a map $\mathcal{G}^{*}$ which is defined like $\mathcal{G}$ an *extension* of $\mathcal{G}$ if $\mathcal{G}^{*}(A) = \mathcal{G}(A)$ for each $A \in \Omega_{\mathcal{G}}$, and if $\mathcal{G}^{*}$ is an extension of $\mathcal{G}$, then let us call the induced relation $\preceq^{\mathcal{G}^{*}}$ a *$\mathcal{G}$-based extension* of the relation $\preceq^{\mathcal{G}}$. Of particular interest to us is the question of whether a partial order $\preceq^{\mathcal{G}}$ with $\Omega_{\mathcal{G}} \neq \mathbf{R}^{m\times n}$ admits a $\mathcal{G}$-based partial order extension $\preceq^{\mathcal{G}^{*}}$ such that $\mathcal{G}^{*}(A) \neq \emptyset$ for all A. This question was affirmatively answered in the special case of the sharp order in [7].

The purpose of this paper is manifold. In §4 we prove that the above latter question can, in general, be answered in the affirmative. Actually, we show that there do exist many $\mathcal{G}$-based extensions. When along with $\preceq^{\mathcal{G}}$ its extension $\preceq^{\mathcal{G}^{*}}$ is also required to be semicomplete, the answer becomes more complicated; this case is discussed in §5. A typical example of a semicomplete $\mathcal{G}$-based partial order $\preceq^{\mathcal{G}}$ not allowing such an extension is given there. Of course, this does not mean that it is always impossible. But even if for a certain $\preceq^{\mathcal{G}}$ there is such a semicomplete partial order extension, this does not mean that there also need be such an extension $\preceq^{\mathcal{G}^{*}}$ for which $\mathcal{G}^{*}(A)$ is a singleton set for all $A \notin \Omega_{\mathcal{G}}$. The sharp order (see Example 4.3 along with Theorem 5.5) can serve as an example to illustrate this fact. In addition, some further interesting results on matrix partial orders are given. Section 3, for instance, contains a useful characterization for a semicomplete map $\mathcal{G}$, which is defined on the set of square $n \times n$ matrices, to induce a partial order $\preceq^{\mathcal{G}}$. The concept of a *property-p relation* discussed in §5 arises from this characterization. It was implicitly studied earlier by Mitra in [6], although its significance is brought out in the present paper. It is shown in §5 that such a relation always possesses a (unique) *maximal* property-p extension, and a method for getting this extension is given. Section 2 contains a few background results from the theory of g-inversion which are used in the subsequent sections and which might also be very useful in a future attempt to solve the open problem stated in §5.

2. Generalized inversion and preliminaries. Let $\mathcal{M}$ and $\mathcal{N}$ be linear subspaces in the n-dimensional real space $\mathbf{R}^{n}$. Then $\mathcal{M}^{\perp}$ will denote the orthogonal complement of $\mathcal{M}$ in $\mathbf{R}^{n}$ (with respect to the usual inner product), and if $\mathcal{M} \cap \mathcal{N} = \{0\}$, then $\mathcal{M} \oplus \mathcal{N}$ will denote the direct sum of $\mathcal{M}$ and $\mathcal{N}$. Next, if $\mathcal{N}$ is a direct complement of $\mathcal{M}$ (i.e., $\mathbf{R}^{n} = \mathcal{M} \oplus \mathcal{N}$), then $P_{\mathcal{M},\mathcal{N}}$ will denote the well-defined projector on $\mathcal{M}$ along $\mathcal{N}$ (see e.g., [10, pp. 106–113]). Notice that $P_{\mathcal{M},\mathcal{N}}$ may be defined by $P_{\mathcal{M},\mathcal{N}}u = u$ if $u \in \mathcal{M}$ and $P_{\mathcal{M},\mathcal{N}}u = 0$ if $u \in \mathcal{N}$. For a given matrix A in the space

$\mathbf{R}^{m\times n}$ of all real $m \times n$ matrices, denote by A^t, $\mathcal{N}(A)$, $\mathcal{R}(A)$, $\mathcal{N}_c(A)$, and $\mathcal{R}_c(A)$, respectively, the transpose of A, the null space of A, the range space of A, the set of all direct complements of $\mathcal{N}(A)$, and the set of all direct complements of $\mathcal{R}(A)$. Let I be the identity matrix and 0 the zero matrix of whatever size is appropriate for the context. Further, we denote by $A\mathcal{M}$ the image of $\mathcal{M}$ under A; i.e., $A\mathcal{M} = \{Au \mid u \in \mathcal{M}\}$.

Now let $A \in \mathbf{R}^{m\times n}$, let $\mathcal{M} \in \mathcal{N}_c(A)$, and let $\mathcal{S} \in \mathcal{R}_c(A)$. Consider the matrix equations

$$\begin{array}{llll} \text{(G1)} & AXA = A, & (\text{G}\mathcal{M}) & XA = P_{\mathcal{M},\mathcal{N}(A)}, \\ \text{(G2)} & XAX = X, & (\text{G}\mathcal{S}) & AX = P_{\mathcal{R}(A),\mathcal{S}}. \end{array} \tag{2.1}$$

Suppose that $\emptyset \neq \eta \subseteq \{1, 2, \mathcal{M}, \mathcal{S}\}$. Then let $A\eta$ denote the set of all those matrices X that satisfy equations (Gi) for all $i \in \eta$. Any $X \in A\eta$ is called an *η-inverse* of A, and is denoted by A^η. $\{1\}$-inverses are usually called *generalized inverses* or *g-inverses* and are also denoted by A^-. $\{1,2\}$-inverses are called *reflexive g-inverses* and are also denoted by A_r^-. For an extensive discussion of the theory of g-inversion, we refer, e.g., to the books by Ben-Israel and Greville [1], Hartung and Werner [3], Pringle and Rayner [9], and Rao and Mitra [10]; for a geometric approach, to Werner [12, Chap. 1] and Rao and Yanai [11]; and for a projector theoretical one to, e.g., the paper by Langenhop [4].

Only for the sake of clarity and for easier reference do we mention the following well-known results (cf. [12]; see also [15], [17]).

THEOREM 2.1. *For given $A \in \mathbf{R}^{m\times n}$, we have the following.*

(i) *The $\{2, \mathcal{M}, \mathcal{S}\}$-inverse of A exists uniquely.*

(ii) *Any $\{\mathcal{M}\}$-inverse of A and likewise any $\{\mathcal{S}\}$-inverse of A is always a $\{1\}$-inverse of A. Conversely, for any $\{1\}$-inverse of A there uniquely exist an $\mathcal{M} \in \mathcal{N}_c(A)$ and an $\mathcal{S} \in \mathcal{R}_c(A)$ such that $X \in A\{\mathcal{M}, \mathcal{S}\}$. Moreover, if $X \in A\{\mathcal{M}, \mathcal{S}\}$, then $XAX = A^{\{2,\mathcal{M},\mathcal{S}\}}$.*

(iii) *If $X \in A\{\mathcal{M}, \mathcal{S}\}$, then $\mathcal{M} = \mathcal{R}(XA) \subseteq \mathcal{R}(X)$, $\mathcal{N}(X) \subseteq \mathcal{S}$, and $X\mathcal{S} \subseteq \mathcal{N}(A)$. Hence, in particular,* $\operatorname{rank}(A) = \operatorname{rank}(AX) = \operatorname{rank}(XA)$. *Moreover, $X = A^{\{2,\mathcal{M},\mathcal{S}\}}$ iff $\mathcal{R}(X) = \mathcal{M}$ and $\mathcal{N}(X) = \mathcal{S}$.*

(iv) *If* $\operatorname{rank}(A) = r < \min\{m, n\}$ *then, for each s with $r \leq s \leq \min\{m, n\}$, there exist g-inverses A^- such that* $\operatorname{rank}(A^-) = s$. *Moreover,* $\operatorname{rank}(A) = \operatorname{rank}(A^-)$ *iff A^- is a reflexive g-inverse.*

(v) *The $\{2, \mathcal{R}(A^t), \mathcal{N}(A^t)\}$-inverse of A coincides with the Moore–Penrose inverse of A and is henceforth denoted by $A^\dagger$.*

When A is square, $\operatorname{ind}(A)$, the *index* of A, denotes the smallest positive integer k for which $\operatorname{rank}(A^k) = \operatorname{rank}(A^{k+1})$ or, equivalently, $\mathcal{R}(A^k) = \mathcal{R}(A^{k+1})$. Now, let A be any square matrix of index 1. Then $\mathcal{R}(A) \in \mathcal{N}_c(A)$ and $\mathcal{N}(A) \in \mathcal{R}_c(A)$, so that by Theorem 2.1(i) the $\{2, \mathcal{R}(A), \mathcal{N}(A)\}$-inverse exists and is unique. This g-inverse is called the *group-inverse* of A and is denoted by $A^\#$. It is the unique $\{1,2\}$-inverse X of A satisfying $AX = XA$. Recall that a square matrix A has $\operatorname{ind}(A) = k$ iff there exist a *nilpotent* matrix N_A of degree k (i.e., $N_A^k = 0$ whereas $N_A^{k-1} \neq 0$) and a *core* matrix C_A (i.e., $\operatorname{ind}(C_A) = 1$) such that

$$A = C_A + N_A, \quad C_A N_A = 0, \quad N_A C_A = 0; \tag{2.2}$$

see, for instance, [1, pp. 175–177] or [16, p. 246]. This decomposition is called the *core-nilpotent* decomposition of A and is uniquely determined. In the literature, core matrices are also often called *group matrices* or GP matrices.

From Chipman we have the following definition: Two matrices A and B of the same column number, say n, are said to be *complementary* to one another if $\mathbf{R}^n =$

$\mathcal{R}(A^t) \oplus \mathcal{R}(B^t)$. In the literature, complementary matrices have been studied because of their importance in statistics (see, for instance, Chipman [2], Pringle and Rayner [9], Hartung and Werner [3], and Werner [14]). In [15] (see also [14], [13], [17]) the following weaker or stronger versions of that notion are studied.

(a) B is said to be *weakly complementary* to A if $\mathcal{R}(A^t) \cap \mathcal{R}(B^t) = \{0\}$.

(b) B is said to be *(weakly) bicomplementary* to A if B and B^t are (weakly) complementary to A and A^t, respectively.

A pair of weakly bicomplementary matrices is also often said to be a pair of *disjoint* matrices (also written $A+B = A \oplus B$); cf. Mitra [5]. The connections between these concepts and the concept of generalized inversion are discussed in detail in [15] and [5]. Below we cite only those results that are of interest to us in this paper.

THEOREM 2.2 (see [15, p. 369]). *For $A \in \mathbf{R}^{n\times n}$, let $\mathcal{M} \in \mathcal{N}_c(A)$ and let $\mathcal{S} \in \mathcal{R}_c(A)$. Further, let H be a matrix of basis vectors for $\mathcal{M}^\perp$, and let T be a matrix of basis vectors for $\mathcal{S}$. If we define*

$$B = TH^t,$$

then B satisfies $\mathcal{N}(B) = \mathcal{M}$ and $\mathcal{R}(B) = \mathcal{S}$, and so B is bicomplementary to A.

Theorem 2.3 is also well known (cf. [15, pp. 359–364] in combination with [6, p. 240]).

THEOREM 2.3. *For given $A, B \in \mathbf{R}^{m\times n}$, the following conditions are equivalent:*

(i) $A + B = A \oplus B$;

(ii) $\mathcal{R}(A + B) = \mathcal{R}(A) \oplus \mathcal{R}(B)$;

(iii) $\operatorname{rank}(A + B) = \operatorname{rank}(A) + \operatorname{rank}(B)$;

(iv) $\mathcal{N}(A + B) = \mathcal{N}(A) \cap \mathcal{N}(B)$, $\mathcal{M} = [\mathcal{M} \cap \mathcal{N}(A)] \oplus [\mathcal{M} \cap \mathcal{N}(B)]$ *for each* $\mathcal{M} \in \mathcal{N}_c(A + B)$;

(v) $(A + B)\{1\} \subseteq A\{1\}$;

(vi) $(A + B)\{\mathcal{M}, \mathcal{S}\} \subseteq A\{\mathcal{M} \cap \mathcal{N}(B), \mathcal{S} \oplus \mathcal{R}(B)\}$ *for each* $\mathcal{M} \in \mathcal{N}_c(A + B)$ *and* $\mathcal{S} \in \mathcal{R}_c(A + B)$;

(vii) $A \preceq^- A + B$;

(viii) $A^t + B^t = A^t \oplus B^t$.

THEOREM 2.4 (see [15, p. 362]). *If $A + B = A \oplus B$, then*

$$(A + B)^{\{2,\mathcal{M},\mathcal{S}\}} = A^{\{2,\mathcal{M}\cap\mathcal{N}(B),\mathcal{S}\oplus\mathcal{R}(B)\}} + B^{\{2,\mathcal{M}\cap\mathcal{N}(A),\mathcal{S}\oplus\mathcal{R}(A)\}}$$

for every $\mathcal{M} \in \mathcal{N}_c(A + B)$ and every $\mathcal{S} \in \mathcal{R}_c(A + B)$.

3. $\mathcal{G}$-based partial order characterizations. In what follows, let $\mathcal{G}$ be a map on $\mathbf{R}^{m\times n}$ satisfying (1.1) for each $m \times n$ matrix A, and let $\Omega_{\mathcal{G}}$ be its support. For A, $B \in \Omega_{\mathcal{G}}$, it is convenient to put

$$\mathcal{G}(A \mid B) := \{GAG \mid G \in \mathcal{G}(B)\}, \tag{3.1}$$

$$\mathcal{G}_r(B) := \{B_r^- \mid B_r^- \in \mathcal{G}(B)\}, \tag{3.2}$$

and

$$\mathcal{G}_r(A \mid B) := \{GAG \mid G \in \mathcal{G}_r(B)\}. \tag{3.3}$$

Observe that $\mathcal{G}$ is semicomplete iff $\mathcal{G}(A \mid A) = \mathcal{G}_r(A)$ for each $A \in \Omega_{\mathcal{G}}$.

Mitra [6, p. 242] has shown that $\preceq^{\mathcal{G}}$ as defined in (1.2) need not always correspond to a partial order. It is therefore quite natural to ask for sufficient and/or necessary

conditions under which (1.2) defines a partial order. In [6, p. 243], Mitra derived the following sufficient condition for $\preceq^{\mathcal{G}}$ to correspond to a partial order.

THEOREM 3.1. *Let*

$$A \preceq^{\mathcal{G}} B,\ B \text{ not maximal} \implies \mathcal{G}(A \mid B) \subseteq \mathcal{G}(A). \tag{3.4}$$

Then $\preceq^{\mathcal{G}}$ defines a partial order.

Note that a matrix B is called *maximal* relative to $\preceq^{\mathcal{G}}$ if there is no matrix $C \neq B$ such that $B \preceq^{\mathcal{G}} C$. Since $B \preceq^{\mathcal{G}} C$ implies $B \preceq^{-} C$, by Theorem 2.3(iii) $\operatorname{rank}(B) < \min\{m, n\}$ whenever B is not maximal. Further, note that if $B \notin \Omega_{\mathcal{G}}$ and/or B is of full rank (i.e., $\operatorname{rank}(B) = \min\{m, n\}$) then B is maximal relative to $\preceq^{\mathcal{G}}$.

In context with Theorem 3.1 it is further pertinent to mention that the only time when condition (3.4) is invoked by Mitra is in proving that the implication

$$A \preceq^{\mathcal{G}} B,\ B \preceq^{\mathcal{G}} C \implies A \preceq^{\mathcal{G}} C$$

holds true, that is, in proving that the relation $\preceq^{\mathcal{G}}$ is transitive. But there $B \in \Omega_{\mathcal{G}}$ so that, without loss of generality, B can be assumed not to be maximal, for otherwise $B = C$, in which case the implication is trivial. In the same paper, Mitra showed [6, p. 243] that condition (3.4), although sufficient, is in general not necessary for $\preceq^{\mathcal{G}}$ to define a partial order.

In this paper we are especially interested in semicomplete maps. Our next theorem will show, in particular, that if $\mathcal{G}$ is a semicomplete map on the set of square $n \times n$ matrices, then, for $\preceq^{\mathcal{G}}$ to be a partial order, condition (3.4) is not only sufficient but also necessary. The proof follows from [6, Thms. 2.3, 2.4, and 2.5]. In passing, we mention that a different possibility for proving this theorem would be to make use of Theorems 2.2, 2.3, and 2.4 in this paper.

THEOREM 3.2. *Let $\mathcal{G}$ be a semicomplete map on $\mathbf{R}^{n\times n}$. Then the following conditions are equivalent:*

(i) *$\preceq^{\mathcal{G}}$ is a partial order;*
(ii) *$A \preceq^{\mathcal{G}} B$, B not maximal $\implies \mathcal{G}_r(A \mid B) \subseteq \mathcal{G}_r(A)$;*
(iii) *$A \preceq^{\mathcal{G}} B$, B not maximal $\implies \mathcal{G}(A \mid B) \subseteq \mathcal{G}(A)$.*

Next we give Theorem 3.3.

THEOREM 3.3. *Let $\mathcal{G}$ as defined by* (1.1) *be a map on $\mathbf{R}^{m\times n}$ and assume that*

$$\mathcal{G}(A) \neq \emptyset \ \Rightarrow\ \mathcal{G}(A) \cap A\{1,2\} \neq \emptyset. \tag{3.5}$$

Let the relation $\preceq^{\mathcal{G}}$ be defined as in (1.2). *If $A \in \Omega_{\mathcal{G}}$ then A is maximal if and only if* $\operatorname{rank}(A) = \min\{m, n\}$.

Proof. Let $A \in \Omega_{\mathcal{G}}$ and $\operatorname{rank}(A) \neq \min\{m, n\}$. Here we exactly follow the steps in the proof of Theorem 2.5 in [6] to arrive at a matrix $C \neq A$ which dominates A under $\preceq^{\mathcal{G}}$; the modifications required to prove this are obvious. □

Note that a semicomplete map trivially satisfies condition (3.5). Hence Theorem 3.3 holds for a semicomplete map.

That the characterizations in Theorem 3.2 are not necessarily true for a map $\mathcal{G}$ that is defined on the set of nonsquare $m \times n$ matrices (i.e., $m \neq n$) is shown by our next numerical example. Although Example 5 in [6] might also be used for this purpose, the example given below is easier to understand.

Example 3.4. Consider the matrices

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad G = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix},$$

and the set $\mathcal{A}$ consisting of all those real matrices

$$H_{a,b} := \begin{pmatrix} 1 & 0 & a & b \\ 0 & 0 & 0 & 0 \\ 1 & 0 & a & b \end{pmatrix}$$

for which $a \neq 0$ and/or $b \neq 0$. By checking the corresponding defining equations (G1) and (G2) of (2.1) it is seen that $G \in B\{1,2\}$ and $\mathcal{A} \subseteq A\{1,2\}$. Define the map $\mathcal{G}$ on $\mathbf{R}^{4\times 3}$ by

$$\mathcal{G}(C) = \begin{cases} \mathcal{A} & \text{if } C = A, \\ \{G\} & \text{if } C = B, \\ \emptyset & \text{otherwise.} \end{cases}$$

As is evident, $\mathcal{G}$ is semicomplete. Check that $A \preceq^{\mathcal{G}} B$. Since $\Omega_{\mathcal{G}} = \{A, B\}$, the relation $\preceq^{\mathcal{G}}$ is transitive iff

$$A \preceq^{\mathcal{G}} B,\ B \preceq^{\mathcal{G}} C \Longrightarrow A \preceq^{\mathcal{G}} C$$

holds true. Notice that C satisfies $B \preceq^{\mathcal{G}} C$ iff

$$C = C_{c,d} := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ c & 0 & -c \\ d & 0 & -d \end{pmatrix}$$

for some real scalars c and d. Since $A \preceq^{\mathcal{G}} C_{c,d}$ holds irrespective of c and d, it is now clear that the relation $\preceq^{\mathcal{G}}$ defines a partial order. However, observe that $GAG = H_{0,0}$ fails to belong to $\mathcal{A}$. □

In this context it should be mentioned that if the semicomplete map $\mathcal{G}$ is such that for each A its image $\mathcal{G}(A)$ is a singleton or empty then, even in the nonsquare case (i.e., $m \neq n$), condition (3.4) turns out to be necessary and sufficient for $\preceq^{\mathcal{G}}$ to define a partial order. The following result is only slightly more general than Theorem 2.6 in [6] insofar as it allows $\mathcal{G}(A)$ to be empty; however, observe that the proof given in [6, p. 250] actually includes our result.

THEOREM 3.5. *Let $\mathcal{G}$ as defined by* (1.1) *be a semicomplete map with* card $\mathcal{G}(A) \leq 1$ *for each $A \in \mathbf{R}^{m\times n}$. Then for $\preceq^{\mathcal{G}}$ to correspond to a partial order, it is necessary and sufficient that*

$$A \preceq^{\mathcal{G}} B,\ B \text{ not maximal} \Longrightarrow \mathcal{G}(A \mid B) \subseteq \mathcal{G}(A). \tag{3.4}$$

4. $\mathcal{G}$-based partial order extensions. We begin this section with the following result.

THEOREM 4.1. *Let $\preceq^{\mathcal{G}}$ as defined by* (1.2) *correspond to a partial order, and let $\Omega_{\mathcal{G}}$ be a proper subset of $\mathbf{R}^{m\times n}$. The relation $\preceq^{\mathcal{G}}$ can then be extended to a $\mathcal{G}$-based partial order $\preceq^{\mathcal{G}^*}$ that supports the whole of $\mathbf{R}^{m\times n}$.*

Proof. Define the map $\mathcal{G}^*$ by

$$\mathcal{G}^*(A) = \begin{cases} \mathcal{G}(A) & \text{if } A \in \Omega_{\mathcal{G}}, \\ \{A^-_{\text{max}}\} & \text{otherwise,} \end{cases} \tag{4.1}$$

where A^-_{max} is an arbitrary but fixed full-rank g-inverse of A. Theorem 2.1(iv) tells us that such a g-inverse of A with rank $r := \min\{m, n\}$ does always exist. We next prove that each matrix $A \notin \Omega_{\mathcal{G}}$ is maximal relative to $\preceq^{\mathcal{G}^*}$. To that end let $A \notin \Omega_{\mathcal{G}}$.

Moreover, let B be any matrix such that $A \preceq^{\mathcal{G}^*} B$. Clearly, $A \preceq^{\mathcal{G}^*} B$ iff $(B-A)A^-_{\max} = 0$ and $A^-_{\max}(B-A) = 0$. Since $A^-_{\max}$ is a full-rank matrix, this, of course, happens iff $B - A = 0$. So we arrive at $B = A$, and it is now clear that A is maximal relative to $\preceq^{\mathcal{G}^*}$. With this observation in mind, it is evident that $\preceq^{\mathcal{G}^*}$ inherits the transitivity property from $\preceq^{\mathcal{G}}$. As $\preceq^{\mathcal{G}^*}$ supports each matrix, the proof is complete. □

The proof of the preceding theorem has shown that there are many trivial ways to extend a $\mathcal{G}$-based partial order on the whole of $\mathbf{R}^{m\times n}$. We call an extension $\preceq^{\mathcal{G}^*}$ a *trivial extension* of $\preceq^{\mathcal{G}}$ if

$$A \preceq^{\mathcal{G}^*} B \iff A \preceq^{\mathcal{G}} B$$

for all pairs of matrices with $A \neq B$. In fact, this means, in an obvious sense, that $\preceq^{\mathcal{G}}$ and $\preceq^{\mathcal{G}^*}$ are *equivalent* relations.

Now it seems quite natural to ask the following: Given a $\mathcal{G}$-based partial order that excludes from its support a chunk of $\mathbf{R}^{m\times n}$, does it always admit a *nontrivial* $\mathcal{G}$-based partial order extension that supports the whole of $\mathbf{R}^{m\times n}$?

Concerning the sharp order $\preceq^{\#}$, Mitra [7, Thm. 2.1] has already given an affirmative answer to this question. In order to discuss that result we need some further notation. Recall that if A is a square $n \times n$ matrix, then

$$A = C_A + N_A$$

stands for the uniquely determined core-nilpotent decomposition of A; see (2.2) in §2. Let us write $A \preceq^{\dagger} B$ if

$$A \preceq^{-} B \quad \text{and} \quad C_A \preceq^{\#} C_B.$$

Note that the relation $\preceq^{\dagger}$ is an *extension* of the sharp order $\preceq^{\#}$ in the sense that

$$A \preceq^{\#} B \implies A \preceq^{\dagger} B.$$

Since the minus order and the sharp order are partial orders, $\preceq^{\dagger}$ is also a partial order. Moreover, observe that by Theorem 2.1 in [7] the relation $\preceq^{\dagger}$ is equivalent to the $\mathcal{G}$-based relation $\preceq^{\mathcal{G}^0}$, where the map $\mathcal{G}^0$ is defined by

$$\mathcal{G}^0(A) := \begin{cases} \{A^{\#}\} & \text{if } \operatorname{ind}(A) = 1, \\ \{A^- \mid \mathcal{R}(C_A) \subseteq \mathcal{R}(A^-),\ \mathcal{N}(A^-) \subseteq \mathcal{N}(C_A)\} & \text{otherwise.} \end{cases} \tag{4.2}$$

It is pertinent to prove here that this map $\mathcal{G}^0$ fails to be semicomplete whenever $n \geq 3$. For this purpose, let $n \geq 3$ and let $A \in \mathbf{R}^{n\times n}$ be a matrix that is neither core nor nilpotent. Then $A = C_A + N_A$, $C_A \neq 0$, $N_A \neq 0$, $1 < \operatorname{rank}(A) = \operatorname{rank}(C_A) + \operatorname{rank}(N_A) < n$. So it is possible to choose an $\mathcal{M} \in \mathcal{N}_c(A)$ such that $\mathcal{R}(C_A) \not\subseteq \mathcal{M}$ (see §2). In addition, choose $\mathcal{S} \in \mathcal{R}_c(A)$. From Theorem 2.2 we then know that there exists a matrix B such that $\mathcal{R}(B) = \mathcal{S}$, $\mathcal{N}(B) = \mathcal{M}$, and B is bicomplementary to A. Note that the matrix $A + B$ is hence, in particular, nonsingular. From Theorem 2.3 along with Theorems 2.4 and 2.1 we get

$$(A+B)^{-1}A(A+B)^{-1} = A^{\{2,\mathcal{N}(B),\mathcal{R}(B)\}}. \tag{4.3}$$

But although trivially $(A+B)^{-1} \in \mathcal{G}^0(A)$, $(A+B)^{-1}A(A+B)^{-1}$ does not belong to $\mathcal{G}^0(A)$ because $\mathcal{R}(C_A) \not\subseteq \mathcal{N}(B) = \mathcal{M}$. But now it is clear that $\mathcal{G}^0$ is not a semicomplete map if $n \geq 3$. The only exceptions are the cases $n = 1$ and $n = 2$. Every matrix of order 1×1 is a core matrix. Here the map is clearly semicomplete. A matrix A of

order 2×2 is either a core matrix or is nilpotent. If A is nilpotent then $\mathcal{G}^0(A) = A\{1\}$, so that the map $\mathcal{G}^0$ is also trivially seen to be semicomplete in that case.

For A, let P_A denote the well-defined projector onto $\mathcal{R}(C_A)$ along $\mathcal{N}(C_A)$; that is, let $P_A = C_A C_A^{\#}$. Since

$$C_A^{\#} + (I - P_A)A^{-}(I - P_A) \in \mathcal{G}^0(A) \quad \text{if} \quad \text{ind}(A) > 1$$

(see [7, Lem. 2.3]), clearly $\mathcal{G}^0(A) \neq \emptyset$ for each A. Combining observations now shows that $\preceq^{\mathcal{G}^0}$ is indeed a $\mathcal{G}$-based partial order extension of the sharp order that supports the whole of $\mathbf{R}^{n\times n}$. This is shown in Mitra [7].

The theorem that follows is somewhat different; it gives a $\mathcal{G}$-based extension $\preceq^{\mathcal{G}_*}$ of $\preceq^{\#}$ and various equivalent descriptions of the underlying map $\mathcal{G}_*$.

THEOREM 4.2. *For square $n \times n$ matrices, let $\mathcal{G}^0$ and $\mathcal{G}_*$ be defined by* (4.2) *and*

$$\mathcal{G}_*(A) := \begin{cases} \{A^{\#}\} & \text{if } \text{ind}(A) = 1, \\ \{A^{-} \mid A^{-}C_A A^{-} = C_A^{\#}\} & \text{otherwise,} \end{cases} \tag{4.4}$$

respectively. Then we have the following.

(i) *For noncore square matrices A, $\mathcal{G}_*(A)$ allows the following equivalent descriptions:*

$$\mathcal{G}_*(A) = \{A^{-} \mid A^{-}C_A = C_A A^{-}\}, \tag{4.5}$$

$$\mathcal{G}_*(A) = \{A^{-} \mid \mathcal{R}(C_A) \subseteq \mathcal{R}(A^{-}A),\ \mathcal{N}(AA^{-}) \subseteq \mathcal{N}(C_A)\}, \tag{4.6}$$

$$\mathcal{G}_*(A) = \{C_A^{\#} + (I - P_A)Z(I - P_A) \mid Z \in A\{1\}\}, \quad \textit{where} \quad P_A = C_A C_A^{\#}. \tag{4.7}$$

(ii) *The relation $\preceq^{\mathcal{G}_*}$ is a semicomplete partial order extension of the sharp order $\preceq^{\#}$ and supports the whole of $\mathbf{R}^{n\times n}$. Moreover, if $n \geq 3$ then $\mathcal{G}_*$ is properly finer than $\mathcal{G}^0$. Precisely, $\mathcal{G}_*(A) \subseteq \mathcal{G}^0(A)$ for each matrix A, and $\mathcal{G}_*(A) \neq \mathcal{G}^0(A)$ whenever A is neither core nor nilpotent.*

Proof. (i): Let A be a noncore square matrix. It is easily checked that (4.7) is a subset of $A\{1\}$. We now show that $A^{-} \in$ the set (4.7) $\Rightarrow A^{-} \in$ the set (4.4) $\Rightarrow A^{-} \in$ the set (4.5) $\Rightarrow A^{-} \in$ the set (4.6) $\Rightarrow A^{-} \in$ the set (4.7), thus establishing equivalence. Let $A^{-} = C_A^{\#} + (I - P_A)Z(I - P_A)$. Then $A^{-}C_A A^{-} = [C_A^{\#} + (I - P_A)Z(I - P_A)]C_A C_A^{\#} C_A[C_A^{\#} + (I - P_A)Z(I - P_A)] = C_A^{\#} C_A C_A^{\#} C_A C_A^{\#} = C_A^{\#}$. This implies $A^{-}C_A = A^{-}C_A A^{-}C_A = C_A^{\#} C_A = C_A C_A^{\#} = C_A A^{-}C_A A^{-} = C_A A^{-}$ using Theorem 2.3(v) since

$$A = C_A \oplus N_A,$$

which in turn shows that

$$C_A = C_A A^{-}C_A = A^{-}C_A^2 = A^{-}AC_A.$$

This is equivalent to $\mathcal{R}(C_A) \subseteq \mathcal{R}(A^{-}A)$. Similarly,

$$C_A = C_A A^{-}C_A = C_A^2 A^{-} = C_A AA^{-}$$

or, equivalently, $\mathcal{N}(AA^{-}) \subseteq \mathcal{N}(C_A)$. From the pair of equivalences just established it is seen that $A^{-} \in$ the set (4.6) implies that the matrix $X = A^{-}$ satisfies the simultaneous system of equations

$$XC_A^2 = C_A = C_A^2 X. \tag{4.8}$$

The matrix $X = C_A^{\#}$ is a particular solution of (4.8). Using Lemma 2.3.1 in [10], we thus have $X = C_A^{\#} + (I - P_A)Z(I - P_A)$, Z arbitrary as the expression for the general solution of (4.8). But then it is immediate that in order that $X \in A\{1\}$ we must have $Z \in N_A\{1\}$. Since $(I - P_A)X(I - P_A) = (I - P_A)Z(I - P_A)$, it is now clear that (4.6) implies (4.7).

(ii): From part (i) it is clear that $\preceq^{\mathcal{G}_*}$ is a $\mathcal{G}$-based extension of $\preceq^{\#}$ that supports the whole of $\mathbf{R}^{n\times n}$. Semicompleteness of $\mathcal{G}_*$ follows from (4.6) by observing that $\mathcal{R}(A^-AA^-A) = \mathcal{R}(A^-A)$ and $\mathcal{N}(AA^-AA^-) = \mathcal{N}(AA^-)$. Next note that the proof of Theorem 2.1 in [7] can be used word for word to establish that $\preceq^{\dagger}$ is equivalent to $\preceq^{\mathcal{G}_*}$. The relation $\preceq^{\mathcal{G}_*}$ therefore corresponds to a $\mathcal{G}$-based partial order extension of the sharp order and is equivalent to the $\mathcal{G}$-based relation $\preceq^{\mathcal{G}^0}$. Comparing (4.2) with (4.6) shows that $\mathcal{G}_*(A) \subseteq \mathcal{G}^0(A)$ holds for each matrix A. That $\mathcal{G}_*$ is properly finer than $\mathcal{G}^0$ whenever $n \geq 3$ is an easy consequence of the lines directly following (4.2). To see this, let A be neither core nor nilpotent and consider the matrix $A + B$ constructed there. As seen above, $(A+B)^{-1} \in \mathcal{G}^0(A)$. That $(A+B)^{-1}$ fails to belong to $\mathcal{G}^*(A)$ follows from (4.3) since $\mathcal{R}(C_A) \not\subseteq \mathcal{N}(B)$. Note that $\mathcal{G}_*(A) = \mathcal{G}^0(A)$ if A is core or nilpotent. This completes the proof. □

Notice that $\mathcal{G}_*$ as defined by (4.4) is semicomplete and that $\operatorname{card}\mathcal{G}_*(A) > 1$ whenever $\operatorname{ind}(A) > 1$. It is therefore interesting to mention that the following example will show that for $n \geq 3$ there does not exist any $\mathcal{G}$-based semicomplete partial order extension $\preceq^{\tilde{\mathcal{G}}}$ of the sharp order $\preceq^{\#}$ such that $\tilde{\mathcal{G}}(A)$ is a singleton set for all A.

Example 4.3. First, let $n = 3$. Consider the nilpotent matrix

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

By virtue of Theorem 3.2 it is sufficient to show that for each reflexive g-inverse A_r^- of A we can always find a matrix B of rank 2 and index 1 such that $B^{\#}AB^{\#} \neq A_r^-$, although $AA_r^- = BA_r^-$ and $A_r^-A = A_r^-B$.

Observe that

$$A\{1,2\} = \left\{ \begin{pmatrix} 1 \\ c \\ d \end{pmatrix} \begin{pmatrix} a & b & 1 \end{pmatrix} \mid a, b, c, d \in \mathbf{R} \right\}$$

is an efficient parametrization of the set of all reflexive g-inverses of the nilpotent matrix A. This can be easily seen, for instance, by means of (G1) and (G2) in (2.1). Consider an arbitrary but fixed reflexive g-inverse

$$A_r^- = \begin{pmatrix} 1 \\ c \\ d \end{pmatrix} \begin{pmatrix} a & b & 1 \end{pmatrix}$$

of A and put

$$x := \begin{pmatrix} 1 \\ f \\ -(a + fb) \end{pmatrix} \quad \text{and} \quad y^t := \begin{pmatrix} \dfrac{-fd - ca}{f - c} & \dfrac{d + a + fb}{f - c} & 1 \end{pmatrix},$$

where $f := 1 - cc^{\dagger}$. Set

$$B := A + xy^t.$$

Since

$$f = 0 \iff c \neq 0 \quad \text{and} \quad f = 1 \iff c = 0, \tag{4.9}$$

$fc = 0$ and $f^2 = f$. This in turn implies $y^t x = 0$. As $A_r^- x = 0$ and $y^t A_r^- = 0$, clearly $A_r^- A = A_r^- B$ and $A A_r^- = B A_r^-$. Next notice that $B^3 = Axy^t A + xy^t Axy^t = A + xy^t = B$. Therefore, $\operatorname{ind}(B) = 1$ and $B^\# = B$ (note §2). Recalling (4.9), the desired result now follows from

$$\begin{aligned} B^\# A B^\# &= BAB \\ &= xy^t \\ &= \begin{pmatrix} 1 \\ \mathbf{f} \\ -(c+fb) \end{pmatrix} \begin{pmatrix} \dfrac{-fd-ca}{f-c} & \dfrac{d+a+fb}{f-c} & \mathbf{1} \end{pmatrix} \\ &\neq \begin{pmatrix} 1 \\ \mathbf{c} \\ d \end{pmatrix} \begin{pmatrix} a & b & \mathbf{1} \end{pmatrix} \\ &= A_r^-. \end{aligned}$$

If $n > 3$, then the proof follows along similar lines and is thus omitted. □

5. Semicomplete $\mathcal{G}$-based partial order extensions. In this final section, let $\mathcal{G}$ as defined by (1.1) be a semicomplete map on $\mathbf{R}^{m\times n}$ and let this map induce a partial order $\preceq^{\mathcal{G}}$.

Recall that each *full-rank* matrix A (i.e., $\operatorname{rank}(A) = \min\{m, n\}$) is maximal relative to $\preceq^{\mathcal{G}}$. It is pertinent to mention that modifying $\mathcal{G}(\cdot)$ for one or more full-rank matrices will lead to a new map but that this modified map continues to induce exactly the same relation $\preceq^{\mathcal{G}}$ as $\mathcal{G}$ (at least concerning all pairs (A, B) of matrices with $A \neq B$). Without loss of generality, let us henceforth assume that $\mathcal{G}(A) = A\{1\}$ for each full-rank matrix A.

Let us further assume that $\Omega_{\mathcal{G}} \neq \mathbf{R}^{m\times n}$. Then $A \notin \Omega_{\mathcal{G}}$ for some matrix A with $\operatorname{rank}(A) < \min\{m, n\}$. For each matrix $B \in \mathbf{R}^{m\times n}$, let $\mathcal{UP}_{\mathcal{G}}(B) := \{A \mid A \preceq^{\mathcal{G}} B\}$ denote the set of all those matrices that are *upstream* of B. Moreover, whenever $B \notin \Omega_{\mathcal{G}}$ is such that $\mathcal{UP}_{\mathcal{G}}(B) \neq \emptyset$, put

$$\mathcal{G}_0(B) := \{B^- \mid B^- A B^- \in \mathcal{G}(A) \text{ for each } A \in \mathcal{UP}_{\mathcal{G}}(B)\}. \tag{5.1}$$

Finally, define a new map $\mathcal{G}^*$ (in respect to $\mathcal{G}$) by

$$\mathcal{G}^*(B) := \begin{cases} \mathcal{G}(B) & \text{if } B \in \Omega_{\mathcal{G}}; \\ \mathcal{G}_0(B) & \text{if } B \notin \Omega_{\mathcal{G}},\ \mathcal{UP}_{\mathcal{G}}(B) \neq \emptyset; \\ B\{1\} & \text{otherwise.} \end{cases} \tag{5.2}$$

For square matrices (i.e., $m = n$), Theorem 3.2 could now tend to make one believe that $\mathcal{G}^*$ as defined by (5.2) does induce, in any case, a (nontrivial) semicomplete partial order extension $\preceq^{\mathcal{G}^*}$ of $\preceq^{\mathcal{G}}$ which, unlike $\preceq^{\mathcal{G}}$, supports the whole of $\mathbf{R}^{n\times n}$. That this, however, is erroneous is seen by the following example.

Example 5.1. Consider the matrices

$$A_1 := \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad A_2 := \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and

$$B := \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

By checking the corresponding defining equations (G1) and (G2) of (2.1) it is seen that

$$G_1 := \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad G_2 := \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & -1 & 0 \end{pmatrix}$$

are reflexive g-inverses of A_1 and A_2, respectively. Define the map

$$\mathcal{G}: \quad \mathbf{R}^{4\times 4} \to \mathcal{P}(\mathbf{R}^{4\times 4})$$

by

$$\mathcal{G}(A) = \begin{cases} \{G_1\} & \text{if } A = A_1, \\ \{G_2\} & \text{if } A = A_2, \\ \emptyset & \text{otherwise.} \end{cases}$$

It is obvious that $\mathcal{G}$ is semicomplete and that the induced order $\preceq^{\mathcal{G}}$ defines a partial order with support $\Omega_{\mathcal{G}} = \{A_1, A_2\}$. Now let $\tilde{\mathcal{G}}$ be a semicomplete extension of $\mathcal{G}$ such that the induced order $\preceq^{\tilde{\mathcal{G}}}$ corresponds to a partial order.

Suppose that $\preceq^{\tilde{\mathcal{G}}}$ supports $\mathbf{R}^{4\times 4}$; that is, $\Omega_{\tilde{\mathcal{G}}} = \mathbf{R}^{4\times 4}$. Then $B \in \Omega_{\tilde{\mathcal{G}}}$ or, equivalently, $\tilde{\mathcal{G}}_r(B) \neq \emptyset$. On the one hand, by Theorem 3.3, B cannot be maximal with respect to $\preceq^{\tilde{\mathcal{G}}}$ because B is singular. Hence, by Theorem 3.2, $\tilde{\mathcal{G}}_r(A \mid B) \subseteq \tilde{\mathcal{G}}_r(A)$ for each $A \preceq^{\tilde{\mathcal{G}}} B$. For $i = 1, 2$, clearly $\tilde{\mathcal{G}}_r(A_i) = \mathcal{G}_r(A_i) = \{G_i\}$. Since $BG_1 = A_1G_1$ and $G_1B = G_1A_1$, $A_1 \preceq^{\tilde{\mathcal{G}}} B$. Likewise, it is seen that $A_2 \preceq^{\tilde{\mathcal{G}}} B$. Consequently,

$$B_r^- A_i B_r^- = G_i \text{ for } i = 1, 2 \text{ and for each } B_r^- \in \tilde{\mathcal{G}}_r(B).$$

On the other hand, recall that $\tilde{\mathcal{G}}_r(B) \subseteq B\{1, 2\}$. It is easy to check that

$$B\{1\} = \left\{ \begin{pmatrix} 1 & 0 & 0 & x_1 \\ 0 & 1 & 0 & x_2 \\ 0 & 0 & 1 & x_3 \\ y_1 & y_2 & y_3 & z \end{pmatrix} \mid x_i, y_i, z \in \mathbf{R} \quad (i = 1, 2, 3) \right\}.$$

Therefore,

$$B\{1, 2\} = \left\{ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ y_1 & y_2 & y_3 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & x_1 \\ 0 & 1 & 0 & x_2 \\ 0 & 0 & 1 & x_3 \end{pmatrix} \mid x_i, y_i \in \mathbf{R} \quad (i = 1, 2, 3) \right\}$$

because $B\{1, 2\} = B\{1\}\{B\}B\{1\}$ (recall Theorem 2.1). From this we get

$$\{B_r^- A_1 B_r^- \mid B_r^- \in B\{1, 2\}\} = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ y_1 & y_2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & x_1 \\ 0 & 1 & 0 & x_2 \end{pmatrix} \mid x_i, y_i \in \mathbf{R} \quad (i = 1, 2) \right\} \tag{5.3}$$

and

$$\{B_r^- A_2 B_r^- \mid B_r^- \in B\{1,2\}\} = \left\{ \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ y_2 & y_3 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 & x_2 \\ 0 & 0 & 1 & x_3 \end{pmatrix} \mid x_i, y_i \in \mathbf{R} \quad (i = 2,3) \right\}. \tag{5.4}$$

Invoking (5.3) and (5.4), respectively, now yields

$$B_r^- A_1 B_r^- = G_1 \text{ iff } y_1 = 1, y_2 = 0, x_1 = 1, x_2 = 1$$

and

$$B_r^- A_2 B_r^- = G_2 \text{ iff } y_2 = 1, y_3 = -1, x_2 = 0, x_3 = 0.$$

Since it is impossible to simultaneously set y_2 equal to 0 and equal to 1, there does not exist any reflexive g-inverse B_r^- of B simultaneously satisfying

$$B_r^- A_1 B_r^- = G_1 \quad \text{and} \quad B_r^- A_2 B_r^- = G_2.$$

But, in fact, $\tilde{\mathcal{G}}_r(B) \subseteq B\{1,2\}$. Consequently, $\tilde{\mathcal{G}}_r(B) = \emptyset$, which is a contradiction to $\Omega_{\tilde{\mathcal{G}}} = \mathbf{R}^{4\times 4}$. As desired, the relation $\preceq^{\mathcal{G}}$ thus fails to possess a $\mathcal{G}$-based semicomplete partial order extension $\preceq^{\tilde{\mathcal{G}}}$ that supports the whole of $\mathbf{R}^{4\times 4}$. □

Example 5.1 thus exhibits that we cannot always expect $\mathcal{G}^*$ to support the whole set of $m \times n$ matrices. Nevertheless, we will next show that $\mathcal{G}^*$ is semicomplete whenever $\mathcal{G}$ is semicomplete.

THEOREM 5.2. *Let $\mathcal{G}$ as defined by* (1.1) *be a semicomplete map, and let $\mathcal{G}^*$ be defined according to* (5.2). *Then $\mathcal{G}^*$ is also semicomplete.*

Proof. Let $B \in \Omega_{\mathcal{G}^*}$. Then $\mathcal{G}^*(B) \neq \emptyset$. We must show that $B^- BB^- \in \mathcal{G}^*(B)$ whenever $B^- \in \mathcal{G}^*(B)$. If $B \in \Omega_{\mathcal{G}}$, then $\mathcal{G}^*(B) = \mathcal{G}(B)$, and the desired result follows from the semicompleteness of $\mathcal{G}$. Next, assume that $B \notin \Omega_{\mathcal{G}}$. Then either $\mathcal{UP}_{\mathcal{G}}(B) = \emptyset$ or $\mathcal{UP}_{\mathcal{G}}(B) \neq \emptyset$. In the former case $\mathcal{G}^*(B) = B\{1\}$, so that $B^- BB^- \in \mathcal{G}^*(B)$ holds trivially for each $B^- \in \mathcal{G}^*(B)$ (note Theorem 2.1(ii)). In the latter case $\mathcal{G}^*(B) = \mathcal{G}_0(B)$. But then, by the definition of $\mathcal{G}_0(B)$, $B^- \in \mathcal{G}^*(B)$ iff $B^- AB^- \in \mathcal{G}(A)$ for each $A \in \mathcal{UP}_{\mathcal{G}}(B)$. Since $A \in \mathcal{UP}_{\mathcal{G}}(B)$ iff $A \preceq^{\mathcal{G}} B$, then $A \preceq^- B$ and so, in view of Theorem 2.3(ii) and (v), $\mathcal{R}(A) \subseteq \mathcal{R}(B)$ and $\mathcal{N}(B) \subseteq \mathcal{N}(A)$ or, equivalently, $BB^- A = A = AB^- B$. Consequently, $B^- BB^- AB^- BB^- = B^- AB^- \in \mathcal{G}(A)$. Since $B_r^- := B^- BB^-$ is a reflexive g-inverse of B and $B_r^- AB_r^- \in \mathcal{G}(A)$ is satisfied, the proof is done. □

For the $\mathcal{G}$-based sharp order $\preceq^{\#}$ (set $\mathcal{G}(A) = \{A^{\#}\}$ if $\mathrm{ind}(A) = 1$ and $\mathcal{G}(A) = \emptyset$ if $\mathrm{ind}(A) > 1$), Theorem 5.5 will even show that the associated relation $\preceq^{\mathcal{G}^*}$ corresponds to a partial order. In order to prove this, we need the following result.

THEOREM 5.3. *For square matrices, let the semicomplete $\mathcal{G}$-based relations $\preceq^{\mathcal{G}}$ and $\preceq^{\tilde{\mathcal{G}}}$ both correspond to partial orders. Further, let $\mathcal{G}^*$ be defined in respect to $\mathcal{G}$ according to* (5.2). *If $\preceq^{\tilde{\mathcal{G}}}$ is a $\mathcal{G}$-based extension of $\preceq^{\mathcal{G}}$, then*

$$\tilde{\mathcal{G}}(A) \subseteq \mathcal{G}^*(A) \tag{5.5}$$

for each A. Thus $\preceq^{\tilde{\mathcal{G}}}$ is finer than $\preceq^{\mathcal{G}^*}$; that is,

$$A \preceq^{\tilde{\mathcal{G}}} B \implies A \preceq^{\mathcal{G}^*} B. \tag{5.6}$$

Note that (5.5) *and* (5.6) *are not equivalent conditions. In general* (5.5) $\Rightarrow$ (5.6). *The reverse implication, however, need not always be true.*

Proof. By the definition of a $\mathcal{G}$-based extension, clearly $\tilde{\mathcal{G}}(B) = \mathcal{G}(B)$ whenever $\mathcal{G}(B) \neq \emptyset$. Since in that case also $\mathcal{G}^*(B) = \mathcal{G}(B)$, we trivially arrive at $\tilde{\mathcal{G}}(B) \subseteq \mathcal{G}^*(B)$.

Let us next consider the case when $\tilde{\mathcal{G}}(B) \neq \emptyset$ but $\mathcal{G}(B) = \emptyset$. Note that, by Theorem 3.3, B is maximal relative to $\preceq^{\tilde{\mathcal{G}}}$ iff B is nonsingular. First, let B be not maximal and suppose that there does exist some proper predecessor of B relative to $\preceq^{\tilde{\mathcal{G}}}$, say A. Then $A \neq B$ and $A \preceq^{\tilde{\mathcal{G}}} B$. Since $\tilde{\mathcal{G}}$ is semicomplete and $\preceq^{\tilde{\mathcal{G}}}$ corresponds to a partial order, we know from Theorem 3.2 that $\tilde{\mathcal{G}}(A \mid B) \subseteq \tilde{\mathcal{G}}(A)$ for each $A \preceq^{\tilde{\mathcal{G}}} B$. Therefore, in particular, $\tilde{\mathcal{G}}(A \mid B) \subseteq \mathcal{G}(A)$ for each A with $A \preceq^{\mathcal{G}} B$, thus showing that $\tilde{\mathcal{G}}(B) \subseteq \mathcal{G}_0(B) = \mathcal{G}^*(B)$ when $\mathcal{UP}_{\mathcal{G}}(B) \neq \emptyset$. When $\mathcal{UP}_{\mathcal{G}}(B) = \emptyset$, then trivially $\tilde{\mathcal{G}}(B) \subseteq B\{1\} = \mathcal{G}^*(B)$. Second, let B be not maximal and suppose $\mathcal{UP}_{\tilde{\mathcal{G}}}(B) = \{B\}$. Then $\mathcal{UP}_{\mathcal{G}}(B) = \emptyset$, so that $\mathcal{G}^*(B) = B\{1\}$. Since $\tilde{\mathcal{G}}(B) \subseteq B\{1\}$, again $\tilde{\mathcal{G}}(B) \subseteq \mathcal{G}^*(B)$. Third, let B be maximal relative to $\preceq^{\tilde{\mathcal{G}}}$. Since $\tilde{\mathcal{G}}(B) \neq \emptyset$, B is nonsingular. Consequently, $\tilde{\mathcal{G}}(B) = \{B^{-1}\}$. Recall that in the beginning of this section we saw that modifying $\mathcal{G}(C)$ in case of a nonsingular matrix C has no effect on the induced relation; to avoid unnecessary considerations we thus agreed to assume that $\mathcal{G}(C) = \{C^{-1}\}$ for each nonsingular matrix C. Hence $\mathcal{G}^*(B) = \{B^{-1}\}$, so that trivially $\tilde{\mathcal{G}}(B) = \mathcal{G}^*(B)$.

To complete the proof we finally must consider the case when $\tilde{\mathcal{G}}(B) = \emptyset$. Needless to say, the inclusion $\tilde{\mathcal{G}}(B) \subseteq \mathcal{G}^*(B)$ holds trivially in this case. □

This theorem admits the following interesting corollary; its proof is straightforward and thus is omitted. Observe that Example 5.1 is in accordance with this corollary.

COROLLARY 5.4. *For square $n \times n$ matrices, let $\preceq^{\mathcal{G}}$ denote a $\mathcal{G}$-based semicomplete partial order excluding from its support at least one singular matrix. Let $\mathcal{G}^*$ be defined by* (5.2). *For a $\mathcal{G}$-based semicomplete partial order extension of $\preceq^{\mathcal{G}}$ which supports all $n \times n$ matrices to exist it is then necessary that $\mathcal{G}^*(A) \neq \emptyset$ for each singular matrix A.*

Theorem 5.3 now enables us to prove the following preannounced result regarding the sharp order.

THEOREM 5.5. *Let $\mathcal{G}^*$ be defined according to* (5.2) *in respect to the usual semicomplete map $\mathcal{G}$ which belongs to the $\mathcal{G}$-based sharp order $\preceq^{\#}$. Then $\mathcal{G}^*$ induces a semicomplete partial order extension of $\preceq^{\#}$ that supports the set of all square matrices. Moreover, $\mathcal{G}^*$ is equal to the map $\mathcal{G}_*$ which was introduced in Theorem* 4.2.

Proof. Let $\mathcal{G}_*$ denote the map introduced in Theorem 4.2 by (4.4). Let A be any square matrix, and let $A = C_A + N_A$ again be its core-nilpotent decomposition. Then $C_A \preceq^{\#} A$ since $C_A^{\#} N_A = N_A C_A^{\#} = 0$. Consequently, $C_A \in \mathcal{UP}_{\mathcal{G}}(A)$, which in turn implies $\mathcal{G}^*(A) \subseteq \mathcal{G}_*(A)$. To prove the converse inclusion, note that, by Theorem 4.2, $\preceq^{\mathcal{G}_*}$ is semicomplete. Hence, in view of Theorem 5.3, $\mathcal{G}_*(A) \subseteq \mathcal{G}^*(A)$. Combining observations results in $\mathcal{G}^*(A) = \mathcal{G}_*(A)$, and our claims follow from Theorem 4.2. □

In context with Theorems 5.5, 5.3, and 4.2 it is pertinent to mention the following. Consider the maps $\mathcal{G}^0$ and $\mathcal{G}_*$ defined by (4.2) and (4.4), respectively. From Theorem 4.2 it is known that if $n \geq 3$ then $\mathcal{G}_*$ is properly finer than $\mathcal{G}^0$. Theorem 5.5 now tells us that $\mathcal{G}_* = \mathcal{G}^*$. Since $\preceq^{\mathcal{G}^0}$ is equivalent to the partial order $\preceq^{\dagger}$ (see §4), this seems to contradict Theorem 5.3. Fortunately, however, from the lines directly following (4.2) we already know that $\mathcal{G}^0$ fails to be a semicomplete map whenever $n \geq 3$.

For what follows it is convenient to call $\mathcal{G}$ as defined by (1.1) a *property-p* map if $\mathcal{G}$ is semicomplete and condition (3.4) of §3 is satisfied. If $\mathcal{G}$ is a property-p map, the induced relation $\preceq^{\mathcal{G}}$ as defined by (1.2) is called a *property-p* relation. Since a name should give some aid in visualizing the notion, it is natural to ask the following: Where does the name *"property-p"* come from? The answer is (at least implicitly) already

given by Theorem 3.1. For notice that, according to this theorem, each property-p relation is a *p*artial order. In other words, to possess this property "p" is a sufficient condition for relation $\preceq^{\mathcal{G}}$ to define a *p*artial order. In the case of square matrices (i.e., when $m = n$), we even know from Theorem 3.2 that a semicomplete $\mathcal{G}$-based relation $\preceq^{\mathcal{G}}$ corresponds to a partial order iff it is a property-p relation.

Motivated by Theorem 5.5 one might conjecture that, even in the nonsquare case (i.e., when $m \neq n$), the relation $\preceq^{\mathcal{G}^*}$ does always correspond to a partial order, provided $\mathcal{G}$ is a property-p map. The next part of this paper is devoted to establishing that this is indeed the case. For that purpose we need the following lemma.

LEMMA 5.6. *Let $\mathcal{G}$ be a property-p map, and let $\mathcal{G}^*$ be defined in respect to $\mathcal{G}$ according to* (5.2). *If $A \preceq^{\mathcal{G}} B$ and $B \preceq^{\mathcal{G}^*} C$, then $A \preceq^{\mathcal{G}} C$.*

Proof. If $A = B$ and/or $B = C$, then trivially $A \preceq^{\mathcal{G}} C$. Henceforth, let $A \neq B$ and let $B \neq C$. Then $A \in \mathcal{UP}_{\mathcal{G}}(B)$ and $\mathcal{G}^*(B) \neq \emptyset$. Since $\mathcal{G}$ is a property-p map, by Theorem 3.1 clearly $A \preceq^{\mathcal{G}} C$ whenever $B \in \Omega_{\mathcal{G}}$. Now let $B \notin \Omega_{\mathcal{G}}$. Then $\mathcal{G}^*(B) = \mathcal{G}_0(B)$. Since $B \preceq^{\mathcal{G}^*} C$, $(C - B)B^- = 0$ and $B^-(C - B) = 0$ for some suitable $B^- \in \mathcal{G}_0(B)$. This in turn implies

$$(C - B)B^-AB^- = 0, \tag{5.7a}$$

$$B^-AB^-(C - B) = 0. \tag{5.7b}$$

By the definition of $\mathcal{G}_0(B)$, $B^-AB^- \in \mathcal{G}(A)$. That $B\{1\} \subseteq A\{1\}$, $\mathcal{R}(A) \subseteq \mathcal{R}(B)$, and $\mathcal{N}(B) \subseteq \mathcal{N}(A)$ follows from $A \preceq^{\mathcal{G}} B$ by means of Theorem 2.3. But then $BB^-AB^- = AB^-$ because BB^- is a projector onto $\mathcal{R}(B)$. Moreover, $AB^-AB^- = AB^-$. So $BB^-AB^- = AB^-AB^-$ or, equivalently,

$$(B - A)B^-AB^- = 0. \tag{5.8a}$$

Since B^-B is a projector along $\mathcal{N}(B)$, we likewise get $B^-AB^-B = B^-AB^-A$ or, equivalently,

$$B^-AB^-(B - A) = 0. \tag{5.8b}$$

Combining (5.7) with (5.8) yields

$$(C - A)B^-AB^- = 0, \qquad B^-AB^-(C - A) = 0.$$

Since $B^-AB^- \in \mathcal{G}(A)$, $A \preceq^{\mathcal{G}} C$, as claimed. □

THEOREM 5.7. *Let $\mathcal{G}$ be a property-p map, and let $\mathcal{G}^*$ be defined in respect to $\mathcal{G}$ according to* (5.2). *Then $\mathcal{G}^*$ is also a property-p map. Although the relation $\preceq^{\mathcal{G}^*}$ thus defines a partial order it need not necessarily support each matrix.*

Proof. That $\preceq^{\mathcal{G}^*}$ does not necessarily support each matrix follows from Example 5.1. That $\mathcal{G}^*$ is semicomplete is the result of Theorem 5.2. In view of Theorem 3.1 we therefore only have to prove that the implication

$$A \preceq^{\mathcal{G}^*} B,\ B \text{ not maximal} \implies \mathcal{G}^*(A \mid B) \subseteq \mathcal{G}^*(A)$$

holds true. So let us assume that $A \preceq^{\mathcal{G}^*} B$ and that B is not maximal relative to $\preceq^{\mathcal{G}^*}$. Recall that B is not maximal iff $B \in \Omega_{\mathcal{G}^*}$ and $\operatorname{rank}(B) < \min\{m, n\}$. We consider the following four exhaustive cases.

Case 1: $A \in \Omega_{\mathcal{G}}$, $B \in \Omega_{\mathcal{G}}$. This case is trivial because $\mathcal{G}$ is a property-p map.

Case 2: $A \in \Omega_{\mathcal{G}}$, $B \notin \Omega_{\mathcal{G}}$. Then $A \in \mathcal{UP}_{\mathcal{G}}(B)$, so that $\mathcal{G}^*(B) = \mathcal{G}_0(B)$. Since B is not maximal, $\mathcal{G}_0(B) \neq \emptyset$. Now let $B^- \in \mathcal{G}_0(B)$. By the definition of $\mathcal{G}_0(B)$,

$B^-AB^- \in \mathcal{G}(A)$. Since $A \in \Omega_{\mathcal{G}}$, $\mathcal{G}^*(A) = \mathcal{G}(A)$. Combining these observations results in $\mathcal{G}^*(A \mid B) \subseteq \mathcal{G}^*(A)$.

Case 3: $A \notin \Omega_{\mathcal{G}}$, $\mathcal{UP}_{\mathcal{G}}(A) = \emptyset$. Then $\mathcal{G}^*(A) = A\{1\}$. Since $A \preceq^{\mathcal{G}^*} B$, $B\{1\} \subseteq A\{1\}$ by Theorem 2.3(v). Then, in view of Theorem 2.1, $B^-AB^- \in A\{1,2\} \subseteq A\{1\}$ for each B^-. Observing that $\mathcal{G}^*(B) \subseteq B\{1\}$ thus yields $\mathcal{G}^*(A \mid B) \subseteq \mathcal{G}^*(A)$.

Case 4: $A \notin \Omega_{\mathcal{G}}$, $\mathcal{UP}_{\mathcal{G}}(A) \neq \emptyset$. Then $\mathcal{G}^*(A) = \mathcal{G}_0(A) \neq \emptyset$ because $A \in \Omega_{\mathcal{G}^*}$. So by Lemma 5.6, $\mathcal{UP}_{\mathcal{G}}(A) \subseteq \mathcal{UP}_{\mathcal{G}}(B)$. Therefore,

$$\mathcal{G}^*(B) = \begin{cases} \mathcal{G}(B) & \text{if } B \in \Omega_{\mathcal{G}}, \\ \mathcal{G}_0(B) & \text{otherwise.} \end{cases}$$

Since B is not maximal, necessarily $\mathcal{G}^*(B) \neq \emptyset$. By the definition of $\mathcal{G}_0$ and since $\mathcal{G}$ is a property-p map,

$$\mathcal{G}^*(C \mid B) \subseteq \mathcal{G}(C) \quad \text{for each} \quad C \in \mathcal{UP}_{\mathcal{G}}(A). \tag{5.9}$$

Recall that

$$\mathcal{G}_0(A) := \{A^- \mid A^-CA^- \in \mathcal{G}(C) \quad \text{for each} \quad C \in \mathcal{UP}_{\mathcal{G}}(A)\}.$$

Now let $C \in \mathcal{UP}_{\mathcal{G}}(A)$. Then, by (5.9), $B^-CB^- \in \mathcal{G}(C)$. Observe that $C \preceq^{\mathcal{G}} A \preceq^{\mathcal{G}^*} B$ implies, by Theorem 2.3, $B\{1\} \subseteq A\{1\}$, $\mathcal{R}(C) \subseteq \mathcal{R}(A)$, and $\mathcal{N}(A) \subseteq \mathcal{N}(C)$. Therefore, $B^-AB^- \in A\{1\}$ and $B^-AB^-CB^-AB^- = B^-CB^-$. So $(B^-AB^-)C(B^-AB^-) \in \mathcal{G}(C) = \mathcal{G}^*(C)$, which implies $B^-AB^- \in \mathcal{G}^*(A)$, by the definition of $\mathcal{G}^*(A)$. Hence again $\mathcal{G}^*(A \mid B) \subseteq \mathcal{G}^*(A)$, and the proof is complete. □

Theorem 5.3 also admits a version that includes the possibly nonsquare case. Since the proof is nearly identical, it is omitted.

THEOREM 5.8. *Let $\mathcal{G}$ and $\tilde{\mathcal{G}}$ be property-p maps on the set of $m \times n$ matrices. Further, let $\mathcal{G}^*$ be defined in respect to $\mathcal{G}$ according to* (5.2). *If $\preceq^{\tilde{\mathcal{G}}}$ is a $\mathcal{G}$-based extension of $\preceq^{\mathcal{G}}$, then $\preceq^{\tilde{\mathcal{G}}}$ is finer than $\preceq^{\mathcal{G}^*}$.*

In other words, if $\mathcal{G}$ is a property-p map then $\preceq^{\mathcal{G}^*}$ as defined via (5.2) represents the *maximal* possible $\mathcal{G}$-based extension of $\preceq^{\mathcal{G}}$ in the set of all property-p relations. This shows, in particular, that if $\preceq^{\mathcal{G}^*}$ does not support each matrix, then it is impossible to find a property-p relation that is an extension of $\preceq^{\mathcal{G}}$ and supports the whole set of matrices (recall the convention regarding the full-rank matrices). Example 3.4 has shown that in the nonsquare case there are semicomplete $\mathcal{G}$-based partial orders which fail to be property-p relations. The question of how to obtain a maximal $\mathcal{G}$-based partial order extension in such a case remains unanswered in this paper. It is expected, however, that the geometry of g-inversion (see §2) might be helpful in finding an answer.

Our next theorem will provide us with a sufficient condition on $\preceq^{\mathcal{G}}$ under which its maximal possible $\mathcal{G}$-based extension $\preceq^{\mathcal{G}^*}$ supports each matrix. Let us call D a *maximal element* of $\mathcal{UP}_{\mathcal{G}}(A)$ (in respect to $\preceq^{\mathcal{G}}$) if $D \in \mathcal{UP}_{\mathcal{G}}(A)$ and there is no matrix $C \in \mathcal{UP}_{\mathcal{G}}(A)$ such that $C \neq D$ and $D \preceq^{\mathcal{G}} C$. If $\mathcal{UP}_{\mathcal{G}}(A)$ possesses a unique maximal element C, C is called the *greatest element* of $\mathcal{UP}_{\mathcal{G}}(A)$.

THEOREM 5.9. *Let $\mathcal{G}$ be a property-p map such that, for each matrix A, $\mathcal{UP}_{\mathcal{G}}(A)$ possesses a greatest element whenever $\mathcal{UP}_{\mathcal{G}}(A) \neq \emptyset$. Again, let $\mathcal{G}^*$ be defined in respect to $\mathcal{G}$ according to* (5.2). *The property-p relation $\preceq^{\mathcal{G}^*}$ then supports each matrix and is the maximal possible (partial order) extension of $\preceq^{\mathcal{G}}$ in the class of property-p relations.*

Proof. Recalling Theorem 5.7, we have only to show that, in the framework of our theorem, $\mathcal{G}^*$ supports each matrix. By the definition of $\mathcal{G}^*$, clearly $\mathcal{G}^*(A) = \emptyset$ only

if $A \notin \Omega_{\mathcal{G}}$ is such that $\mathcal{UP}_{\mathcal{G}}(A) \neq \emptyset$. Let $A \notin \Omega_{\mathcal{G}}$, and let $\mathcal{UP}_{\mathcal{G}}(A) \neq \emptyset$. Notice that $\mathcal{UP}_{\mathcal{G}}(A)$ has a greatest element, say D. So

$$C \preceq^{\mathcal{G}} D \preceq^{\mathcal{G}} A \tag{5.10}$$

for each $C \in \mathcal{UP}_{\mathcal{G}}(A)$. In particular, $D \preceq^{\mathcal{G}} A$. Hence $D \preceq^{-} A$ which, in view of Theorem 2.3, also implies $(A-D) \preceq^{-} A$. But then $D_r^-(A-D) = 0$ and $(A-D)D_r^- = 0$ for some $D_r^- \in \mathcal{G}(D)$ as well as $(A-D)_r^- D = 0$ and $D(A-D)_r^- = 0$ for some $\{1,2\}$-inverse $(A-D)_r^-$ of $(A-D)$. With these observations in mind it is easy to check that $G := D_r^- + (A-D)_r^-$ is a $\{1,2\}$-inverse of A and that $GDG = D_r^-$. Put

$$\tilde{\mathcal{G}}(A) := \{A^- \mid A^-DA^- \in \mathcal{G}(D)\}.$$

Since $G \in \tilde{\mathcal{G}}(A)$, $\tilde{\mathcal{G}}(A) \neq \emptyset$. In order to complete the proof it thus suffices to show that $\tilde{\mathcal{G}}(A) = \mathcal{G}^*(A)$. Trivially, $\tilde{\mathcal{G}}(A) \supseteq \mathcal{G}^*(A)$. So let $A^- \in \tilde{\mathcal{G}}(A)$ and let $C \preceq^{\mathcal{G}} A$. Notice that (5.10) implies $A\{1\} \subseteq D\{1\}$, $\mathcal{R}(C) \subseteq \mathcal{R}(D)$, and $\mathcal{N}(D) \subseteq \mathcal{N}(C)$ (recall Theorem 2.3). But then $A^-CA^- = A^-DA^-CA^-DA^-$. By observing $A^-DA^- \in \mathcal{G}(D)$, and since $\mathcal{G}(C \mid D) \subseteq \mathcal{G}(C)$, we now obtain $A^-CA^- \in \mathcal{G}(C)$, yielding that $A^- \in \mathcal{G}^*(A)$. □

At this point it is interesting to mention that Theorem 5.5 is in accordance with Theorem 5.9. For observe that, according to Lemma 2.1 in [7], $B \preceq^{\#} A$ implies $B \preceq^{\#} C_A$, so that the core part C_A of A is the *greatest* predecessor of A in respect to the sharp order. Theorem 5.5 tells us that $\mathcal{G}^*(A)$ can be defined, equivalently, by $\mathcal{G}_*(A)$ from (4.4), that is, in terms of the greatest (and so uniquely determined maximal) predecessor of A. Since this might be advantageous computationally it is pertinent to mention that for each property-p map $\mathcal{G}$ the crucial part $\mathcal{G}_0(B)$ in the definition of $\mathcal{G}^*(B)$ (see (5.1)) can always be redefined in a similar manner as

$$\mathcal{G}_0(B) := \{B^- \mid B^-CB^- \in \mathcal{G}(C) \quad \text{for each maximal element } C \text{ from } \mathcal{UP}_{\mathcal{G}}(B)\}. \tag{5.11}$$

The equivalence of these definitions can be seen basically as the last part in the proof of Theorem 5.9.

In context with the previous theorem it is further natural to ask the following: Is the phenomenon observed in Example 5.1 universally true whenever there is multiplicity of *maximal* elements for at least one set $\mathcal{UP}_{\mathcal{G}}(A)$? In other words, does failure of uniqueness always correspond to a situation where along with $\preceq^{\mathcal{G}}$ all its $\mathcal{G}$-based property-p extensions also have poor support? That this is not the case is illustrated in our final example.

Example 5.10. Consider the matrices

$$A_1 := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad A_2 := \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Put $B := A_1 + A_2$. Observe that these matrices are all of index 1. Check $A_1^{\#} = A_1$ and $A_2^{\#} = A_2$. Define the map $\mathcal{G}(\cdot)$ on the set of 3×3 matrices according to

$$\mathcal{G}(A) = \begin{cases} \{A_1\} & \text{if } A = A_1, \\ \{A_2\} & \text{if } A = A_2, \\ \emptyset & \text{otherwise.} \end{cases}$$

Notice that $\preceq^{\mathcal{G}}$ defines a partial ordering. Checking $A_i \preceq^{\mathcal{G}} B$ $(i = 1, 2)$ yields $\mathcal{UP}_{\mathcal{G}}(B) = \{A_1, A_2\}$. The sharp order $\preceq^{\#}$ is obviously a possible partial order extension of $\preceq^{\mathcal{G}}$. It thus follows from Theorem 4.2 that there is a partial order extension of $\preceq^{\mathcal{G}}$ that supports each matrix, although $\mathcal{UP}_{\mathcal{G}}(B)$ does not possess a greatest element. □

Acknowledgments. The authors would like to thank three referees for their comments on this paper.

REFERENCES

[1] A. BEN-ISRAEL AND T. N. E. GREVILLE, *Generalized Inverses: Theory and Applications*, Wiley, New York, 1974; Krieger, New York, 1980.

[2] J. S. CHIPMAN, *On least squares with insufficient observations*, J. Amer. Statist. Assoc., 59 (1964), pp. 1078–1111.

[3] J. HARTUNG AND H. J. WERNER, *Hypothesenprüfung im Restringierten Linearen Modell—Theorie und Anwendungen*, Vandenhoeck & Ruprecht, Göttingen, 1984.

[4] C. E. LANGENHOP, *On generalized inverses of matrices*, SIAM J. Appl. Math., 15 (1967), pp. 1239–1246.

[5] S. K. MITRA, *Fixed rank solutions of linear matrix equations*, Sankhyā Ser. A, 34 (1972), pp. 387–392.

[6] ———, *Matrix partial orders through generalized inverses: Unified theory*, Linear Algebra Appl., 148 (1991), pp. 237–263.

[7] ———, *On $\mathcal{G}$-based extensions of the sharp order*, Linear and Multilinear Algebra, 31 (1992), pp. 147–151.

[8] S. K. MITRA AND R. E. HARTWIG, *Partial orders based on outer inverses*, Linear Algebra Appl., 176 (1992), pp. 3–20.

[9] R. M. PRINGLE AND A. A. RAYNER, *Generalized Inverse Matrices with Applications to Statistics*, Griffin, London, UK, 1971.

[10] C. R. RAO AND S. K. MITRA, *Generalized Inverse of Matrices and its Applications*, Wiley, New York, 1971.

[11] C. R. RAO AND H. YANAI, *Generalized inverse of linear transformations: A geometric approach*, Linear Algebra Appl., 66 (1985), pp. 87–98.

[12] H. J. WERNER, *G-Inverse und monotone Matrizen*, Dissertation, University of Bonn, Bonn, Germany, 1977.

[13] ———, *On extensions of Cramer's rule for solutions of restricted linear systems*, Linear and Multilinear Algebra, 15 (1984), pp. 319–330.

[14] ———, *Weak complementarity and missing observations*, Methods Oper. Res., 50 (1985), pp. 421–439.

[15] ———, *Generalized inversion and weak bi-complementarity*, Linear and Multilinear Algebra, 19 (1986), pp. 357–372.

[16] ———, *Some recent results on Drazin-monotonicity of property-n matrices*, Linear and Multilinear Algebra, 21 (1987), pp. 243–251.

[17] ———, *C. R. Rao's IPM method: A geometric approach*, in New Perspectives in Theoretical and Applied Statistics, M. L. Puri, J. P. Vilaplana, and W. Wertz, eds., John Wiley, New York, 1987, pp. 367–382.

[18] ———, *A closed form formula for the intersection of two complex matrices under the star order*, Linear Algebra Appl., 140 (1990), pp. 13–30.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 4, pp. 851–868, October 1996

009

SOME NONINTERIOR CONTINUATION METHODS FOR LINEAR COMPLEMENTARITY PROBLEMS*

CHRISTIAN KANZOW†

Abstract. We introduce some new path-following methods for the solution of the linear complementarity problem. We call these methods noninterior continuation methods since, in contrast to interior-point methods, not all iterates have to stay in the positive orthant. This is possible since we reformulate certain perturbed complementarity problems as a nonlinear system of equations. However, similar to interior-point methods, we also try to follow the central path. We present some conditions which guarantee the existence of this central path, prove a global convergence result for some implementable noninterior continuation methods, and report some numerical results obtained with these methods. We also prove global error bound results for the perturbed linear complementarity problems.

Key words. linear complementarity problems, path-following methods, interior-point methods, global error bounds, P_0-matrix, R_0-matrix

AMS subject classification. 90C33

1. Introduction.

The linear complementarity problem, denoted by LCP(q, M), is to find a vector $z = (x, y) \in \Re^{2n}$ such that

$$x \geq 0, y \geq 0, x^T y = 0, y = Mx + q,$$

where $M \in \Re^{n \times n}$ is a given matrix and $q \in \Re^n$ is a given vector. This problem serves as a unified formulation of linear and quadratic programming problems as well as of two-person (noncooperative) matrix-games and has several important applications in economics and engineering sciences; see Cottle, Pang, and Stone [3], Harker [15], and Isac [17] for some examples.

There exist several methods for solving LCP(q, M); the interested reader is referred to the excellent books of Murty [31] and Cottle, Pang, and Stone [3]. Here we focus on the interior-point approach. This approach solves (approximately) a sequence of certain perturbed linear complementarity problems, PLCP(q, M, μ) for short; these perturbed problems depend on a positive parameter $\mu > 0$ and consist of finding a vector $z(\mu) = (x(\mu), y(\mu)) \in \Re^{2n}$ satisfying the conditions

$$x > 0, y > 0, x_i y_i = \mu \ (i \in I), y = Mx + q.$$

Here and throughout the paper, the index set I is an abbreviation for the set $\{1, \ldots, n\}$. Under certain conditions, the PLCP(q, M, μ) are uniquely solvable for all $\mu > 0$, and the pair $(\mu, z(\mu))$ forms a smooth trajectory, usually called the *central path*. By tracing this trajectory as μ tends to zero, one hopes to find a solution of the original LCP(q, M). Actually, this can be shown under suitable assumptions; see, e.g., Meggido [30], Kojima, Meggido, and Noma [23, 25, 26], and Kojima et al. [24].

The aim of this paper is to introduce some new tools to allow us to reformulate the PLCP(q, M, μ) as a nonlinear system of equations and to show that these tools can be used both for the theoretical analysis of a continuation method and for the

* Received by the editors August 22, 1994; accepted for publication (in revised form) by R. Cottle November 13, 1995. This paper combines results of the two earlier reports [20, 21] by the author. A few results from [20, 21] have been removed; on the other hand, some new results have been included, mainly in §§5 and 6 of this paper.

† Institute of Applied Mathematics, University of Hamburg, Bundesstrasse 55, D-20146 Hamburg, Germany (kanzow@math.uni-hamburg.de).

numerical solution of LCP(q, M). These tools are defined in §2. In §3, we prove some results concerning the existence and uniqueness of the solution $z(\mu)$ of PLCP(q, M, μ). We also compare our approach with a recent method of Chen and Harker [2], which is very similar to our approach; actually, we show in §4 that their method can be viewed as a scaled variant of one of our methods. Some global error bound results are given in §5. In §6 we describe in detail the implemented algorithm and present a global convergence result for this method. The suggested algorithm differs from interior-point methods in at least two points: on the one hand, the iterates do not necessarily have to stay in the positive orthant, and on the other hand we can start with an arbitrary vector $(x^0, y^0) \in \Re^{2n}$ and an arbitrary initial parameter $\mu_0 > 0$. Some promising numerical results are given in §7. We conclude with some final remarks in §8.

Notation. For a vector $x = (x_1, \ldots, x_n)^T \in \Re^n$, the inequalities $x \geq 0$ and $x > 0$ are defined componentwise. If $x, y \in \Re^n$, the vector $z := (x^T, y^T)^T \in \Re^{2n}$ is usually abbreviated by $z = (x, y)$. Given two vectors $x, y \in \Re^n$, the symbol $\min\{x, y\}$ denotes the n-vector having $\min\{x_i, y_i\}$ as its ith component. All norms are Euclidean norms. By $dist(x, \mathcal{S})$ we denote the (Euclidean) distance of a vector $x \in \Re^n$ to a set $\mathcal{S} \subseteq \Re^n$; i.e., $dist(x, \mathcal{S}) = \inf_{y \in \mathcal{S}} \|x - y\|$.

2. Main tools. The noninterior continuation methods to be presented in the following section are based on reformulations of the PLCP as nonlinear systems of equations. The main tools used in these reformulations are functions $\varphi_\mu : \Re^2 \to \Re$ having the property; cf. [20].

$$\varphi_\mu(a, b) = 0 \iff a > 0, b > 0, ab = \mu, \tag{1}$$

where, unless otherwise stated, μ is any fixed positive parameter. In this section, we introduce some functions φ_μ having this property; cf. [20].

LEMMA 2.1. *The function*

$$\varphi_\mu(a, b) := a + b - \sqrt{(a - b)^2 + 4\mu}$$

has the property (1).

Proof. First assume that $a > 0, b > 0$, and $ab = \mu$. Then, we obtain

$$\begin{aligned} \varphi_\mu(a, b) &= a + b - \sqrt{a^2 - 2ab + b^2 + 4ab} \\ &= a + b - \sqrt{(a + b)^2} \\ &= a + b - |a + b| \\ &= 0. \end{aligned}$$

To prove the converse result, assume that $\varphi_\mu(a, b) = 0$; i.e.,

$$a + b = \sqrt{(a - b)^2 + 4\mu} > 0. \tag{2}$$

Squaring both sides of the equation in (2), we get $ab = \mu$. Therefore $\text{sign}(a) = \text{sign}(b)$. Consequently it follows from the inequality in (2) that $a > 0$ and $b > 0$. □

LEMMA 2.2. *The function*

$$\varphi_\mu(a, b) := a + b - \sqrt{a^2 + b^2 + 2\mu}$$

has the property (1).

Proof. If $a > 0, b > 0, ab = \mu$, we get

$$\begin{aligned}\varphi_\mu(a,b) &= a + b - \sqrt{(a+b)^2} \\ &= a + b - |a+b| \\ &= 0.\end{aligned}$$

On the other hand, the condition $\varphi_\mu(a,b) = 0$ can be rewritten as

$$a + b = \sqrt{a^2 + b^2 + 2\mu} > 0,$$

from which $a > 0, b > 0$, and $ab = \mu$ follow in a similar way as in the proof of Lemma 2.1. □

Remark. For the special case $\mu = 0$, the function introduced in Lemma 2.1 reduces to the min function used, e.g., by Pang [32, 33] and Harker and Pang [16], in order to characterize the complementarity problem itself as a (nonsmooth) system of equations. On the other hand, the function φ_μ defined in Lemma 2.2 coincides for $\mu = 0$ with a recently introduced function of Fischer [8] which has subsequently been used by several authors, including Fischer [9, 10, 11], Qi and Jiang [36], Facchinei and Soares [5, 6], Tseng [40], and Kanzow [19, 21].

Remark. Let φ_μ denote the function defined in Lemma 2.1 or 2.2. Then φ_μ is continuously differentiable for all $(a,b) \in \Re^2$, and it is not difficult to see that the partial derivatives have the property

$$\frac{\partial \varphi_\mu}{\partial a}(a,b), \frac{\partial \varphi_\mu}{\partial b}(a,b) \in (0,2)$$

for all $a, b \in \Re$. This property turns out to be important in the following section; see Theorem 3.5.

Remark. We have found two other functions having the property (1); namely,

$$\varphi_\mu(a,b) := \frac{1}{2}\min^2\{0, a+b\} - ab + \mu, \tag{3}$$

$$\varphi_\mu(a,b) := (a-b)^2 - a|a| - b|b| + 2\mu. \tag{4}$$

It is straightforward to see that these functions satisfy condition (1). However, they have at least two main disadvantages when comparing them with the functions of Lemmas 2.1 and 2.2. On the one hand, the functions (3) and (4) are more nonlinear than their counterparts of Lemmas 2.1 and 2.2; on the other hand, they do not have the property mentioned in the previous remark. We note that the functions (3) and (4) also reduce to some known functions for $\mu = 0$; see, e.g., Mangasarian [28], Kanzow [18, 19], and Kanzow and Kleinmichel [22].

There might exist several other functions φ_μ having the property (1). It is our feeling, however, that the two functions given in Lemmas 2.1 and 2.2 are the most interesting ones, both from a theoretical and a numerical point of view.

3. Continuation methods. Let $\mu > 0$. Throughout this section, let $\varphi_\mu : \Re^2 \to \Re$ denote the function given in Lemma 2.1 or 2.2. If we introduce the nonlinear operator $F_{\varphi_\mu} : \Re^{2n} \to \Re^{2n}$ defined by

$$F_{\varphi_\mu}(z) := F_{\varphi_\mu}(x,y) := \begin{pmatrix} Mx + q - y \\ \varphi_\mu(x,y) \end{pmatrix},$$

where

$$\varphi_\mu(x,y) := (\varphi_\mu(x_1,y_1), \ldots, \varphi_\mu(x_n,y_n))^T \in \Re^n,$$

we obtain as a direct consequence of Lemmas 2.1 and 2.2 the following characterization of PLCP(q, M, μ).

THEOREM 3.1. *A vector $z(\mu) := (x(\mu), y(\mu)) \in \Re^{2n}$ is a solution of PLCP(q, M, μ) if and only if $z(\mu)$ satisfies the nonlinear equation $F_{\varphi_\mu}(z) = 0$.*

We now investigate the following (theoretical) continuation method. A variant of this algorithm that can be implemented is described in §6.

ALGORITHM 3.2.

(S.0): Choose $z^0 := (x^0, y^0) \in \Re^{2n}$, set $k := 0$, and let $\{\mu_k\} \subseteq \Re$ be a strictly decreasing sequence with $\lim_{k\to\infty} \mu_k = 0$.

(S.1): Terminate if $z^k = (x^k, y^k)$ solves LCP(q, M).

(S.2): Find a solution $z^{k+1} := z(\mu_{k+1})$ of the nonlinear system of equations

$$F_{\varphi_{\mu_{k+1}}}(z) = 0$$

(or, equivalently, of PLCP(q, M, μ_{k+1})).

(S.3): Set $k := k + 1$ and go to (S.1).

In order for Algorithm 3.2 to be well defined, one has to guarantee that the perturbed problems have a (unique) solution z^{k+1}. A well-known condition which guarantees the existence and uniqueness of such a solution is that M is positive semidefinite and that there exists a strictly feasible vector $\hat{x}$ (i.e., $\hat{x} > 0$ and $M\hat{x} + q > 0$); see, e.g., Meggido [30] and Kojima, Mizuno, and Noma [26]. These assumptions have been weakened by Kojima et al. [24, Condition 2.1], where the following conditions are assumed to hold:

CONDITION KMNY.

(a) M is a P_0-matrix (see below).

(b) There exists a strictly feasible vector for LCP(q, M).

(c) The level sets $\mathcal{L}_t(q, M) := \{(x, y) \in \Re^n \mid x \geq 0, y \geq 0, y = Mx + q, x^T y \leq t\}$ are bounded for all $t \geq 0$.

Here we show that Algorithm 3.2 is well defined if M is a P_0- and R_0-matrix. (At the end of this section, we show how this assumption is related to Condition KMNY.) We first restate the definitions of these matrix classes as well as the definition of P-matrices.

DEFINITION 3.3. *A matrix $M \in \Re^{n\times n}$ is said to be a*

(a) *P_0-matrix* $:\Longleftrightarrow \forall x \in \Re^n, x \neq 0, \exists i \in I : x_i \neq 0,$ *and* $x_i[Mx]_i \geq 0$;

(b) *P-matrix* $:\Longleftrightarrow \forall x \in \Re^n, x \neq 0, \exists i \in I : x_i \neq 0,$ *and* $x_i[Mx]_i > 0$;

(c) *R_0-matrix* $:\Longleftrightarrow$ *LCP$(0, M)$ has $z^* := (0, 0) \in \Re^{2n}$ as its unique solution.*

It is not difficult to see that every P-matrix is both a P_0-matrix and an R_0-matrix. Moreover, it is well known that positive definite matrices are P-matrices and positive semidefinite matrices are P_0-matrices; see, e.g., Murty [31].

Our analysis of Algorithm 3.2 is similar to the one of Chen and Harker [2]. We stress, however, that we do not have to assume that all diagonal elements of M are positive. This contrasts favourably with Chen and Harker's method and is, in fact, the main advantage of our approach. We note that this positiveness assumption is often not satisfied; e.g., if we formulate a (linear or) quadratic programming problem as an LCP, the corresponding matrix M has zero entries on its diagonal (in case of a linear programming problem, all diagonal entries are zero).

Now we come to the convergence analysis of Algorithm 3.2. Consider the set of paths

$$\mathcal{P}(\bar{\mu}) := \{(\mu, z(\mu)) \mid 0 < \mu \leq \bar{\mu}, z(\mu) \text{ solves PLCP}(q, M, \mu)\},$$

where $\bar{\mu} > 0$ is some given constant. The following theorem is due to Chen and Harker [2, Thm. 3.2] and follows from the implicit function theorem and simple continuity arguments.

THEOREM 3.4. *If the Jacobian matrix $\nabla F_{\varphi_\mu}(z(\mu))$ is nonsingular for all $z(\mu), 0 < \mu \leq \bar{\mu}$, then $\mathcal{P}(\bar{\mu})$ consists solely of continuously differentiable paths. If, in addition, $\mathcal{P}(\bar{\mu})$ is bounded, then every limit point of $z(\mu)$ as μ approaches zero is a solution of LCP(q, M).*

We next show that the Jacobian matrices $\nabla F_{\varphi_\mu}(z)$ are nonsingular for P_0-matrix LCPs.

THEOREM 3.5. *Assume that $M \in \Re^{n\times n}$ is a P_0-matrix. Then the Jacobian matrix $\nabla F_{\varphi_\mu}(z)$ is nonsingular for all $z \in \Re^{2n}$ and all $\mu > 0$.*

Proof. Let $z = (x, y) \in \Re^{2n}$ and $\mu > 0$ be fixed. Then, using the notation

$$D_a := D_a(z) := \operatorname{diag}\left(\ldots, \frac{\partial \varphi_\mu}{\partial a}(x_i, y_i), \ldots\right),$$

$$D_b := D_b(z) := \operatorname{diag}\left(\ldots, \frac{\partial \varphi_\mu}{\partial b}(x_i, y_i), \ldots\right),$$

the Jacobian matrix is given by

$$\nabla F_{\varphi_\mu}(z) = \nabla F_{\varphi_\mu}(x, y) = \begin{pmatrix} M & -I \\ D_a & D_b \end{pmatrix}. \tag{5}$$

Let $\nabla F_{\varphi_\mu}(z)p = 0$ with $p = \left(p^{(1)}, p^{(2)}\right), p^{(i)} \in \Re^n, i = 1, 2$. Then

$$Mp^{(1)} - p^{(2)} = 0, \tag{6}$$

$$D_a p^{(1)} + D_b p^{(2)} = 0. \tag{7}$$

From the second remark in §2, we have that the diagonal matrices D_a and D_b are positive definite. Thus, (7) can be rewritten as

$$p^{(2)} = -D_b^{-1} D_a p^{(1)}, \tag{8}$$

where $D_b^{-1} D_a$ is also a positive definite diagonal matrix. Inserting (8) into (6) yields

$$(M + D_b^{-1} D_a) p^{(1)} = 0. \tag{9}$$

Since M is a P_0-matrix and $D_b^{-1} D_a$ is positive definite, we directly obtain from Definition 3.3 that the matrix $M + D_b^{-1} D_a$ is a P-matrix. In particular, this matrix is nonsingular. Consequently, we have

$$p^{(1)} = 0.$$

Thus, we get

$$p^{(2)} = 0$$

from (8), which proves the theorem. □

The following theorem can be shown in a similar way as the corresponding theorem of Chen and Harker [2, Thm. 3.7].

THEOREM 3.6. *Let $M \in \Re^{n\times n}$ be a P_0-matrix, let $z^0 \in \Re^{2n}$, and assume that the level sets*

$$\mathcal{L}(\mu) := \{z \in \Re^{2n} \mid \|F_{\varphi_\mu}(z)\| \leq \|F_{\varphi_\mu}(z^0)\|\}$$

are uniformly bounded for all $0 < \mu \leq \bar{\mu}$ *for some* $\bar{\mu} > 0$. *Then the PLCP*(q, M, μ) *have a unique solution* $z(\mu)$ *for all* $0 < \mu \leq \bar{\mu}$.

The following is a (necessary and) sufficient condition for the above level sets to be bounded.

THEOREM 3.7. *Assume that* $M \in \Re^{n \times n}$ *is an* R_0*-matrix,* $z^0 \in \Re^{2n}$, *and* $\bar{\mu} > 0$. *Then the level sets* $\mathcal{L}(\mu)$ *as defined in Theorem* 3.6 *are uniformly bounded for all* $0 \leq \mu \leq \bar{\mu}$.

Proof. Assume there exists an unbounded sequence $\{z^k\} = \{(x^k, y^k)\} \subseteq \Re^{2n}$ such that $z^k \in \mathcal{L}(\mu_k)$ where $\{\mu_k\}$ is a sequence with $0 < \mu_k \leq \bar{\mu}$ for all k. Since the mapping $\mu \to F_{\varphi_\mu}(z^0)$ is continuous and $[0, \bar{\mu}]$ is a compact interval, the maximum $\alpha := \max_{\mu \in [0,\bar{\mu}]} \|F_{\varphi_\mu}(z^0)\|$ exists. We therefore have

$$\|Mx^k + q - y^k\| \leq \|F_{\varphi_{\mu_k}}(z^k)\| \leq \|F_{\varphi_{\mu_k}}(z^0)\| \leq \alpha.$$

Let $z^* := (x^*, y^*)$ be an accumulation point of the bounded sequence $\left\{\frac{(x^k, y^k)}{\|(x^k, y^k)\|}\right\}$. Then we obtain

$$0 \leq \frac{\|Mx^k + q - y^k\|}{\|(x^k, y^k)\|} \leq \frac{\alpha}{\|z^k\|} \to 0;$$

i.e., $z^* = (x^*, y^*)$ satisfies the equation

$$Mx^* = y^*. \tag{10}$$

Similarly, if φ_μ denotes one of the functions defined in Lemma 2.1 or 2.2, we get from the definition of these functions and the boundedness of the sequence $\{\mu_k\}$ that

$$\frac{\varphi_{\mu_k}(x_i^k, y_i^k)}{\|(x^k, y^k)\|} \to \varphi_0(x_i^*, y_i^*)$$

and, on the other hand,

$$\frac{|\varphi_{\mu_k}(x_i^k, y_i^k)|}{\|(x^k, y^k)\|} \leq \frac{\alpha}{\|z^k\|} \to 0,$$

so that

$$\varphi_0(x_i^*, y_i^*) = 0$$

holds for all $i \in I$. This is equivalent to $x^* \geq 0, y^* \geq 0$, and $(x^*)^T y^* = 0$. In view of (10) this means that z^* is a solution of LCP$(0, M)$. However, since $\|z^*\| = 1$ and therefore $z^* \neq 0$, we have a contradiction to the assumed R_0-property of the matrix M. □

Using similar arguments as in Fischer [9] and Tseng [40], one can show that even the single level set $\mathcal{L}(\mu)$ for $\mu = 0$ is bounded for all $q \in \Re^n$ if and only if M is an R_0-matrix. Hence, the assumption of M being an R_0-matrix is also necessary for Theorem 3.7 to be true.

We can summarize the above results as follows: If $M \in \Re^{n \times n}$ is a P_0- and R_0-matrix, then all PLCP(q, M, μ) have a unique solution, the sequence $\{z^k\}$ generated by Algorithm 3.2 is bounded, and every limit point of this sequence is a solution of LCP(q, M). In particular, we obtain again by our constructive approach the result of Aganagic and Cottle [1] that the LCP has a nonempty solution set if M is a P_0- and R_0-matrix.

We now turn back to the condition KMNY. Of course, the assumption of M being a P_0- and R_0-matrix is not directly related to condition KMNY since the former is independent of the specific vector q, whereas the latter depends on this vector. However, based on the previous results, we are now able to prove the following equivalence theorem.

THEOREM 3.8. *M is a P_0- and R_0-matrix if and only if condition KMNY is satisfied for all $q \in \Re^n$.*

Proof. First assume that M is a P_0- and R_0-matrix. Then condition KMNY (a) is obviously satisfied. Let $\mu > 0$. Since M is a P_0- and R_0-matrix, there exists a (unique) solution $z(\mu) = (x(\mu), y(\mu)) \in \Re^{2n}$ of the PLCP(q, M, μ). Obviously, $x(\mu)$ is then a strictly feasible vector for LCP(q, M); i.e., condition KMNY (b) is also satisfied. Finally, the validity of condition KMNY (c) follows from Proposition 3.9.23 in [3]. Conversely, assume that condition KMNY is satisfied for all $q \in \Re^n$. We only have to show that M is an R_0-matrix. Assume this is not true. Then there exists a nonzero solution $z = (x, y) \in \Re^{2n}$ of LCP$(0, M)$. Then the vector τz is also a solution of LCP$(0, M)$ for all $\tau \geq 0$. In particular, the level set $\mathcal{L}_t(0, M)$ is not bounded for any $t \geq 0$, which contradicts condition KMNY (c). □

We note that the main results of this section are now a simple consequence of Theorem 3.8 and of Theorem 4.4 in [24]. However, the proof of Theorem 3.8 (namely the existence of a strictly feasible vector) is based on the previous results of this section which will also play a central role in the following sections. On the other hand, Kojima et al. [24] were able to prove, under their condition KMNY, that the entire sequence $z(\mu)$ converges to a solution of LCP(q, M) as μ approaches zero. Hence we get the following corollary from Theorem 3.8 which improves on our Theorem 3.4.

COROLLARY 3.9. *If M is a P_0- and R_0-matrix, then the PLCP(q, M, μ) have a unique solution $z(\mu)$ for all $\mu > 0$, and the entire sequence $z(\mu)$ converges to a solution of LCP(q, M) as μ tends to zero.*

A similar result holds for Chen and Harker's method [2] which also improves on their Theorem 3.2.

4. The method of Chen and Harker. First note that it would also have been possible to use the tools of §2 to characterize the problem

$$x > 0, Mx + q > 0, x_i[Mx + q]_i = \mu \ (i \in I), \tag{11}$$

which is obviously equivalent to PLCP(q, M, μ). Using the functions φ_μ of Lemmas 2.1 and 2.2, respectively, we obtain the following characterizations of (11):

$$x_i + [Mx + q]_i - \sqrt{(x_i - [Mx + q]_i)^2 + 4\mu} = 0 \ (i \in I) \text{ and} \tag{12}$$

$$x_i + [Mx + q]_i - \sqrt{x_i^2 + [Mx + q]_i^2 + 2\mu} = 0 \ (i \in I). \tag{13}$$

On the other hand, Chen and Harker [2] give the following reformulation of (11):

$$m_{ii}x_i + [Mx + q]_i - \sqrt{(m_{ii}x_i - [Mx + q]_i)^2 + 4m_{ii}\mu} = 0 \ (i \in I). \tag{14}$$

This formulation is very similar to our characterization (12), but it has the disadvantage that all diagonal entries of M must be positive (otherwise (14) is not equivalent to (11)). If this is the case, however, it is also possible to reformulate problem (11) as follows:

$$m_{ii}x_i + [Mx + q]_i - \sqrt{(m_{ii}x_i)^2 + [Mx + q]_i^2 + 2m_{ii}\mu} = 0 \ (i \in I). \tag{15}$$

This formulation can be regarded as the counterpart to (13). Note that the characterizations (12) and (14) coincide if all diagonal elements of M are equal to one. Under this assumption, the reformulations (13) and (15) are also equivalent.

Now let $S \in \Re^{n\times n}$ be any positive definite diagonal matrix. Then

$$Sy \geq 0 \Longleftrightarrow y \geq 0 \tag{16}$$

holds for any vector $y \in \Re^n$. Hence the LCP(q, M) is equivalent to

$$x \geq 0, S(Mx+q) \geq 0, x^T(Mx+q) = 0, \tag{17}$$

and interior-point-like methods for problem (17) try to solve the corresponding perturbed problem

$$x > 0, S(Mx+q) > 0, x_i[Mx+q]_i = \mu \ \ (i \in I). \tag{18}$$

If M has positive diagonal entries and if we define $S := \operatorname{diag}(\ldots, 1/m_{ii}, \ldots)$, then the matrix $M_S := SM$ has unit entries on its diagonal. Consequently, Chen and Harker's characterization for this scaled problem is exactly our characterization (12) for this problem; i.e., Chen and Harker's method can be regarded as a scaled variant of one of our methods. Therefore, it is quite natural to apply our tools of §2 to the scaled problem (17)/(18). In §7, we consider three scaling matrices:

$$\begin{aligned} S_0 &:= I_n \qquad \text{(no scaling)}, \\ S_1 &:= \operatorname{diag}(s_1, \ldots, s_n), s_i := \begin{cases} 1/m_{ii} & \text{if } m_{ii} \neq 0, \\ 1 & \text{if } m_{ii} = 0, \end{cases} \\ S_2 &:= \operatorname{diag}(\sigma_1, \ldots, \sigma_n), \sigma_i := 1/\|M_{i.}\|, \end{aligned}$$

where $M_{i.}$ denotes the ith row of the matrix M. (Numerically the conditions $m_{ii} = 0$ and $m_{ii} \neq 0$ should, of course, be replaced by something like $|m_{ii}| < \tau$ and $|m_{ii}| \geq \tau$ for a small constant $\tau > 0$.)

We note, however, that the above scaling is somewhat strange and has its disadvantages; for example, we solve LCP(q_S, M_S) instead of LCP(q, M), where $q_S := Sq, M_S := SM$, and S is the scaling matrix, but some matrix-classes are not invariant under this (one-sided) scaling; see, e.g., Todd [37] for a brief discussion. Nevertheless, this scaling seems to be natural in view of the correspondence of our methods to Chen and Harker's characterization of PLCP(q, M, μ) and leads to a substantial improvement of the numerical results presented in §7.

5. Global error bounds for PLCP(q, M, μ). Consider the LCP in the following formulation: find a vector $x \in \Re^n$ satisfying the conditions

$$x \geq 0, Mx+q \geq 0, x^T(Mx+q) = 0.$$

The corresponding perturbed problem is exactly problem (11). Obviously, these problems are equivalent to LCP(q, M) and PLCP(q, M, μ), respectively, as defined in §1 (just take $y = Mx + q$), so we denote these problems also by LCP(q, M) and PLCP(q, M, μ), and the solution set of PLCP(q, M, μ) will again be denoted by $\mathcal{S}(\mu)$.

Although there are several existing error bound results for LCP(q, M)(see, e.g., [27, 29]), there is, to our knowledge, no error bound result for the PLCP(q, M, μ). Such a result, however, will play an important role in the algorithm to be described in the following section, namely in the termination rule of the inner iteration. We first give the relevant definitions.

DEFINITION 5.1. *Let $\mu > 0$, let $r : \Re^n \to \Re$, and assume that PLCP(q, M, μ) has a nonempty solution set $\mathcal{S}(\mu)$.*

(a) *The continuous function r is called a* residual *of PLCP(q, M, μ) if $r(x) \geq 0$ for all $x \in \Re^n$ and $r(x) = 0$ if and only if x solves PLCP(q, M, μ).*

(b) *A residual r is a* lower local error bound *for PLCP(q, M, μ) if there exist constants $\tau_1 > 0$ and $c_1 > 0$ such that*

$$\tau_1 r(x) \leq dist(x, \mathcal{S}(\mu))$$

for all $x \in \Re^n$ with $r(x) \leq c_1$.

(c) *A residual r is a* lower global error bound *for PLCP(q, M, μ) if there exists a constant $\tau_1 > 0$ such that*

$$\tau_1 r(x) \leq dist(x, \mathcal{S}(\mu))$$

for all $x \in \Re^n$.

(d) *A residual r is an* upper local error bound *for PLCP(q, M, μ) if there exist constants $\tau_2 > 0$ and $c_2 > 0$ such that*

$$dist(x, \mathcal{S}(\mu)) \leq \tau_2 r(x)$$

for all $x \in \Re^n$ with $r(x) \leq c_2$.

(e) *A residual r is an* upper global error bound *for PLCP(q, M, μ) if there exists a constant $\tau_2 > 0$ such that*

$$dist(x, \mathcal{S}(\mu)) \leq \tau_2 r(x)$$

for all $x \in \Re^n$.

Let φ_μ denote one of the functions defined in Lemma 2.1 or 2.2. For $x \in \Re^n$, let $y = Mx + q$ and define $r_{\varphi_\mu}(x) := \|\varphi_\mu(x, y)\| = \|F_{\varphi_\mu}(x, y)\|$, where $\varphi_\mu(x, y) := (\varphi_\mu(x_1, y_1), \ldots, \varphi_\mu(x_n, y_n)) \in \Re^n$ and where F_{φ_μ} denotes the operator as introduced at the beginning of §3. In view of Lemmas 2.1 and 2.2, the two possible functions r_{φ_μ} are residuals for PLCP(q, M, μ). In this section we prove that these two functions provide both lower and upper global error bounds for PLCP(q, M, μ) if the matrix M is a P_0- and R_0-matrix.

We first show that r_{φ_μ} provide lower global error bounds for PLCP(q, M, μ). Note that this result holds for an arbitrary matrix $M \in \Re^{n \times n}$.

LEMMA 5.2. *Let $\mu > 0$ and assume that PLCP(q, M, μ) has a nonempty solution set $\mathcal{S}(\mu)$. Then there exists a constant $\tau_1 > 0$ such that $\tau_1 r_{\varphi_\mu}(x) \leq dist(x, \mathcal{S}(\mu))$ for all $x \in \Re^n$.*

Proof. We first note that r_{φ_μ} are globally Lipschitz-continuous functions; this follows immediately from the second remark in §2 and the integral mean value theorem. Let κ_1 denote the Lipschitz constant of r_{φ_μ}. Let $x \in \Re^n$ be arbitrary, and let $x(\mu)$ be the closest solution of PLCP(q, M, μ) to x. Then we have

$$r_{\varphi_\mu}(x) = |r_{\varphi_\mu}(x) - r_{\varphi_\mu}(x(\mu))| \leq \kappa_1 \|x - x(\mu)\| = \kappa_1 dist(x, \mathcal{S}(\mu)).$$

The assertion therefore follows by taking $\tau_1 = 1/\kappa_1$. □

Next we note that the two possible functions r_{φ_μ} provide upper local error bounds.

LEMMA 5.3. *Let $\mu > 0, M \in \Re^{n \times n}$ be a P_0-matrix, and $\mathcal{S}(\mu)$ be nonempty. Then there exist constants $\tau_2 > 0$ and $c_2 > 0$ such that $dist(x, \mathcal{S}(\mu)) \leq \tau_2 r_{\varphi_\mu}(x)$ for all $x \in \Re^n$ with $r_{\varphi_\mu}(x) \leq c_2$.*

Proof. Since M is a P_0-matrix, the Jacobian $\nabla F_{\varphi_\mu}(z)$ is nonsingular for all $z = (x, y) \in \Re^{2n}, y = Mx + q$. In particular, $\nabla F_{\varphi_\mu}(z(\mu))$ is nonsingular for $z(\mu) = (x(\mu), y(\mu))$, where $x(\mu) \in \mathcal{S}(\mu)$ and $y(\mu) = Mx(\mu) + q$. Using Lemma 4.1.16 from

[4] and the continuity of r_{φ_μ}, we therefore obtain the existence of $\kappa_2 > 0$ and $c_2 > 0$ such that

$$r_{\varphi_\mu}(x) = \|F_{\varphi_\mu}(z)\| = \|F_{\varphi_\mu}(z) - F_{\varphi_\mu}(z(\mu))\| \geq \kappa_2 \|z - z(\mu)\| \geq \kappa_2 \|x - x(\mu)\| \\ = \kappa_2 dist(x, \mathcal{S}(\mu))$$

for all $x \in \Re^n$ such that $r_{\varphi_\mu}(x) \leq c_2$, where $x(\mu)$ denotes the closest solution of x in $\mathcal{S}(\mu)$. From this the assertion follows by taking $\tau_2 := 1/\kappa_2$. □

We are now in a position to prove the main result of this section.

THEOREM 5.4. *Let $\mu > 0$ and let $M \in \Re^{n \times n}$ be a P_0- and R_0-matrix. Then there exists a constant $\tau_2 > 0$ such that $dist(x, \mathcal{S}(\mu)) \leq \tau_2 r_{\varphi_\mu}(x)$ for all $x \in \Re^n$; i.e., r_{φ_μ} provide upper global error bounds for PLCP(q, M, μ).*

Proof. Based on Lemma 5.3, the proof is similar to the one of Theorem 2.1 by Mangasarian and Ren [29]. For simplicity, we assume that φ_μ is the function from Lemma 2.2. (The proof is analogous for the function φ_μ of Lemma 2.1.) By Theorems 3.6 and 3.7, PLCP(q, M, μ) has a unique solution $x(\mu)$. Assume the theorem is false. Then there exists a vector $x^k \in \Re^n$ such that

$$\|x^k - x(\mu)\| > k r_{\varphi_\mu}(x^k) \tag{19}$$

for all k. Since r_{φ_μ} is an upper local error bound by Lemma 5.3, one can prove as in [29] that there exist $k_0 > 0$ and $\varepsilon > 0$ such that $r_{\varphi_\mu}(x^k) > \varepsilon$ for all $k \geq k_0$. From this and (19) we get $\|x^k\| \to \infty$. Let x^* be an accumulation point of the bounded sequence $\{\frac{x^k}{\|x^k\|}\}$. Note that $\|x^*\| = 1$ and therefore $x^* \neq 0$. Dividing both sides of (19) by $\|x^k\|$, we obtain

$$1 = \lim_{k \to \infty} \frac{\|x^k - x(\mu)\|}{\|x^k\|} \geq \lim_{k \to \infty} k \frac{r_{\varphi_\mu}(x^k)}{\|x^k\|}. \tag{20}$$

From the definition of φ_μ, the unboundedness of the sequence $\{x^k\}$ and the fact that $\frac{x^k}{\|x^k\|} \to x^*$ on a subsequence, we get

$$\varphi_\mu(x_i^k, y_i^k)/\|x^k\| \to x_i^* + [Mx^*]_i - \sqrt{(x_i^*)^2 + [Mx^*]_i^2} \quad (i \in I) \tag{21}$$

on this subsequence, where $y^k := Mx^k + q$. In view of (20), we see that the right-hand side of (21) is equal to 0 for all $i \in I$. This, however, means that $\varphi_0(x_i^*, [Mx^*]_i) = 0$ for all $i \in I$; i.e., x^* solves LCP$(0, M)$ (cf. the first remark in §2). Since $x^* \neq 0$, this contradicts the assumption that M is an R_0-matrix. □

We note that the constants τ_1 and τ_2 in Lemmas 5.2 and 5.3 and Theorem 5.4 depend on the matrix M and the value of the perturbation parameter $\mu > 0$ but not on the particular choice of the vector q.

6. Implemented algorithm and its convergence. The following algorithm is an implementable version of Algorithm 3.2. Instead of solving the nonlinear systems $F_{\varphi_\mu}(z) = 0$ of step (S.2) of Algorithm 3.2 exactly, we try to solve them inexactly using just one step of Newton's method for a fixed value of μ. If this step is successful (in a certain sense as defined below, cf. step (S.5) of Algorithm 6.1), we reduce the perturbation parameter μ; otherwise, we perform another Newton step for the same value of μ. Note that Newton's method is globalized by a line search for the merit function $\frac{1}{2}\|F_{\varphi_\mu}\|^2$.

ALGORITHM 6.1.

(S.0): (Initial Data)
Let $\varphi_\mu : \Re^2 \to \Re$ be any of the functions defined in §2. Choose $x^0 \in \Re^n$, set $y^0 := Mx^0 + q, z^0 := (x^0, y^0)$, let $\mu_0 > 0, \beta \in (0,1), \sigma \in (0, \frac{1}{2}), \epsilon \geq 0, \gamma > 0, \eta \in (0,1)$, and set $k := 0$.

(S.1): (Termination Criterion)
If $err(z^k) := \| \min\{x^k, y^k\} \| \leq \epsilon$, stop: z^k is an (approximate) solution of LCP(q, M).

(S.2): (Computation of a Search Direction)
Compute $\Delta z^k \in \Re^{2n}$ as the solution of the linear system

$$\nabla F_{\varphi_{\mu_k}}(z^k)\Delta z = -F_{\varphi_{\mu_k}}(z^k).$$

(S.3): (Computation of a Steplength)
Let $t^k = \beta^{m_k}$, where m_k is the smallest nonnegative integer m satisfying the Armijo condition

$$\|F_{\varphi_{\mu_k}}(z^k + \beta^m \Delta z^k)\|^2 \leq (1 - \sigma\beta^m)\|F_{\varphi_{\mu_k}}(z^k)\|^2.$$

(S.4): (New Iterate)
Set $z^{k+1} := z^k + t^k \Delta z^k$.

(S.5): (Updating Rule for μ_k)
Define the vector $\varphi_{\mu_k}(x^{k+1}, y^{k+1}) := (\varphi_{\mu_k}(x_1^{k+1}, y_1^{k+1}), \ldots, \varphi_{\mu_k}(x_n^{k+1}, y_n^{k+1}))^T \in \Re^n$. If $\|\varphi_{\mu_k}(x^{k+1}, y^{k+1})\| \leq \gamma\mu_k$, then $\mu_{k+1} := \eta\mu_k$, else $\mu_{k+1} := \mu_k$.

(S.6): (Loop)
Set $k := k + 1$, and go to (S.1).

Remark. From our specific choice of the starting vector $z^0 = (x^0, y^0)$ and a simple induction argument, it follows that the relation $y^k = Mx^k + q$ holds for all iterations k. Note that such a relation does not hold in commonly used (infeasible) interior-point methods. Moreover, this motivates our termination criterion in step (S.1) and, in view of the global error bound results of the previous section, also the updating rule for μ_k in step (S.5).

Remark. There is now a strong theoretical foundation of the termination criterion used in step (S.1) of the above algorithm. From error bound results of Mangasarian and Ren [29] and Luo and Tseng [27] it follows that $err(z)$ is a global error bound of LCP(q, M) for all $q \in \Re^n$ if and only if $M \in \Re^{n \times n}$ is an R_0-matrix.

Remark. The linear system in step (S.2) of our algorithm has been solved as follows: Let $\varphi_{\mu_k}(x^k, y^k) := (\ldots, \varphi_{\mu_k}(x_i^k, z_i^k), \ldots)^T \in \Re^n$, and let D_a^k and D_b^k denote the diagonal matrices diag$(\ldots, \frac{\partial \varphi_{\mu_k}}{\partial a}(x_i^k, y_i^k), \ldots)$ and diag$(\ldots, \frac{\partial \varphi_{\mu_k}}{\partial b}(x_i^k, y_i^k), \ldots)$, respectively. Then, compute Δx^k from the $n \times n$ system

$$(D_a^k + D_b^k M)\Delta x^k = -\varphi_{\mu_k}(x^k, y^k) - D_b^k(Mx^k + q - y^k)$$

and set

$$\Delta y^k = M(x^k + \Delta x^k) + q - y^k.$$

We now show, using the theory developed in §3 for the theoretical Algorithm 3.2, that Algorithm 6.1 is also globally convergent under the same assumptions.

THEOREM 6.2. *Let $\varphi_\mu : \Re^2 \to \Re$ denote one of the functions defined in Lemma 2.1 or 2.2. Assume that $M \in \Re^{n \times n}$ is a P_0- and R_0-matrix, and let $\epsilon = 0$ (ϵ being the constant from the termination criterion). Let $\{z^k\}$ be any infinite sequence generated by Algorithm 6.1. Then*

(a) *the sequence $\{z^k\}$ is well defined;*

(b) *the sequence* $\{z^k\}$ *has at least one accumulation point*;
(c) *any accumulation point of* $\{z^k\}$ *is a solution of LCP*(q, M).

Proof. (a) Since M is a P_0-matrix, the linear systems in step (S.2) are uniquely solvable for all k by Theorem 3.5. Moreover, it is well known that the search direction $\Delta z^k \in \Re^{2n}$ obtained in this step is always a descent direction for the merit function $\frac{1}{2}\|F_{\varphi_{\mu_k}}\|^2$. Hence a steplength $t^k > 0$ can always be computed in step (S.3). Therefore Algorithm 6.1 is well defined.

(b) Since $\mu_k \in [0, \mu_0]$ for all k and since our algorithm is a descent method for any fixed μ_k, the entire sequence $\{z^k\}$ remains bounded in view of Theorem 3.7 and the assumed R_0-matrix property of M. Thus the sequence $\{z^k\}$ has at least one accumulation point.

(c) We first show that the sequence $\{\mu_k\}$ generated by Algorithm 6.1 converges to zero. Assume the contrary; i.e., assume there exists an iteration index $\hat{k}$ such that $\mu_{\hat{k}} = \mu_{\hat{k}+l}$ for all $l = 1, 2, 3, \ldots$. This means that Algorithm 6.1 eventually reduces to a damped Newton method for the (single) system of nonlinear equations $F_{\varphi_{\mu_{\hat{k}}}}(z) = 0$. In view of Theorems 3.5–3.7, however, it is well known that this method converges to the unique solution $z(\mu_{\hat{k}})$ of this system. In particular, the condition in step (S.5), i.e.,

$$\|\varphi_{\mu_k}(x^{k+1}, y^{k+1})\| = \|\varphi_{\mu_{\hat{k}}}(x^{k+1}, y^{k+1})\| \le \gamma\mu_k = \gamma\mu_{\hat{k}},$$

is satisfied for a finite value $k \ge \hat{k}$, so that the parameter μ_k will be reduced in this step. Hence $\{\mu_k\}$ converges to zero.

Now let $\bar{z} = (\bar{x}, \bar{y})$ be one of the accumulation points of $\{z^k\}$. (Because of (b), such an accumulation point exists.) Let $\{z^{k+1}\}_{k\in K}$ denote a subsequence converging to $\bar{z}$. For each sufficiently large $k \in K$, let $l(k)$ denote the largest index for which the condition $\|\varphi_{\mu_k}(x^{k+1}, y^{k+1})\| \le \gamma\mu_{l(k)}$ is satisfied; i.e.,

$$l(k) := \max\{j \in \{1, \ldots, k\} \mid \|\varphi_{\mu_k}(x^{k+1}, y^{k+1})\| \le \gamma\mu_j\}.$$

We note that this index $l(k)$ is well defined for all k sufficiently large and that $\lim_{k\in K} \mu_{l(k)} = 0$ since μ_k is reduced infinitely many times according to the first part of this proof. (Note that the index $l(k)$ is not necessarily in the index set K.) Taking into account the definition of $l(k)$ as well as the continuity of the two possible functions φ_μ in their arguments and in the perturbation parameter μ, we have

$$\|\varphi_0(\bar{x}, \bar{y})\| = \lim_{k\in K} \|\varphi_{\mu_k}(x^{k+1}, y^{k+1})\| \le \gamma \lim_{k\in K} \mu_{l(k)} = 0;$$

i.e., $\varphi_0(\bar{x}, \bar{y}) = 0$. This, however, is equivalent to $\bar{x}_i \ge 0, \bar{y}_i \ge 0$, and $\bar{x}_i\bar{y}_i = 0$ $(i \in I)$. According to the first remark after Algorithm 6.1, we also have $\bar{y} = M\bar{x} + q$. Hence $\bar{z} = (\bar{x}, \bar{y})$ is a solution of LCP(q, M). □

The implemented version of Algorithm 6.1 differs in two points from the above description, namely in steps (S.3) and (S.5). Instead of step (S.3), we employ the following nonmonotone Armijo rule (see Grippo, Lampariello, and Lucidi [13]):

(S.3') Define $p_k := \min\{k, p\}$ and let $t^k := \beta^{m_k}$, where m_k is the smallest nonnegative integer m satisfying the nonmonotone Armijo condition

$$\|F_{\varphi_{\mu_k}}(z^k + \beta^m \Delta z^k)\|^2 \le \max_{j=k-p_k+1,\ldots,k} \|F_{\varphi_{\mu_j}}(z^j)\|^2 - \sigma\beta^m \|F_{\varphi_{\mu_k}}(z^k)\|^2.$$

Here, p is any fixed nonnegative integer. In our numerical experiments, this nonmonotone line search gives better results than the standard (monotone) Armijo rule. This,

we think, is an interesting aspect since usually the nonmonotone line search is only preferred for highly nonlinear objective functions, whereas our merit function is not too nonlinear.

The second modification is in the updating rule for the perturbation parameter μ_k. The updating rule used is the following one. It is very similar to the one presented by Chen and Harker [2] and works very well in our numerical experiments.

UPDATING RULE FOR μ_k:

(a) Let $u_{k+1} := err(z^{k+1})^2/n$. If $u_{k+1} \geq 1$, then $\mu_{k+1} := \sqrt{u_{k+1}}$, else $\mu_{k+1} := u_{k+1}$.

(b) If $\mu_{k+1} > \mu_k$, set $\mu_{k+1} := \mu_k$.

(c) If $\|\varphi_{\mu_k}(x^{k+1}, y^{k+1})\| < 10^{-4}$, set $\mu_{k+1} := 10^{-1}\mu_{k+1}$.

(d) If $\mu_{k+1} < 10^{-16}$, then $\mu_{k+1} := 10^{-16}$.

The main part of this (heuristic) updating rule is part (a). Part (b) guarantees that the sequence $\{\mu_k\}$ is nonincreasing; part (c) means that the parameter μ_k is reduced very fast if we are already close to a solution of LCP(q, M). Part (d), of course, is just a safeguard.

7. Numerical results. In this section, we present numerical results for the following methods:

1: This is the algorithm described in §6 with φ_μ being the function defined in Lemma 2.1.

2: This is the corresponding algorithm with φ_μ taken from Lemma 2.2.

3: Similar to methods 1 and 2 but with φ_μ from (3).

4: The same as method 3, but with φ_μ as defined in (4).

Furthermore, we will consider the three scaling strategies as suggested at the end of §4. We compare our results with the following two algorithms:

CH 1: This is Chen and Harker's method; cf. (14).

CH 2: This is the modification of Chen and Harker's method as suggested in (15). (The body for the algorithms CH 1 and CH 2 is, of course, the same as for the methods 1–4; the main difference is only in step (S.2) of Algorithm 6.1, where different linear systems are to be solved.) The algorithms have been implemented in MATLAB and tested on a 486 PC-type computer. The parameters used are as follows:

$$\beta = 0.5, \sigma = 10^{-4}, \epsilon = 10^{-6}, p = 10, \tau = 10^{-8}.$$

The initial barrier parameter μ_0 is $\|q\|/n$.

The first two test examples are well scaled, and we therefore present numerical results only for the choice $S = S_0$ of the scaling matrix.

EXAMPLE 7.1 (see Murty [31]). n variable,

$$M = \begin{pmatrix} 1 & 2 & 2 & \cdots & 2 \\ 0 & 1 & 2 & \cdots & 2 \\ 0 & 0 & 1 & \cdots & 2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}, q = (-1, \ldots, -1)^T.$$

This example of an LCP is a standard test problem for which both Lemke's complementary pivot algorithm and Cottle and Danzig's principal pivoting method are known to run in exponential time; see Murty [31, Chap. 6]. Its solution is $x^* = (0, \ldots, 0, 1)^T, y^* = (1, \ldots, 1, 0)^T$. Obviously, the matrix M in this example is a P-matrix and therefore a P_0- and R_0-matrix. Consequently, methods 1 and 2 satisfy

TABLE 1
Number of iterations for Example 7.1.

Method	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$
1	7	7	7	5	5	5
2	7	7	7	6	5	5
3	10	10	10	11	12	12
4	10	10	10	11	12	12

TABLE 2
Number of iterations for Example 7.2.

Method	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$
1	8	8	8	8	8	7
2	8	8	9	9	9	10
3	11	14	15	17	18	20
4	11	13	14	16	18	20
CH 1	8	11	11	11	11	11
CH 2	9	10	10	10	10	10

the assumptions of the global convergence Theorem 6.2. The starting vector chosen is $x^0 = (1, \ldots, 1)^T$. The number of iterations needed by our algorithm is indicated in Table 1 for several values of the dimension n. The methods CH 1 and CH 2 need the same number of iterations as the methods 1 and 2, respectively, since the diagonal entries of M are all equal to one, so that the corresponding characterizations of the complementarity problem coincide.

EXAMPLE 7.2 (see Fathi [7]). n variable,

$$M = \begin{pmatrix} 1 & 2 & 2 & \cdots & 2 \\ 2 & 5 & 6 & \cdots & 6 \\ 2 & 6 & 9 & \cdots & 10 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 2 & 6 & 9 & \cdots & 4(n-1)+1 \end{pmatrix}, q = (-1, \ldots, -1)^T.$$

This LCP(q, M) is a standard test problem too. Again, the complementary pivot algorithm and the principal pivoting method are known to run in exponential time. Moreover, some Newton-type algorithms are also known to have some difficulties with this example; see, e.g., Harker and Pang [16] and Kanzow [19]. Its solution is $x^* = (1, 0, \ldots, 0)^T, y^* = (0, 1, \ldots, 1)^T$. The matrix M of this example is positive definite and therefore a P-matrix. Our numerical results obtained with the methods 1–4, CH 1, and CH 2 are summarized in Table 2. The starting vector chosen is the same as in Example 7.1.

From Tables 1 and 2, we see that methods 1 and 2 as well as methods CH 1 and CH 2 lead to better results than methods 3 and 4. This is not an unexpected behaviour since methods 3 and 4 are based on the more nonlinear functions φ_μ from (3) and (4), hence Newton's method leads to more difficulties in solving the corresponding nonlinear systems of equations $F_{\varphi_\mu}(z) = 0$. In the following randomly generated examples, we will therefore concentrate on these four methods.

EXAMPLE 7.3 (see Harker and Pang [16]). The matrix M is computed as follows: let $A, B \in \Re^{n \times n}$ and $q, \eta \in \Re^n$ be randomly generated such that $a_{ij}, b_{ij} \in (-5, 5), q_i \in (-500, 500)$, and $\eta_i \in (0.0, 0.3)$ and that B is skew-symmetric. Define $M = A^T A + B + diag(\eta)$. Then, M is a P-matrix. For several values of n, 10 examples have been generated in this way. The maximum, average, and minimum numbers of iterations

TABLE 3
Number of iterations for Example 7.3.

Method		1			2			CH 1	CH 2
n		S_0	S_1	S_2	S_0	S_1	S_2		
$n = 50$	Max.	19.0	12.0	12.0	20.0	12.0	12.0	16.0	16.0
	Avg.	15.0	10.4	10.3	16.8	10.8	10.7	14.7	14.8
	Min.	12.0	9.0	9.0	13.0	10.0	9.0	13.0	13.0
$n = 100$	Max.	20.0	13.0	13.0	20.0	12.0	13.0	16.0	18.0
	Avg.	17.3	11.2	11.4	18.2	11.2	11.4	15.2	16.3
	Min.	16.0	9.0	10.0	17.0	9.0	10.0	14.0	15.0
$n = 150$	Max.	22.0	13.0	14.0	22.0	13.0	14.0	18.0	21.0
	Avg.	18.6	11.4	11.6	19.6	11.4	11.6	16.4	18.2
	Min.	17.0	11.0	11.0	18.0	11.0	11.0	16.0	16.0
$n = 200$	Max.	22.0	14.0	15.0	21.0	14.0	15.0	18.0	19.0
	Avg.	18.9	11.7	11.8	19.6	12.0	12.1	16.5	18.1
	Min.	17.0	11.0	11.0	19.0	11.0	11.0	14.0	16.0

TABLE 4
Number of iterations for Example 7.4.

Method		1			2			CH 1	CH 2
n		S_0	S_1	S_2	S_0	S_1	S_2		
$n = 50$	Max.	19.0	13.0	13.0	21.0	13.0	13.0	17.0	18.0
	Avg.	16.0	11.1	11.1	17.6	11.3	11.6	14.3	14.9
	Min.	13.0	9.0	10.0	15.0	10.0	10.0	13.0	13.0
$n = 100$	Max.	20.0	12.0	12.0	24.0	12.0	12.0	17.0	18.0
	Avg.	19.0	10.8	11.0	20.9	10.3	10.5	15.6	16.7
	Min.	16.0	9.0	10.0	18.0	9.0	10.0	13.0	15.0
$n = 150$	Max.	23.0	12.0	13.0	26.0	13.0	14.0	18.0	19.0
	Avg.	20.6	11.3	11.2	23.1	11.3	11.5	16.9	18.3
	Min.	19.0	10.0	10.0	21.0	10.0	10.0	15.0	17.0
$n = 200$	Max.	24.0	13.0	12.0	25.0	13.0	13.0	19.0	20.0
	Avg.	21.1	11.4	11.2	23.0	11.6	11.9	17.2	18.4
	Min.	18.0	10.0	10.0	21.0	10.0	10.0	17.0	18.0

needed by the algorithms are summarized in Table 3. In all test runs, $x^0 = (0, \ldots, 0)^T$ has been chosen as starting vector.

EXAMPLE 7.4 (see Harker and Pang [16] "hard examples"). In this example, M is computed in the same way as in the previous example, and $q \in \Re^n$ is randomly generated with entries $q_i \in (-500, 0)$. Table 4 contains our numerical results, which we have obtained using the starting vector $x^0 = (0, \ldots, 0)^T$.

EXAMPLE 7.5 (see Pardalos et al. [34]). This example is a P_0-matrix LCP. The matrix M, the vector q, as well as the starting point x^0 are randomly generated. A detailed description is given in Pardalos et al. [34]. We note that at least half of the diagonal entries of M are zero, so that neither Chen and Harker's method nor the suggested modification are applicable to this example. Moreover, there are substantial difficulties in solving this example for the potential reduction algorithm of Pardalos et al. [34]. (Their algorithm needs hundreds of iterations in order to solve this problem.) On the other hand, as is shown in Table 5, our algorithms do not have any problems with this example.

We can summarize the above results as follows: All problems have been solved using only a small number of iterations. The number of iterations needed is almost independent of the problem dimension. Methods 1 and 2 lead to better results than do methods 3 and 4, and it is difficult to decide which of these two methods is best, although for most examples method 1 has a slightly better behaviour. In their un-

TABLE 5
Number of iterations for Example 7.5.

Method		1			2		
n		S_0	S_1	S_2	S_0	S_1	S_2
$n = 100$	Max.	17.0	17.0	17.0	17.0	16.0	20.0
	Avg.	14.8	10.7	8.3	16.5	13.7	14.2
	Min.	14.0	9.0	6.0	15.0	12.0	12.0
$n = 200$	Max.	19.0	12.0	13.0	19.0	15.0	21.0
	Avg.	16.4	10.7	9.2	17.5	13.8	16.1
	Min.	14.0	9.0	6.0	16.0	12.0	13.0

scaled versions, these two algorithms, however, are inferior to both Chen and Harker's method CH 1 and its modification CH 2. On the other hand, simple diagonal scalings of the original data lead to substantial improvement of the results obtained with the methods 1 and 2; these methods are by far the most successful ones. Both scaling techniques, S_1 and S_2, lead to similar results, so the slightly cheaper scaling S_1 might be preferable. We emphasize that we have not introduced the diagonal scaling in order to improve our results. (In fact, there might be better scaling techniques, though this is not a simple problem.) Instead the considerations of §4 show that it is more natural to compare our algorithms with the one of Chen and Harker after rescaling the original problem.

Finally, we note that in almost all iterations the full stepsize $t^k = 1$ has been accepted. (In just 10 of the 820 test runs made for our numerical results we observed a steplength less than 1; this happened 6 times in Example 7.2, 1 time in Example 7.3, and 3 times in Example 7.4.) This is not the case if the nonmonotone line search rule is replaced by a monotone one. Moreover, it is very unlikely that a practical interior-point method for LCP(q, M) will almost always take a full step, since the condition that all iterates have to stay in the positive orthant will usually truncate the steplength. Taking the stepsize $t^k = 1$ very often (in a controlled way), however, is now known to be very successful in combination with Newton-type methods; see, e.g., the highly promising and extensive numerical results reported by Grippo, Lampariello, and Lucidi [13, 14] and Toint [38, 39]. Another advantage of our noninterior continuation methods is the fact that we can start at an arbitrary vector $(x^0, y^0) \in \Re^{2n}$, in particular, the very natural choice $y^0 = Mx^0 + q$ is allowed even if y^0 has nonpositive or negative components. This contrasts favourably with interior-point methods.

8. Final remarks. In this paper, we have introduced some new continuation methods for the solution of LCPs. As with interior-point methods, we try to follow the central path; however, we also allow negative iterates. One of our methods is closely related to a recently proposed method of Chen and Harker [2], but here we do not need their assumption that all diagonal entries of M are positive.

There are several questions which remain to be answered: a theoretical justification of the updating rule for the perturbation parameter μ is missing. We also have not given a local or global rate of convergence result. Moreover, it is currently not known to the author whether or not the noninterior continuation methods have a polynomial complexity bound. The numerical results of the previous section seem to indicate that such a complexity bound exists since the number of iterations is almost independent of the problem dimensions.

Of course, the main tools introduced in §2 are also applicable to (convex) constrained optimization problems and could lead to noninterior continuation methods for, e.g., linear and quadratic programming problems. It would be interesting to com-

pare these methods with standard interior-point methods as well as with the shifted barrier method of Freund [12] and the modified barrier method of Polyak [35]. The methods of Freund and Polyak also allow negative components of the iteration vectors and are, in this respect, related to our algorithms.

For the LCP, we plan to compare the numerical behaviour of the noninterior continuation methods suggested in this paper with some interior-point methods.

Acknowledgements. I would like to thank Ph. L. Toint and M. J. Todd for their helpful comments on an earlier version of this paper. I am also grateful to the referees for their many constructive suggestions.

REFERENCES

[1] M. AGANAGIC AND R. W. COTTLE, *A note on Q-matrices*, Math. Programming, 16 (1979), pp. 374–377.

[2] B. CHEN AND P. T. HARKER, *A non-interior-point continuation method for linear complementarity problems*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1168–1190.

[3] R. W. COTTLE, J.-S. PANG, AND R. E. STONE, *The Linear Complementarity Problem*, Academic Press, New York, 1992.

[4] J. E. DENNIS, JR. AND R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[5] F. FACCHINEI AND J. SOARES, *A new merit function for nonlinear complementarity problems and a related algorithm*, SIAM J. Optim., 7 (1997), to appear.

[6] F. FACCHINEI AND J. SOARES, *Testing a new class of algorithms for nonlinear complementarity problems*, in Variational Inequalities and Network Equilibrium Problems, F. Giannessi and A. Maugeri, eds., Plenum Press, New York, 1995, pp. 69–83.

[7] Y. FATHI, *Computational complexity of LCPs associated with positive definite matrices*, Math. Programming, 17 (1979), pp. 335–344.

[8] A. FISCHER, *A special Newton-type optimization method*, Optimization, 24 (1992), pp. 269–284.

[9] ———, *A Newton-type method for positive semidefinite linear complementarity problems*, J. Optim. Theory Appl., 86 (1995), pp. 585–608.

[10] ———, *On the local superlinear convergence of a Newton-type method for LCP under weak conditions*, Optimization Methods and Software, 6 (1995), pp. 83–107.

[11] ———, *On an NCP-function and its use for the solution of complementarity problems*, in Recent Advances in Nonsmooth Optimization, D.-Z. Du, L. Qi, and R. S. Womersley, eds., World Scientific Publishers, Singapore, 1995, pp. 88–105.

[12] R. M. FREUND, *Theoretical efficiency of a shifted-barrier-function algorithm for linear programming*, Linear Algebra Appl., 152 (1991), pp. 19–41.

[13] L. GRIPPO, F. LAMPARIELLO, AND S. LUCIDI, *A nonmonotone line search technique for Newton's method*, SIAM J. Numer. Anal., 23 (1986), pp. 707–716.

[14] ———, *A class of nonmonotone stabilization methods in unconstrained optimization*, Numer. Math., 59 (1991), pp. 779–805.

[15] P. T. HARKER, *Lectures on Computation of Equilibria with Equation-Based Methods*, CORE Lecture Series, Louvain-la-Neuve, Belgium, 1993.

[16] P. T. HARKER AND J.-S. PANG, *A damped Newton method for the linear complementarity problem*, in Computational Solution of Nonlinear Systems of Equations, Lectures in Applied Mathematics, Vol. 26, E. L. Allgower and K. Georg, eds., American Mathematical Society, Providence, RI, 1990, pp. 265–284.

[17] G. ISAC, *Complementarity Problems*, Lecture Notes in Mathematics 1528, Springer-Verlag, Berlin, 1992.

[18] C. KANZOW, *Some equation-based methods for the nonlinear complementarity problem*, Optimization Methods and Software, 3 (1994), pp. 327–340.

[19] ———, *Global convergence properties of some iterative methods for linear complementarity problems*, SIAM J. Optim., 6 (1996), pp. 326–341.

[20] ———, *Some Tools Allowing Interior-Point Methods to Become Noninterior*, Preprint 79, Institute of Applied Mathematics, University of Hamburg, Hamburg, Germany, February 1994.

[21] ———, *Noninterior Continuation Methods for Mixed Linear Complementarity Problems*, Preprint 85, Institute of Applied Mathematics, University of Hamburg, Hamburg, Germany, August 1994.

[22] C. KANZOW AND H. KLEINMICHEL, *A class of Newton-type methods for equality and inequality constrained optimization*, Optimization Methods and Software, 5 (1995), pp. 173–198.

[23] M. KOJIMA, N. MEGGIDO, AND T. NOMA, *Homotopy continuation methods for nonlinear complementarity problems*, Math. Oper. Res., 16 (1991), pp. 754–774.

[24] M. KOJIMA, N. MEGGIDO, T. NOMA, AND A. YOSHISE, *A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems*, Lecture Notes in Computer Science 538, Springer-Verlag, Berlin, 1991.

[25] M. KOJIMA, S. MIZUNO, AND T. NOMA, *A new continuation method for complementarity problems with uniform P-functions*, Math. Programming, 43 (1989), pp. 107–113.

[26] ———, *Limiting behaviour of trajectories generated by a continuation method for monotone complementarity problems*, Math. Oper. Res., 15 (1990), pp. 662–675.

[27] X.-D. LUO AND P. TSENG, *Conditions for a projection-type error bound for the linear complementarity problem to be global*, Linear Algebra Appl., to appear.

[28] O. L. MANGASARIAN, *Equivalence of the complementarity problem to a system of nonlinear equations*, SIAM J. Appl. Math., 31 (1976), pp. 89–92.

[29] O. L. MANGASARIAN AND J. REN, *New improved error bounds for the linear complementarity problem*, Math. Programming, 66 (1994), pp. 241–255.

[30] N. MEGGIDO, *Pathways to the optimal set in linear programming*, in Progress in Mathematical Programming. Interior Point and Related Methods, N. Meggido, ed., Springer-Verlag, New York, 1988.

[31] K. G. MURTY, *Linear Complementarity, Linear and Nonlinear Programming*, Sigma Series in Applied Mathematics 3, Heldermann-Verlag, Berlin, 1988.

[32] J.-S. PANG, *Newton's method for B-differentiable equations*, Math. Oper. Res., 15 (1990), pp. 311–341.

[33] ———, *A B-differentiable equation-based, globally and locally quadratically convergent algorithm for nonlinear programs, complementarity and variational inequality problems*, Math. Programming, 51 (1991), pp. 101–131.

[34] P. M. PARDALOS, Y. YE, C.-G. HAN, AND J. A. KALISKI, *Solution of P_0-matrix linear complementarity problems using a potential reduction algorithm*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1048–1060.

[35] R. POLYAK, *Modified barrier functions (theory and methods)*, Math. Programming, 54 (1992), pp. 177–222.

[36] L. QI AND H. JIANG, *Semismooth Karush–Kuhn–Tucker Equations and Convergence Analysis of Newton and Quasi-Newton Methods for Solving These Equations*, Technical Report AMR 94/5, School of Mathematics, University of New South Wales, Sydney, Australia, March 1994 (revised November 1994).

[37] M. J. TODD, *The two faces of the linear complementarity problem*, SIAG/OPT Views-and-News, 5 (Fall 1994), pp. 3–5.

[38] PH. L. TOINT, *An assessment of nonmonotone linesearch techniques for unconstrained optimization*, SIAM J. Sci. Comput., 17 (1996), pp. 561–578.

[39] ———, *A Non-Monotone Trust-Region Algorithm for Nonlinear Optimization Subject to Convex Constraints*, Report 94/24, Department of Mathematics, Facultés Universitaires ND de la Paix, Namur, Belgium, November 1994.

[40] P. TSENG, *Growth behaviour of a class of merit functions for the nonlinear complementarity problem*, J. Optim. Theory Appl., 89 (1996), pp. 17–37.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 4, pp. 869–885, October 1996

010

ON TRIDIAGONALIZING AND DIAGONALIZING SYMMETRIC MATRICES WITH REPEATED EIGENVALUES*

CHRISTIAN H. BISCHOF† AND XIAOBAI SUN‡

Abstract. We describe a divide-and-conquer tridiagonalization approach for matrices with repeated eigenvalues. Our algorithm hinges on the fact that, under easily constructively verifiable conditions, a symmetric matrix with band width b and k distinct eigenvalues must be block diagonal with diagonal blocks of size at most bk. A slight modification of the usual orthogonal band-reduction algorithm allows us to reveal this structure, which then leads to potential parallelism in the form of independent diagonal blocks. Compared to the usual Householder reduction algorithm, the new approach exhibits improved data locality, significantly more scope for parallelism, and the potential to reduce arithmetic complexity by close to 50% for matrices that have only two numerically distinct eigenvalues. The actual improvement depends to a large extent on the number of distinct eigenvalues and a good estimate thereof. However, at worst the algorithms behave like a successive band-reduction approach to tridiagonalization. Moreover, we provide a numerically reliable and effective algorithm for computing the eigenvalue decomposition of a symmetric matrix with two numerically distinct eigenvalues. Such matrices arise, for example, in invariant subspace decomposition approaches to the symmetric eigenvalue problem.

Key words. tridiagonalization, eigenvalue decomposition, repeated eigenvalues

AMS subject classifications. 15A23, 15A18, 65F15, 65F25

1. Introduction.

Let A be an $n \times n$ symmetric matrix. Our goal is to compute an orthogonal–tridiagonal decomposition of A, $AQ = QT$, where Q is orthogonal and T is tridiagonal. Reduction to tridiagonal form is a standard preprocessing step in dense eigensolvers based on QR iteration, bisection, or Cuppen's method [16]. The conventional tridiagonalization procedure [16, p. 419] reduces A one column at a time through Householder transformation at a cost of $O(4n^3/3)$ flops for the reduction of A, and an additional $O(4n^3/3)$ flops if the orthogonal matrix is accumulated at the same time. This algorithm mainly employs matrix–vector multiplications and symmetric rank-one updates, which require more memory references than matrix–matrix operations [9, 8, 14].

The block tridiagonalization algorithm in [5, 15] combines sets of p successive symmetric rank-one updates into one symmetric rank-p update at the cost of $O(2pn^2)$ extra flops. As a result, this algorithm exhibits improved data locality and hence is likely to be preferable on cache-based architectures. This block algorithm has been incorporated into the LAPACK library of portable linear algebra codes for high-performance architectures [1, 2]. Parallel versions for distributed memory machines of the standard algorithm and the block algorithm are described in [12] and [13], respectively. A different approach to tridiagonalization is the so-called successive

* Received by the editors March 16, 1992; accepted for publication (in revised form) by C. Van Loan November 14, 1995. This work was supported by Advanced Research Projects Agency contracts DM28E04120 and P-95006 and the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Computational and Technology Research, U. S. Department of Energy, under contract W-31-109-Eng-38. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

† Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439-4801 (bischof@mcs.anl.gov).

‡ Department of Computer Science, Duke University, Durham, NC 27708-0129 (xiaobai@cs.duke.edu). The work of this author was partially performed as a postdoctoral associate at Argonne National Laboratory.

band-reduction (SBR) method, which completes the tridiagonal reduction through a sequence of band reductions [10, 7]. This approach leads to algorithms that exhibit an even greater degree of memory locality, among other desirable features.

In this paper we show that if the number k (say) of distinct eigenvalues of a symmetric matrix A is small, then there is considerable scope for further savings in tridiagonalization algorithms. As will be demonstrated, A can be cheaply reduced to a block diagonal banded form through a slightly modified SBR approach. The final tridiagonal form is then achieved by applying the algorithm recursively on the subblocks on the diagonal. Compared to the conventional approach, this approach has the following advantages.

Improved data locality. The tridiagonalization process can employ mainly matrix–matrix operations both in the reduction of A and in the update of the transformation matrix Q (see also [10, 7]).

Enhanced scope for parallelism. In the traditional algorithm, the scope for the exploitation of parallelism in the reduction of A is limited to the application of the rank-one update (for the unblocked algorithm) or the rank-p update (for the blocked algorithm), and the scope for parallelism decreases as subproblems become smaller. In contrast, our algorithm generates independent subproblems during the reduction of A, which can be worked on independently, and the number of independent subproblems increases as the iteration proceeds. Thus, there is a shift from data parallelism (updates of large matrices) to functional parallelism (several independent subproblems), but at any stage, there is plenty of parallelism to exploit.

Reduced complexity. Depending on the number of distinct eigenvalues, we may almost halve the number of floating-point operations. In addition, the need for data movement is reduced.

One particular situation where repeated eigenvalues arise is in the context of invariant subspace methods for eigenvalue problems [3, 19, 6, 4], where a matrix with only two distinct predetermined eigenvalues is generated either by repeated application of incomplete beta functions [19] or the matrix sign function [4]. In exact arithmetic, our tridiagonalization procedure would result in a block diagonal matrix with diagonal blocks of order no larger than 2. Hence the eigenvalue decomposition could be computed easily by independently diagonalizing the 2×2 blocks on the diagonal. In the presence of roundoff errors, the computed tridiagonal matrix may not have this desirable structure. However, we can prove that such a tridiagonal matrix can be diagonalized as reliably as with any other method by two "clean up sweeps," where each sweep solves at most $n/2$ independent 2×2 eigenvalue problems.

The paper is organized as follows. We show in §2 that, under certain easily constructively verifiable conditions, a banded symmetric matrix with band width b and k distinct eigenvalues is block diagonal with diagonal blocks of order at most bk. In §3, we present a reduction algorithm to achieve the desired banded block diagonal structure through a slight modification of the conventional band-reduction procedure. This approach is then employed to develop a divide-and-conquer tridiagonalization algorithm. An inexpensive algorithm for decoupling invariant subspaces of matrices with eigenvalue clusters at 0 and 1 is given and verified in §4. Numerical experiments with a Matlab implementation are reported in §5. Lastly, we summarize our results.

2. The structure of band matrices with repeated eigenvalues. A tridiagonal matrix whose off diagonal entries are all nonzero is called *unreduced.* It is well known [18, p. 66] that an unreduced tridiagonal matrix does not have multiple eigenvalues. Consequently, if an $n\times n$ tridiagonal matrix has only $k \ll n$ distinct

eigenvalues, it must be block diagonal, and the largest block cannot be larger than $k \times k$. The generalization of this fact to banded matrices underpins the algorithm we propose, yet it is not as straightforward as it might seem.

Assuming that A is an $n \times n$ symmetric matrix, we define the *ith row-band-width* of A, denoted by band_row(i), as

$$\text{band\_row}(i) \stackrel{\text{def}}{=} \max_j \{i - j \mid j = i \text{ or } j < i \text{ and } a_{ij} \neq 0\}, \quad 1 \leq i \leq n. \tag{1}$$

That is, band_row(i) is the distance of the first nonzero element in row i from the ith diagonal element. Further, we say that A is *nonincreasing in row-band-width from b* if

$$a(b, 1) \neq 0 \text{ and band\_row}(i) \leq \text{band\_row}(i-1), \quad b+1 < i \leq n. \tag{2}$$

In particular, a banded matrix that is all zero below the bth subdiagonal and all nonzero on the bth subdiagonal is nonincreasing in row-band-width from b.

With these definitions, we can now prove the following theorem.

THEOREM 2.1. *Let T be a symmetric matrix with k distinct eigenvalues. If T is block diagonal, with each diagonal block nonincreasing in band width from at most b, then the size of the largest block cannot exceed kb.*

Proof. Assume that T has a diagonal block D of size $p > kb$. By assumption, D is nonincreasing in band width from b; that is, D has $p-b$ rows with their first nonzero elements in different columns to the left of the diagonal. Thus, for any λ, rank$(D-\lambda I)$ is not less than $p-b$.

On the other hand, since $p > kb$ and D has at most k distinct eigenvalues, D has an eigenvalue μ with multiplicity greater than b. Hence, rank$(D-\mu I)$ is less than $p-b$. The contradiction verifies the result of the theorem. □

The following example shows the necessity of the "nonincreasing band-width" restriction in Theorem 2.1. Let

$$Q^T = \begin{pmatrix} \xi & \eta & \mu & -\nu & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \beta & 0 & \gamma & \delta \\ 0 & 0 & \nu & \mu & 0 & \alpha & 0 & 0 \end{pmatrix},$$

where $\nu^2 + \mu^2 + \alpha^2 = 1$, $\xi^2 + \eta^2 = \alpha^2$, and $\beta^2 + \gamma^2 + \delta^2 = 1$. Then Q has orthonormal columns and $A = QQ^T$ is symmetric with only 0 and 1 as eigenvalues. In fact,

$$A = \begin{bmatrix} \times & \times & \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \times & 0 & 0 & \times & 0 & 0 \\ \times & \times & 0 & \times & 0 & \times & 0 & 0 \\ 0 & 0 & 0 & 0 & \times & 0 & \times & \times \\ 0 & 0 & \times & \times & 0 & \times & 0 & 0 \\ 0 & 0 & 0 & 0 & \times & 0 & \times & \times \\ 0 & 0 & 0 & 0 & \times & 0 & \times & \times \end{bmatrix}. \tag{3}$$

We see that A is banded with semi-band-width $b = 3$, but it is *not* block diagonal with blocks of size at most $2b \times 2b = 6 \times 6$ since the "nonincreasing band-width condition" is violated by $a(5, 2) = a(7, 4) = 0$.

3. A divide-and-conquer tridiagonalization approach. The example in the previous section showed that the standard Householder band-reduction algorithm will not necessarily reveal the block diagonal structure. For example, if we had applied the standard algorithm for reduction to band width 3 to the matrix of example (3), the matrix would have remained unchanged. Fortunately, a minor modification of the standard algorithm enforces nonincreasing row-band-width, and hence the prerequisites of Theorem 2.1.

Let us consider the conventional reduction approach, where the matrix is reduced one column at a time to semi-band-width b. In each reduction, the pivot row is always b rows below the diagonal, no matter whether the reduction of the previous column was skipped (i.e., the transformation was an identity) or not. For example, reducing the matrix A in (3) to semi-band-width 3, row number 4 is the pivot row for the reduction of the second column and, since $a(4:8,2) = 0$, this reduction is skipped. We then proceed to column 3, using row 5 as pivot row, and the row-band-width increases. If instead we employ a Householder transformation acting on $a(4:8,3)$ to eliminate $a(5:8,3)$, keeping row 4 as pivot row, we obtain

$$\tilde{A} = \begin{bmatrix} \times & \times & \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \times & 0 & \times & 0 & 0 & 0 \\ \times & \times & 0 & \times & \times & 0 & 0 & 0 \\ 0 & 0 & \times & \times & \times & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & 0 & \times & \times & \times \end{bmatrix}.$$

Now $\tilde{A}$ is decoupled into two diagonal blocks of size at most 6×6.

This example shows that nonincreasing band width can easily be obtained if we do not increase the pivot row when the previous reduction is skipped. For computational purposes, we define the row-band-width with respect to a threshold τ:

$$\text{(4) band\_row}(i,\tau) \stackrel{\text{def}}{=} \max_j \{i - j \mid j = i \text{ or } j < i \text{ and } \|a(i:n,j)\|_2 > \tau\}, \quad 1 \le i \le n.$$

That is, given a tolerance threshold τ, a column $a(i:n)$ is considered numerically zero if its 2-norm is at most τ. The Matlab function `bred` in Figure 1 shows the conventional band-reduction algorithm augmented with

(1) a threshold criterion for the generation of a Householder vector, and

(2) a modified pivot row selection strategy, which does not change the pivot row if a transformation is skipped.

The subroutines `gen_hh`, `pre_hh`, `post_hh`, and `sym_hh` generate a Householder vector and apply it from the left, right, and symmetrically, respectively. Note that for simplicity the algorithm presented here does not exploit the symmetry of A. However, if we wish to do so, we can have `sym_hh` work only with a triangular part of A and omit the `post_hh` (`pre_hh`) call when working only with the lower (upper) triangle. We also note that all the algorithms presented in this paper are available via anonymous ftp from the `pub/prism` directory at `ftp.super.org`.

If no transformations are skipped, the procedure is identical to the conventional band-reduction procedure; otherwise, it may terminate earlier when the reduction reaches the last column of the first diagonal block, and the problem is decoupled. Since we drop pivot columns whose norm is $O(\tau)$, the decomposition will be accurate up to a residual of order τ.

```
function [A, block1, Q] = bred( A, b, tau, Q );

%    Given a symmetric matrix A, a bandwidth b, and a threshold tau, bred
%    computes an orthogonal-banded matrix decomposition,
%           A_input * W = W * A_output + O(tau)
%    where O(tau) denotes a matrix with a two-norm of order tau, and
%    W is an orthogonal matrix.
%    The output matrix A_output will be a 2x2 block diagonal matrix,
%    where the first diagonal block A_output(1:block1,1:block1)
%    is banded with bandwidth nonincreasing from b, and the second block
%    may be empty.

[ m, n ] = size(A); if (m~=n) error('nonsquare A'); end
piv_row  = min(b+1,n);       % current pivot row
if (piv_row == n) block1 = n; return; end;
for j = 1:n-b
          % matrix is decoupled, stop
       if (piv_row == j), break, end
          % row and column sets involved in current transformation
       rows = (piv_row : n); cols = (j+1:piv_row-1);
          % generate HH matrix to annihilate A(piv_row+1:n,j)
       [ v, beta, gamma ] = gen_hh( A( rows, j), tau );
          % update jth row and column of A
       A( rows, j) = zeros(size(rows')); A(piv_row, j) = gamma;
       A( j, rows) = zeros(size(rows));  A(j, piv_row) = gamma;
          % if the reduction is not "skipped", perform symmetric
          % update of A, update Q if required, and shift the pivot row
       if ( beta ~= 0)
         if( cols~= [] )
            A(rows, cols) =  pre_hh( beta, v, A(rows, cols) );
            A(cols, rows) = post_hh( beta, v, A(cols, rows) );
         end
         A( rows, rows ) = symm_hh( beta, v, A(rows, rows) );
         if( Q ~= [] ), Q(:, rows) = post_hh( beta, v, Q(:, rows ) ); end
       end % beta
          % increase pivot row if A(piv_row,j) is not negligible
       if (abs(A(j,piv_row)) > tau), piv_row = piv_row + 1; end
end % j-loop
if (j == n - b)
   if (piv_row == j+1), block1 = piv_row - 1; else, block1 = n; end
else
   block1 = piv_row-1;
end
return; end
```

FIG. 1. *Nonincreasing row-band-width preserving band-reduction algorithm.*

For simplicity we omitted an optimization in Figure 1—if the reduction of the first column of A results in a band width $\tilde{b}$, say, where $\tilde{b} < b$, due to the small size of entries $a(\tilde{b}+1:n,1)$, we can directly pursue a reduction of the trailing block to nonincreasing band width $\tilde{b}$ in the same fashion as shown above.

If the parameter b is chosen such that $kb < n$, where k is the number of distinct eigenvalues of A, Theorem 2.1 predicts a decoupling of the problem with the leading block being of size no larger than kb. In particular, if b is chosen such that $kb = n/2$, we can expect `bred` to generate two decoupled subproblems of about the same size. We can then recursively divide the problem until the transformed matrix becomes tridiagonal (i.e., $b = 1$). Figure 2 is a serial implementation of tridiagonalization based on this approach. Note that the various subproblems can be dealt with independently and simultaneously. The subroutine `blk_diag`, which is called in `tri_sbr`, is shown in Figure 3 and reduces a matrix to block diagonal form with a given band width.

For example, if we reduce a 12×12 matrix A with only two eigenvalues to band width 3, then no diagonal block can be larger than 6×6. So, if $a(4,1)$, $a(5,2)$, and $a(6,3)$ are all nonzero after the reductions in the first three columns have been completed, then the next three columns must already be reduced, and the (partially reduced) matrix A is of the form

$$\begin{bmatrix}
\times & \times & \times & \times & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\times & \times & \times & \times & \times & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\times & \times & \times & \times & \times & \times & 0 & 0 & 0 & 0 & 0 & 0 \\
\times & \times & \times & \times & \times & \times & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \times & \times & \times & \times & \times & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \times & \times & \times & \times & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \times & \times & \times & \times & \times & \times \\
0 & 0 & 0 & 0 & 0 & 0 & \times & \times & \times & \times & \times & \times \\
0 & 0 & 0 & 0 & 0 & 0 & \times & \times & \times & \times & \times & \times \\
0 & 0 & 0 & 0 & 0 & 0 & \times & \times & \times & \times & \times & \times \\
0 & 0 & 0 & 0 & 0 & 0 & \times & \times & \times & \times & \times & \times \\
0 & 0 & 0 & 0 & 0 & 0 & \times & \times & \times & \times & \times & \times
\end{bmatrix}.$$

As a result, we do not need to perform the reductions that would otherwise have occurred in columns 4 through 6. Compared to the conventional approach, the complexity of the algorithm for the case $k = 2$ is $O(0.55\,n^3)$ for the reduction of A and $O(1.25\,n^3)$ for the update of Q, as compared to $O(4n^3/3)$ for both these operations in the usual approach. The savings for Q are minor since updates at later stages still involve vectors of length n, whereas only diagonal subblocks are affected in A. In addition, we can work in parallel on independent problems. If the estimate k of the number of distinct eigenvalues is inaccurate, the algorithm becomes either the standard eigenvalue algorithm (for $k > n/2$) or the SBR tridiagonalization procedure suggested in [10], but in either case, it will return numerically accurate results.

4. Invariant subspace splitting. The computational cost and the degree of parallelism in the algorithm depend on k, the number of distinct eigenvalues. One particularly intriguing case is matrices that have only two eigenvalues. It is intriguing because they arise in eigensolvers based on variant subspace decompositions [3, 19, 4]. We may assume without loss of generality that the eigenvalues are at 1 and 0 (any other two eigenvalues can be mapped to 0 and 1 by shifting and scaling). The following corollary is a special case of Theorem 2.1.

```
function [A, Q] = tri_sbr( A, k, tau, Q )
%
%  produces an orthogonal-tridiagonal decomposition of
%  a symmetric matrix A such that
%            A_old*Q = Q*A_new + O(tau)
%  where A_new is tridiagonal and Q is orthogonal.
%
%  The number k is a guess at the number of numerically distinct
%  eigenvalues of A.
%
%  Matrices are successively reduced to smaller bandwidth in an
%  attempt to exploit the divide-and-conquer nature becoming
%  apparent in the successive bandreduction algorithm when the number
%  k chosen is a good guess at the actual number of numerically distinct
%  eigenvalues.

[m, n] = size(A); if( m ~= n ) error('non-square A'); end

b = max( floor(n/(2*k)), 1 );

[A, block1, Q] = bred( A, b, tau, Q );

if (block1 == n) % If problem didn't decouple, just reduce to
                 % tridiagonal form
   [A,blkvec,Q] = blk_diag(A,1,tau,Q); return;
else
 if( b > 1 )     % first subproblem is not tridiagonal yet
   sub = 1:block1; V = eye(block1);
   [ A(sub,sub), V ] = tri_sbr( A( sub, sub), k, tau, V );
   Q(:,sub) = Q(:,sub) * V;
 end;
 if( n-block1 > 2 )  % second subproblem is nontrivial
   sub = (block1+1):n; V = eye(n-block1);
   [ A(sub, sub), V ] = tri_sbr( A(sub, sub), k, tau, V );
   Q(:,sub) = Q(:,sub) * V;
 end
end

return;
end
```

FIG. 2. *Divide-and-conquer tridiagonalization.*

COROLLARY 4.1. *Let A be a matrix with two distinct eigenvalues, and let $A = Q^T T Q$ be a tridiagonalization of A. Then T is block diagonal with diagonal blocks of size at most 2×2.*

Corollary 4.1 implies that one can determine the range space, $\mathcal{R}(A)$, and the null space, $\mathcal{N}(A)$, in essence via a tridiagonalizing of A. Let $AQ = QT$ be the orthogonal–tridiagonal decomposition of A. For a 1×1 diagonal block $T(j,j)$,

$$Q(:,j) \in \mathcal{R}(A) \text{ if } T(j,j) = 1, \quad \text{and} \quad Q(:,j) \in \mathcal{N}(A) \text{ if } T(j,j) = 0.$$

```
function [ A, blkvec, Q ] = blk_diag( A, b, tau, Q )
%
% Given a symmetric matrix A, a desired bandwidth b, and a threshold tau,
%          [ A, bvec, Q ] = blk_diag( A, b, tau, Q )
% produces an orthogonal-block-diagonal decomposition
%           A_input * W = W * A_output + O(tau)
% where O(tau) denotes a matrix whose norm is of order tay, and
% W is an orthogonal matrix.
%
% A_output will be a block diagonal matrix with each block banded with
% nonincreasing bandwidth b. The i-th diagonal block starts
% at (blkvec(i), blkvec(i)).
%
% If Q is not the empty matrix on input, Q is postmultiplied by W,
%    i.e.,, Q_output = Q_input * W.

[m, n] = size(A); if( m ~= n ) error('non-square A'); end

j = 1; blkvec = [];
while( j < n )
  blkvec = [ blkvec j ]; rows = j:n; cols = j:n;
  [A(rows, cols), dj, Q(:,cols) ] = bred( A(rows,cols), b, tau, Q(:,cols));
  j = j + dj;
end

return; end
```

FIG. 3. *Reduction to block diagonal form.*

Since the eigenvalues of A and T are the same, a 2×2 diagonal block $T(j{:}j{+}1, j{:}j{+}1)$ must have eigenvalues 0 and 1. Because the trace is the sum of the eigenvalues and the off diagonal entry is nonzero, we have

$$T(j{:}j{+}1, j{:}j{+}1) = \begin{pmatrix} 1-\gamma & \mu \\ \mu & \gamma \end{pmatrix},$$

where $\mu \neq 0$ and $0<\gamma<1$. Since

$$\begin{pmatrix} 1-\gamma & \mu \\ \mu & \gamma \end{pmatrix} = \frac{1}{\gamma} \begin{pmatrix} \mu & \gamma \\ \gamma & -\mu \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \mu & \gamma \\ \gamma & -\mu \end{pmatrix}^T,$$

we conclude that

$$Q(:, j{:}j{+}1) \begin{pmatrix} \mu \\ \gamma \end{pmatrix} \in \mathcal{R}(A), \quad \text{and} \quad Q(:, j{:}j{+}1) \begin{pmatrix} \gamma \\ -\mu \end{pmatrix} \in \mathcal{N}(A).$$

One can see that the separation of the range and null subspaces of A, and in fact its eigenvalue decomposition, can be effected by diagonalizing (potentially in parallel) the 2×2 subproblems still occurring in the block tridiagonal decomposition.

In the presence of rounding errors, a computed tridiagonal matrix may not, however, exhibit the block structure we could expect from Corollary 4.1 due to perturbations in the eigenvalues. That is, $\lambda(T) \subset \{[-\nu, \nu] \cup [1-\nu, 1+\nu]\}$, and a repeated eigenvalue numerically manifests itself as an eigenvalue cluster.

Example 4.2. The matrix

$$T = \begin{pmatrix} 1 & e_1 & & & & & \\ e_1 & 0 & e_2 & & & & \\ & e_2 & 1 & e_3 & & & \\ & & e_3 & 0 & e_4 & & \\ & & & e_4 & \ddots & & \\ & & & & & 1 & e_{n-1} \\ & & & & & e_{n-1} & 0 \end{pmatrix},$$

where $e_j = O(\sqrt{\epsilon})$, has eigenvalues $\lambda(T) \subset \{[-\nu, \nu] \cup [1-\nu, 1+\nu]\}$ with $\nu = O(\epsilon)$.

Hence, it seems as if for numerically relevant computations, we now would be faced with computing the eigenvalue decomposition of a tridiagonal matrix. This is not the case, however. By exploiting the special structure of the tridiagonal matrix, we can diagonalize it in two "sweeps" which compute the eigendecomposition of all "even" or "odd" 2×2 blocks on the diagonal (simultaneously), respectively. As we show in what follows, the fill-ins generated by these sweeps are of the same order as the perturbation in the eigenvalues and hence can be considered negligible.

LEMMA 4.3. *Let T be a symmetric tridiagonal matrix with*

$$\lambda(T) \subset [-\nu, \nu] \cup [1-\nu, 1+\nu],$$

where $\nu \stackrel{\text{def}}{=} \max_{\lambda\in\lambda(T)}\{\min(|\lambda-1|, |\lambda|)\} \ll 1$. Then $\|T^2-T\|_2 \le \bar{\nu}$, where

$$\bar{\nu} \stackrel{\text{def}}{=} \nu + \nu^2. \tag{5}$$

Proof. Let Q be orthogonal and $E = \operatorname{diag}(E_1, E_0)$ be diagonal, respectively, such that

$$T = Q \begin{pmatrix} I + E_1 & \\ & E_0 \end{pmatrix} Q^T.$$

Then $\|E\|_2 = \nu$, and

$$T^2 = T + Q\left(\begin{pmatrix} E_1 & \\ & -E_2 \end{pmatrix} + E^2\right) Q^T.$$

Thus, $\|T^2-T\|_2 \le \| \, |E|+E^2\|_2 = \bar{\nu}$. □

The next lemma gives bounds on the elements of the Givens rotation we will choose to diagonalize a 2×2 block and minimize the size of fill-ins.

LEMMA 4.4. *Let $G = \left(\begin{smallmatrix} c & s \\ -s & c \end{smallmatrix}\right)$ be a Givens rotation that diagonalizes a 2×2 symmetric matrix $\left(\begin{smallmatrix} \alpha_1 & \beta \\ \beta & \alpha_2 \end{smallmatrix}\right)$. Assume that without loss of generality $\beta > 0$ and define $\sigma \ge 0$ by*

$$\sigma^2 \stackrel{\text{def}}{=} \left(\frac{\alpha_1 - \alpha_2}{2}\right)^2 + \beta^2. \tag{6}$$

Then s and c can be chosen such that

$$0 \le |s| \le \frac{\beta}{\sqrt{2}\,\sigma} \quad \text{and} \quad \frac{1}{\sqrt{2}} \le c \le 1.$$

Proof. Let $c = \cos(\theta)$ and $s = \sin(\theta)$. Since we want to eliminate the off diagonal elements in $G\left(\begin{smallmatrix}\alpha_1 & \beta \\ \beta & \alpha_2\end{smallmatrix}\right)G^T$, we obtain

$$0 = (c^2 - s^2)\beta + 2cs\left(\frac{\alpha_2 - \alpha_1}{2}\right) = \beta\cos(2\theta) - \left(\frac{\alpha_2 - \alpha_1}{2}\right)\sin(2\theta).$$

If we choose

$$\cos(2\theta) = \frac{|\alpha_1 - \alpha_2|}{2\sigma}, \tag{7}$$

with σ as defined in (6), then

$$s^2 = \frac{1 - \cos(2\theta)}{2} = \frac{\beta^2}{2\sigma(\sigma + |\alpha_1 - \alpha_2|/2)},$$

and

$$c^2 = \frac{1 + \cos(2\theta)}{2} = \frac{\sigma + |\alpha_1 - \alpha_2|/2}{2\sigma},$$

as claimed. □

In the following theorem we now show that, employing these Givens rotations, we can limit the size of the fill-in entries generated when applying these rotations to a tridiagonal matrix with eigenvalue clusters around 0 and 1.

THEOREM 4.5. *Let T and $\bar{\nu}$ be as in Lemma 4.3. Let $G = \mathrm{diag}(I, \left(\begin{smallmatrix} c & s \\ -s & c\end{smallmatrix}\right), I)$ be the Givens rotation that diagonalizes one 2×2 diagonal block of T; i.e.,*

$$G \cdot \begin{pmatrix} \ddots & \overline{\beta} & & \\ \overline{\beta} & \alpha_1 & \beta & \\ & \beta & \alpha_2 & \underline{\beta} \\ & & \underline{\beta} & \ddots \end{pmatrix} \cdot G^T = \begin{pmatrix} \ddots & * & \gamma & \\ * & \tilde{\alpha_1} & 0 & \delta \\ \gamma & 0 & \tilde{\alpha_2} & * \\ & \delta & * & \ddots \end{pmatrix},$$

where we assume that $\beta > 0$ without loss of generality. If $\beta > \sqrt{7}\,\bar{\nu}$ and c and s are chosen as suggested in Lemma 4.4, then

$$\gamma \leq \sqrt{7}\,\bar{\nu} \quad \textit{and} \quad \delta \leq \sqrt{7}\,\bar{\nu}.$$

Proof. Comparing corresponding entries in T^2 and T and invoking Lemma 4.3, we know that there exist $\bar{\epsilon}$, $\underline{\epsilon}$, and ϵ_o, $|\bar{\epsilon}|, |\underline{\epsilon}|, |\epsilon_o| \leq \bar{\nu}$, such that

$$\beta(\alpha_1 + \alpha_2) = \beta + \epsilon_o, \tag{8}$$

$$\overline{\beta}^2 + \alpha_1^2 + \beta^2 = \alpha_1 + \bar{\epsilon}, \tag{9}$$

$$\beta^2 + \alpha_2^2 + \underline{\beta}^2 = \alpha_2 + \underline{\epsilon}, \tag{10}$$

$$\overline{\beta}\beta \leq \bar{\nu}, \quad \underline{\beta}\beta \leq \bar{\nu}. \tag{11}$$

Using these identities, we have

$$\beta^2 - \alpha_1\alpha_2 = \frac{\alpha_1 + \alpha_2}{2}(1 - (\alpha_1 + \alpha_2)) + \frac{(\underline{\epsilon} + \bar{\epsilon}) - (\underline{\beta}^2 + \overline{\beta}^2)}{2},$$

and hence we can express σ^2 defined as in (6) as

$$\begin{aligned}\sigma^2 &= \left(\frac{\alpha_1+\alpha_2}{2}\right)^2 + (\beta^2 - \alpha_1\alpha_2) \\ &= \frac{1}{4}\left(1 - \frac{\epsilon_o^2}{\beta^2}\right) + \frac{\overline{\epsilon}+\underline{\epsilon}}{2} - \frac{(\overline{\beta}\beta)^2 + (\underline{\beta}\beta)^2}{2\beta^2}.\end{aligned}$$

Thus,

$$\sigma^2 \geq \frac{1}{4} - \bar{\nu} - \frac{5}{4}\left(\frac{\bar{\nu}}{\beta}\right)^2.$$

Now let $\tau \geq 1$ be chosen such that $\beta > \tau\bar{\nu}$. Then

$$\sigma^2 \geq \frac{1}{4} - \bar{\nu} - \frac{5}{4}\left(\frac{1}{\tau}\right)^2. \tag{12}$$

Equations (11) together with $s \leq \frac{\beta}{\sqrt{2}\sigma}$ imply that

$$\gamma = s\overline{\beta} \leq \frac{\bar{\nu}}{\sqrt{2}\sigma} \quad \text{and} \quad \delta = s\underline{\beta} \leq \frac{\bar{\nu}}{\sqrt{2}\sigma}.$$

Using (12), it is now easy to show that $\tau \geq \sqrt{7}$ implies $\frac{1}{\sqrt{2}\sigma} \leq \tau$ and hence the result of the theorem. □

As a consequence of Theorem 4.5, we are then able to compute the eigenvalue decomposition of a 2×2 diagonal block in a tridiagonal matrix T with eigenvalue clusters at 0 and 1 such that the generated fill-in is negligible compared to the eigenvalue perturbation. Thus, the diagonalization of T can be done by two sweeps of (potentially concurrent) 2×2 eigenvalue problems, as shown in Figure 4. In the first sweep, we diagonalize an "odd–even" 2×2 problem if the off diagonal entry is not too small, and set the fill entries to zero, or otherwise just zero the off diagonal entry. In the second sweep, we diagonalize the "even–odd" blocks. Since no more rotations follow, there is no need to zero out fill-in entries.

Theorem 4.5 shows that the Frobenius norm of the fill-in matrix introduced by the algorithm `rr_diag` shown in Figure 4 is bounded by $3\sqrt{n}\bar{\nu}$, which is of the same order as the perturbation in eigenvalues. The subroutine `diag2`, which is not shown here, computes the diagonalizing rotations as outlined in Lemma 4.4. Hence, Algorithm `rr_diag` is as numerically reliable as any other approach for diagonalizing T, albeit much cheaper due to its exploitation of the special structure of T.

5. Numerical experiments. In this section we report on some numerical experiments with the algorithms presented in this paper. All experiments were performed with Matlab Version 4.2a on a Sun Sparcstation iPX. For the reader wishing to experiment on his or her own, the Matlab files employed to generate these results can be retrieved via anonymous ftp from the `pub/prism` directory at `ftp.super.org`.

First, we apply the band-reduction algorithm `bred` of Figure 1 recursively to the trailing subblock of a 200×200 matrix with two eigenvalue clusters of size 50, each at $\lambda = \{-1, -2, 0, 1\}$. The radius of each cluster is $\epsilon\, 1.0e^3$, where ϵ is the machine precision. The drop threshold `tau` in `bred` is set to $\sqrt{7}\,\epsilon\, 1.0e^3$, and at each step the band width is chosen so as to decouple the problem in the middle. The succession of matrices generated is shown in Figure 5. The caption of each picture shows the

```
function [Q, D] = rr_diag( A, Q, tau )
%
% Given a tridiagonal matrix A with eigenvalues 1 and 0, with
% lambda(A) contained in [1-tau,1+tau] or [-tau,tau]
% rr_diag computes an approximate eigendecomposition
%        D = Q' * A * Q
% where
%       || D - Q'* A * Q||_Frobenius <= sqrt(7*n)*tau*(1+tau)

[m,n] = size(A); if( m~=n ) error('non-square A'); end
drop_threshold = sqrt(7)*tau*(1+tau);

for j = 1:2:floor(n/2)*2          % diagonalize all (odd-even)
  k = j:j+1;                      % diagonal 2x2 matrices
  if (abs(A(j+1,j)) > drop_threshold)
    [G A(k,k)] = diag2( A(k,k) );
    if( j+2 <= n )
      A(j+2,k) = A(j+2,k)*G; A(k,j+2) = G'*A(k,j+2);
      A(j+2,j) = 0; A(j,j+2) = 0; % zero out negligible fill-ins
    end
    if( j-1 >= 1 )
      A(j-1,k) = A(j-1,k)*G; A(k,j-1) = G'*A(k,j-1);
      A(j-1,j+1) = 0; A(j+1,j-1) = 0;
    end
    Q(:,k) = Q(:, k)*G;
  end
end
for j = 2:2:floor((n-1)/2)*2      % diagonalize all (even-odd)
  k = j:j+1;                      % diagonal 2x2 matrices
  if (abs(A(j+1,j)) > drop_threshold)
    [G A(k,k)] = diag2( A(k,k) );
    if( j+2 <= n )
      A(j+2,k) = A(j+2,k)*G; A(k,j+2) = G'*A(k,j+2);
   % no more need to zero fill-ins
    end
    if( j-1 >= 1 )
      A(j-1,k) = A(j-1,k)*G; A(k,j-1) = G'*A(k,j-1);
    end
    Q(:,k) = Q(:, k)*G;
  end
end
D = diag(diag(A));
return; end
```

FIG. 4. *Diagonalization of a tridiagonal matrix with eigenvalue clusters at* 0 *and* 1.

current matrix size being worked on and the band width to which it is to be reduced. At each step, we compute the residual

$$\delta \stackrel{\text{def}}{=} \|A_{original} * Q - Q * A_{current}\|_2.$$

We observe that $\delta \approx 7.2e^{-13}$, which, given a machine precision $\epsilon = 2.2e^{-16}$, is consis-

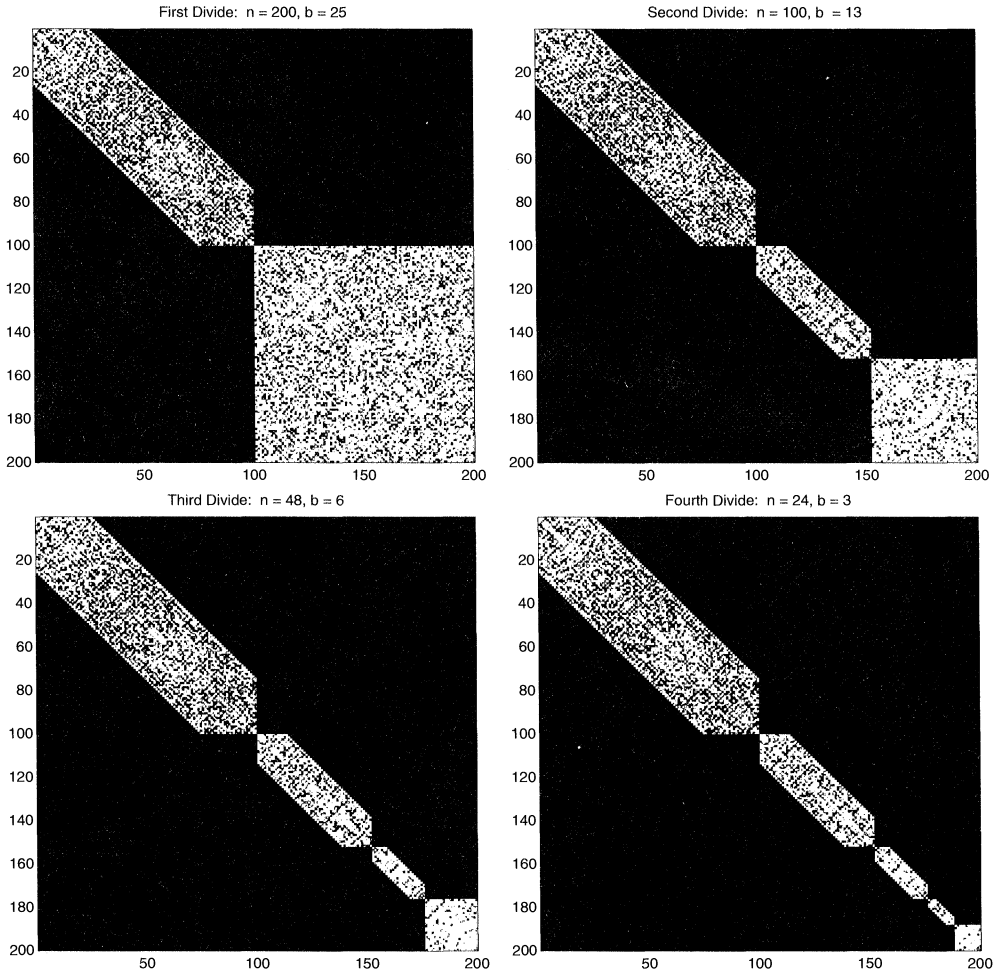


FIG. 5. *Band reduction applied to trailing subblock of a* 200×200 *matrix with four distinct eigenvalue clusters.*

tent with our theory.

The same experiment, employing a matrix with 100 eigenvalues at 0 and 1 each and using the same eigenvalue perturbation and drop threshold, is shown in Figure 6. Note that it is sufficient to reduce the matrix to half the band width chosen in Figure 5 to achieve decoupling. We observe that $\delta \approx 2.7e^{-13}$. We also note that in both cases, the first, third, and fourth splits occurred at row (and column) 100, 176, and 188, respectively. The second split occurred at row 152 for Figure 5 and at row 150 for Figure 6.

To test the behavior of our rank-revealing tridiagonalization (RRDG), we compare it with the standard eigenvalue decomposition (EIG) and the QR factorization with column pivoting (QR); the results are presented in Table 1 and Table 2. Our test matrices are

1. tridiagonal matrices with eigenvalue clusters of radius $p\,\epsilon$ generated by inserting random off diagonal perturbations of the order $\sqrt{p\,\epsilon}$ in the matrix shown in Example 4.2, and

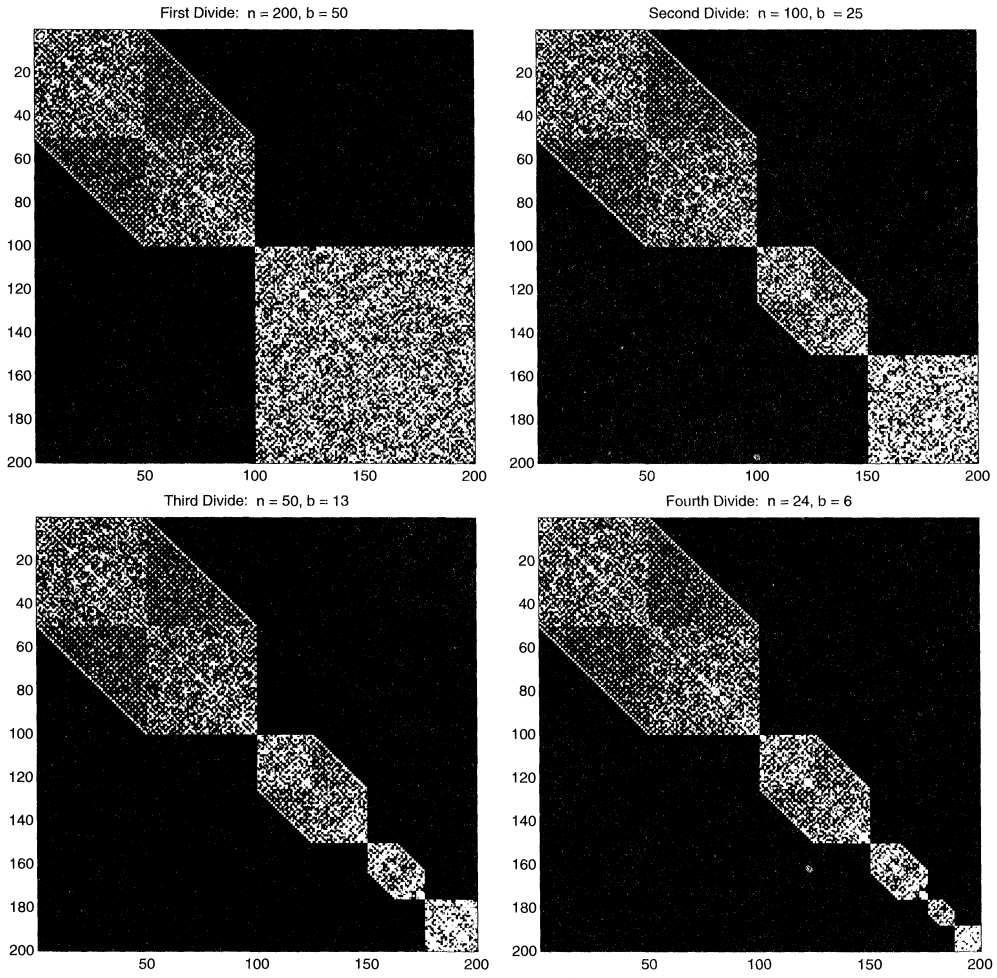


FIG. 6. *Band reduction applied to trailing subblock of a* 200×200 *matrix with two distinct eigenvalue clusters.*

2. matrices generated by symmetrically multiplying the matrices from Example 4.2 with orthogonal matrices generated via the QR factorization of a random matrix.

In the first case, we call `rr_diag`, listed in Figure 4. In the second case, we precede the call to `rr_diag` by a call to `tri_sbr`, as shown in Figure 2. The drop threshold for the divide-and-conquer tridiagonalization is set to $\sqrt{7}\,p\,\epsilon$, which is the same threshold as that employed in the two final diagonalization sweeps. For each of $p = 1, 10, 100$, we run 50 test cases each with matrix sizes 125, 250, and 375. RRDG and EIG both compute an eigenvalue decomposition $Q^T A Q = D$ with D diagonal. We compute $\tilde{D} \stackrel{\text{def}}{=} \text{round}(D)$, i.e., round each diagonal entry to the nearest integer, and we report both the relative eigenvalue residual $\|Q^T A - \tilde{D} Q\|_F / \sqrt{n/2}$ (in Table 1) as well as the relative orthogonality residual $\|Q^T Q - I\|_F / \sqrt{n}$ (in Table 2). Note that $\sqrt{n/2}$ is an estimate of $\|A\|_F$. In the case of the QR factorization with pivoting, which computes $AP = QR$ for a permutation matrix P and an upper triangular matrix R, we compute

TABLE 1
Relative residual in subspace splitting.

Tridiagonal Matrices				Full Matrices			
n	$\text{RRDG}_{\max}$	$\text{EIG}_{\max}$	$\text{QR}_{\max}$	n	$\text{RRDG}_{\max}$	$\text{EIG}_{\max}$	$\text{QR}_{\max}$
$p = 1$				$p = 1$			
125	5.3e−16	1.6e−15	1.7e−15	125	5.3e−14	1.7e−14	1.4e−14
250	5.0e−16	1.6e−15	3.8e−15	250	1.5e−13	3.3e−14	3.7e−14
375	4.9e−16	1.5e−15	5.6e−15	375	2.4e−14	3.8e−14	5.5e−14
$p = 10$				$p = 10$			
125	3.5e−15	4.2e−15	2.2e−15	125	5.0e−15	6.0e−15	1.6e−14
250	3.3e−15	4.9e−15	5.1e−15	250	5.5e−15	3.0e−14	4.0e−14
375	3.4e−15	4.5e−15	4.3e−15	375	6.1e−15	4.1e−14	4.8e−14
$p = 100$				$p = 100$			
125	3.3e−14	3.3e−14	2.7e−15	125	4.6e−14	3.5e−14	1.4e−14
250	3.2e−14	3.2e−14	6.8e−15	250	4.5e−14	5.2e−14	3.9e−14
375	3.2e−14	4.4e−14	6.6e−15	375	4.2e−14	3.2e−14	4.9e−14
$p = 1000$				$p = 1000$			
125	3.3e−13	3.3e−13	2.5e−15	125	4.6e−13	3.5e−13	1.6e−14
250	3.2e−13	3.2e−13	4.1e−15	250	4.4e−13	3.4e−13	3.6e−14
375	3.2e−13	3.2e−13	6.2e−15	375	4.2e−13	3.2e−13	4.2e−14

TABLE 2
Relative residual in orthogonality.

Tridiagonal Matrices				Full Matrices			
n	$\text{RRDG}_{\max}$	$\text{EIG}_{\max}$	$\text{QR}_{\max}$	n	$\text{RRDG}_{\max}$	$\text{EIG}_{\max}$	$\text{QR}_{\max}$
$p = 1$				$p = 1$			
125	2.3e−16	1.2e−15	1.1e−15	125	2.1e−15	1.2e−14	1.7e−15
250	2.2e−16	1.3e−15	1.3e−15	250	3.0e−15	2.4e−14	2.4e−15
375	2.1e−16	1.2e−15	1.3e−15	375	3.6e−15	2.7e−14	2.8e−15
$p = 10$				$p = 10$			
125	3.0e−16	2.8e−15	1.1e−15	125	1.4e−15	1.1e−14	1.7e−15
250	2.8e−16	3.0e−15	1.4e−15	250	1.9e−15	2.1e−14	2.3e−15
375	2.8e−16	2.8e−15	1.6e−15	375	3.4e−15	2.9e−14	2.9e−15
$p = 100$				$p = 100$			
125	3.4e−16	1.1e−14	1.3e−15	125	1.4e−15	1.1e−14	1.7e−15
250	3.2e−16	2.0e−14	1.4e−15	250	1.9e−15	2.2e−14	2.4e−15
375	3.1e−16	1.9e−14	1.7e−15	375	2.3e−15	2.6e−14	2.9e−15
$p = 1000$				$p = 1000$			
125	3.2e−16	1.0e−14	1.2e−15	125	1.4e−15	1.3e−14	1.8e−15
250	3.1e−16	2.3e−14	1.4e−15	250	1.9e−15	2.4e−14	2.4e−15
375	3.2e−16	3.3e−14	1.6e−15	375	2.3e−15	3.3e−14	2.9e−15

the rank

$$r \stackrel{\text{def}}{=} \max_i |r_{ii}| > \sqrt{7}\, p\, \epsilon$$

and $\tilde{A} \stackrel{\text{def}}{=} Q^T * A * Q$. We then report

$$|\|\tilde{A}(1:r, 1:r)\|_F - \|A\|_F| / \sqrt{n/2},$$

which should be small since $Q(1:r,:)$ is a basis for the range space of A. For each case, we report the worst residual.

We see that the divide-and-conquer tridiagonalization, followed by the two clean up sweeps over the resulting tridiagonal matrix, performs just as well as a full-fledged eigenvalue decomposition. In both cases, the residual in the subspace splitting is of $O(p\,\epsilon)$, as expected. The residual for QR factorization does not include the perturbation at the eigenvalue 1 as the other two approaches do and therefore is smaller in all cases. In any case, the computed orthogonal matrices are orthogonal up to machine precision. The Q computed by the `eig` function in Matlab is slightly less orthogonal

since `eig` involves more transformations and as a result accumulates more rounding errors. Note that all three approaches are worse for a full matrix in the case $p = 1$. This is due to the fact that the roundoff errors in the orthogonal reductions are of the same order of machine precision. When p is bigger, the roundoff errors are dominated by the perturbation in the eigenvalues, and hence RRDG and EIG behave about the same for tridiagonal and full matrices.

6. Conclusions. This paper introduced an algorithm for reducing a symmetric matrix with repeated eigenvalues to tridiagonal form. The algorithm progresses through a series of band reductions, each band-reduction stage forcing a decoupling of the band matrix into independent subblocks. Compared to the usual Householder tridiagonalization procedure, this approach can save up to 50% of the floating-point operations. We also developed a robust and inexpensive numerical procedure for diagonalizing the resulting tridiagonal matrix in the case where the matrix has only two eigenvalue clusters around 0 and 1. This case arises in eigenvalue decomposition algorithms based on invariant subspace approaches. Taken together, these two algorithms allow for a very efficient diagonalization of such matrices.

The algorithm can be generalized immediately to the reduction of unsymmetric matrices to Hessenberg form. The same irreducibility argument underlying Theorem 2.1 goes through for Hessenberg matrices. We also note that in exact arithmetic, conjugate transposed eigenvalue pairs would end up in the same block. However, since one triangle of a Hessenberg matrix is still full, the potential for computational savings is greatly reduced.

We mention that, apart from its divide-and-conquer nature and the resulting potential for parallelism, as well as its reduced operation count, our divide-and-conquer algorithm has another attractive feature. Since our algorithm, at least in the early stages, reduces matrices to banded form with a relatively wide band, it is easy to block the Householder transformations using the WY representation [11] or the compact WY representation [20], as has been described, for example, in [17]. In this fashion, one can easily capitalize on the favorable memory transfer characteristics of block algorithms.

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, PA, 1992.

[2] ———, *LAPACK User's Guide Release* 2.0, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1994.

[3] L. Auslander and A. Tsao, *A Divide-and-Conquer Algorithm for the Eigenproblem Via Complementary Invariant Subspace Decomposition*, Tech. report SRC-TR-89-003, Supercomputing Research Center, Institute for Defense Analysis, Bowie, MD, 1989.

[4] Z. Bai and J. Demmel, *Design of a parallel nonsymmetric eigenroutine toolbox, Part* I, in Proc. of the Sixth SIAM Conference on Parallel Processing for Scientific Computing, 1993, pp. 391–398.

[5] C. Bischof, G. Corliss, and A. Griewank, *Structured second- and higher-order derivatives through univariate Taylor series*, Optim. Meth. Software, 2 (1993), pp. 211–232.

[6] C. Bischof, S. Huss-Lederman, X. Sun, and A. Tsao, *The PRISM project: Infrastructure and algorithms for parallel eigensolvers*, in Proc. of the Scalable Parallel Libraries Conference, Washington, DC, 1994, IEEE Computer Society, pp. 123–131.

[7] C. Bischof, X. Sun, and B. Lang, *Parallel tridiagonalization through two-step band reduction*, in Proc. of Scalable High Performance Computing Conference, Knoxville, TN, 1994, IEEE Computer Society Press, pp. 23–27.

[8] C. H. BISCHOF, *Fundamental Linear Algebra Computations on High-Performance Computers*, Informatik Fachberichte, Vol. 250, Springer-Verlag, Berlin, 1990.

[9] C. H. BISCHOF AND J. J. DONGARRA, *A project for developing a linear algebra library for high-performance computers*, in Parallel and Vector Supercomputing: Methods and Algorithms, Graham Carey, ed., John Wiley, Somerset, NJ, 1989.

[10] C. H. BISCHOF AND X. SUN, *A Framework for Band Reduction and Tridiagonalization of Symmetric Matrices*, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 1992, Preprint MCS-P298-0392.

[11] C. H. BISCHOF AND C. F. VAN LOAN, *The WY representation for products of Householder matrices*, SIAM J. Sci. Stat. Comput., 8 (1987), pp. s2–s13.

[12] H. CHANG, S. UTKU, M. SALAMA, AND D. RAPP, *A parallel Householder tridiagonalization stratagem using scattered square decomposition*, Parallel Comput., 6 (1988), pp. 297–311.

[13] J. DONGARRA AND R. VAN DE GEIJN, *Reduction to condensed form for the eigenvalue problem on distributed-memory architectures*, Parallel Comput., 18 (1992), pp. 973–982.

[14] J. DONGARRA AND S. HAMMARLING, *Evolution of Numerical Software for Dense Linear Algebra*, Oxford University Press, Oxford, UK, 1989.

[15] J. J. DONGARRA, S. J. HAMMARLING, AND D. C. SORENSEN, *Block Reduction of Matrices to Condensed Form for Eigenvalue Computations*, Tech. report MCS–TM–99, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 1987.

[16] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

[17] R. G. GRIMES AND H. D. SIMON, *Solution of large, dense symmetric generalized eigenvalue problems using secondary storage*, ACM Trans. Math. Software, 14, 3 (1988), pp. 241–256.

[18] A. S. HOUSEHOLDER, *The Theory of Matrices in Numerical Analysis*, Dover, New York, 1964.

[19] S. LEDERMAN, A. TSAO, AND T. TURNBULL, *A Parallelizable Eigensolver for Real Diagonalizable Matrices with Real Eigenvalues*, Tech. report TR-91-042, Supercomputing Research Center, Institute for Defense Analysis, Bowie, MD, 1991.

[20] R. SCHREIBER AND C. VAN LOAN, *A storage efficient WY representation for products of Householder transformations*, SIAM J. Sci. Stat. Comput., 10 (1989), pp. 53–57.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 4, pp. 886–905, October 1996

011

AN APPROXIMATE MINIMUM DEGREE ORDERING ALGORITHM*

PATRICK R. AMESTOY†, TIMOTHY A. DAVIS‡, AND IAIN S. DUFF§

Abstract. An approximate minimum degree (AMD) ordering algorithm for preordering a symmetric sparse matrix prior to numerical factorization is presented. We use techniques based on the quotient graph for matrix factorization that allow us to obtain computationally cheap bounds for the minimum degree. We show that these bounds are often equal to the actual degree. The resulting algorithm is typically much faster than previous minimum degree ordering algorithms and produces results that are comparable in quality with the best orderings from other minimum degree algorithms.

Key words. approximate minimum degree ordering algorithm, quotient graph, sparse matrices, graph algorithms, ordering algorithms

AMS subject classifications. 65F50, 65F05

1. Introduction.

When solving large sparse symmetric linear systems of the form $\mathbf{Ax} = \mathbf{b}$, it is common to precede the numerical factorization by a symmetric reordering. This reordering is chosen so that pivoting down the diagonal in order on the resulting permuted matrix $\mathbf{PAP}^T = \mathbf{LL}^T$ produces much less fill-in and work than computing the factors of $\mathbf{A}$ by pivoting down the diagonal in the original order. This reordering is computed using only information on the matrix structure without taking account of numerical values and so may not be stable for general matrices. However, if the matrix $\mathbf{A}$ is positive definite [21], a Cholesky factorization can safely be used. This technique of preceding the numerical factorization with a symbolic analysis can also be extended to unsymmetric systems, although the numerical factorization phase must allow for subsequent numerical pivoting [1, 2, 16]. The goal of the preordering is to find a permutation matrix $\mathbf{P}$ so that the subsequent factorization has the least fill-in. Unfortunately, this problem is NP-complete [31], so heuristics are used.

The minimum degree ordering algorithm is one of the most widely used heuristics, since it produces factors with relatively low fill-in on a wide range of matrices. Because of this, the algorithm has received much attention over the past three decades. The algorithm is a symmetric analogue of Markowitz's method [26] and was first proposed by Tinney and Walker [30] as algorithm S2. Rose [27, 28] developed a graph theoretical model of Tinney and Walker's algorithm and renamed it the minimum degree algorithm, since it performs its pivot selection by choosing from a graph a node of minimum degree. Later implementations have dramatically improved the time and memory requirements of Tinney and Walker's method, while maintaining the basic idea of selecting a node or set of nodes of minimum degree. These improvements have

* Received by the editors December 19, 1994; accepted for publication (in revised form) by J. W. H. Liu November 15, 1995.

† ENSEEIHT-IRIT, Toulouse, France (amestoy@enseeiht.fr).

‡ Computer and Information Sciences Department, University of Florida, Gainesville, FL 32611-6120 (davis@cise.ufl.edu). Technical reports and matrices are available via the World Wide Web at http://www.cise.ufl.edu/~davis or by anonymous ftp at ftp.cise.ufl.edu:pub/tech-reports. Support for this project was provided by National Science Foundation grants ASC-9111263 and DMS-9223088. Portions of this work were supported by a postdoctoral grant from CERFACS.

§ Rutherford Appleton Laboratory, Chilton, Didcot, Oxon. 0X11 0QX, England, and European Center for Research and Advanced Training in Scientific Computation (CERFACS), Toulouse, France (isd@letterbox.rl.ac.uk). Technical reports, information on the Harwell Subroutine Library, and matrices are available via the World Wide Web at http://www.cis.rl.ac.uk/struct/ARCD/NUM.html or by anonymous ftp at seamus.cc.rl.ac.uk/pub.

reduced the memory complexity so that the algorithm can operate within the storage of the original matrix, and have reduced the amount of work needed to keep track of the degrees of nodes in the graph (which is the most computationally intensive part of the algorithm). This work includes that of Duff, Erisman, and Reid [10]; Duff and Reid [13, 14, 15]; George and McIntyre [23]; Eisenstat, et al. [17]; Eisenstat, Schultz, and Sherman [18]; George and Liu [19]–[22]; and Liu [25]. More recently, several researchers have relaxed this heuristic by computing upper bounds on the degrees, rather than the exact degrees, and selecting a node of minimum upper bound on the degree. This work includes that of Gilbert, Moler, and Schreiber [24] and Davis and Duff [8, 7]. Davis and Duff use degree bounds in the unsymmetric-pattern multifrontal method (UMFPACK), an unsymmetric Markowitz-style algorithm. In this paper, we describe an approximate minimum degree ordering algorithm based on the symmetric analogue of the degree bounds used in UMFPACK.

Section 2 presents the original minimum degree algorithm of Tinney and Walker in the context of the graph model of Rose. Section 3 discusses the quotient graph (or element graph) model and the use of this model to reduce the time taken by the algorithm. In this context, we present our notation for the quotient graph and present a small example matrix and its graphs. We then use the notation to describe our approximate degree bounds in §4. The approximate minimum degree (AMD) algorithm and its time complexity are presented in §5. In §6, we first analyse the performance and accuracy of our approximate degree bounds on a set of test matrices from a wide range of disciplines. The AMD algorithm is then compared with other established codes that compute minimum degree orderings.

Throughout this paper, we will use the superscript k to denote a graph, set, or other structure obtained after the first k pivots have been chosen and eliminated. For simplicity, we will drop the superscript when the context is clear.

2. Elimination graphs. The nonzero pattern of a symmetric n-by-n matrix $\mathbf{A}$ can be represented by a graph $G^0 = (V^0, E^0)$, with nodes $V^0 = \{1, \ldots, n\}$ and edges E^0. An edge (i, j) is in E^0 if and only if $a_{ij} \neq 0$ and $i \neq j$. Since $\mathbf{A}$ is symmetric, G^0 is undirected.

The elimination graph $G^k = (V^k, E^k)$ describes the nonzero pattern of the submatrix still to be factorized after the first k pivots have been chosen and eliminated. It is undirected, since the matrix remains symmetric as it is factorized. At step k, the graph G^k depends on G^{k-1} and the selection of the kth pivot. To find G^k, the kth pivot node p is selected from V^{k-1}. Edges are added to E^{k-1} to make the nodes adjacent to p in G^{k-1} a *clique* (a fully connected subgraph). This addition of edges (fill-in) means that we cannot know the storage requirements in advance. The edges added correspond to fill-in caused by the kth step of factorization. A fill-in is a nonzero entry $\mathbf{L}_{ij}$, where $(\mathbf{PAP}^T)_{ij}$ is zero. The pivot node p and its incident edges are then removed from the graph G^{k-1} to yield the graph G^k. Let $\mathrm{Adj}_{G^k}(i)$ denote the set of nodes adjacent to i in the graph G^k. When the kth pivot is eliminated, the graph G^k is given by

$$V^k = V^{k-1} \setminus \{p\}$$

and

$$E^k = \left(E^{k-1} \cup (\mathrm{Adj}_{G^{k-1}}(p) \times \mathrm{Adj}_{G^{k-1}}(p))\right) \cap (V^k \times V^k).$$

The minimum degree algorithm selects node p as the kth pivot such that the degree of p, $t_p \equiv |\mathrm{Adj}_{G^{k-1}}(p)|$, is minimum (where $|\ldots|$ denotes the size of a set or

the number of nonzeros in a matrix, depending on the context). The minimum degree algorithm is a nonoptimal greedy heuristic for reducing the number of new edges (fill-ins) introduced during the factorization. We have already noted that the optimal solution is NP-complete [31]. By minimizing the degree, the algorithm minimizes the upper bound on the fill-in caused by the kth pivot. Selecting p as pivot creates at most $(t_p^2 - t_p)/2$ new edges in G.

3. Quotient graphs. In contrast to the elimination graph, the quotient graph models the factorization of $\mathbf{A}$ using an amount of storage that never exceeds the storage for the original graph G^0 [21]. The quotient graph is also referred to as the generalized element model [13, 14, 15, 29]. An important component of a quotient graph is a clique. It is a particularly economic structure since a clique is represented by a list of its members rather than by a list of all the edges in the clique. Following the generalized element model, we refer to nodes removed from the elimination graph as *elements* (George and Liu refer to them as eliminated nodes). We use the term *variable* to refer to uneliminated nodes.

The quotient graph $\mathcal{G}^k = (V^k, \overline{V}^k, E^k, \overline{E}^k)$ implicitly represents the elimination graph G^k, where $\mathcal{G}^0 = G^0$, $V^0 = V$, $\overline{V}^0 = \emptyset$, $E^0 = E$, and $\overline{E}^0 = \emptyset$. For clarity, we drop the superscript k in the following. The nodes in $\mathcal{G}$ consist of variables (the set V) and elements (the set $\overline{V}$). The edges are divided into two sets: edges between variables $E \subseteq V \times V$ and between variables and elements $\overline{E} \subseteq V \times \overline{V}$. Edges between elements are not required since we could generate the elimination graph from the quotient graph without them. The sets $\overline{V}^0$ and $\overline{E}^0$ are empty.

We use the following set notation ($\mathcal{A}$, $\mathcal{E}$, and $\mathcal{L}$) to describe the quotient graph model and our approximate degree bounds. Let $\mathcal{A}_i$ be the set of variables adjacent to variable i in $\mathcal{G}$, and let $\mathcal{E}_i$ be the set of elements adjacent to variable i in $\mathcal{G}$ (we refer to $\mathcal{E}_i$ as element list i). That is, if i is a variable in V, then

$$\mathcal{A}_i \equiv \{j : (i,j) \in E\} \subseteq V,$$

$$\mathcal{E}_i \equiv \{e : (i,e) \in \overline{E}\} \subseteq \overline{V},$$

and

$$\mathrm{Adj}_{\mathcal{G}}(i) \equiv \mathcal{A}_i \cup \mathcal{E}_i \subseteq V \cup \overline{V}.$$

The set $\mathcal{A}_i$ refers to a subset of the nonzero entries in row i of the original matrix $\mathbf{A}$ (thus the notation $\mathcal{A}$). That is, $\mathcal{A}_i^0 \equiv \{j : a_{ij} \neq 0\}$, and $\mathcal{A}_i^k \subseteq \mathcal{A}_i^{k-1}$ for $1 \leq k \leq n$. Let $\mathcal{L}_e$ denote the set of variables adjacent to element e in $\mathcal{G}$. That is, if e is an element in $\overline{V}$, then we define

$$\mathcal{L}_e \equiv \mathrm{Adj}_{\mathcal{G}}(e) = \{i : (i,e) \in \overline{E}\} \subseteq V.$$

The edges E and $\overline{E}$ in the quotient graph are represented using the sets $\mathcal{A}_i$ and $\mathcal{E}_i$ for each variable in $\mathcal{G}$ and the sets $\mathcal{L}_e$ for each element in $\mathcal{G}$. We will use $\mathcal{A}$, $\mathcal{E}$, and $\mathcal{L}$ to denote three sets containing all $\mathcal{A}_i$, $\mathcal{E}_i$, and $\mathcal{L}_e$, respectively, for all variables i and all elements e. George and Liu [21] show that the quotient graph takes no more storage than the original graph ($|\mathcal{A}^k| + |\mathcal{E}^k| + |\mathcal{L}^k| \leq |\mathcal{A}^0|$ for all k).

The quotient graph $\mathcal{G}$ and the elimination graph G are closely related. If i is a variable in G, it is also a variable in $\mathcal{G}$, and

$$\mathrm{Adj}_G(i) = \left(\mathcal{A}_i \cup \bigcup_{e \in \mathcal{E}_i} \mathcal{L}_e \right) \setminus \{i\}, \tag{3.1}$$

where the "\" is the standard set subtraction operator.

When variable p is selected as the kth pivot, element p is formed (variable p is removed from V and added to $\overline{V}$). The set $\mathcal{L}_p = \mathrm{Adj}_G(p)$ is found using equation (3.1). The set $\mathcal{L}_p$ represents a permuted nonzero pattern of the kth column of $\mathbf{L}$ (thus the notation $\mathcal{L}$). If $i \in \mathcal{L}_p$, where p is the kth pivot, and variable i will become the mth pivot (for some $m > k$), then the entry $\mathbf{L}_{mk}$ will be nonzero.

Equation (3.1) implies that $\mathcal{L}_e \setminus \{p\} \subseteq \mathcal{L}_p$ for all elements e adjacent to variable p. This means that all variables adjacent to an element $e \in \mathcal{E}_p$ are adjacent to the element p and these elements $e \in \mathcal{E}_p$ are no longer needed. They are *absorbed* into the new element p and deleted [15], and reference to them is replaced by reference to the new element p. The new element p is added to the element lists $\mathcal{E}_i$ for all variables i adjacent to element p. Absorbed elements $e \in \mathcal{E}_p$ are removed from all element lists.

The sets $\mathcal{A}_p$ and $\mathcal{E}_p$, and $\mathcal{L}_e$ for all e in $\mathcal{E}_p$, are deleted. Finally, any entry j in $\mathcal{A}_i$, where both i and j are in $\mathcal{L}_p$, is redundant and is deleted. The set $\mathcal{A}_i$ is thus disjoint with any set $\mathcal{L}_e$ for $e \in \mathcal{E}_i$. In other words, $\mathcal{A}_i^k$ is the pattern of those entries in row i of $\mathbf{A}$ that are not modified by steps 1 through k of the Cholesky factorization of $\mathbf{PAP}^T$. The net result is that the new graph $\mathcal{G}$ takes the same, or less, storage than before the kth pivot was selected.

The following equations summarize how the sets $\mathcal{L}$, $\mathcal{E}$, and $\mathcal{A}$ change when pivot p is chosen and eliminated. The new element p is added, old elements are absorbed, and redundant entries are deleted:

$$\mathcal{L}^k = \left(\mathcal{L}^{k-1} \setminus \bigcup_{e \in \mathcal{E}_p} \mathcal{L}_e \right) \cup \mathcal{L}_p,$$

$$\mathcal{E}^k = \left(\mathcal{E}^{k-1} \setminus \bigcup_{e \in \mathcal{E}_p} e \right) \cup \{p\},$$

$$\mathcal{A}^k = \left(\mathcal{A}^{k-1} \setminus (\mathcal{L}_p \times \mathcal{L}_p) \right) \cup \left(V^k \times V^k \right).$$

3.1. Quotient graph example. We illustrate the sequence of elimination graphs and quotient graphs of a 10-by-10 sparse matrix in Figures 3.1 and 3.2. The example is ordered so that a minimum degree algorithm recommends pivoting down the diagonal in the natural order (that is, the permutation matrix is the identity). In Figures 3.1 and 3.2, variables and elements are shown as thin-lined and heavy-lined circles, respectively. In the matrices in these figures, diagonal entries are numbered and unmodified original nonzero entries (entries in $\mathcal{A}$) are shown as solid squares. The solid squares in row i form the set $\mathcal{A}_i$. The variables in current unabsorbed elements (sets $\mathcal{L}_e$) are indicated by solid circles in the columns of $\mathbf{L}$ corresponding to the unabsorbed elements. The solid circles in row i form the set $\mathcal{E}_i$. Entries that do not correspond to edges in the quotient graph are shown as an $\times$. Figure 3.1 shows the elimination graph, quotient graph, and the matrix prior to elimination (in the left column) and after the first three steps (from left to right). Figure 3.2 continues the example for the next four steps.

Consider the transformation of the graph $\mathcal{G}^2$ to the graph $\mathcal{G}^3$. Variable 3 is selected as pivot. We have $\mathcal{L}_3 = \mathcal{A}_3 = \{5, 6, 7\}$ (a simple case of equation (3.1)). The

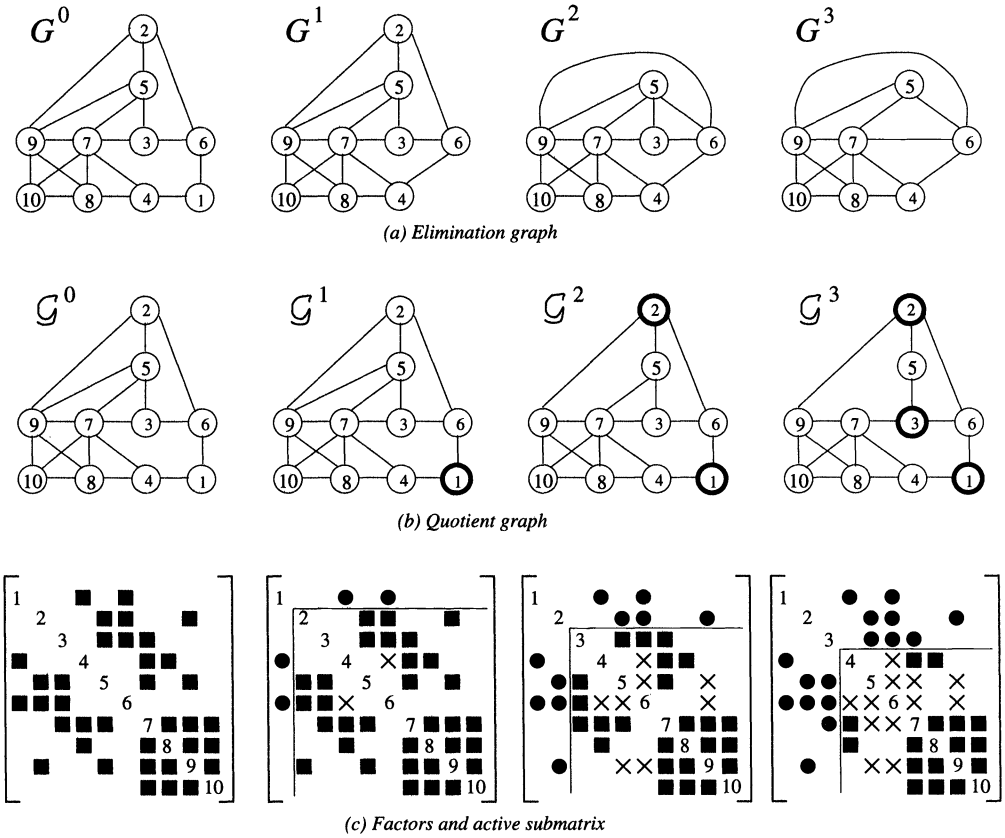


FIG. 3.1. *Elimination graph, quotient graph, and matrix for the first three steps.*

new element 3 represents the pairwise adjacency of variables 5, 6, and 7. The explicit edge (5, 7) is now redundant and is deleted from $\mathcal{A}_5$ and $\mathcal{A}_7$.

Also consider the transformation of the graph $\mathcal{G}^4$ to the graph $\mathcal{G}^5$. Variable 5 is selected as pivot. The set $\mathcal{A}_5$ is empty and $\mathcal{E}_5 = \{2, 3\}$. Following equation (3.1),

$$\begin{aligned}\mathcal{L}_5 &= (\mathcal{A}_5 \cup \mathcal{L}_2 \cup \mathcal{L}_3) \setminus \{5\} \\ &= (\emptyset \cup \{5, 6, 9\} \cup \{5, 6, 7\}) \setminus \{5\} \\ &= \{6, 7, 9\},\end{aligned}$$

which is the pattern of column 5 of $\mathbf{L}$ (excluding the diagonal). Since the new element 5 implies that variables 6, 7, and 9 are pairwise adjacent, elements 2 and 3 do not add any information to the graph. They are removed, having been "absorbed" into element 5. Additionally, the edge (7, 9) is redundant and is removed from $\mathcal{A}_7$ and $\mathcal{A}_9$. In $\mathcal{G}^4$ we have

$$\begin{aligned}\mathcal{A}_6 &= \emptyset, & \mathcal{E}_6 &= \{2, 3, 4\}, \\ \mathcal{A}_7 &= \{9, 10\}, & \mathcal{E}_7 &= \{3, 4\}, \\ \mathcal{A}_9 &= \{7, 8, 10\}, & \mathcal{E}_9 &= \{2\}.\end{aligned}$$

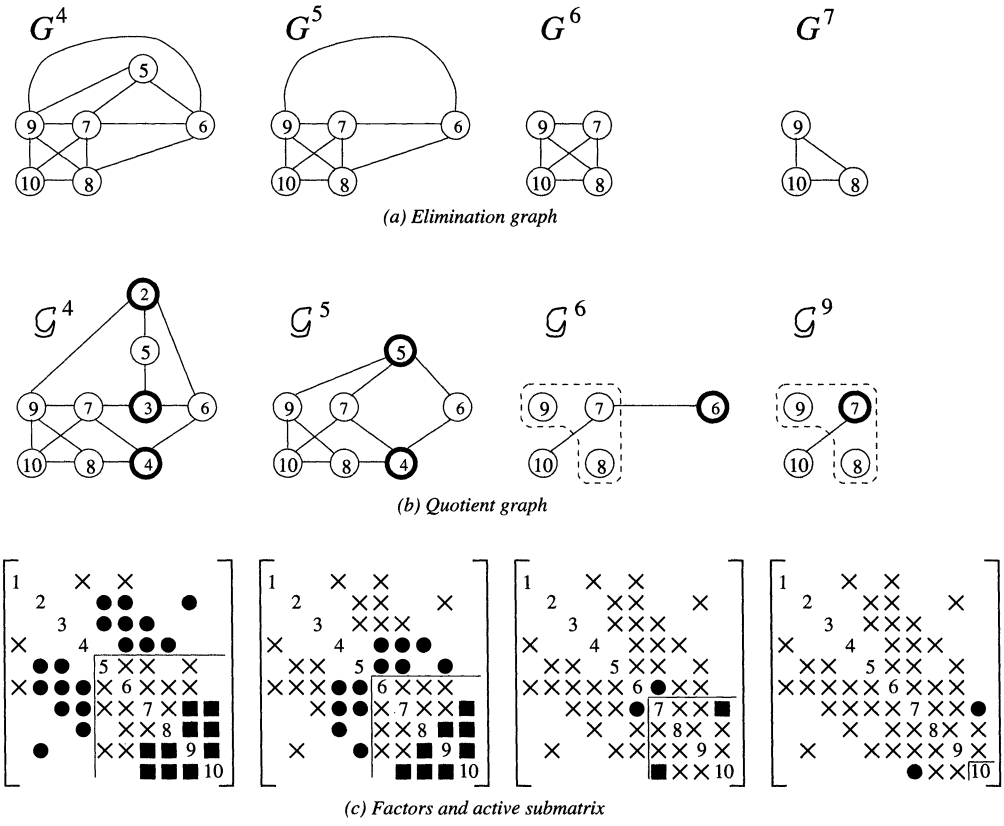


FIG. 3.2. *Elimination graph, quotient graph, and matrix for steps* 4 *to* 7.

After these transformations, we have in $\mathcal{G}^5$,

$$\begin{aligned} \mathcal{A}_6 &= \emptyset, & \mathcal{E}_6 &= \{4,5\}, \\ \mathcal{A}_7 &= \{10\}, & \mathcal{E}_7 &= \{4,5\}, \\ \mathcal{A}_9 &= \{8,10\}, & \mathcal{E}_9 &= \{5\}, \end{aligned}$$

and the new element in $\mathcal{G}^5$,

$$\mathcal{L}_5 = \{6,7,9\}.$$

3.2. Indistinguishable variables and external degree. Two variables i and j are *indistinguishable* in G if $\mathrm{Adj}_G(i) \cup \{i\} = \mathrm{Adj}_G(j) \cup \{j\}$. They will have the same degree until one is selected as pivot. If i is selected, then j can be selected next without causing any additional fill-in. Selecting i and j together is called *mass elimination* [23]. Variables i and j are replaced in $\mathcal{G}$ by a *supervariable* containing both i and j, labeled by its *principal* variable (i, say) [13, 14, 15]. Variables that are not supervariables are called *simple* variables. In practice, new supervariables are constructed at step k only if both i and j are in $\mathcal{L}_p$ (where p is the pivot selected at step k). In addition, rather than checking the graph G for indistinguishability, we use the quotient graph $\mathcal{G}$ so that two variables i and j are found to be indistinguishable if $\mathrm{Adj}_{\mathcal{G}}(i) \cup \{i\} = \mathrm{Adj}_{\mathcal{G}}(j) \cup \{j\}$. This comparison is faster than determining if two variables are indistinguishable in G, but may miss some identifications because,

although indistinguishability in $\mathcal{G}$ implies indistinguishability in G, the reverse is not true.

We denote the set of simple variables in the supervariable with principal variable i as $\mathbf{i}$ and define $\mathbf{i} = \{i\}$ if i is a simple variable. When p is selected as pivot at the kth step, all variables in $\mathbf{p}$ are eliminated. The use of supervariables greatly reduces the number of degree computations performed, which is the most costly part of the algorithm. Nonprincipal variables and their incident edges are removed from the quotient graph data structure when they are detected. The set notation $\mathcal{A}$ and $\mathcal{L}$ refers either to a set of supervariables or to the variables represented by the supervariables, depending on the context. In degree computations and when used in representing elimination graphs, the sets refer to variables; otherwise they refer to supervariables.

In Figure 3.2, detected supervariables are circled by dashed lines. Nonprincipal variables are left inside the dashed supervariables. These are, however, removed from the quotient graph. The last quotient graph in Figure 3.2 represents the selection of pivots 7, 8, and 9, and thus the right column of the figure depicts G^7, $\mathcal{G}^9$, and the matrix after the ninth pivot step.

The *external* degree $d_i \equiv t_i - |\mathbf{i}| + 1$ of a principal variable i is

$$d_i = |\mathrm{Adj}_G(i) \setminus \mathbf{i}| = |\mathcal{A}_i \setminus \mathbf{i}| + \left| \left(\bigcup_{e \in \mathcal{E}_i} \mathcal{L}_e \right) \setminus \mathbf{i} \right|, \tag{3.2}$$

since the set $\mathcal{A}_i$ is disjoint from any set $\mathcal{L}_e$ for $e \in \mathcal{E}_i$. At most, $(d_i^2 - d_i)/2$ fill-ins occur if all variables in $\mathbf{i}$ are selected as pivots. We refer to t_i as the true degree of variable i. Selecting the pivot with minimum external degree tends to produce a better ordering than selecting the pivot with minimum true degree [25] (also see §6.2).

3.3. Quotient-graph-based minimum degree algorithm. A minimum degree algorithm based on the quotient graph is shown in Algorithm 1. It includes element absorption, mass elimination, supervariables, and external degrees. Supervariable detection is simplified by computing a hash function on each variable, so that not all pairs of variables need be compared [3]. Algorithm 1 does not include two important features of Liu's *multiple minimum degree* (MMD) algorithm: incomplete update [17, 18] and multiple elimination [25]. With multiple elimination, an independent set of pivots with minimum degree is selected before any degrees are updated. If a variable is adjacent to two or more pivot elements, its degree is computed only once. A variable j is *outmatched* if $\mathrm{Adj}_G(i) \subseteq \mathrm{Adj}_G(j)$. With incomplete degree update, the degree update of the outmatched variable j is avoided until variable i is selected as pivot. These two features further reduce the amount of work needed for the degree computation in MMD. We will discuss their relationship to the AMD algorithm in the next section.

The time taken to compute d_i using equation (3.2) by a quotient-graph-based minimum degree algorithm is

$$\Theta\left(|\mathcal{A}_i| + \sum_{e \in \mathcal{E}_i} |\mathcal{L}_e| \right), \tag{3.3}$$

ALGORITHM 1 (minimum degree algorithm, based on quotient graph).

$V = \{1 \ldots n\}$
$\overline{V} = \emptyset$
for $i = 1$ **to** n **do**
  $\mathcal{A}_i = \{j : a_{ij} \neq 0 \text{ and } i \neq j\}$
  $\mathcal{E}_i = \emptyset$
  $d_i = |\mathcal{A}_i|$
  $\mathbf{i} = \{i\}$
end for
$k = 1$
while $k \leq n$ **do**
  mass elimination:
  select variable $p \in V$ that minimizes d_p
  $\mathcal{L}_p = \left(\mathcal{A}_p \cup \bigcup_{e \in \mathcal{E}_p} \mathcal{L}_e\right) \setminus \mathbf{p}$
  for each $\mathbf{i} \in \mathcal{L}_p$ **do**
    remove redundant entries:
    $\mathcal{A}_i = (\mathcal{A}_i \setminus \mathcal{L}_p) \setminus \mathbf{p}$
    element absorption:
    $\mathcal{E}_i = (\mathcal{E}_i \setminus \mathcal{E}_p) \cup \{p\}$
    compute external degree:
    $d_i = |\mathcal{A}_i \setminus \mathbf{i}| + \left|\left(\bigcup_{e \in \mathcal{E}_i} \mathcal{L}_e\right) \setminus \mathbf{i}\right|$
  end for
  supervariable detection, pairs found via hash function:
  for each pair $\mathbf{i}$ and $\mathbf{j} \in \mathcal{L}_p$ **do**
    if $\mathbf{i}$ and $\mathbf{j}$ are indistinguishable **then**
      remove the supervariable $\mathbf{j}$:
      $\mathbf{i} = \mathbf{i} \cup \mathbf{j}$
      $d_i = d_i - |\mathbf{j}|$
      $V = V \setminus \{j\}$
      $\mathcal{A}_j = \emptyset$
      $\mathcal{E}_j = \emptyset$
    end if
  end for
  convert variable p to element p:
  $\overline{V} = (\overline{V} \cup \{p\}) \setminus \mathcal{E}_p$
  $V = V \setminus \{p\}$
  $\mathcal{A}_p = \emptyset$
  $\mathcal{E}_p = \emptyset$
  $k = k + |\mathbf{p}|$
end while

which is $\Omega(|\mathrm{Adj}_{G^k}(i)|)$ if all variables are simple.[1] This degree computation is the most costly part of the minimum degree algorithm. When supervariables are present, in the best case the time taken is proportional to the degree of the variable in the

[1] Asymptotic complexity notation is defined in [6]. We write $f(n) = \Theta(g(n))$ if there exist positive constants c_1, c_2, and n_0 such that $0 \leq c_1 g(n) \leq f(n) \leq c_2 g(n)$ for all $n > n_0$. Similarly, $f(n) = \Omega(g(n))$ if there exist positive constants c and n_0 such that $0 \leq cg(n) \leq f(n)$ for all $n > n_0$, and $f(n) = O(g(n))$ if there exist positive constants c and n_0 such that $0 \leq f(n) \leq cg(n)$ for all $n > n_0$.

ALGORITHM 2 (computation of $|\mathcal{L}_e \setminus \mathcal{L}_p|$ for all $e \in \overline{V}$).
assume $w(k) < 0$ for $k = 1, \ldots, n$
for each supervariable $\mathbf{i} \in \mathcal{L}_p$ **do**
for each element $e \in \mathcal{E}_i$ **do**
if $(w(e) < 0)$ **then** $w(e) = |\mathcal{L}_e|$
$w(e) = w(e) - |\mathbf{i}|$
end for
end for

"compressed" elimination graph, where all nonprincipal variables and their incident edges are removed.

4. Approximate degree. Having now discussed the data structures and the standard minimum degree implementations, we now consider our approximation for the minimum degree and indicate its lower complexity.

We assume that p is the kth pivot and that we compute the bounds only for supervariables $\mathbf{i} \in \mathcal{L}_p$. Rather than computing the exact external degree d_i, our AMD algorithm computes an upper bound [8, 7],

$$\overline{d}_i^k = \min \left\{ \begin{array}{l} n - k, \\ \overline{d}_i^{k-1} + |\mathcal{L}_p \setminus \mathbf{i}|, \\ |\mathcal{A}_i \setminus \mathbf{i}| + |\mathcal{L}_p \setminus \mathbf{i}| + \displaystyle\sum_{e \in \mathcal{E}_i \setminus \{p\}} |\mathcal{L}_e \setminus \mathcal{L}_p| \end{array} \right\}. \tag{4.1}$$

The first two terms ($n - k$, the size of the active submatrix, and $\overline{d}_i^{k-1} + |\mathcal{L}_p \setminus \mathbf{i}|$, the worst case fill-in) are usually not as tight as the third term in equation (4.1). Algorithm 2 computes $|\mathcal{L}_e \setminus \mathcal{L}_p|$ for *all* elements e in the entire quotient graph. The set $\mathcal{L}_e$ splits into two disjoint subsets: the *external* subset $\mathcal{L}_e \setminus \mathcal{L}_p$ and the *internal* subset $\mathcal{L}_e \cap \mathcal{L}_p$. If Algorithm 2 scans element e, the term $w(e)$ is initialized to $|\mathcal{L}_e|$ and then decremented once for each variable i in the internal subset $\mathcal{L}_e \cap \mathcal{L}_p$, and, at the end of Algorithm 2, we have $w(e) = |\mathcal{L}_e| - |\mathcal{L}_e \cap \mathcal{L}_p| = |\mathcal{L}_e \setminus \mathcal{L}_p|$. If Algorithm 2 does not scan element e, the term $w(e)$ is less than zero. Combining these two cases, we obtain

$$|\mathcal{L}_e \setminus \mathcal{L}_p| = \left\{ \begin{array}{ll} w(e) & \text{if } w(e) \geq 0, \\ |\mathcal{L}_e| & \text{otherwise} \end{array} \right\} \text{ for all } e \in \overline{V}. \tag{4.2}$$

Algorithm 2 is followed by a second loop to compute our upper bound degree $\overline{d}_i$ for each supervariable $\mathbf{i} \in \mathcal{L}_p$, using equations (4.1) and (4.2). The total time for Algorithm 2 is

$$\Theta\Big(\sum_{\mathbf{i} \in \mathcal{L}_p} |\mathcal{E}_i| \Big).$$

The second loop to compute the upper bound degree takes time

$$\Theta\Big(\sum_{\mathbf{i} \in \mathcal{L}_p} (|\mathcal{A}_i| + |\mathcal{E}_i|) \Big). \tag{4.3}$$

The total asymptotic time is thus given by expression (4.3).

Multiple elimination [25] improves the minimum degree algorithm by updating the degree of a variable only once for each set of independent pivots. Incomplete degree update [17, 18] skips the degree update of outmatched variables. We cannot take full advantage of the incomplete degree update since it avoids the degree update for some supervariables adjacent to the pivot element. With our technique (Algorithm 2), we must scan the element lists for all supervariables $\mathbf{i}$ in $\mathcal{L}_p$. If the degree update of one of the supervariables is to be skipped, its element list must still be scanned so that the external subset terms can be computed for the degree update of other supervariables in $\mathcal{L}_p$. The only advantage of multiple elimination or incomplete degree update would be to skip the second loop that computes the upper bound degree for outmatched variables or supervariables for which the degree has already been computed.

If the total time in expression (4.3) is amortized across the computation of all supervariables $\mathbf{i} \in \mathcal{L}_p$, then the time taken to compute $\overline{d}_i$ is

$$\Theta(|\mathcal{A}_i| + |\mathcal{E}_i|) = O(|\mathcal{A}_i^0|),$$

which is $\Theta(|\mathrm{Adj}_{\mathcal{G}^k}(i)|)$ if all variables are simple. Computing our bound takes time proportional to the degree of the variable in the *quotient graph* $\mathcal{G}$. This is much faster than the time taken to compute the exact external degree (see expression (3.3)).

4.1. Accuracy of our approximate degrees. Gilbert, Moler, and Schreiber [24] also use approximate external degrees that they can compute in the same time as our degree bound $\overline{d}_i$. In our notation, their bound $\widehat{d}_i$ is

$$\widehat{d}_i = |\mathcal{A}_i \setminus \mathbf{i}| + \sum_{e \in \mathcal{E}_i} |\mathcal{L}_e \setminus \mathbf{i}|.$$

Since many pivotal variables are adjacent to two or fewer elements when selected, Ashcraft, Eisenstat, and Lucas [4] have suggested a combination of $\widehat{d}_i$ and d_i,

$$\widetilde{d}_i = \begin{cases} d_i & \text{if } |\mathcal{E}_i| = 2, \\ \widehat{d}_i & \text{otherwise.} \end{cases}$$

Computing $\widetilde{d}_i$ takes the same time as $\overline{d}_i$ or $\widehat{d}_i$, except when $|\mathcal{E}_i| = 2$. In this case, it takes $O(|\mathcal{A}_i| + |\mathcal{L}_e|)$ time to compute $\widetilde{d}_i$, whereas computing $\overline{d}_i$ or $\widehat{d}_i$ takes $\Theta(|\mathcal{A}_i|)$ time. In the Yale sparse matrix package [17] the $|\mathcal{L}_e \setminus \mathcal{L}_p|$ term for the $\mathcal{E}_i = \{e, p\}$ case is computed by scanning $\mathcal{L}_e$ once. It is then used to compute d_i for all $i \in \mathcal{L}_p$ for which $\mathcal{E}_i = \{e, p\}$. This technique can also be used to compute $\widetilde{d}_i$, and thus the time it takes to compute $\widetilde{d}_i$ is $O(|\mathcal{A}_i| + |\mathcal{L}_e|)$ and not $\Theta(|\mathcal{A}_i| + |\mathcal{L}_e|)$.

THEOREM 4.1. *The relationship between external degree and the three approximate degree bounds now follows. The equality* $d_i = \overline{d}_i = \widetilde{d}_i = \widehat{d}_i$ *holds when* $|\mathcal{E}_i| \leq 1$. *The inequality* $d_i = \overline{d}_i = \widetilde{d}_i \leq \widehat{d}_i$ *holds when* $|\mathcal{E}_i| = 2$. *Finally, the inequality* $d_i \leq \overline{d}_i \leq \widetilde{d}_i = \widehat{d}_i$ *holds when* $|\mathcal{E}_i| > 2$. *Consequently, the inequality* $d_i \leq \overline{d}_i \leq \widetilde{d}_i \leq \widehat{d}_i$ *holds for all values of* $|\mathcal{E}_i|$.

Proof. The bound $\widehat{d}_i$ is equal to the exact degree when variable i is adjacent to at most one element ($|\mathcal{E}_i| \leq 1$). The accuracy of the $\widehat{d}_i$ bound is unaffected by the size of $\mathcal{A}_i$, since entries that fall within the pattern $\mathcal{L}$ of an element are removed from $\mathcal{A}$. Thus, if there is just one element (the current element p, say), the bound $\widehat{d}_i$ is tight. If $|\mathcal{E}_i|$ is two (the current element p and a prior element e, say), we have

$$\widehat{d}_i = |\mathcal{A}_i \setminus \mathbf{i}| + |\mathcal{L}_p \setminus \mathbf{i}| + |\mathcal{L}_e \setminus \mathbf{i}| = d_i + |(\mathcal{L}_e \cap \mathcal{L}_p) \setminus \mathbf{i}|.$$

The bound $\widehat{d}_i$ counts entries in the set $(\mathcal{L}_e \cap \mathcal{L}_p) \setminus \mathbf{i}$ twice, and so $\widehat{d}_i$ will be an overestimate in the possible (even likely) case that a variable $j \neq i$ exists that is adjacent to both e and p. Combined with the definition of $\widetilde{d}$, we have $d_i = \widetilde{d}_i = \widehat{d}_i$ when $|\mathcal{E}_i| \leq 1$, $d_i = \widetilde{d}_i \leq \widehat{d}_i$ when $|\mathcal{E}_i| = 2$, and $d_i \leq \widetilde{d}_i = \widehat{d}_i$ when $|\mathcal{E}_i| > 2$.

If $|\mathcal{E}_i| \leq 1$ our bound $\overline{d}_i$ is exact for the same reason that $\widehat{d}_i$ is exact. If $|\mathcal{E}_i|$ is two we have

$$\overline{d}_i = |\mathcal{A}_i \setminus \mathbf{i}| + |\mathcal{L}_p \setminus \mathbf{i}| + |\mathcal{L}_e \setminus \mathcal{L}_p| = d_i.$$

No entry is in both $\mathcal{A}_i$ and any element $\mathcal{L}$, since these redundant entries are removed from $\mathcal{A}_i$. Any entry in $\mathcal{L}_p$ does not appear in the external subset $(\mathcal{L}_e \setminus \mathcal{L}_p)$. Thus, no entry is counted twice, and $d_i = \overline{d}_i$ when $|\mathcal{E}_i| \leq 2$. Finally, consider both $\overline{d}_i$ and $\widehat{d}_i$ when $|\mathcal{E}_i| > 2$. We have

$$\overline{d}_i = |\mathcal{A}_i \setminus \mathbf{i}| + |\mathcal{L}_p \setminus \mathbf{i}| + \sum_{e \in \mathcal{E}_i \setminus \{p\}} |\mathcal{L}_e \setminus \mathcal{L}_p|$$

and

$$\widehat{d}_i = |\mathcal{A}_i \setminus \mathbf{i}| + |\mathcal{L}_p \setminus \mathbf{i}| + \sum_{e \in \mathcal{E}_i \setminus \{p\}} |\mathcal{L}_e \setminus \mathbf{i}|.$$

Since these degree bounds are used only when computing the degree of a supervariable $\mathbf{i} \in \mathcal{L}_p$, we have $\mathbf{i} \subseteq \mathcal{L}_p$. Thus, $\overline{d}_i \leq \widehat{d}_i$ when $|\mathcal{E}_i| > 2$. □

Note that if a variable i is adjacent to two elements or less then our bound is equal to the exact external degree. This is very important, since most variables of minimum degree are adjacent to two elements or less.

4.2. Degree computation example. We illustrate the computation of our approximate external degree bound in Figures 3.1 and 3.2. Variable 6 is adjacent to three elements in $\mathcal{G}^3$ and $\mathcal{G}^4$. All other variables are adjacent to two or less elements. In $\mathcal{G}^3$, the bound $\overline{d}_6$ is tight, since the two sets $|\mathcal{L}_1 \setminus \mathcal{L}_3|$ and $|\mathcal{L}_2 \setminus \mathcal{L}_3|$ are disjoint.

In graph $\mathcal{G}^4$, the current pivot element is $p = 4$. We compute

$$\begin{aligned}
\overline{d}_6 &= |\mathcal{A}_i \setminus \mathbf{i}| + |\mathcal{L}_p \setminus \mathbf{i}| + \left(\sum_{e \in \mathcal{E}_i \setminus \{p\}} |\mathcal{L}_e \setminus \mathcal{L}_p| \right) \\
&= |\emptyset \setminus \{6\}| + |\{6,7,8\} \setminus \{6\}| + (|\mathcal{L}_2 \setminus \mathcal{L}_4| + |\mathcal{L}_3 \setminus \mathcal{L}_4|) \\
&= |\{7,8\}| + (|\{5,6,9\} \setminus \{6,7,8\}| + |\{5,6,7\} \setminus \{6,7,8\}|) \\
&= |\{7,8\}| + (|\{5,9\}| + |\{5\}|) \\
&= 5.
\end{aligned}$$

The exact external degree of variable 6 is $d_6 = 4$, as can be seen in the elimination graph G^4 on the left of Figure 3.2(a). Our bound is one more than the exact external degree, since the variable 5 appears in both $\mathcal{L}_2 \setminus \mathcal{L}_4$ and $\mathcal{L}_3 \setminus \mathcal{L}_4$, but is one less than the bound $\widehat{d}_i$, which is equal to 6 in this case. Our bound on the degree of variable 6 is again tight after the next pivot step, since elements 2 and 3 are absorbed into element 5.

5. The AMD algorithm. The AMD algorithm is identical to Algorithm 1, except that the external degree d_i is replaced with $\overline{d}_i$ throughout. The bound on

the external degree $\overline{d}_i$ is computed using Algorithm 2 and equations (4.1) and (4.2). In addition to the *natural* absorption of elements in $\mathcal{E}_p$, any element with an empty external subset ($|\mathcal{L}_e \setminus \mathcal{L}_p| = 0$) is also absorbed into element p, even if e is not adjacent to p. This *aggressive* element absorption improves the degree bounds by reducing $|\mathcal{E}|$. For many matrices, aggressive absorption rarely occurs. In some cases, however, up to half of the elements are aggressively absorbed. Consider the matrix

$$\begin{bmatrix} a_{11} & 0 & a_{13} & a_{14} \\ 0 & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & 0 \\ a_{41} & a_{42} & 0 & a_{44} \end{bmatrix},$$

where we assume the pivots are chosen down the diagonal in order. The external subset $|\mathcal{L}_1 \setminus \mathcal{L}_2|$ is zero ($\mathcal{L}_1 = \mathcal{L}_2 = \{3, 4\}$). Element 2 aggressively absorbs element 1, even though element 1 is not adjacent to variable 2 (a_{12} is zero).

As in many other minimum degree algorithms, we use a set of n linked lists to assist the search for a variable of minimum degree. A single linked list holds all supervariables with the same degree bound. Maintaining this data structure takes time proportional to the total number of degree computations, or $O(|\mathbf{L}|)$ in the worst case.

Computing the pattern of each pivot element $\mathcal{L}_p$ takes a total of $O(|\mathbf{L}|)$ time overall, since each element is used in the computation of at most one other element, and the total size of all elements constructed is $O(|\mathbf{L}|)$.

The AMD algorithm is based on the quotient graph data structure used in the MA27 minimum degree algorithm [13, 14, 15]. Initially, the sets $\mathcal{A}$ are stored, followed by a small amount of elbow room. When the set $\mathcal{L}_p$ is formed, it is placed in the elbow room (or in place of $\mathcal{A}_p$ if $|\mathcal{E}_p| = 0$). Garbage collection occurs if the elbow room is exhausted. During garbage collection, the space taken by $\mathcal{A}_i$ and $\mathcal{E}_i$ is reduced to exactly $|\mathcal{A}_i| + |\mathcal{E}_i|$ for each supervariable $\mathbf{i}$ (which is less than or equal to $|\mathcal{A}_i^0|$) and the extra space is reclaimed. The space for $\mathcal{A}_e$ and $\mathcal{E}_e$ for all elements $e \in \overline{V}$ is fully reclaimed, as is the space for $\mathcal{L}_e$ of any absorbed elements e. Each garbage collection takes time that is proportional to the size of the workspace (normally $\Theta(|\mathbf{A}|)$). In practice, elbow room of size n is sufficient.

During the computation of our degree bounds, we compute the following hash function for supervariable detection [3],

$$\text{Hash}(i) = \left\{ \left(\sum_{\mathbf{j} \in \mathcal{A}_i} j + \sum_{e \in \mathcal{E}_i} e \right) \bmod (n-1) \right\} + 1,$$

which increases the degree computation time by a small constant factor. We place each supervariable $\mathbf{i}$ in a hash bucket according to Hash(i), taking time $O(|\mathbf{L}|)$ overall. If two or more supervariables are placed in the same hash bucket, then each pair of supervariables $\mathbf{i}$ and $\mathbf{j}$ in the hash bucket are tested for indistinguishability. If the hash function results in no collisions then the total time taken by the comparison is $O(|\mathbf{A}|)$.

Ashcraft [3] uses this hash function as a preprocessing step on the entire matrix (without the mod($n-1$) term and with an $O(|V| \log |V|)$ sort instead of $|V|$ hash buckets). In contrast, we use this function during the ordering and hash only those variables adjacent to the current pivot element.

For example, variables 7, 8, and 9 are indistinguishable in G^5 in Figure 3.2(a). The AMD algorithm would not consider variable 8 at step 5, since it is not adjacent to the pivot element 5 (refer to quotient graph $\mathcal{G}^5$ in Figure 3.2(b)). AMD would not construct $\mathbf{7} = \{7, 9\}$ at step 5, since 7 and 9 are distinguishable in $\mathcal{G}^5$. It would construct $\mathbf{7} = \{7, 8, 9\}$ at step 6, however.

The total number of times the approximate degree $\overline{d}_i$ of variable i is computed during elimination is no more than the number of nonzero entries in row k of $\mathbf{L}$, where variable i is the kth pivot. The total time taken to compute $\overline{d}_i$ using Algorithm 2 and equations (4.1) and (4.2) is $O(|\mathcal{A}_i^0|)$ or, equivalently, $O(|(\mathbf{PAP}^T)_{k*}|)$, the number of nonzero entries in row k of the permuted matrix. The total time taken by the entire AMD algorithm is thus bounded by the degree computation,

$$O\left(\sum_{k=1}^{n} |\mathbf{L}_{k*}| \cdot |(\mathbf{PAP}^T)_{k*}|\right). \tag{5.1}$$

This bound assumes no (or few) supervariable hash collisions and a constant number of garbage collections. In practice these assumptions seem to hold, but the asymptotic time would be higher if they did not. In many problem domains, the number of nonzeros per row of $\mathbf{A}$ is a constant, independent of n. For matrices in these domains, our AMD algorithm takes time $O(|\mathbf{L}|)$ (with the same assumptions).

6. Performance results. In the following sections, we present the results of our experiments with AMD on a wide range of test matrices. We first compare the degree computations discussed above (t, d, $\overline{d}$, $\widetilde{d}$, and $\widehat{d}$), as well as an upper bound on the true degree $\overline{t} \equiv \overline{d} + |\mathbf{i}| - 1$. We then compare the AMD algorithm with other established minimum degree codes (MMD and MA27).

6.1. Test matrices. We tested all degree bounds and codes on all matrices in the Harwell/Boeing collection of type PUA, RUA, PSA, and RSA [11, 12] (at `orion.cerfacs.fr` or `numerical.cc.rl.ac.uk`), all nonsingular matrices in Saad's SPARSKIT2 collection (at `ftp.cs.umn.edu`), all matrices in the University of Florida collection (available from `ftp.cis.ufl.edu` in the directory `pub/umfpack/matrices`), and several other matrices from NASA and Boeing. Of those 378 matrices, we present results below on those matrices requiring 500 million or more floating-point operations for the Cholesky factorization, as well as the ORANI678 matrix in the Harwell/Boeing collection and the EX19 in Saad's collection (a total of 26 matrices). The latter two are best-case and worst-case examples from the set of smaller matrices.

For the unsymmetric matrices in the test set, we first used the maximum transversal algorithm `MC21` from the Harwell Subroutine Library [9] to reorder the matrix so that the permuted matrix has a zero-free diagonal. We then formed the symmetric pattern of the permuted matrix plus its transpose. This is how a minimum degree ordering algorithm is used in MUPS [1, 2]. For these matrices, Table 6.1 lists the statistics for the symmetrized pattern.

Table 6.1 lists the matrix name, the order, the number of nonzeros in lower triangular part, two statistics obtained with an exact minimum degree ordering (using d), and a description. In column 4, we report the percentage of pivots p such that $|\mathcal{E}_p| > 2$. Column 4 shows that there is only a small percentage of pivots selected using an exact minimum degree ordering that have more than two elements in their adjacency list. Therefore, we can expect a good quality ordering with an algorithm based on our approximate degree bound. In column 5, we indicate how often a degree d_i is computed when $|\mathcal{E}_i| > 2$ (as a percentage of the total number of degree

TABLE 6.1
Selected matrices in test set.

Matrix	n	nz	Percentage of $\lvert\mathcal{E}_p\rvert > 2$	Percentage of $\lvert\mathcal{E}_i\rvert > 2$	Description
RAEFSKY3	21,200	733,784	0.00	13.4	fluid/structure interaction, turbulence
VENKAT01	62,424	827,684	0.71	15.7	unstructured 2D Euler solver
BCSSTK32	44,609	985,046	0.20	27.3	structural eng., automobile chassis
EX19	12,005	123,937	1.57	29.4	2D developing pipe flow (turbulent)
BCSSTK30	28,924	1,007,284	0.66	31.8	structural eng., off-shore platform
CT20STIF	52,329	1,323,067	0.77	33.2	structural eng., CT20 engine block
NASASRB	54,870	1,311,227	0.06	35.0	shuttle rocket booster
OLAF	16,146	499,505	0.41	35.2	NASA test problem
RAEFSKY1	3,242	145,517	0.00	38.9	incompressible flow, pressure-driven pipe
CRYSTK03	24,696	863,241	0.00	40.9	structural eng., crystal vibration
RAEFSKY4	19,779	654,416	0.00	41.4	buckling problem for container model
CRYSTK02	13,965	477,309	0.00	42.0	structural eng., crystal vibration
BCSSTK33	8,738	291,583	0.00	42.6	structural eng., auto steering mech.
BCSSTK31	35,588	572,914	0.60	43.1	structural eng., automobile component
EX11	16,614	540,167	0.04	43.3	CFD, 3D cylinder & flat plate heat exch.
FINAN512	74,752	261,120	1.32	46.6	economics, portfolio optimization
RIM	22,560	862,411	2.34	63.2	chemical eng., fluid mechanics problem
BBMAT	38,744	1,274,141	5.81	64.4	CFD, 2D airfoil with turbulence
EX40	7,740	225,136	17.45	64.7	CFD, 3D die swell problem on square die
WANG4	26,068	75,564	15.32	78.3	3D MOSFET semicond. (30x30x30 grid)
LHR34	35,152	608,830	7.69	78.7	chemical eng., light hydrocarbon recovery
WANG3	26,064	75,552	15.29	79.2	3D diode semiconductor (30x30x30 grid)
LHR71	70,304	1,199,704	8.47	81.1	chemical eng., light hydrocarbon recovery
ORANI678	2,529	85,426	6.68	86.9	Australian economic model
PSMIGR1	3,140	410,781	6.65	91.0	US county-by-county migration
APPU	14,000	1,789,392	15.64	94.4	NASA test problem (random matrix)

updates). Table 6.1 is sorted according to this degree update percentage. Column 5 thus reports the percentage of "costly" degree updates performed by a minimum degree algorithm based on the exact degree. For matrices with relatively large values in column 5, significant time reductions can be expected with an approximate degree-based algorithm.

Since any minimum degree algorithm is sensitive to tie breaking issues, we randomly permuted all matrices and their adjacency lists 21 times (except for the random APPU matrix, which we ran only once). All methods were given the same set of 21 randomized matrices. We also ran each method on the original matrix. On some matrices, the original matrix gives better ordering time and fill-in results for all methods than the best result obtained with the randomized matrices. The overall comparisons are not, however, dependent on whether original or randomized matrices are used. We thus report only the median ordering time and fill-in obtained for the randomized matrices.

The APPU matrix is a random matrix used in a NASA benchmark, and is thus not representative of sparse matrices from real problems. We include it in our test set as a pathological case that demonstrates how well AMD handles a very irregular problem. Its factors are about 90% dense. It was not practical to run the APPU matrix 21 times because the exact degree update algorithms took too much time.

6.2. Comparing the exact and approximate degrees. To make a valid comparison between degree update methods, we modified our code for the AMD algorithm so that we could compute the exact external degree (d), our bound ($\overline{d}$), the $\tilde{d}$ bound, the $\widehat{d}$ bound, the exact true degree (t), and our upper bound on the true degree ($\overline{t}$).

TABLE 6.2
Median fill-in results of the degree update methods.

Matrix	Number of nonzeros below diagonal in $\mathbf{L}$, in thousands						
	AMD	d	$\overline{d}$	$\widetilde{d}$	$\widehat{d}$	t	$\overline{t}$
RAEFSKY3	4709	4709	4709	4709	5114	4992	4992
VENKAT01	5789	5771	5789	5798	6399	6245	6261
BCSSTK32	5080	5081	5079	5083	5721	5693	5665
EX19	319	319	319	318	366	343	343
BCSSTK30	3752	3751	3753	3759	4332	4483	4502
CT20STIF	10858	10758	10801	11057	13367	12877	12846
NASASRB	12282	12306	12284	12676	14909	14348	14227
OLAF	2860	2858	2860	2860	3271	3089	3090
RAEFSKY1	1151	1151	1151	1151	1318	1262	1262
CRYSTK03	13836	13836	13836	13836	17550	15507	15507
RAEFSKY4	7685	7685	7685	7685	9294	8196	8196
CRYSTK02	6007	6007	6007	6007	7366	6449	6449
BCSSTK33	2624	2624	2624	2640	3236	2788	2787
BCSSTK31	5115	5096	5132	5225	6194	6079	6057
EX11	6014	6016	6014	6014	7619	6673	6721
FINAN512	4778	4036	6042	11418	11505	8235	8486
RIM	3948	3898	3952	3955	4645	4268	4210
BBMAT	19673	19880	19673	21422	37820	21197	21445
EX40	1418	1386	1417	1687	1966	1526	1530
WANG4	6547	6808	6548	6566	7871	7779	7598
LHR34	3618	3743	3879	11909	27125	4383	4435
WANG3	6545	6697	6545	6497	7896	7555	7358
LHR71	7933	8127	8499	28241	60175	9437	9623
ORANI678	147	147	146	150	150	147	146
PSMIGR1	3020	3025	3011	3031	3176	2966	2975
APPU	87648	87613	87648	87566	87562	87605	87631

The six codes based on d, $\overline{d}$, $\widetilde{d}$, $\widehat{d}$, t, and $\overline{t}$ (columns 3 to 8 of Table 6.2) differ only in how they compute the degree. Since aggressive absorption is more difficult when using some bounds than others, we switched off aggressive absorption for these six codes. The actual AMD code (in column 2 of Table 6.2) uses $\overline{d}$ with aggressive absorption.

Table 6.2 lists the median number of nonzeros below the diagonal in $\mathbf{L}$ (in thousands) for each method. Results 20% higher than the lowest median $|\mathbf{L}|$ in the table (or higher) are underlined. Our upper bound on the true degree ($\overline{t}$) and the exact true degree (t) give nearly identical results. As expected, using minimum degree algorithms based on external degree noticeably improves the quality of the ordering (compare columns 3 and 7, or columns 4 and 8). From the inequality $d \leq \overline{d} \leq \widetilde{d} \leq \widehat{d}$, we would expect a similar ranking in the quality of ordering produced by these methods. Table 6.2 confirms this. The bound $\overline{d}$ and the exact external degree d produce nearly identical results. Comparing the AMD results and the $\overline{d}$ column, aggressive absorption tends to result in slightly lower fill-in, since it reduces $|\mathcal{E}|$ and thus improves the accuracy of our bound. The $\widetilde{d}$ bound is often accurate enough to produce good results, but can fail catastrophically for matrices with a high percentage of approximate pivots (see column 4 in Table 6.1). The less accurate $\widehat{d}$ bound produces notably worse results for many matrices.

Comparing all 378 matrices, the median $|\mathbf{L}|$ when using $\overline{d}$ is never more than 9% higher than the median fill-in obtained when using the exact external degree d (with the exception of the FINAN512 matrix). The fill-in results for d and $\overline{d}$ are identical for nearly half of the 378 matrices. The approximate degree bound $\overline{d}$ thus gives a very reliable estimation of the degree in the context of a minimum degree algorithm.

TABLE 6.3
Median ordering time of the degree update methods.

Matrix	Ordering time, in seconds						
	AMD	d	$\overline{d}$	$\widetilde{d}$	$\widehat{d}$	t	$\overline{t}$
RAEFSKY3	1.05	1.10	1.09	1.05	1.02	1.15	1.09
VENKAT01	4.07	4.95	4.11	4.47	3.88	4.32	3.85
BCSSTK32	4.67	5.64	4.54	4.91	4.35	5.55	4.48
EX19	0.87	1.12	0.89	1.01	0.86	1.09	0.87
BCSSTK30	3.51	5.30	3.55	3.65	3.51	4.38	3.38
CT20STIF	6.62	8.66	6.54	7.07	6.31	8.63	6.45
NASASRB	7.69	11.03	7.73	9.23	7.78	11.78	7.99
OLAF	1.83	2.56	1.90	2.16	1.83	2.33	1.78
RAEFSKY1	0.27	0.34	0.28	0.32	0.25	0.35	0.28
CRYSTK03	3.30	4.84	3.08	3.68	3.14	5.23	3.30
RAEFSKY4	2.32	2.90	2.18	2.45	2.08	3.12	2.07
CRYSTK02	1.49	2.34	1.55	1.64	1.45	2.04	1.52
BCSSTK33	0.91	1.36	1.05	0.99	0.85	1.62	0.91
BCSSTK31	4.55	7.53	4.92	5.68	4.56	7.41	4.92
EX11	2.70	4.06	2.77	3.00	2.60	4.23	2.89
FINAN512	15.03	34.11	14.45	17.79	15.84	46.49	18.58
RIM	5.74	10.38	5.69	6.12	5.72	10.01	5.58
BBMAT	27.80	115.75	27.44	42.17	23.02	129.32	28.33
EX40	1.04	1.56	1.10	1.09	0.95	1.46	1.12
WANG4	5.45	11.45	5.56	6.98	5.21	11.59	5.88
LHR34	19.56	109.10	25.62	45.36	43.70	125.41	24.73
WANG3	5.02	10.45	5.49	6.52	4.81	11.02	5.02
LHR71	46.03	349.58	58.25	129.85	121.96	389.70	60.40
ORANI678	5.49	196.01	8.13	6.97	7.23	199.01	8.45
PSMIGR1	10.61	334.27	10.07	14.20	8.16	339.28	9.94
APPU	41.75	2970.54	39.83	43.20	40.64	3074.44	38.93

The FINAN512 matrix is highly sensitive to tie breaking variations. Its graph consists of two types of nodes: "constraint" nodes and "linking" nodes [5]. The constraint nodes form independent sparse subgraphs, connected together via a tree of linking nodes. This matrix is a pathological worst-case matrix for any minimum degree method. All constraint nodes should be ordered first, but linking nodes have low degree and tend to be selected first, which causes high fill-in. Using a tree dissection algorithm, Berger et al. [5] obtain an ordering with only 1.83 million nonzeros in $\mathbf{L}$.

Table 6.3 lists the median ordering time (in seconds on a SUN SPARCstation 10) for each method. Ordering time twice that of the minimum median ordering time listed in the table (or higher) is underlined. Computing the $\widehat{d}$ bound is often the fastest, since it requires only a single pass over the element lists instead of the two passes required for the $\overline{d}$ bound. It is, however, sometimes slower than $\overline{d}$ because it can generate more fill-in, which increases the ordering time (see expression (5.1)). The ordering time of the two exact degree updates (d and t) increases dramatically as the percentage of "costly" degree updates increases (those for which $|\mathcal{E}_i| > 2$).

Garbage collection has little effect on the ordering time obtained. In the above runs, we gave each method elbow room of size n. Usually a single garbage collection occurred. At most two garbage collections occurred for AMD and at most three for the other methods (aggressive absorption reduces the memory requirements).

6.3. Comparing algorithms. In this section, we compare AMD with two other established minimum degree codes: Liu's MMD code [25] and Duff and Reid's MA27 code [15]. MMD stores the element patterns $\mathcal{L}$ in a fragmented manner and requires no elbow room [20, 21]. It uses the exact external degree d. MMD creates supervariables

TABLE 6.4
Median fill-in results of the four codes.

Matrix	Number of nonzeros below diagonal in **L**, in thousands			
	AMD	MMD	CMMD	MA27
RAEFSKY3	4709	4779	4724	5041
VENKAT01	5789	5768	5811	6303
BCSSTK32	5080	5157	5154	5710
EX19	319	322	324	345
BCSSTK30	3752	3788	3712	4529
CT20STIF	10858	11212	10833	12760
NASASRB	12282	12490	12483	14068
OLAF	2860	2876	2872	3063
RAEFSKY1	1151	1165	1177	1255
CRYSTK03	13836	13812	14066	15496
RAEFSKY4	7685	7539	7582	8245
CRYSTK02	6007	5980	6155	6507
BCSSTK33	2624	2599	2604	2766
BCSSTK31	5115	5231	5216	6056
EX11	6014	5947	6022	6619
FINAN512	4778	8180	8180	8159
RIM	3948	3947	3914	4283
BBMAT	19673	19876	19876	21139
EX40	1418	1408	1401	1521
WANG4	6547	6619	6619	7598
LHR34	3618	4162	4162	4384
WANG3	6545	6657	6657	7707
LHR71	7933	9190	9190	9438
ORANI678	147	147	147	147
PSMIGR1	3020	2974	2974	2966
APPU	87648	87647	87647	87605

only when two variables **i** and **j** have no adjacent variables and exactly two adjacent elements ($\mathcal{E}_i = \mathcal{E}_j = \{e, p\}$, and $\mathcal{A}_i = \mathcal{A}_j = \emptyset$, where p is the current pivot element). A hash function is not required. MMD takes advantage of multiple elimination and incomplete update.

MA27 uses the true degree t and the same data structures as AMD. It detects supervariables whenever two variables are adjacent to the current pivot element and have the same structure in the quotient graph (as does AMD). MA27 uses the true degree as the hash function for supervariable detection and does aggressive absorption. Neither AMD nor MA27 take advantage of multiple elimination or incomplete update.

Structural engineering matrices tend to have many rows of identical nonzero pattern. Ashcraft [3] has found that the total ordering time of MMD can be significantly improved by detecting these initial supervariables before starting the elimination. We implemented the precompression algorithm used in [3] and modified MMD to allow for initial supervariables. We call the resulting code CMMD ("compressed" MMD). Precompression has little effect on AMD, since it finds these supervariables when their degrees are first updated. The AMD algorithm on compressed matrices together with the cost of precompression was never faster than AMD.

Table 6.4 lists the median number of nonzeros below the diagonal in **L** (in thousands) for each code. Results 20% higher than the lowest median $|\mathbf{L}|$ in the table (or higher) are underlined. AMD, MMD, and CMMD find orderings of about the same quality. MA27 is slightly worse because it uses the true degree (t) instead of the external degree (d).

TABLE 6.5
Median ordering time of the four codes.

Matrix	Ordering time, in seconds			
	AMD	MMD	CMMD	MA27
RAEFSKY3	1.05	2.79	1.18	1.23
VENKAT01	4.07	9.01	4.50	5.08
BCSSTK32	4.67	12.47	5.51	6.21
EX19	0.87	0.69	0.83	1.03
BCSSTK30	3.51	7.78	3.71	4.40
CT20STIF	6.62	26.00	9.59	9.81
NASASRB	7.69	22.47	11.28	12.75
OLAF	1.83	5.67	4.41	2.64
RAEFSKY1	0.27	0.82	0.28	0.40
CRYSTK03	3.30	10.63	3.86	5.27
RAEFSKY4	2.32	5.24	2.36	2.91
CRYSTK02	1.49	3.89	1.53	2.37
BCSSTK33	0.91	2.24	1.32	1.31
BCSSTK31	4.55	11.60	7.76	7.92
EX11	2.70	7.45	5.05	3.90
FINAN512	15.03	895.23	897.15	40.31
RIM	5.74	9.09	8.11	10.13
BBMAT	27.80	200.86	201.03	134.58
EX40	1.04	2.13	2.04	1.30
WANG4	5.45	10.79	11.60	9.86
LHR34	19.56	139.49	141.16	77.83
WANG3	5.02	10.37	10.62	8.23
LHR71	46.03	396.03	400.40	215.01
ORANI678	5.49	124.99	127.10	124.66
PSMIGR1	10.61	186.07	185.74	229.51
APPU	41.75	5423.23	5339.24	2683.27

Considering the entire set of 378 matrices, AMD produces a better median fill-in than MMD, CMMD, and MA27 for 62%, 61%, and 81% of the matrices, respectively. AMD never generates more than 7%, 7%, and 4% more nonzeros in $\mathbf{L}$ than MMD, CMMD, and MA27, respectively. We have shown empirically that AMD produces an ordering at least as good as these other three methods for this large test set.

If the apparent slight difference in ordering quality between AMD and MMD is statistically significant, we conjecture that it has more to do with earlier supervariable detection (which affects the external degree) than with the differences between the external degree and our upper bound.

Table 6.5 lists the median ordering time (in seconds on a SUN SPARCstation 10) for each method. The ordering time for CMMD includes the time taken by the precompression algorithm. Ordering time twice that of the minimum median ordering time listed in the table (or higher) is underlined. On certain classes of matrices, typically those from structural analysis applications, CMMD is significantly faster than MMD. AMD is the fastest method for all but the EX19 matrix. For the other 352 matrices in our full test set, the differences in ordering time among these various methods is typically less. If we compare the ordering time of AMD with the other methods on all matrices in our test set requiring at least a tenth of a second of ordering time, then AMD is slower than MMD, CMMD, and MA27 for only 6, 15, and 8 matrices, respectively. For the full set of matrices, AMD is never more than 30% slower than these other methods. The best and worst cases for the relative run time of AMD for the smaller matrices are included in Table 6.5 (the EX19 and ORANI678 matrices).

7. Summary. We have described a new upper bound for the degree of nodes in the elimination graph that can be easily computed in the context of a minimum degree algorithm. We have demonstrated that this upper bound for the degree is more accurate than all previously used degree approximations. We have experimentally shown that we can replace an exact degree update by our approximate degree update and obtain almost identical fill-in.

An AMD algorithm based on external degree approximation has been described. We have shown that the AMD algorithm is highly competitive with other ordering algorithms. It is typically faster than other minimum degree algorithms and produces comparable results to MMD (which is also based on external degree) in terms of fill-in. AMD typically produces better results, in terms of fill-in and computing time, than the MA27 minimum degree algorithm (based on true degrees).

Acknowledgments. We would like to thank John Gilbert for outlining the $\overline{d}_i \leq \widehat{d}_i$ portion of the proof to Theorem 4.1, Joseph Liu for providing a copy of the MMD algorithm, and Cleve Ashcraft and Stan Eisenstat for their comments on a draft of this paper.

REFERENCES

[1] P. R. AMESTOY, *Factorization of Large Sparse Matrices Based on a Multifrontal Approach in a Multiprocessor Environment*, INPT Ph.D. thesis TH/PA/91/2, CERFACS, Toulouse, France, 1991.

[2] P. R. AMESTOY, M. DAYDÉ, AND I. S. DUFF, *Use of level 3 BLAS in the solution of full and sparse linear equations*, in High Performance Computing: Proc. of the International Symposium on High Performance Computing, Montpellier, France, 1989, North–Holland, Amsterdam, pp. 19–31.

[3] C. ASHCRAFT, *Compressed graphs and the minimum degree algorithm*, SIAM J. Sci. Comput., 16 (1995), pp. 1404–1411.

[4] C. ASHCRAFT, S. C. EISENSTAT, AND R. F. LUCAS, personal communication.

[5] A. BERGER, J. MULVEY, E. ROTHBERG, AND R. VANDERBEI, *Solving Multistage Stochastic Programs Using Tree Dissection*, Tech. report SOR-97-07, Program in Statistics and Operations Research, Princeton University, Princeton, NJ, 1995.

[6] T. H. CORMEN, C. E. LEISERSON, AND R. L. RIVEST, *Introduction to Algorithms*, MIT Press, Cambridge, MA, and McGraw–Hill, New York, 1990.

[7] T. A. DAVIS AND I. S. DUFF, *An Unsymmetric-Pattern Multifrontal Method for Sparse LU Factorization*, SIAM J. Matrix Anal. Appl., to appear.

[8] ———, *Unsymmetric-Pattern Multifrontal Methods for Parallel Sparse LU Factorization*, Tech. report TR-91-023, CISE Department, University of Florida, Gainesville, FL, 1991.

[9] I. S. DUFF, *On algorithms for obtaining a maximum transversal*, ACM Trans. Math. Software, 7 (1981), pp. 315–330.

[10] I. S. DUFF, A. M. ERISMAN, AND J. K. REID, *Direct Methods for Sparse Matrices*, Oxford University Press, London, UK, 1986.

[11] I. S. DUFF, R. G. GRIMES, AND J. G. LEWIS, *Sparse matrix test problems*, ACM Trans. Math. Software, 15 (1989), pp. 1–14.

[12] ———, *Users' Guide for the Harwell-Boeing Sparse Matrix Collection (Release 1)*, Tech. report RAL-92-086, Rutherford Appleton Laboratory, Didcot, Oxon, UK, 1992.

[13] I. S. DUFF AND J. K. REID, *A comparison of sparsity orderings for obtaining a pivotal sequence in Gaussian elimination*, J. Inst. Math. Appl., 14 (1974), pp. 281–291.

[14] ———, *MA27—A Set of Fortran Subroutines for Solving Sparse Symmetric Sets of Linear Equations*, Tech. report AERE R10533, HMSO, London, UK, 1982.

[15] ———, *The multifrontal solution of indefinite sparse symmetric linear equations*, ACM Trans. Math. Software, 9 (1983), pp. 302–325.

[16] ———, *The multifrontal solution of unsymmetric sets of linear equations*, SIAM J. Sci. Comput., 5 (1984), pp. 633–641.

[17] S. C. EISENSTAT, M. C. GURSKY, M. H. SCHULTZ, AND A. H. SHERMAN, *Yale sparse matrix package, I: The symmetric codes*, Internat. J. Numer. Meth. Eng., 18 (1982), pp. 1145–1151.

[18] S. C. EISENSTAT, M. H. SCHULTZ, AND A. H. SHERMAN, *Algorithms and data structures for sparse symmetric Gaussian elimination*, SIAM J. Sci. Comput., 2 (1981), pp. 225–237.

[19] A. GEORGE AND J. W. H. LIU, *A fast implementation of the minimum degree algorithm using quotient graphs*, ACM Trans. Math. Software, 6 (1980), pp. 337–358.

[20] ———, *A minimal storage implementation of the minimum degree algorithm*, SIAM J. Numer. Anal., 17 (1980), pp. 282–299.

[21] ———, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

[22] ———, *The evolution of the minimum degree ordering algorithm*, SIAM Rev., 31 (1989), pp. 1–19.

[23] A. GEORGE AND D. R. MCINTYRE, *On the application of the minimum degree algorithm to finite element systems*, SIAM J. Numer. Anal., 15 (1978), pp. 90–111.

[24] J. R. GILBERT, C. MOLER, AND R. SCHREIBER, *Sparse matrices in MATLAB: Design and implementation*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 333–356.

[25] J. W. H. LIU, *Modification of the minimum-degree algorithm by multiple elimination*, ACM Trans. Math. Software, 11 (1985), pp. 141–153.

[26] H. M. MARKOWITZ, *The elimination form of the inverse and its application to linear programming*, Management Sci., 3 (1957), pp. 255–269.

[27] D. J. ROSE, *Symmetric Elimination on Sparse Positive Definite Systems and the Potential Flow Network Problem*, Ph.D. thesis, Applied Mathematics Department, Harvard University, 1970.

[28] ———, *A graph-theoretic study of the numerical solution of sparse positive definite systems of linear equations*, in Graph Theory and Computing, R. C. Read, ed., Academic Press, New York, 1973, pp. 183–217.

[29] B. SPEELPENNING, *The Generalized Element Method*, Tech. report UIUCDCS-R-78-946, Department of Computer Science, University of Illinois, Urbana, IL, 1978.

[30] W. F. TINNEY AND J. W. WALKER, *Direct solutions of sparse network equations by optimally ordered triangular factorization*, Proc. IEEE, 55 (1967), pp. 1801–1809.

[31] 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. 17, No. 4, pp. 906–926, October 1996

012

ON THE SOLUTION OF A NONLINEAR MATRIX EQUATION ARISING IN QUEUEING PROBLEMS*

DARIO BINI† AND BEATRICE MEINI†

Abstract. By extending the cyclic reduction technique to infinite block matrices we devise a new algorithm for computing the solution G_0 of the matrix equation $G = \sum_{i=0}^{+\infty} G^i A_i$ arising in a wide class of queueing problems. Here A_i, $i = 0, 1, \ldots$, are $k \times k$ nonnegative matrices such that $\sum_{i=0}^{+\infty} A_i$ is column stochastic. Our algorithm, which under mild conditions generates a sequence of matrices converging quadratically to G_0, can be fully described in terms of simple operations between matrix power series, i.e., power series in z having matrix coefficients. Such operations, like multiplication and reciprocation modulo z^m, can be quickly computed by means of FFT-based fast polynomial arithmetic; here m is the degree where the power series are numerically cut off in order to reduce them to polynomials. These facts lead to a dramatic reduction of the complexity of solving the given matrix equation; in fact, $O(k^3 m + k^2 m \log m)$ arithmetic operations are sufficient to carry out each iteration of the algorithm. Numerical experiments and comparisons performed with the customary techniques show the effectiveness of our algorithm. For a problem arising from the modelling of metropolitan networks, our algorithm was about 30 times faster than the algorithms customarily used in the applications. Cyclic reduction applied to quasi-birth–death (QBD) problems, i.e., problems where $A_i = O$ for $i > 2$, leads to an algorithm similar to the one of [Latouche and Ramaswami, *J. Appl. Probab.*, 30 (1993), pp. 650–674], but which has a lower computational cost.

Key words. queueing problems, M/G/1-type matrices, cyclic reduction, Toeplitz matrices

AMS subject classifications. 15A51, 15A24, 60J10, 60K25, 65U05

1. Introduction.

Let A_i, $\widetilde{A}_{i+1}$, $i = 0, 1, 2, \ldots$, be two sequences of $k \times k$ nonnegative matrices such that $\sum_{i=0}^{+\infty} A_i$ and $\sum_{i=1}^{+\infty} \widetilde{A}_i$ are column stochastic. A matrix $A = \{a_{ij}\}_{ij}$ is column stochastic if $0 \le a_{i,j} \le 1$, $\sum_i a_{i,j} = 1$ for any j; here and hereafter a stochastic matrix will be a column-stochastic matrix. Consider the infinite stochastic matrix P defined by

$$P = \begin{pmatrix} \widetilde{A}_1 & A_0 & & \bigcirc & \\ \widetilde{A}_2 & A_1 & A_0 & & \\ \widetilde{A}_3 & A_2 & A_1 & A_0 & \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{1.1}$$

Matrices of structure (1.1) are called M/G/1-type matrices [15] and arise in the mathematical modelling of a wide and important class of queueing problems occurring in many applications, where the matrix P^T is the transition matrix associated with a Markov chain [7]. Matrices of this kind have the block Toeplitz structure, except on the first block column, and are in lower block Hessenberg form. We recall that a block Toeplitz matrix is a block matrix having (i, j)-blocks that depend on $i - j$ and that a matrix is in lower block Hessenberg form if its blocks are null for $j > i + 1$.

The computation of the probability invariant vector associated with P, that is, a nonnegative vector $\boldsymbol{\pi}$ such that

$$P\boldsymbol{\pi} = \boldsymbol{\pi}, \ ||\boldsymbol{\pi}|| = 1, \tag{1.2}$$

*Received by the editors April 17, 1995; accepted for publication (in revised form) by C. Meyer November 22, 1995.

†Dipartimento di Matematica, Università di Pisa, Pisa, Italy (bini@dm.unipi.it, meini@dm.unipi.it).

plays an important role in the analysis of the queue. Here, for a vector $\mathbf{v} = (v_i)$ we denote $||\mathbf{v}|| = \sum_i |v_i|$.

If matrix (1.1) is irreducible, the existence of a solution of (1.2) is equivalent to the positive recurrence of P [7]. In this case the solution $\boldsymbol{\pi}$ is unique and positive. For a complete and detailed study of matrices of type M/G/1 we refer the reader to [15].

Properties of the solution $\boldsymbol{\pi}$ and numerical techniques for its computation are strongly related to the minimal nonnegative solution of the matrix equation

$$G = \sum_{i=0}^{+\infty} G^i A_i. \tag{1.3}$$

In particular, it can be proved that if P is irreducible and positive recurrent then the only nonnegative solution G_0 of (1.3) is a stochastic matrix [15]. The matrix G_0 has very important properties; in fact, it is possible to express the solution $\boldsymbol{\pi} = \{\boldsymbol{\pi}_i\}_{i\geq 0}$ of (1.2) in terms of G_0 by means of the Ramaswami formula [15]:

$$\boldsymbol{\pi}_i = (I - A_1^*)^{-1} \left[\widetilde{A}_{i+1}^* \boldsymbol{\pi}_0 + \sum_{j=1}^{i-1} A_{i+1-j}^* \boldsymbol{\pi}_j \right],$$

where $\boldsymbol{\pi}_i$ is the k-dimensional vector obtained by partitioning the vector $\boldsymbol{\pi}$ according to structure (1.1) of P, and

$$A_i^* = \sum_{j=i}^{+\infty} G_0^{j-i} A_j, \quad \widetilde{A}_{i+1}^* = \sum_{j=i}^{+\infty} G_0^{j-i} \widetilde{A}_{j+1}, \quad i \geq 1. \tag{1.4}$$

The Ramaswami formula, due also to its numerical stability, is an effective tool for solving problem (1.2) once the solution G_0 of (1.3) has been computed. Thus, the availability of efficient numerical methods for computing the matrix G_0 is crucial in order to solve (1.2).

Different algorithms for computing the solution of (1.3) have been proposed and analyzed by several authors. Most of them are based on functional iteration techniques obtained by manipulating the matrix equation (1.3). For instance, in [16] the iteration

$$X_{j+1} = \sum_{i=0}^{+\infty} X_j^i A_i, \quad X_0 \geq O \tag{1.5a}$$

is considered. Similar techniques, based on the recurrences

$$X_{j+1} = A_0 (I - A_1)^{-1} + \sum_{i=2}^{+\infty} X_j^i A_i (I - A_1)^{-1} \tag{1.5b}$$

or

$$X_{j+1} = A_0 \left(I - \sum_{i=1}^{+\infty} X_j^{i-1} A_i \right)^{-1}, \tag{1.5c}$$

are introduced and analyzed in [15], [9], [14] in order to speed up the convergence. However, the convergence of these numerical schemes still remains linear. In [10] a

sort of Newton's iteration is introduced to arrive at a quadratic convergence with an extreme increase of the computational cost. In [12] the approximation of G_0 is reduced to solving nested finite systems of linear equations associated with the matrix P by means of a doubling technique. In this way the solution of the matrix P is approximated with the solution of the problem obtained by cutting the infinite block matrix P to a suitable finite block size n.

In this paper we propose a new method for the numerical solution of (1.3) that is quadratically convergent and numerically stable and avoids the necessity of cutting the infinite matrix (1.1) to a finite size. Moreover, this method, based on the use of FFT, requires a small number of arithmetic operations.

Our method relies on the technique of the successive state reduction (cyclic reduction [6]), introduced and analyzed in [2] for the computation of the vector $\boldsymbol{\pi}$ and here extended to the computation of the solution G_0 of (1.3).

In the case where the matrix P of (1.1) is block tridiagonal, i.e., $A_i = \widetilde{A}_i = O$ for $i > 2$, our method is very similar to that of Latouche and Ramaswami [11], where a modified cyclic reduction scheme was applied to compute the matrix G_0 associated with a QBD process, and, moreover, is slightly faster than the Latouche and Ramaswami algorithm.

Our method generates a sequence $P^{(j)}$ of matrices of type M/G/1, defined by the blocks $\{A_i^{(j)}\}_i$ and $\{\widetilde{A}_i^{(j)}\}_i$, such that the solution of the equation $G = \sum_{i=0}^{+\infty} G^i A_i^{(j)}$ is $G_0^{2^j}$. We associate with the sequence $\{P^{(j)}\}_j$ a sequence of $k \times k$ stochastic matrices $\{R^{(j)}\}_j$ which, under mild conditions, due to the quadratic convergence of the matrices $\{G_0^{2^j}\}_j$ to the limit $G' = \lim_j G_0^{2^j}$, tends quadratically to the matrix G' as j tends to $+\infty$. This allows us to devise a fast algorithm for the approximation of G'. The matrix G_0 is then approximated in a back-substitution stage by means of a suitable relation that involves G' and the matrices $A_i^{(j)}$. Moreover, we show that under further mild conditions the blocks $A_i^{(j)}$, $i > 1$, tend to the null matrix as j tends to $+\infty$.

The formulae that relate the sequence $\{A_i^{(j)}\}_i$ with the sequence $\{A_i^{(j+1)}\}_i$ are expressed in functional form in terms of the matrix power series $\varphi^{(j)}(z) = \sum_{i=0}^{+\infty} A_i^{(j)} z^i$ in the following way (compare [2]):

$$\varphi^{(j+1)}(z) = z\varphi_{odd}^{(j)}(z) + \varphi_{even}^{(j)}(z)\big(I - \varphi_{odd}^{(j)}(z)\big)^{-1}\varphi_{even}^{(j)}(z),$$

where $\varphi_{even}^{(j)}(z) = \sum_{i=0}^{+\infty} A_{2i}^{(j)} z^i$, $\varphi_{odd}^{(j)}(z) = \sum_{i=0}^{+\infty} A_{2i+1}^{(j)} z^i$. The computation of the coefficients of $\varphi^{(j+1)}(z)$, given $\varphi^{(j)}(z)$, is performed by applying the above functional relation modulo z^m, where m (cutting level) is an integer that can be dynamically computed and is such that the matrix $\sum_{i=0}^{m} A_i^{(j)}$ is "numerically stochastic"; i.e., $\mathbf{e}^T(I - \sum_{i=0}^{m} A_i^{(j)}) < \epsilon \mathbf{e}^T$, where ϵ is the machine precision. In this way the computation is reduced to computing products and reciprocals of matrix polynomials modulo z^m. This can be efficiently obtained by extending to matrix polynomials the FFT-based techniques which have been devised for scalar polynomial arithmetic [3], [2].

The computational cost of each step of our algorithm is $O(k^3 m + k^2 m \log m)$ arithmetic operations versus $O(k^3 m^2)$ operations needed if the customary polynomial arithmetic were used. The cost of the algorithm presented in [12] is $O(k^3 n \log^2 n)$, where n is the size at which the infinite matrix P is cut in order to reduce the infinite problem to a finite one (typically $n \geq m$). The cost of each step of the linearly convergent algorithms based on (1.5a), (1.5b), and (1.5c) is $O(k^3 m)$ operations.

We implemented our algorithm in Fortran 77 and performed comparisons with the iterative formula based on (1.5c) (which is the fastest among (1.5) (compare [14]))

for $X_0 = 0$ and $X_0 = I$, and with the algorithm of [12] on a problem arising from the mathematical modelling of a metropolitan network [1]. Our method was about 30 times faster than the method based on (1.5c) with $X_0 = I$, and about 170 times faster in the case $X_0 = 0$. The ratio between the number of iterations required by (1.5c) and our method was 167 for $X_0 = I$, and 945 for $X_0 = 0$. From the numerical computations that we performed, our method turned out to be much superior, especially in the cases of "long queues" where either the initial cutting level m of the queue is very large or the number of the nonnegligible components of the vector $\boldsymbol{\pi}$ is very large. In fact, in these cases the advantages of the quadratic convergence and the use of FFT are strongly evident. Another interesting feature of our algorithm that emerged from the numerical experiments is that the cutting level m_j at the jth recursive step of cyclic reduction has an almost decreasing behaviour. This contributes to a speedier computation.

The paper is organized as follows. In §2 we describe the cyclic reduction technique applied to matrices of type M/G/1 and prove the main convergence result. In §3 we give further convergence properties. In §4 we present two different algorithms for the computation of G_0 based on cyclic reduction. The first one, which is slightly faster, requires the storage of a certain number of blocks that are used in the back-substitution stage; the second one performs the back-substitution implicitly in the first stage and does not need to store auxiliary matrices. In §5 we describe our implementations by giving a detailed description of our algorithm together with the FFT-based techniques for manipulating matrix power series. Finally, in §6 we report the results of the numerical experiments and the comparisons with the known algorithms.

2. Cyclic reduction and its convergence properties. In this section we recall the method of cyclic reduction introduced and analyzed in [2] for the computation of the invariant vector $\boldsymbol{\pi}$ and prove some convergence results that are needed to extend the method to the computation of the matrix G_0.

Here and hereafter we assume that the stochastic matrix P is irreducible and positive recurrent; i.e.,

$$\rho = \mathbf{e}^T \sum_{i=1}^{+\infty} i A_i \mathbf{a} < 1, \tag{2.1}$$

where $\sum_{i=0}^{+\infty} A_i \mathbf{a} = \mathbf{a}$, $||\mathbf{a}|| = 1$, so that the matrix equation (1.3) has only one nonnegative solution, which is stochastic [15], and the vector $\boldsymbol{\pi}$ is positive. Under these hypotheses it is possible to prove that [15]

$$r\left(\sum_{i=1}^{+\infty} A_i^*\right) < 1, \tag{2.2}$$

where $r(A)$ denotes the spectral radius of the matrix A and A_i^*, $i = 1, 2, \ldots$, are defined in (1.4).

Let G_0 be the solution of (1.3) and set $H_0 = -A_0$, $H_1 = I - A_1$, $H_i = -A_i$, $i \geq 2$, $\widetilde{H}_1 = I - \widetilde{A}_1$, $\widetilde{H}_i = -\widetilde{A}_i$, $i \geq 2$, $W = \sum_{i=0}^{+\infty} G_0^i \widetilde{A}_{i+1}$. Then from (1.3) we may easily obtain the matrix equation

$$(I, G_0, G_0^2, \ldots) \begin{pmatrix} \widetilde{H}_1 & H_0 & & \bigcirc & \\ \widetilde{H}_2 & H_1 & H_0 & & \\ \widetilde{H}_3 & H_2 & H_1 & H_0 & \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix} = (I - W, O, O, \ldots). \tag{2.3}$$

Performing an odd–even permutation of the block rows and block columns in the matrix equation (2.3) we find that

$$(2.4)\qquad (G_0, G_0^3, \dots \mid I, G_0^2, \dots)\begin{pmatrix} T_1^{(0)} & W^{(0)} \\ Z^{(0)} & T_2^{(0)} \end{pmatrix} = (O, O, \dots \mid I - W, O, \dots),$$

where

$$T_1^{(0)} = \begin{pmatrix} H_1 & & & \bigcirc \\ H_3 & H_1 & & \\ H_5 & H_3 & H_1 & \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix},\quad T_2^{(0)} = \begin{pmatrix} \widetilde{H}_1 & & & \bigcirc \\ \widetilde{H}_3 & H_1 & & \\ \widetilde{H}_5 & H_3 & H_1 & \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix},$$

$$Z^{(0)} = \begin{pmatrix} H_0 & & & \bigcirc \\ H_2 & H_0 & & \\ H_4 & H_2 & H_0 & \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix},\quad W^{(0)} = \begin{pmatrix} \widetilde{H}_2 & H_0 & & & \bigcirc \\ \widetilde{H}_4 & H_2 & H_0 & & \\ \widetilde{H}_6 & H_4 & H_2 & H_0 & \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}.$$

Applying one step of block Gaussian elimination to the 2×2 block matrix defined in (2.4) yields

$$(2.5)\qquad (I, G_0^2, G_0^4, \dots)(T_2^{(0)} - Z^{(0)}{T_1^{(0)}}^{-1}W^{(0)}) = (I - W, O, O \dots).$$

It is interesting to point out (compare [2]) that the Schur complement $Q^{(1)} = T_2^{(0)} - Z^{(0)}{T_1^{(0)}}^{-1}W^{(0)}$ of $T_2^{(0)}$, which appears in (2.5), is such that $P^{(1)} = I - Q^{(1)}$ is a lower block Hessenberg matrix which, except for the first block column, has the block Toeplitz structure of the matrix $P = P^{(0)} = I - Q^{(0)}$ of (1.1) (Toeplitz-like structure). In other words, $Q^{(1)}$ is univocally determined by the blocks $\widetilde{H}_i^{(1)}$, $i \geq 1$, $H_i^{(1)}$, $i \geq 0$, defining the first two block columns of $Q^{(1)}$; in this way (2.5) can be rewritten as

$$(2.6)\qquad (I, G_0^2, G_0^4, \dots)\begin{pmatrix} \widetilde{H}_1^{(1)} & H_0^{(1)} & & \bigcirc & \\ \widetilde{H}_2^{(1)} & H_1^{(1)} & H_0^{(1)} & & \\ \widetilde{H}_3^{(1)} & H_2^{(1)} & H_1^{(1)} & H_0^{(1)} & \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix} = (I - W, O, O, \dots).$$

We may recursively apply the same reduction (cyclic reduction) to the matrix equation (2.6) and thus obtain a sequence of matrices $\{Q^{(j)}\}$ such that $P^{(j)} = I - Q^{(j)}$ has the same lower block Hessenberg and Toeplitz-like structure of (1.1). Each matrix $Q^{(j)}$ is univocally determined by the blocks $\widetilde{H}_i^{(j)}$, $i \geq 1$, $H_i^{(j)}$, $i \geq 0$, defining its first two block columns. The matrices $P^{(j)}$, $j = 0, 1, \dots$, have further properties, as shown by the following lemma.

LEMMA 2.1. *All the matrices $P^{(j)} = I - Q^{(j)}$ recursively generated by applying cyclic reduction to the stochastic, irreducible, and positive recurrent matrix $P^{(0)}$ of* (1.1) *are stochastic, irreducible, and positive recurrent.*

Proof. It is sufficient to prove the lemma for $P^{(1)}$; then the proof can be completed by induction. Since $\mathbf{e}^T P^{(0)} = \mathbf{e}^T$, i.e., $\mathbf{e}^T T_1^{(0)} + \mathbf{e}^T Z^{(0)} = \mathbf{0}$ and $\mathbf{e}^T W^{(0)} + \mathbf{e}^T T_2^{(0)} = \mathbf{0}$, from (2.5) we have $\mathbf{e}^T(I - P^{(1)}) = \mathbf{e}^T(T_2^{(0)} - Z^{(0)}{T_1^{(0)}}^{-1}W^{(0)}) = \mathbf{e}^T T_2^{(0)} + \mathbf{e}^T T_1^{(0)}{T_1^{(0)}}^{-1}W^{(0)} = \mathbf{e}^T T_2^{(0)} + \mathbf{e}^T W^{(0)} = \mathbf{0}$; that is, $P^{(1)}$ is stochastic. Now, in order

to prove the irreducibility of $P^{(1)}$, we first show that a reducible stochastic matrix M having a positive probability invariant vector $\mathbf{p}$ such that $M\mathbf{p} = \mathbf{p}$, $||\mathbf{p}|| = 1$, has at least a probability invariant vector having some null components. Indeed, without loss of generality, we may assume that $M = \begin{pmatrix} M_{1,1} & O \\ M_{2,1} & M_{2,2} \end{pmatrix}$, where $M_{1,1}$ and $M_{2,2}$ are square (possibly infinite) matrices and $M_{1,1}$ is irreducible. Let us partition the vector $\mathbf{p}$ as $\mathbf{p}^T = (\mathbf{p}_1^T, \mathbf{p}_2^T)^T$ according to the block structure of M. From the condition $M\mathbf{p} = \mathbf{p}$ we deduce that $M_{1,1}\mathbf{p}_1 = \mathbf{p}_1$; hence

$$\mathbf{e}^T M_{1,1}\mathbf{p}_1 = \mathbf{e}^T \mathbf{p}_1. \tag{2.7}$$

Since $\mathbf{p}_1 > \mathbf{0}$ and $\mathbf{e}^T M_{1,1} \leq \mathbf{e}^T$ (due to the stochasticity of M), we find that if a component of $\mathbf{e}^T M_{1,1}$ were less than 1, we would have $\mathbf{e}^T M_{1,1}\mathbf{p}_1 < \mathbf{e}^T \mathbf{p}_1$, which would contradict (2.7). Therefore $\mathbf{e}^T M_{1,1} = \mathbf{e}^T$; i.e., $M_{1,1}$ is stochastic and $M_{2,1} = O$. Thus it follows that the vector $\mathbf{y} = (\mathbf{p}_1^T, \mathbf{0})^T / ||\mathbf{p}_1||$ is such that $M\mathbf{y} = \mathbf{y}$, $||\mathbf{y}|| = 1$. Now, since the vector made up of the even block components of $\boldsymbol{\pi}$ is a positive probability invariant vector of $P^{(1)}$, we deduce that, if $P^{(1)}$ were reducible, then there would exist a vector $\mathbf{x} \geq \mathbf{0}$ having some null components such that $P^{(1)}\mathbf{x} = \mathbf{x}$, $||\mathbf{x}|| = 1$. In this way the vector $\mathbf{y} = (\mathbf{y}_1^T, \mathbf{y}_2^T)^T$ defined by $\mathbf{y}_1 = -T_1^{(0)^{-1}} W^{(0)}\mathbf{x}$, $\mathbf{y}_2 = \mathbf{x}$, is such that $P^{(0)}\Pi^T\mathbf{y} = \Pi^T\mathbf{y}$, where Π is the odd–even permutation matrix. This follows from the relation $\Pi Q^{(0)}\Pi^T = \begin{pmatrix} T_1^{(0)} & W^{(0)} \\ Z^{(0)} & T_2^{(0)} \end{pmatrix} = \begin{pmatrix} I & O \\ Z^{(0)}T_1^{(0)^{-1}} & I \end{pmatrix} \begin{pmatrix} T_1^{(0)} & W^{(0)} \\ O & Q^{(1)} \end{pmatrix}$, $Q^{(1)} = T_2^{(0)} - Z^{(0)}T_1^{(0)^{-1}} W^{(0)}$ (compare (2.4)). This contradicts the assumptions since the positive recurrent matrix $P^{(0)}$ cannot have a probability invariant vector with some null components. In order to prove that $P^{(1)}$ is positive recurrent it is sufficient to show that there exists a positive invariant vector of $P^{(1)}$ with finite norm (compare [7, p. 152]). This vector is given by the even blocks of $\boldsymbol{\pi}$.

The following lemma relates the matrices $A_i^{(j)}$ generated by the cyclic reduction to the matrix $G_0^{2^j}$.

LEMMA 2.2. (i) *For every positive integer j, the matrix $G_j = G_0^{2^j}$ is the minimal nonnegative solution of the matrix equation*

$$G = \sum_{i=0}^{+\infty} G^i A_i^{(j)}, \tag{2.8}$$

where $A_0^{(j)} = -H_0^{(j)}$, $A_1^{(j)} = I - H_1^{(j)}$, $A_i^{(j)} = -H_i^{(j)}$, $i \geq 2$, and $A_i^{(0)} = A_i$, $i \geq 0$.

(ii) *There exists $\lim_j G_j = G'$. Moreover, $G' = \mathbf{g}\,\mathbf{e}^T$, where $\mathbf{g} = (g_1, \ldots, g_k)^T$, is the nonnegative vector such that $\mathbf{e}^T\mathbf{g} = 1$, $G_0\mathbf{g} = \mathbf{g}$.*

(iii) *The sequence of matrices $\{E_j\}_j$, where E_j is such that $G_j = G' + E_j$ for $j \geq 0$, converges quadratically to zero.*

Proof. By extending (2.3)–(2.6) to the jth step of cyclic reduction it follows that $G_j = G_0^{2^j}$ is a nonnegative solution of (2.8). From the irreducibility and positive recurrence of the stochastic matrix $P^{(j)} = I - Q^{(j)}$, obtained at each step j of cyclic reduction, the nonnegative solution G_j of (2.8) is unique. The convergence results stated by Lemma 2.2 follow since G_0 is a stochastic matrix and the only eigenvalue having the largest modulus is 1 [15]. Indeed, let $G_0 = \mathbf{g}\,\mathbf{e}^T + E_0$ so that $\mathbf{e}^T E_0 = \mathbf{0}$, $E_0\mathbf{g} = \mathbf{0}$, and the spectral radius of E_0 is strictly less than 1. Then $G_j = \mathbf{g}\,\mathbf{e}^T + E_j$, where $E_j = E_0^{2^j}$; hence E_j tends quadratically to zero and $\lim_j G_j = \mathbf{g}\,\mathbf{e}^T$.

In [2] a functional formulation of the cyclic reduction method was given. More precisely, the matrix power series

$$\varphi^{(j)}(z) = \sum_{i=0}^{+\infty} A_i^{(j)} z^i, \qquad \widetilde{\varphi}^{(j)}(z) = \sum_{i=0}^{+\infty} \widetilde{A}_{i+1}^{(j)} z^i, \tag{2.9}$$

which converge for $|z| \leq 1$, have been introduced. It has been observed that the equations that relate the blocks $A_i^{(j+1)}$ with the blocks $A_i^{(j)}$, $i \geq 0$, and the blocks $\widetilde{A}_i^{(j+1)}$ with the blocks $\widetilde{A}_i^{(j)}$, $i \geq 1$, can be formally expressed in functional form as

$$\varphi^{(j+1)}(z) = z\varphi_{odd}^{(j)}(z) + \varphi_{even}^{(j)}(z)\big(I - \varphi_{odd}^{(j)}(z)\big)^{-1}\varphi_{even}^{(j)}(z), \tag{2.10a}$$

$$\widetilde{\varphi}^{(j+1)}(z) = \widetilde{\varphi}_{odd}^{(j)}(z) + \varphi_{even}^{(j)}(z)(I - \varphi_{odd}^{(j)}(z))^{-1}\widetilde{\varphi}_{even}^{(j)}(z), \tag{2.10b}$$

where

$$\begin{cases} \varphi_{even}^{(j)}(z) = \sum_{i=0}^{+\infty} A_{2i}^{(j)} z^i, \quad \varphi_{odd}^{(j)}(z) = \sum_{i=0}^{+\infty} A_{2i+1}^{(j)} z^i, \\ \widetilde{\varphi}_{even}^{(j)}(z) = \sum_{i=0}^{+\infty} \widetilde{A}_{2(i+1)}^{(j)} z^i, \quad \widetilde{\varphi}_{odd}^{(j)}(z) = \sum_{i=0}^{+\infty} \widetilde{A}_{2i+1}^{(j)} z^i. \end{cases} \tag{2.11}$$

These functional relations not only allow us to express in compact form the recurrence which this method is based on, but also provide the basic tool on which the efficient computation of the matrices $A_i^{(j)}$ relies. For the computational and numerical issues we refer the reader to §§4 and 5.

From the functional relations it is immediate to obtain the explicit expression for $A_0^{(j+1)}$; i.e.,

$$A_0^{(j+1)} = A_0^{(j)}(I - A_1^{(j)})^{-1}A_0^{(j)}. \tag{2.12}$$

THEOREM 2.3. *We have*

$$A_0^{(j)}\left(I - \sum_{i=1}^{+\infty} A_i^{(j)}\right)^{-1} = G' - \left(E_{1j} - \sum_{i=1}^{+\infty} E_{ij}A_i^{(j)}\right)\left(I - \sum_{i=1}^{+\infty} A_i^{(j)}\right)^{-1}, \tag{2.13}$$

where $E_{ij} = G_j^i - G' = G_0^{i2^j} - G'$. *Moreover, if the entries of the matrix* $(I - \sum_{i=1}^{+\infty} A_i^{(j)})^{-1}$ *are bounded above by a constant, then the sequence of matrices* $R^{(j)} = A_0^{(j)}(I - \sum_{i=1}^{+\infty} A_i^{(j)})^{-1}$ *converges quadratically to the matrix* G'.

Proof. Since the matrices $P^{(j)}$ are irreducible and positive recurrent, the matrix

$$S^{(j)} = \sum_{l=1}^{+\infty}\sum_{i=l}^{+\infty} G_j^{i-l}A_i^{(j)} \tag{2.14}$$

has spectral radius less than 1 (compare (2.2), (1.4), and [15]). Therefore, since $\sum_{i=1}^{+\infty} A_i^{(j)} \leq S^{(j)}$, for the Perron–Frobenius theorem [18] it holds that $r(\sum_{i=1}^{+\infty} A_i^{(j)}) \leq r(S^{(j)}) < 1$; that is, the matrix $I - \sum_{i=1}^{+\infty} A_i^{(j)}$ is nonsingular. By replacing G_j^i with

$G' + E_{ij}$ in (2.8) we arrive at (2.13). The quadratic convergence holds for Lemma 2.2, part (iii).

The matrix $(I - \sum_{i=1}^{+\infty} A_i^{(j)})^{-1}$ is bounded above in particular if there exists $P' = \lim_j P^{(j)}$ and the matrix P' is positive recurrent. In fact, if $P' = \lim_j P^{(j)}$ is positive recurrent, then there exists $S' = \lim_j S^{(j)}$ and $1 > r(S') \geq r(\sum_{i=1}^{+\infty} A_i')$, since $S' \geq \sum_{i=1}^{+\infty} A_i'$, where $A_i' = \lim_j A_i^{(j)}$.

It is interesting to point out that, in the case where the matrix P is block tridiagonal (as in the case of QBD processes), all the matrices $P^{(j)}$ generated by the cyclic reduction are block tridiagonal and the functions $\varphi^{(j)}(z)$ and $\widetilde{\varphi}^{(j)}(z)$ are matrix polynomials of degree 2 and 1, respectively. Thus the functional relations (2.10) become

$$\begin{aligned} \varphi^{(j+1)}(z) &= zA_1^{(j)} + (A_0^{(j)} + zA_2^{(j)})(I - A_1^{(j)})^{-1}(A_0^{(j)} + zA_2^{(j)}), \\ \widetilde{\varphi}^{(j+1)}(z) &= \widetilde{A}_1^{(j)} + (A_0^{(j)} + zA_2^{(j)})(I - A_1^{(j)})^{-1}\widetilde{A}_2^{(j)}. \end{aligned} \tag{2.15}$$

Similar relations were obtained by Latouche and Ramaswami in [11].

3. Further convergence properties. Under suitable assumptions, further useful convergence results can be proven. From (2.10a) we easily arrive at the following results.

LEMMA 3.1. *For every integer $j \geq 0$, the matrix $I - A_1^{(j)}$ is nonsingular and the nonnegative matrix sequence $\{A_1^{(j)}\}_j$ is nondecreasing. Moreover, there exists* $\lim_j A_1^{(j)} = A_1'$.

Proof. By following the same argument used in the proof of Theorem 2.3, since $r(S^{(j)}) < 1$ (compare (2.14)), it follows that for any j the matrix $I - A_1^{(j)}$ is nonsingular. Let us prove that the sequence of matrices $\{A_1^{(j)}\}_j$ is monotonically convergent. Comparing the coefficients of z in both sides of (2.10a), it follows that, for every $j \geq 0$, $A_1^{(j+1)} = A_1^{(j)} + C^{(j)}$, where $C^{(j)}$ is the coefficient of z in the series $\varphi_{even}^{(j)}(z)\big(I - \varphi_{odd}^{(j)}(z)\big)^{-1}\varphi_{even}^{(j)}(z)$. Since the coefficients of $\varphi_{even}^{(j)}(z)$ and $(I - \varphi_{odd}^{(j)}(z))^{-1}$ are nonnegative, we find that $C^{(j)} \geq 0$; thus $A_1^{(j+1)} \geq A_1^{(j)}$. Moreover, since $A_1^{(j)}$ has nonnegative entries not greater than 1, it follows that there exists $\lim_j A_1^{(j)} = A_1'$.

LEMMA 3.2. *For every convergent subsequence $\{A_0^{(\sigma_j)}\}_j$ of the sequence of matrices $\{A_0^{(j)}\}_j$, it holds that* $\lim_j A_0^{(\sigma_j)} \neq O$.

Proof. Let $\{A_0^{(\sigma_j)}\}_j$ be a convergent subsequence of $\{A_0^{(j)}\}_j$. We prove that if $\lim_j A_0^{(\sigma_j)} = O$, then $\lim_j G_{\sigma_j} = O$, which would contradict part (ii) of Lemma 2.2. In order to arrive at the condition $\lim_j G_{\sigma_j} = O$, we prove that for any $\epsilon > 0$ there exists a sequence of nonnegative integers $m(j)$ such that for any j it holds that

$$\mathbf{e}^T G_j - \epsilon \mathbf{e}^T \leq \mathbf{e}^T - \mathbf{e}^T (M^{(j)})^{m(j)}, \tag{3.1}$$

where $M^{(j)} = \sum_{h=1}^{+\infty} A_h^{(j)}$. First we introduce the matrix sequence defined for $i, j \geq 0$,

$$\begin{cases} X_{0,j} = O, \\ X_{i+1,j} = \sum_{n=0}^{+\infty} X_{i,j}^n A_n^{(j)}, \ i \geq 0 \end{cases} \tag{3.2}$$

such that $G_j = \lim_i X_{i,j}$. Then we observe that

$$\mathbf{e}^T X_{i,j} \leq \mathbf{e}^T - \mathbf{e}^T M^{(j)^i}, \tag{3.3}$$

as it can be easily proven by induction on i (we leave it to the reader). Since $0 \leq \mathbf{e}^T M^{(j)^i} \leq \mathbf{e}^T$, there exists an increasing sequence of nonnegative numbers $i_n = i_n(j)$ such that the subsequence $\{M^{(j)^{i_n}}\}_n$ is convergent. Therefore, from (3.3), we find that the matrix $G_j = \lim_n X_{i_n,j}$ satisfies the inequality

$$\mathbf{e}^T G_j \leq \mathbf{e}^T - \mathbf{e}^T \lim_n M^{(j)^{i_n}}.$$

Hence, fixed $\epsilon > 0$, there exists a positive integer $\nu = \nu(j)$ such that (3.1) holds with $m(j) = i_{\nu(j)}(j)$. If the sequence $\{A_0^{(\sigma_j)}\}_j$ converges to the null matrix, then the sequence of matrices $\{M^{(\sigma_j)}\}_j$ converges to a stochastic matrix since $\{\sum_{h=0}^{+\infty} A_h^{(\sigma_j)}\}_j$ is stochastic. Thus, by replacing j with σ_j in (3.1) and taking limits for $j \to +\infty$, we conclude that $\lim_j G_{\sigma_j} = O$.

Remark 3.3. We observe that from Lemma 3.2 it follows that, if $\sum_{i=1}^{+\infty} A_i^{(j)}$ is irreducible and no accumulation point of $\sum_{i=1}^{+\infty} A_i^{(j)}$ is reducible, then the matrix $(I - \sum_{i=1}^{+\infty} A_i^{(j)})^{-1}$ is bounded above for any j. In order to prove this we show that, under these assumptions, $r(\sum_{i=1}^{+\infty} A_i^{(j)}) \leq \gamma < 1$ for any j. Indeed, let $\gamma = \limsup_j r(\sum_{i=1}^{+\infty} A_i^{(j)})$ and let σ_j be an increasing sequence of nonnegative numbers such that $\gamma = \lim_j r(\sum_{i=1}^{+\infty} A_i^{(\sigma_j)}) = r(\lim_j \sum_{i=1}^{+\infty} A_i^{(\sigma_j)})$ and $\lim_j A_0^{(\sigma_j)} = A_0'$ (observe that σ_j exists since $\sum_{i=1}^{+\infty} A_i^{(j)}$ and $A_0^{(j)}$ belong to a compact set for any j). In this way, since $\sum_{i=0}^{+\infty} A_i^{(j)}$ is stochastic for any j, we have that $A' + A_0'$ is stochastic, where $A' = \lim_j \sum_{i=1}^{+\infty} A_i^{(\sigma_j)}$. Since A' is irreducible and, for Lemma 3.2, $A_0' \geq O$, $A_0' \neq O$, from the Perron–Frobenius theorem [18] we deduce that $1 = r(A' + A_0') > r(A') = \gamma$.

In light of Lemma 3.1, if $A_1^{(j)}$ is irreducible for $j = j_0$, then the conditions of Remark 3.3 on the irreducibility of $\sum_{i=1}^{+\infty} A_i^{(j)}$ and its accumulation points are satisfied.

LEMMA 3.4. *If the matrix G_0 is irreducible, then the sequence*

$$\theta_j = \mathbf{e}^T A_0^{(j)} (I - A_1^{(j)})^{-1} \mathbf{g} \tag{3.4}$$

is such that $\lim_j \theta_j = 1$.

Proof. Since $0 \leq \theta_j \leq 1$ for every integer j, it follows that the sequence θ_j has convergent subsequences whose limits belong to the set $[0, 1]$. We first prove that 0 cannot be a limit. Let $\{\theta_{\sigma_j}\}_j$ be a convergent subsequence. Assume without loss of generality that the sequence $\{A_0^{(\sigma_j)}\}_j$ is also convergent (the matrices $A_0^{(j)}$ belong to a compact set) and set $A_0' = \lim_j A_0^{(\sigma_j)}$. If $\lim_j \theta_{\sigma_j} = 0$, then from (3.4) it holds that $\lim_j A_0^{(\sigma_j)} = 0$ since the vector $\mathbf{g}$ is strictly positive, due to the irreducibility of G_0, and the nonnegative matrix $(I - A_1^{(j)})^{-1}$ has diagonal entries greater than 1. This contradicts Lemma 3.2. Now we show that any number θ such that $0 < \theta < 1$ cannot be the limit of any subsequence $\{\theta_{\sigma_j}\}_j$. From (2.8) it follows that

$$A_0^{(\sigma_j)} = G_{\sigma_j} - G_{\sigma_j} \sum_{i=1}^{+\infty} G_{\sigma_j}^{i-1} A_i^{(\sigma_j)};$$

by substituting $G_{\sigma_j} = \mathbf{g}\, \mathbf{e}^T + E_{\sigma_j}$ (compare Lemma 2.2) we obtain

$$A_0^{(\sigma_j)} = \mathbf{g}\, \mathbf{e}^T A_0^{(\sigma_j)} + E_{\sigma_j} \sum_{i=1}^{+\infty} G_{\sigma_j}^{i-1} A_i^{(\sigma_j)}.$$

Hence, by taking limits for $j \to \infty$, we find that $A_0' = \mathbf{g}\,\mathbf{e}^T A_0' = \mathbf{g}\,\mathbf{v}^T$, where $\mathbf{v}^T = \mathbf{e}^T A_0'$. From the recursive equation (2.12) it follows that the sequence $\mathbf{e}^T A_0^{(\sigma_j+1)}$ is convergent and

$$\lim_j \mathbf{e}^T A_0^{(\sigma_j+1)} = \theta \mathbf{v}^T.$$

It can be easily shown by induction on h that for every integer $h \geq 0$ there exists

$$\lim_j \mathbf{e}^T A_0^{(\sigma_j+h)} = \theta^{2^h-1} \mathbf{v}^T.$$

Hence, for every nonnegative integer h there exists a nonnegative integer j_h such that

$$\forall j \geq j_h \qquad ||\mathbf{e}^T A_0^{(\sigma_j+h)} - \theta^{2^h-1}\mathbf{v}^T|| < \theta^{2^h-1}. \tag{3.5}$$

Consider the strictly increasing sequence of nonnegative integers:

$$\begin{cases} \nu_0 = \sigma_{j_0}, \\ \nu_h = h + \min\{\sigma_j : j \geq j_h, \sigma_j > \nu_{h-1}\}, \quad h \geq 1. \end{cases}$$

It is readily seen from (3.5) that the sequence $\mathbf{e}^T A_0^{(\nu_h)}$ is such that $\mathbf{e}^T A_0^{(\nu_h)} \leq \mathbf{u}^T \theta^{2^h-1}$ for every nonnegative integer h, where $\mathbf{u}^T$ is a constant nonnegative vector. Hence the sequence of matrices $A_0^{(\nu_h)}$ converges to the null matrix, giving a contradiction, again for Lemma 3.2. Therefore, the unique accumulation point of the sequence $\{\theta_j\}_j$ is 1; hence the sequence $\{\theta_j\}_j$ converges to 1.

THEOREM 3.5. *If the matrix G_0 is irreducible, then the sequence of matrices $P^{(j)}$ converges to a stochastic matrix P' having the M/G/1 structure* (1.1). *Moreover,*

$$\lim_j A_0^{(j)}(I - A_1^{(j)})^{-1} = G'$$

and the matrix P', defined by the blocks $\widetilde{A}_i'$, $i \geq 1$, of its first block column and by the block entries A_i', $i \geq 0$, of its second block column, is such that

$$\widetilde{A}_i' = A_i' = O, \ i = 2, 3, \ldots.$$

Proof. To prove the convergence of the block entries $\widetilde{A}_i^{(j)}$, $i \geq 1$, we follow [2]. For this purpose, denote $\boldsymbol{\pi}^{(j)} = (\boldsymbol{\pi}_0^T, \boldsymbol{\pi}_{2^j}^T, \boldsymbol{\pi}_{2\cdot 2^j}^T, \ldots)^T$, where $\boldsymbol{\pi} = \boldsymbol{\pi}^{(0)}$ is the probability invariant vector defined in (1.2). In this way, due to the cyclic odd–even permutation, $\boldsymbol{\pi}^{(j)}$ is the probability invariant vector associated with $P^{(j)}$. For the positive recurrence of $P = P^{(0)}$, it follows that $\boldsymbol{\pi}_0 > 0$. Moreover, since $||\boldsymbol{\pi}|| = 1$, we have $\lim_j \boldsymbol{\pi}_{i\cdot 2^j} = \mathbf{0}$ for $i \geq 1$. Thus, from the condition $(I - P^{(j)})\boldsymbol{\pi}^{(j)} = \mathbf{0}$, we deduce that $\lim_j \widetilde{A}_i^{(j)} = O$ for $i \geq 2$. Let us now analyze the convergence of the block entries $A_i^{(j)}$, $i \geq 0$. Let $\{A_0^{(\sigma_j)}\}_j$ be a convergent subsequence of $\{A_0^{(j)}\}_j$. Then for Lemma 3.4 we have

$$\lim_j \mathbf{e}^T A_0^{(\sigma_j)}(I - A_1^{(\sigma_j)})^{-1}\mathbf{g} = 1.$$

Hence, for the stochasticity of the matrix $\sum_{i=0}^{+\infty} A_i^{(\sigma_j)}$, it holds that

$$\lim_j \mathbf{e}^T \sum_{i=2}^{+\infty} A_i^{(\sigma_j)}(I - A_1^{(\sigma_j)})^{-1}\mathbf{g} = 0.$$

Since the vector $\mathbf{g}$ is positive and the diagonal entries of the matrices $(I - A_1^{(\sigma_j)})^{-1}$ are greater than or equal to 1, it holds that

$$\lim_j \mathbf{e}^T \sum_{i=2}^{+\infty} A_i^{(\sigma_j)} = \mathbf{0}.$$

Therefore, the unique accumulation points of the sequence $\{P^{(j)}\}_j$ are stochastic matrices P' such that $A_i' = O$, $i = 2, 3, \ldots$. Moreover, for the convergence of the sequence G_j and from (2.8) we obtain that $A_0' = G' - G'A_1'$ and $G' = \lim_j A_0^{(j)}(I - A_1^{(j)})^{-1}$.

4. Computing the matrix G. The convergence results given by Theorem 2.3 can be used to derive two efficient and stable algorithms for computing the solution G_0 of the nonlinear matrix equation (1.3). In the following, the matrix $B = |A|$, where $A = \{a_{ij}\}_{ij}$, is the matrix defined by the entries $b_{ij} = |a_{ij}|$.

It is readily seen that, for every $j \geq 0$, it holds that

$$G_j = \left(\sum_{i=0}^{+\infty} G_j^{2i} A_{2i}^{(j)}\right)\left(I - \sum_{i=0}^{+\infty} G_j^{2i} A_{2i+1}^{(j)}\right)^{-1}, \tag{4.1}$$

where G_j is the unique nonnegative solution of the matrix equation (2.8). Observe that, for every j, the matrix $I - \sum_{i=0}^{+\infty} G_j^{2i} A_{2i+1}^{(j)}$ is nonsingular, since $\sum_{i=0}^{+\infty} G_j^{2i} A_{2i+1}^{(j)} \leq S^{(j)}$ (compare (2.14), (2.2), (1.4), and [15]). Moreover, by replacing G_j^2 with G_{j+1} (compare Lemma 2.2) in (4.1), we arrive at the following recursive relation for the sequence of matrices G_j:

$$G_j = \left(\sum_{i=0}^{+\infty} G_{j+1}^i A_{2i}^{(j)}\right)\left(I - \sum_{i=0}^{+\infty} G_{j+1}^i A_{2i+1}^{(j)}\right)^{-1}, \quad j \geq 0. \tag{4.2}$$

Equations (4.2), together with the convergence properties stated by Theorem 2.3, allow us to derive the following algorithm for computing the matrix G_0 either in the hypotheses of Theorem 3.5 or in the case where the entries of the matrices $(I - \sum_{i=1}^{+\infty} A_i^{(j)})^{-1}$ are bounded by a constant (compare Remark 3.3).

ALGORITHM 4.1.

1. Apply cyclic reduction by means of (2.10a) and compute the matrices $A_i^{(j)}$, $j = 1, 2, \ldots, q$, until one of the following conditions is satisfied:

$$|R^{(q)} - R^{(q-1)}| < \epsilon E, \ \mathbf{e}^T(I - A_0^{(q)}(I - A_1^{(q)})^{-1}) < \epsilon \mathbf{e}^T,$$

where, at each step j, $R^{(j)} = A_0^{(j)}(I - \sum_{i=1}^{+\infty} A_i^{(j)})^{-1}$, $\epsilon > 0$ is fixed, and E is the $k \times k$ matrix having all the entries equal to 1.

2. Compute an approximation of G_0 by replacing G_q with $R^{(q)}$ in (4.2) for $j = q-1$, and then apply back substitution in (4.2) for $j = q-1, q-2, \ldots, 0$, until the approximation of G_0 is obtained.

In order to apply the back-substitution stage of Algorithm 4.1, one must store the blocks $A_i^{(j)}$ for $j \leq q$ computed at stage 1. This may require a large amount of memory in the case where the matrices $A_i^{(j)}$ are negligible only for large values of i.

This problem is overcome by the next algorithm, where back substitution is implicitly performed in the cyclic reduction stage and no storage is required.

Observe that the solution G_0 of (1.3) satisfies the matrix equation

$$G_0(I, G_0, G_0^2, \ldots) \begin{pmatrix} \widehat{H}_1 & H_0 & & \bigcirc & \\ \widehat{H}_2 & H_1 & H_0 & & \\ \widehat{H}_3 & H_2 & H_1 & H_0 & \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix} = (A_0, O, O, \ldots), \tag{4.3}$$

where $H_0 = -A_0$, $H_1 = I - A_1$, $H_i = -A_i$, $i \geq 2$, $\widehat{H}_1 = I - A_1$, $\widehat{H}_i = -A_i$, $i \geq 2$. By recursively applying cyclic reduction we obtain (as can easily be seen by extending (2.3)–(2.6) to equation (4.3)) the sequence of matrix equations

$$G_0(I, G_j, G_j^2, \ldots) \begin{pmatrix} \widehat{H}_1^{(j)} & H_0^{(j)} & & \bigcirc & \\ \widehat{H}_2^{(j)} & H_1^{(j)} & H_0^{(j)} & & \\ \widehat{H}_3^{(j)} & H_2^{(j)} & H_1^{(j)} & H_0^{(j)} & \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix} = (A_0, O, O, \ldots), \quad j \geq 1, \tag{4.4}$$

where $G_j = G_0^{2^j}$ and the block entries $H_i^{(j)}$ coincide with the block entries of the second block column of the matrix $Q^{(j)}$ of §2. The matrices $\widehat{A}_1^{(j)} = I - \widehat{H}_1^{(j)}$, $\widehat{A}_i^{(j)} = -\widehat{H}_i^{(j)}$, $i \geq 2$, $A_0^{(j)} = -H_0^{(j)}$, $A_1^{(j)} = I - H_1^{(j)}$, $A_i^{(j)} = -H_i^{(j)}$, $i \geq 2$, can still be represented by the functional equations (2.10b) and (2.10a), respectively, where in (2.10b) the functions $\widetilde{\varphi}^{(j)}(z)$ must be replaced by $\widehat{\varphi}^{(j)}(z) = \sum_{i=0}^{+\infty} \widehat{A}_{i+1}^{(j)} z^i$, with $\widehat{\varphi}^{(0)}(z) = \sum_{i=0}^{+\infty} A_{i+1} z^i$, so that (2.10b) is replaced by

$$\widehat{\varphi}^{(j+1)}(z) = \widehat{\varphi}_{odd}^{(j)}(z) + \varphi_{even}^{(j)}(z)(I - \varphi_{odd}^{(j)}(z))^{-1} \widehat{\varphi}_{even}^{(j)}(z) \tag{4.5}$$

for $\widehat{\varphi}_{odd}^{(j)}(z) = \sum_{i=0}^{+\infty} \widehat{A}_{2i+1}^{(j)} z^i$ and $\widehat{\varphi}_{even}^{(j)}(z) = \sum_{i=0}^{+\infty} \widehat{A}_{2(i+1)}^{(j)} z^i$. Since from (4.5) it holds that $\mathbf{e}^T \widehat{\varphi}^{(j)}(1) = \mathbf{e}^T \widehat{\varphi}^{(0)}(1) = \mathbf{e}^T - \mathbf{e}^T A_0$, then we have $\mathbf{e}^T \sum_{i=1}^{+\infty} \widehat{A}_i^{(j)} = \mathbf{e}^T - \mathbf{e}^T A_0$ for every $j \geq 0$. Moreover, if the matrix $\sum_{i=0}^{+\infty} G_j^i \widehat{H}_{i+1}^{(j)} = I - \sum_{i=0}^{+\infty} G_j^i \widehat{A}_{i+1}^{(j)}$ is nonsingular, then it holds that

$$G_0 = A_0 \left(I - \sum_{i=0}^{+\infty} G_j^i \widehat{A}_{i+1}^{(j)} \right)^{-1}. \tag{4.6}$$

Suppose that, for every $j \geq 0$, the matrix $I - \sum_{i=0}^{+\infty} G_j^i \widehat{A}_{i+1}^{(j)}$ is nonsingular. Observe that if the matrix A_0 has no null columns, then the matrix $I - \sum_{i=0}^{+\infty} G_j^i \widehat{A}_{i+1}^{(j)}$ is nonsingular for any j. Then the following algorithm for the computation of the matrix G_0 can be carried out.

ALGORITHM 4.2.

1. Apply cyclic reduction to (4.3) by means of (2.10a) and (4.5) for $\varphi^{(0)}(z) = \sum_{i=0}^{+\infty} A_i z^i$, $\widehat{\varphi}^{(0)}(z) = \sum_{i=0}^{+\infty} A_{i+1} z^i$, and compute the matrices $A_i^{(j)}$, $\widehat{A}_i^{(j)}$, $j = 1, 2, \ldots, q$, until one of the following conditions is satisfied:

$$|R^{(q)} - R^{(q-1)}| < \epsilon E, \ \mathbf{e}^T (I - A_0^{(q)} (I - A_1^{(q)})^{-1}) < \epsilon \mathbf{e}^T, \tag{4.7}$$

where, at each step j, $R^{(j)} = A_0^{(j)}(I - \sum_{i=1}^{+\infty} A_i^{(j)})^{-1}$, $\epsilon > 0$ is fixed, and E is the $k \times k$ matrix having all the entries equal to 1.

2. Compute an approximation of G_0 by replacing G_q with $R^{(q)}$ in (4.6) for $j = q$.

We observe that both Algorithms 4.1 and 4.2 involve only multiplications and additions of nonnegative matrices and inversions of M-matrices [18], i.e., matrices having nonpositive off-diagonal entries and nonnegative inverses. These computations, if the diagonal adjustment technique [8] is used, can be reduced to performing additions of positive numbers, multiplications, and divisions. This makes such computations strongly numerically stable; i.e., the relative rounding errors in the results can be bounded componentwise.

By following the proof of Theorem 3.5 to show the convergence to zero of the blocks $\widetilde{A}_i^{(j)}$, $i \geq 2$, for $j \to +\infty$, we now prove the following theorem.

THEOREM 4.1. *For the matrices* $\widehat{A}_i^{(j)} = -\widehat{H}_i^{(j)}$, $i \geq 2$, *of* (4.4) *it holds that* $\lim_j \widehat{A}_i^{(j)} = O$ *for any* $i \geq 2$.

Proof. The probability invariant vector $\boldsymbol{\pi}$ of (1.2) satisfies the following equation:

$$\begin{pmatrix} I - A_1 & -A_0 & & \bigcirc & \\ -A_2 & I - A_1 & -A_0 & & \\ -A_3 & -A_2 & I - A_1 & -A_0 & \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} \boldsymbol{\pi}_1 \\ \boldsymbol{\pi}_2 \\ \boldsymbol{\pi}_3 \\ \vdots \end{pmatrix} = \begin{pmatrix} \widetilde{A}_2 \\ \widetilde{A}_3 \\ \widetilde{A}_4 \\ \vdots \end{pmatrix} \boldsymbol{\pi}_0,$$

where the block matrix in the left-hand side is the same block matrix of (4.3). Thus, applying j steps of cyclic reduction to the above system, we find that

$$\begin{pmatrix} \widehat{H}_1^{(j)} & H_0^{(j)} & & \bigcirc & \\ \widehat{H}_2^{(j)} & H_1^{(j)} & H_0^{(j)} & & \\ \widehat{H}_3^{(j)} & H_2^{(j)} & H_1^{(j)} & H_0^{(j)} & \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} \boldsymbol{\pi}_1 \\ \boldsymbol{\pi}_{2^j+1} \\ \boldsymbol{\pi}_{2\cdot 2^j+1} \\ \vdots \end{pmatrix} = \begin{pmatrix} B_1^{(j)} \\ B_2^{(j)} \\ B_3^{(j)} \\ \vdots \end{pmatrix} \boldsymbol{\pi}_0,$$

where $B_i^{(j)}$ is such that $B_i^{(j)} \geq O$ for every $i \geq 1$. Since $\lim_j \boldsymbol{\pi}_{i\cdot 2^j+1} = \mathbf{0}$, for $i \geq 1$, we have

$$\lim_j \left(\begin{pmatrix} B_1^{(j)} \\ B_2^{(j)} \\ B_3^{(j)} \\ \vdots \end{pmatrix} \boldsymbol{\pi}_0 - \begin{pmatrix} \widehat{H}_1^{(j)} \\ \widehat{H}_2^{(j)} \\ \widehat{H}_3^{(j)} \\ \vdots \end{pmatrix} \boldsymbol{\pi}_1 \right) = \lim_j \left(\begin{pmatrix} B_1^{(j)} \\ B_2^{(j)} \\ B_3^{(j)} \\ \vdots \end{pmatrix} \boldsymbol{\pi}_0 + \begin{pmatrix} \widehat{A}_1^{(j)} - I \\ \widehat{A}_2^{(j)} \\ \widehat{A}_3^{(j)} \\ \vdots \end{pmatrix} \boldsymbol{\pi}_1 \right) = \mathbf{0}.$$

Therefore, for the conditions $\boldsymbol{\pi}_0 > 0$, $\boldsymbol{\pi}_1 > 0$, $B_i^{(j)} \geq O$, $\widehat{A}_i^{(j)} \geq O$, we deduce that $\lim_j B_i^{(j)} = \lim_j \widehat{A}_i^{(j)} = O$ for $i \geq 2$.

The result of Theorem 4.1 suggests adding to the above stop conditions (4.7) the following one:

$$\mathbf{e}^T(I - A_0(I - \widehat{A}_1^{(q)})^{-1}) < \epsilon \mathbf{e}^T. \tag{4.8}$$

If this condition is verified then the matrix G_0 can be readily approximated by $A_0(I - \widehat{A}_1^{(q)})^{-1}$.

In the case of QBD processes, where P is a block tridiagonal matrix, the matrix equation (1.3) becomes $G = A_0 + GA_1 + G^2A_2$ and Algorithm 4.2 can be rewritten as follows (compare (2.15)).

ALGORITHM 4.3. *QBD processes.*

1. Apply cyclic reduction to (4.3) by means of the following formulae:

$$\begin{aligned}
A_0^{(j+1)} &= A_0^{(j)}(I - A_1^{(j)})^{-1}A_0^{(j)},\\
A_1^{(j+1)} &= A_1^{(j)} + A_0^{(j)}(I - A_1^{(j)})^{-1}A_2^{(j)} + A_2^{(j)}(I - A_1^{(j)})^{-1}A_0^{(j)},\\
A_2^{(j+1)} &= A_2^{(j)}(I - A_1^{(j)})^{-1}A_2^{(j)},\\
\widehat{A}_1^{(j+1)} &= \widehat{A}_1^{(j)} + A_0^{(j)}(I - A_1^{(j)})^{-1}\widehat{A}_2^{(j)}
\end{aligned}$$

with $A_0^{(0)} = A_0$, $A_1^{(0)} = A_1$, $A_2^{(0)} = A_2, \widehat{A}_1^{(0)} = A_1$ for $j = 1, 2, \ldots, q$, until one of the following two conditions is satisfied: $\mathbf{e}^T(I - A_0^{(q)}(I - A_1^{(q)})^{-1}) < \epsilon\mathbf{e}^T$ or $\mathbf{e}^T(I - A_0^{(q)}(I - \widehat{A}_1^{(q)})^{-1}) < \epsilon\mathbf{e}^T$.

2. Compute an approximation $\widetilde{G}$ of G_0 by means of $\widetilde{G} = A_0(I - \widehat{A}_1^{(q)})^{-1}$.

Observe that in the QBD case it holds that $\widehat{A}_2^{(j)} = A_2^{(j)}$. This leads to a saving of the computational cost.

Each step of Algorithm 4.3 requires six matrix multiplications and one matrix inversion. The algorithm of [11] is similar to Algorithm 4.3 but it requires eight matrix multiplications and one matrix inversion per step.

5. Implementation of Algorithm 4.2 by means of FFT-based techniques. In this section we give a detailed description of the implementation of Algorithm 4.2, which we sketched in §4.

According to the customary way of dealing with infinite sequences of blocks [15], [9], we truncate the sequences $\{A_i^{(j)}\}_i$ and $\{\widehat{A}_i^{(j)}\}_i$ to the indexes m_j and $\widehat{m}_j$, respectively, such that $\mathbf{e}^T(I - \sum_{i=0}^{m_j} A_i^{(j)}) < \epsilon\mathbf{e}^T$ and $\mathbf{e}^T(I - A_0 - \sum_{i=1}^{\widehat{m}_j} \widehat{A}_i^{(j)}) < \epsilon\mathbf{e}^T$, where ϵ is an upper bound to the machine precision of the floating point arithmetic used in the computation and is related to the accuracy of the computed approximation of G_0. In other words, we consider only those blocks that make the matrices $\sum_{i=0}^{m_j} A_i^{(j)}$ and $A_0 + \sum_{i=1}^{\widehat{m}_j} \widehat{A}_i^{(j)}$ numerically stochastic with respect to ϵ. A nonnegative matrix A is said to be *numerically stochastic with respect to* ϵ, or ϵ-*stochastic*, if $\mathbf{e}^T(I - A) < \epsilon\mathbf{e}^T$. In this way relations (2.10a) and (4.5) can be rewritten in terms of matrix polynomials in the variable z, instead of matrix power series, and all the operations involved in (2.10a) and (4.5) are thus reduced to multiplying matrix polynomials and inverting a matrix polynomial modulo z^m for a suitable m.

This kind of truncation seems to be quite natural since it fits in with the floating point arithmetic used in the computation in the sense that only the terms that cannot add information to the problem are neglected. A different meaning has the truncation used in [12], where the user must replace the infinite matrix (1.1) with a finite one. The determination of the cutting level in the latter case constitutes a nontrivial problem. It is not easy to recover an a priori estimate of this cutting level. Even though this value could be dynamically determined by a suitable modification of the algorithm, its value should be at least greater than or equal to m_0; otherwise, some significant input data would be lost.

By extending to matrix power series the known FFT-based techniques for dealing with scalar power series (see [3]), we now introduce the basic tools on which the fast version of our algorithm is based.

The use of FFT techniques allows us to reduce the cost of the jth step of cyclic reduction from $O(k^3 n_j^2)$ operations (needed if the customary way of dealing with matrix power series is used) to $O(k^3 n_j + k^2 n_j \log n_j)$ operations, where $n_j = \max\{m_j, \widehat{m}_j\}$. This computational advantage is paid by a slight deterioration of the numerical features of the algorithm. In fact, customary techniques for manipulating power series related to stochastic matrices are strongly numerically stable; i.e., the relative errors generated by the floating point arithmetic can be bounded entrywise, whereas with the use of FFT it is possible to give very good relative error bounds, but only in terms of norm (weak stability [3], [13]). However, in the many numerical experiments that we performed we noticed no substantial difference between the numerical behaviour of the two diverse implementations.

We describe the FFT-based techniques for matrix polynomials via the following. We first introduce Algorithm 5.1 for matrix polynomial multiplication modulo $z^m - 1$, and then use it for the computation of the product of a block Toeplitz matrix and a block vector. Then we reduce the inversion of a matrix polynomial modulo z^m to inverting a lower block triangular block Toeplitz matrix, and we introduce Algorithm 5.2 for this computation. Finally, we describe Algorithm 5.3, which expands, with full computational details, Algorithm 4.2 for the numerical solution of the matrix equation (1.3).

Throughout this section we will denote by $\mathbf{y} = (y_0, \ldots, y_{d-1}) = \mathrm{DFT}(\mathbf{x})$ the discrete Fourier transform of the vector $\mathbf{x} = (x_0, \ldots, x_{d-1})$ defined by $y_i = \sum_{j=0}^{d-1} \omega^{ij} x_j$, $i = 0, \ldots, d-1$, where $\omega = \cos\frac{2\pi}{d} + \mathbf{i} \sin\frac{2\pi}{d}$, and $\mathbf{i}$ is the imaginary unit such that $\mathbf{i}^2 = -1$. Analogously, $\mathbf{x} = \mathrm{IDFT}(\mathbf{y})$ denotes the inverse discrete Fourier transform such that $\mathbf{x} = \frac{1}{d} \sum_{j=0}^{d-1} \omega^{-ij} y_j$, $i = 0, \ldots, d-1$. It is well known that if $d = 2^M$, where M is a positive integer, the computation of DFT and IDFT can be performed with $O(d \log d)$ real arithmetic operations (hereafter denoted by ops) by means of the base-2 FFT algorithms. More precisely, in the case of real input the cost of DFT and IDFT is $\frac{5}{2} d \log d + O(1)$ ops and $\frac{5}{2} d \log d + d + O(1)$ ops, respectively, if we do not count the cost of computing the dth roots of 1.

Similarly, we extend the above notation to block vectors. Given the block vector $X = (X_0, \ldots, X_{d-1})$, where $X_s = (x_{i,j}^{(s)})$ are $k \times k$ matrices, we define by $Y = (Y_0, \ldots, Y_{d-1}) = \mathrm{DFT}(X)$ the block vector such that $Y_s = (y^{(s)})_{i,j}$ are $k \times k$ matrices and $(y_{i,j}^{(0)}, \ldots, y_{i,j}^{(d-1)}) = \mathrm{DFT}(x_{i,j}^{(0)}, \ldots, x_{i,j}^{(d-1)})$, $i, j = 1, \ldots, k$. Analogously, we set $X = \mathrm{IDFT}(Y)$ if $(x_{i,j}^{(0)}, \ldots, x_{i,j}^{(d-1)}) = \mathrm{IDFT}(y_{i,j}^{(0)}, \ldots, y_{i,j}^{(d-1)})$. In this way the computation of $\mathrm{DFT}(X)$ and $\mathrm{IDFT}(Y)$ is reduced to the computation of k^2 DFT's and IDFT's, respectively, for the cost of $O(k^2 d \log d)$ ops.

Let $P(z) = \sum_{i=0}^{m-1} P_i z^i$, $Q(z) = \sum_{i=0}^{m-1} Q_i z^i$, and $R(z) = \sum_{i=0}^{m-1} R_i z^i$ be matrix polynomials such that $R(z) = P(z)Q(z) \bmod (z^m - 1)$. Then the matrix coefficients of $R(z)$ can be computed, given the matrix coefficients of $P(z)$ and $Q(z)$, by means of the following algorithm.

ALGORITHM 5.1. *Computation of matrix polynomial product modulo* $z^m - 1$.

Input. Positive integers M, k and the $k \times k$ matrices $P_0, \ldots, P_{m-1}, Q_0, \ldots, Q_{m-1}$, coefficients of the matrix polynomials $P(z) = \sum_{i=0}^{m-1} P_i z^i$, $Q(z) = \sum_{i=0}^{m-1} Q_i z^i$, where $m = 2^M$.

Output. The $k \times k$ matrices $R_0, \ldots, R_{m-1}$, coefficients of the matrix polynomial

$R(z) = \sum_{i=0}^{m-1} R_i z^i$ such that $R(z) = P(z)Q(z) \bmod (z^m - 1)$.

Computation.

1. (Evaluation) Compute the entries of the $2m$ matrices $U_s = P(\omega^s)$, $V_s = Q(\omega^s)$, $s = 0, \ldots, m-1$, $\omega = \cos\frac{2\pi}{m} + \mathbf{i}\sin\frac{2\pi}{m}$, in the following way:

$$(U_0, \ldots, U_{m-1}) = \mathrm{DFT}(P_0, \ldots, P_{m-1}),\ (V_0, \ldots, V_{m-1}) = \mathrm{DFT}(Q_0, \ldots, Q_{m-1}).$$

2. Compute the m matrix products $W_s = U_s V_s$, $s = 0, \ldots, m-1$, such that $W_s = P(\omega^s)Q(\omega^s) = R(\omega^s)$.
3. (Interpolation) Compute the entries of the matrices $R_0, \ldots, R_{m-1}$ by means of the equation $(R_0, \ldots, R_{m-1}) = \mathrm{IDFT}(W_0, \ldots, W_{m-1})$.

The cost of Algorithm 5.1 is $O(k^3 m + k^2 m \log m)$ ops. More precisely, for real input matrices the cost is less than $5k^2 m \log m + m(3k^3 + 2k^2)$. Indeed, due to the real input we have $W_s = \bar{W}_{m-s}$, $s = 1, \ldots, m/2 - 1$, where $\bar{W}_{m-s}$ is the complex conjugate of W_{m-s}; moreover, W_0, $W_{m/2}$ are real matrices. Thus the computation at stage 2 is reduced to $m/2 - 1$ complex matrix products and 2 real matrix products. Complex matrix products are computed by means of the algorithm, which uses three multiplications and five additions [5].

Remark 5.1. Observe that if the input polynomials $P(z)$ and $Q(z)$ have degree at most $m/2 - 1$ then $R(z) = P(z)Q(z)$; thus, the matrix polynomial product can be computed by means of Algorithm 5.1.

Remark 5.2. Algorithm 5.1 can also be used for efficiently computing the product of a block Toeplitz matrix and a vector. In fact, if $P(z)$ has degree $m-1$ and $Q(z)$ has degree $m/2 - 1$, then, comparing the terms of degree $m/2, \ldots, m-1$ in the equation $P(z)Q(z) = R(z) \bmod (z^m - 1)$, we arrive at the following equation involving a block Toeplitz matrix:

$$\begin{pmatrix} R_{m/2} \\ \vdots \\ R_{m-1} \end{pmatrix} = \begin{pmatrix} P_{m/2} & \ldots & P_1 \\ \vdots & \ddots & \vdots \\ P_{m-1} & \ldots & P_{m/2} \end{pmatrix} \begin{pmatrix} Q_0 \\ \vdots \\ Q_{m/2-1} \end{pmatrix}.$$

Comparing the coefficients of the term of degree i in both sides of the equation $P(z)Q(z) = I \bmod z^m$ yields the matrix equation involving a lower block triangular block Toeplitz matrix:

$$\begin{pmatrix} P_0 & & & O \\ P_1 & P_0 & & \\ \vdots & \ddots & \ddots & \\ P_{m-1} & \ldots & P_1 & P_0 \end{pmatrix} \begin{pmatrix} Q_0 \\ Q_1 \\ \vdots \\ Q_{m-1} \end{pmatrix} = \begin{pmatrix} I \\ O \\ \vdots \\ O \end{pmatrix}. \tag{5.1}$$

Let us denote by T_m the matrix in (5.1). Assume for simplicity that $m = 2^M$ and M is positive integer and partition T_m as follows:

$$T_m = \begin{pmatrix} T_{m/2} & O \\ H_{m/2} & T_{m/2} \end{pmatrix},$$

where all four blocks are $(m/2) \times (m/2)$ block Toeplitz matrices. Then the matrix T_m^{-1} can easily be written as

$$T_m^{-1} = \begin{pmatrix} T_{m/2}^{-1} & O \\ -T_{m/2}^{-1} H_{m/2} T_{m/2}^{-1} & T_{m/2}^{-1} \end{pmatrix} \tag{5.2}$$

and the first block column $(Q_0^T, \dots, Q_{m-1}^T)^T$ of T_m^{-1}, which solves (5.1), can be computed from the first block column $(Q_0^T, \dots, Q_{m/2-1}^T)^T$ of $T_{m/2}$ by means of two multiplications between an $(m/2) \times (m/2)$ block Toeplitz matrix and a block vector. In this way we arrive at the following algorithm, which extends to the block case the algorithm of Lafon and Sieveking–Kung (compare [4]).

ALGORITHM 5.2. *Solution of the congruence* $P(z)Q(z) = I \bmod z^m$ *or, equivalently, of system* (5.1).

Input. A positive integer M, the $k \times k$ matrices $P_0, P_1, \dots, P_{m-1}$, $m = 2^M$, defining the matrix polynomial $P(z) = \sum_{i=0}^{m-1} P_i z^i$, $\det P_0 \neq 0$.

Output. The matrices $Q_0, Q_1, \dots, Q_{m-1}$ satisfying (5.1) or, equivalently, such that the polynomial $Q(z) = \sum_{i=0}^{m-1} Q_i z^i$ solves the congruence $P(z)Q(z) = I \bmod z^m$.

Computation.

1. Compute $Q_0 = P_0^{-1}$.
2. For $i = 0, \dots, M-1$, given the first column $U = (Q_0^T, \dots, Q_{2^i-1}^T)^T$ of $T_{2^i}^{-1}$, compute the block vector $V = (Q_{2^i}^T, \dots, Q_{2^{i+1}-1}^T)^T$, which defines the remaining blocks of the first column of $T_{2^{i+1}}^{-1}$, by applying equation (5.2) with $m = 2^{i+1}$ in the following way: by applying Algorithm 5.1 and Remark 5.2 compute the products $W = H_{2^i} U$ and $V = T_{2^i}^{-1} W$.

Observe that at the ith stage of Algorithm 5.2 the computation of three DFT's and two IDFT's of order 2^{i+1} must be performed, together with the multiplication of about 2^{i+1} complex $k \times k$ matrices. The overall cost is about $6mk^3 + \frac{25}{2} k^2 m \log m$ ops.

Remark 5.3. In our applications $P(z) = I - \varphi_{odd}^{(j)}(z)$ (compare (2.11)) is such that the power series $P(z)^{-1} = \sum_{i=0}^{+\infty} Q_i z^i$ is convergent for $|z| = 1$. In this way we have $\lim_i Q_i = O$; thus $P(z)^{-1}$ can be numerically truncated to the polynomial $Q(z) = \sum_{i=0}^{m} Q_i z^i$ for a suitable m such that

$$\mathbf{e}^T (P(1)^{-1} - Q(1)) < \epsilon \mathbf{e}^T. \tag{5.3}$$

The value of the integer m such that (5.3) holds can be dynamically adjusted by modifying Algorithm 5.2 in such a way to test, at each step i, inequality (5.3).

In the following, we will denote with the same symbols $\varphi^{(j)}(z)$ and $\widehat{\varphi}^{(j)}(z)$ the polynomials of degree m_j and $\widehat{m}_j$, respectively, obtained by numerically truncating the power series $\varphi^{(j)}(z)$ and $\widehat{\varphi}^{(j)}(z)$ of (2.10a) and (4.5). Now we are ready to give a detailed description of Algorithm 4.2 by means of the following.

ALGORITHM 5.3. *Computation of the solution* G_0 *of* (1.3).

Input. Positive integers m_0, k, an error bound $\epsilon > 0$, and the nonnegative $k \times k$ matrices A_i, $i = 0, 1, \dots, m_0$, such that the matrix $\sum_{i=0}^{m_0} A_i$ is ϵ-stochastic and the associated Markov chain is positive recurrent.

Output. An approximation $\widetilde{G}$ of the solution G_0 of (1.3) together with an error bound δ such that $\mathbf{e}^T |\sum_{i=0}^{+\infty} \widetilde{G}^i A_i - \widetilde{G}| \leq \delta \mathbf{e}^T$.

Computation.

1. (Inizialization) Let $\varphi^{(0)}(z) = \sum_{i=0}^{m_0} A_i z^i$, $\widehat{\varphi}^{(0)}(z) = \sum_{i=0}^{m_0-1} A_{i+1} z^i$, $R^{(0)} = A_0 (I - \sum_{i=1}^{m_0} A_i)^{-1}$, $j = 0$.
2. (Computation of the coefficients of the polynomials $\varphi^{(j)}(z)$ and $\widehat{\varphi}^{(j)}(z)$ of degree m_j and $\widehat{m}_j$, respectively, for $j = 1, 2, \dots$, where m_j, $\widehat{m}_j$ are such that the matrices $\varphi^{(j)}(1)$ and $A_0 + \widehat{\varphi}^{(j)}(1)$ are ϵ-stochastic.)

   Repeat

2.1. By means of Algorithm 5.2 and Remark 5.3 compute an integer q and the coefficients of the polynomial $Q(z) = \sum_{i=0}^{q-1} Q_i z^i$ such that

$$(I - \varphi_{odd}^{(j)}(z))Q(z) = I \bmod z^q$$

and (5.3) is verified for $P(z) = I - \varphi_{odd}^{(j)}(z)$.

2.2. By means of Algorithm 5.1 and Remark 5.1 compute the coefficients of the polynomial

$$S(z) = \varphi_{even}^{(j)}(z)Q(z).$$

2.3. By means of Algorithm 5.1 and Remark 5.1 compute the coefficients of the polynomials

$$T(z) = S(z)\varphi_{even}^{(j)}(z), \quad \widehat{T}(z) = S(z)\widehat{\varphi}_{even}^{(j)}(z).$$

2.4. Set $j = j + 1$.

2.5. Compute the integers m_j, $\widehat{m}_j$ and the coefficients of the polynomials $\varphi^{(j)}(z)$ and $\widehat{\varphi}^{(j)}(z)$ of degree m_j and $\widehat{m}_j$, respectively, such that

$$\varphi^{(j)}(z) = z\varphi_{odd}^{(j-1)}(z) + T(z) \bmod z^{m_j},$$
$$\widehat{\varphi}^{(j)}(z) = \widehat{\varphi}_{odd}^{(j-1)}(z) + \widehat{T}(z) \bmod z^{\widehat{m}_j},$$

and the matrices $\varphi^{(j)}(1)$ and $A_0 + \widehat{\varphi}^{(j)}(1)$ are ϵ-stochastic.

2.6 Compute $R^{(j)} = A_0^{(j)}(I - \sum_{i=1}^{m_0} A_i^{(j)})^{-1}$.
Until one of the following conditions is verified.

(C1) $|R^{(j)} - R^{(j-1)}| < \epsilon E$,

(C2) $\mathbf{e}^T(I - A_0^{(j)}(I - A_1^{(j)})^{-1}) < \epsilon\mathbf{e}^T$,

(C3) $\mathbf{e}^T(I - A_0(I - \widehat{A}_1^{(j)})^{-1}) < \epsilon\mathbf{e}^T$.

3. (Computation of the matrix $\widetilde{G}$.)

3.1. If condition (C1) is verified then $\widetilde{G} = A_0(I - \sum_{i=0}^{\widehat{m}_j - 1} {R^{(j)}}^i \widehat{A}_{i+1}^{(j)})^{-1}$.

3.2. If condition (C2) is verified then $\widetilde{G} = A_0(I - \sum_{i=0}^{\widehat{m}_j - 1} {R^{(j)}}^i \widehat{A}_{i+1}^{(j)})^{-1}$.

3.3. If condition (C3) is verified then $\widetilde{G} = A_0(I - \widehat{A}_1^{(j)})^{-1}$.

4. (Computation of the error bound δ.) Compute $\delta = \max_{j=1,\ldots,k} \sum_{i=1}^{k} |w_{i,j}|$ for $W = (w_{i,j})_{i,j}$, $W = \sum_{i=0}^{m_0} \widetilde{G}^i A_i - \widetilde{G}$.

6. Numerical experiments. We implemented Algorithm 4.2 in Fortran 77 by using the customary arithmetic for matrix polynomials (program CCR: customary cyclic reduction) and by using the FFT-based arithmetic for matrix polynomials, as described in detail by Algorithm 5.3 (program FCR: fast cyclic reduction). The programs were run on a Spark Workstation using the standard double precision IEEE arithmetic. The value of ϵ defining the stopping condition was chosen as $\epsilon = 10^{-13}$. We compared our algorithms with the functional iteration (1.5c), which is the fastest among iterations (1.5) (compare [14]), starting with $X_0 = O$ and $X_0 = I$, respectively, and with the method proposed in [12]. The implementation of functional iteration (1.5c) was performed in Fortran 77 (program FIF: functional iteration formula), while for the algorithm of Latouche and Stewart [12] we adapted to the specific problem the

implementation in C given by Stewart [17] (program Latouche–Stewart (LS)). In this adaptation we did not use FFT.

We tested our programs on a problem analyzed in [1] arising from the mathematical modelling of a metropolitan network. The blocks A_i of (1.3) have dimension $k = 16$. We tested the algorithms for different values of the parameter ρ of (2.1). For $\rho = 0.1$ the initial cutting level m_0 is 168, for $\rho = 0.8$ the initial cutting level m_0 is 240, and for the remaining tested values of $\rho \geq 0.9$ the initial cutting level m_0 is 264. The problems with ρ close to 1 are characterized by a "long queue"; i.e., the components π_i of the invariant probability vector are not negligible even for very large values of i. Programs FCR and CCR provided an approximation to the solution G_0 with a residual $\delta < 10^{-13}$ for $\rho = 0.1$, and $\delta < 10^{-12}$ for the tested values of $\rho \geq 0.8$.

Table 6.1 reports the CPU time and the number of iterations needed by algorithms FCR and CCR, together with the residual $||\mathbf{e}^T|\widetilde{G} - \sum_{i=0}^{m_0} \widetilde{G}^i A_i|||$. Observe that, despite the weak stability property of FCR, the residual values of FCR and CCR are equal. We may observe that the use of FFT leads to a substantial reduction of the CPU time.

TABLE 6.1
Cyclic reduction.

	FCR			CCR		
	Time (s.)	Iterations	Residual	Time (s.)	Iterations	Residual
$\rho = 0.1$	27	9	$8.5 \cdot 10^{-14}$	185	9	$8.5 \cdot 10^{-14}$
$\rho = 0.8$	61	13	$2.7 \cdot 10^{-13}$	484	13	$2.7 \cdot 10^{-13}$
$\rho = 0.9$	62	14	$2.7 \cdot 10^{-13}$	545	14	$2.7 \cdot 10^{-13}$
$\rho = 0.95$	65	16	$1.8 \cdot 10^{-13}$	592	16	$1.8 \cdot 10^{-13}$
$\rho = 0.96$	65	17	$2.0 \cdot 10^{-13}$	592	17	$2.0 \cdot 10^{-13}$
$\rho = 0.97$	66	20	$2.3 \cdot 10^{-13}$	603	20	$2.3 \cdot 10^{-13}$

Table 6.2 reports the time and the number of iterations needed by program FIF, starting with $X_0 = I$ and $X_0 = O$, respectively, to compute an approximation of the matrix G_0 with a residual value less than 10^{-12}.

TABLE 6.2
Functional iterations.

	FIF ($X_0 = I$)			FIF ($X_0 = O$)		
	Time (s.)	Iterations	Residual	Time (s.)	Iterations	Residual
$\rho = 0.1$	8.4	22	$2.3 \cdot 10^{-14}$	8.4	22	$2.6 \cdot 10^{-14}$
$\rho = 0.8$	79	148	$1.1 \cdot 10^{-13}$	116	211	$2.0 \cdot 10^{-13}$
$\rho = 0.9$	220	373	$1.2 \cdot 10^{-13}$	294	495	$2.3 \cdot 10^{-13}$
$\rho = 0.95$	902	1534	$1.2 \cdot 10^{-13}$	1104	1866	$1.2 \cdot 10^{-13}$
$\rho = 0.96$	1862	3157	$1.3 \cdot 10^{-13}$	2561	4343	$1.4 \cdot 10^{-13}$
$\rho = 0.97$	1967	3336	$1.2 \cdot 10^{-13}$	11148	18900	$1.5 \cdot 10^{-13}$

Table 6.3 reports the CPU time and the cut-off level needed by algorithm LS, together with the residual $||\mathbf{e}^T|\widetilde{G} - \sum_{i=0}^{m_0} \widetilde{G}^i A_i|||$. For $\rho = 0.1$ the cut-off level sufficient to obtain a residual value less than 10^{-12} is 600. For $\rho \geq 0.8$ we have higher residual values since we were not able to increase the cut-off level above 3500.

Table 6.4 reports the ratio between the CPU times needed by algorithms CCR, FIF (with $X_0 = I$ and $X_0 = O$), and LS and the CPU time needed by algorithm

TABLE 6.3
Algorithm LS.

	LS		
	Time (s.)	Cut-off level	Residual
$\rho = 0.1$	1019	600	$7.8 \cdot 10^{-13}$
$\rho = 0.8$	12332	3500	$5.3 \cdot 10^{-8}$
$\rho = 0.9$	12332	3500	$6.5 \cdot 10^{-6}$
$\rho = 0.95$	12332	3500	$2.2 \cdot 10^{-5}$
$\rho = 0.96$	12332	3500	$2.4 \cdot 10^{-5}$
$\rho = 0.97$	12332	3500	$2.3 \cdot 10^{-5}$

TABLE 6.4
Ratio of CPU times.

	CCR/FCR	FIF/FCR (X_0=I)	FIF/FCR (X_0=O)	LS/FCR
$\rho = 0.1$	6.8	0.3	0.3	37.7
$\rho = 0.8$	7.9	1.3	1.9	202.2 (*)
$\rho = 0.9$	8.8	3.5	4.7	198.9 (*)
$\rho = 0.95$	9.1	13.9	17.0	189.7 (*)
$\rho = 0.96$	9.1	28.6	39.4	189.7 (*)
$\rho = 0.97$	9.1	29.8	168.9	186.8 (*)

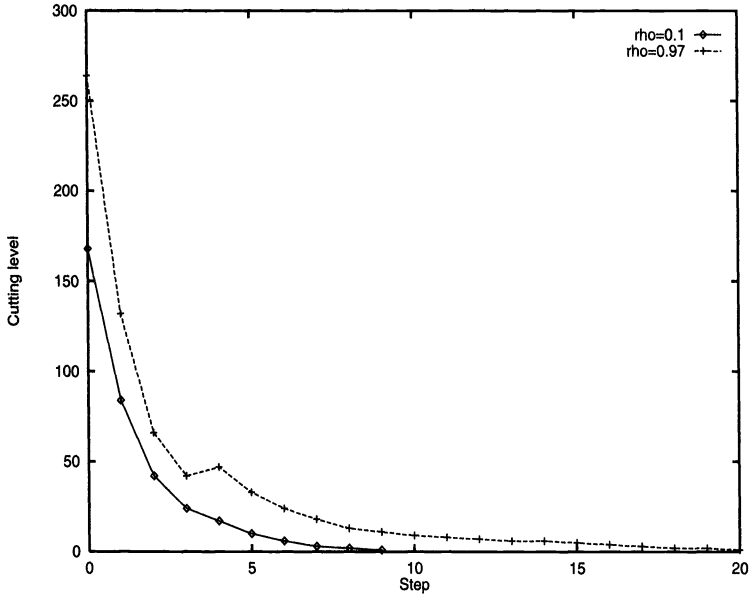


FIG. 6.1.

FCR. The symbol "*" means that the corresponding residual value is higher than the residual value obtained with program FCR and we were not able to decrease it.

We observe that our method is almost not sensible with respect to the values of ρ, whereas the performance of FIF strongly deteriorates when ρ approaches 1. It is interesting to point out the different rates of convergence of FIF in the cases $X_0 = I$ and $X_0 = O$. This behaviour, already pointed out by Latouche [9], received a full theoretical explanation in [14].

The number of steps needed by our algorithm seems to depend on $\log_2 d$, where d is a positive integer such that the components π_i are negligible for $i > d$. It is worth pointing out that our algorithm is particularly suitable for "hard" problems, where the value of ρ is close to 1 and the classical functional iteration formulae (1.5) converge very slowly.

Figure 6.1 shows the size of the cutting level m_j at each step of cyclic reduction for $\rho = 0.1$ and $\rho = 0.97$. The decreasing behaviour leads to a further reduction of the complexity.

REFERENCES

[1] G. Anastasi, L. Lenzini, and B. Meini, *Performance Evaluation of a Worst Case Model of the Metaring MAC Protocol with Global Fairness*, Performance Evaluation, to appear.

[2] D. Bini and B. Meini, *On cyclic reduction applied to a class of Toeplitz-like matrices arising in queueing problems*, in Proc. of the Second International Workshop on Numerical Solution of Markov Chains, Raleigh, NC, 1995, pp. 21–38.

[3] D. Bini and V. Pan, *Matrix and Polynomial Computations, Vol. 1: Fundamental Algorithms*, Birkhäuser, Boston, 1994.

[4] ———, *Polynomial division and its computational complexity*, J. Complexity, 6 (1986), pp. 179–203.

[5] A. Borodin and I. Munro, *The Computational Complexity of Algebraic and Numeric Problems*, American Elsevier, New York, 1975.

[6] B. L. Buzbee, G. H. Golub, and C. W. Nielson, *On direct methods for solving Poisson's equation*, SIAM J. Numer. Anal., 7 (1970), pp. 627–656.

[7] E. Çinlar, *Introduction to Stochastic Processes*, Prentice–Hall, Englewood Cliffs, NJ, 1975.

[8] W. K. Grassman, M. I. Taksar, and D. P. Heyman, *Regenerative analysis and steady state distribution for Markov chains*, Oper. Res., 33 (1985), pp. 1107–1116.

[9] G. Latouche, *Algorithms for Evaluating the Matrix G in Markov Chains of PH/G/1 Type*, Bellcore Tech. report, Moorestown, NJ, 1992.

[10] ———, *Newton's iteration for non-linear equations in Markov chains*, IMA J. Numer. Anal., 14 (1994), pp. 583–598.

[11] G. Latouche and V. Ramaswami, *A logarithmic reduction algorithm for quasi-birth–death processes*, J. Appl. Probab., 30 (1993), pp. 650–674.

[12] G. Latouche and G. W. Stewart, *Numerical methods for M/G/1 type queues*, in Proc. of the Second International Workshop on Numerical Solution of Markov Chains, Raleigh, NC, 1995, pp. 571–581.

[13] E. Linzer, *On the stability of transform-based circular deconvolution*, SIAM J. Numer. Anal., 29 (1992), pp. 1482–1492.

[14] B. Meini, *New Convergence Results on Functional Iteration Techniques for the Numerical Solution of M/G/1 Type Markov Chains*, Numer. Math., to appear.

[15] M.F. Neuts, *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, Dekker, New York, 1989.

[16] V. Ramaswami, *Nonlinear matrix equations in applied probability—solution techniques and open problems*, SIAM Rev., 30 (1988), pp. 256–263.

[17] G. W. Stewart, *Implementing an Algorithm for Solving Block Hessenberg Systems*, Tech. report CS-TR-3295, Department of Computer Science, University of Maryland, College Park, MD, 1993.

[18] R. Varga, *Matrix Iterative Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1963.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 4, pp. 927–949, October 1996

013

GROUP INVARIANCE AND CONVEX MATRIX ANALYSIS*

A. S. LEWIS†

Abstract. Certain interesting classes of functions on a real inner product space are invariant under an associated group of orthogonal linear transformations. This invariance can be made explicit via a simple decomposition. For example, rotationally invariant functions on $\mathbf{R}^2$ are just even functions of the Euclidean norm, and functions on the Hermitian matrices (with trace inner product) which are invariant under unitary similarity transformations are just symmetric functions of the eigenvalues. We develop a framework for answering geometric and analytic (both classical and nonsmooth) questions about such a function by answering the corresponding question for the (much simpler) function appearing in the decomposition. The aim is to understand and extend the foundations of eigenvalue optimization, matrix approximation, and semidefinite programming.

Key words. convexity, group invariance, nonsmooth analysis, semidefinite program, eigenvalue optimization, Fenchel conjugate, subdifferential, spectral function, unitarily invariant norm, Schur convex, extreme point, von Neumann's lemma

AMS subject classifications. Primary, 15A45, 90C25; Secondary, 49J52, 65K05

1. Introduction.

Why is there such a strong parallel between, on the one hand, semidefinite programming and other eigenvalue optimization problems, and on the other hand, ordinary linear programming and related problems? Why are there close analogies between many important matrix norms on the one hand, and associated vector norms on the other? This paper aims to explain the simple algebraic symmetries which drive these parallels.

A simple example may be illustrative. Suppose that we wish to understand convex functions $f : \mathbf{R}^n \to \mathbf{R}$ which are "orthogonally invariant." By this we mean that $f(x) = f(Ux)$ for any point x in $\mathbf{R}^n$ and any orthogonal matrix U. What can we say about such functions?

We might observe first that, since f is determined by its behaviour on the half-line $\{\beta e^1 \mid \beta \geq 0\}$, where $e^1 = (1, 0, 0, \ldots, 0)$, we can write $f(x) = h(\|x\|)$, where the function $h : \mathbf{R}_+ \to \mathbf{R}$ is defined by $h(\beta) = f(\beta e^1)$. What conditions on h are equivalent to the convexity of f? Clearly h must be convex (being the restriction of f to a half-line), but this is not sufficient.

After some more thought we might arrive at the following answer: h must be convex and nondecreasing at the origin. But this obscures the essential symmetry of f. A simple trick allows us to preserve this in our answer. Instead of examining the restriction of f to the half-line $\mathbf{R}_+ e^1$ we consider the restriction to the whole subspace $\mathbf{R}e^1$. We then arrive at the following much more satisfactory answer: $h(\|\cdot\|)$ is convex if and only if the function $h : \mathbf{R} \to \mathbf{R}$ is *even* and convex.

This easy example illustrates the fundamental technique of this paper—analyzing the consequences of the symmetries of a function by analyzing its symmetries on a "transversal" (or defining) subspace. von Neumann's famous 1937 characterization of unitarily invariant matrix norms [27] is precisely of this mold. One statement of this result is that a unitarily invariant matrix function f (one satisfying $f(x) = f(uxv)$

* Received by the editors March 17, 1995; accepted for publication (in revised form) by M. L. Overton November 20, 1995. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada.

† Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada (aslewis@orion.uwaterloo.ca).

for any unitary u and v) is a norm exactly when its restriction to the subspace of real diagonal matrices is a symmetric gauge function.

What algebraic structure underlies von Neumann's result? There are three essential ingredients: first, a real inner product space X (in this case $X = \mathbf{C}^{m\times n}$ with $\langle x, w\rangle = \operatorname{Re}\operatorname{tr} x^* w$); second, a (closed) group $\mathcal{G}$ of orthogonal linear transformations (in this case those of the form $x \mapsto uxv$ for unitary u and v); third, a map γ from X to a transversal subspace (in this case $\gamma(x)$ is the diagonal matrix with diagonal entries the singular values of x arranged in nonincreasing order). The map γ should be $\mathcal{G}$-invariant and should satisfy the following conditions.

AXIOM 1.1 (decomposition). *Any element x of X can be decomposed as $x = A\gamma(x)$ for some operator A in $\mathcal{G}$.*

AXIOM 1.2 (angle contraction). *Any elements x and w in X satisfy the inequality $\langle x, w\rangle \leq \langle \gamma(x), \gamma(w)\rangle$.*

In von Neumann's case, Axiom 1.1 is just the singular value decomposition, and Axiom 1.2 is "von Neumann's lemma" (see, for example, [7]).

This structure $(X, \mathcal{G}, \gamma)$ (which we call a *normal decomposition system*) is the focus of this paper. Our aim is to analyze $\mathcal{G}$-invariant functions on X via their restriction on the range of γ. For this to be of much interest we would hope that the range of γ has lower dimension than X. Our other main example, of fundamental interest in matrix optimization, also has this property:

$$\begin{aligned} X &= \{n \times n \text{ symmetric matrices}\}, \quad \text{with } \langle x, w\rangle = \operatorname{tr} xw, \\ \mathcal{G} &= \{x \mapsto u^T x u \mid u \text{ orthogonal}\}, \quad \text{and} \\ \gamma(x) &= \operatorname{Diag}\lambda(x), \quad \text{where} \\ &\qquad \lambda_1(x) \geq \lambda_2(x) \geq \cdots \geq \lambda_n(x) \quad \text{are the eigenvalues of } x. \end{aligned}$$

In a later paper [22] a broad family of examples generated from the theory of semisimple Lie groups will be discussed. In this paper we concentrate on outlining how the idea of a normal decomposition system provides a simple yet powerful unifying framework in which to study a wide variety of important results. Examples include Schur convexity (see, for example, [23]), the convexity of eigenvalue functions [10, 6, 11, 3, 13, 18], calculations of Fenchel conjugates and subdifferentials of convex eigenvalue functions [26, 5, 12, 34, 31, 28, 29, 30, 15, 16, 1, 17, 18, 24, 33], von Neumann's original result [27] and generalizations (for example, [4, 19]), subdifferentials of unitarily invariant norms [37, 38, 39, 40, 41, 8, 7, 9, 19], and characterizations of extreme, exposed, and smooth points of unit balls [2, 40, 41, 8, 7, 9, 19].

This paper concentrates on convexity and its ramifications.

2. Group invariant normal forms. Underlying all the work in this paper is a rather simple algebraic structure. We therefore begin by fixing our notation and formally defining this structure.

We will work in a real inner product space X. For simplicity we will assume that X is finite dimensional, although many of our results extend easily. The *adjoint* of a linear operator $A : X \to X$ is the linear operator $A^* : X \to X$ defined by $\langle A^* w, x\rangle = \langle w, Ax\rangle$ for all points w and x in X. We denote the identity operator by $\mathrm{id} : X \to X$, and if $A^*A = \mathrm{id}$ then we say that A is *orthogonal*. In fact, A is orthogonal if and only if it is norm preserving: $\|Ax\| = \|x\|$ for all x in X, where the norm is defined by $\|x\| = \sqrt{\langle x, x\rangle}$. In this case, $A^{-1} = A^*$.

We denote the group of all orthogonal linear operators on X (with composition) by $O(X)$, which we endow with the natural topology. Thus A_r approaches A in $O(X)$

if and only if $A_r x$ approaches Ax in X for all x in X. Given a subgroup $\mathcal{G}$ of $O(X)$, a function f on X is *$\mathcal{G}$-invariant* if $f(Ax) = f(x)$ for all x in X and A in $\mathcal{G}$. We can now describe our fundamental structure—this structure is an underlying assumption throughout the paper.

DEFINITION 2.1. *Given a real inner product space X and a closed subgroup $\mathcal{G}$ of the orthogonal group $O(X)$, the map $\gamma : X \to X$ induces a* $\mathcal{G}$-invariant normal form *on X if*

(a) *γ is $\mathcal{G}$-invariant,*

(b) *for any point x in X there is an operator A in $\mathcal{G}$ with $x = A\gamma(x)$, and*

(c) *any points x and w in X satisfy the inequality $\langle x, w\rangle \leq \langle \gamma(x), \gamma(w)\rangle$.*

In this case $(X, \mathcal{G}, \gamma)$ is called a normal decomposition system.

Notice two immediate consequences of this definition: the map γ must be *idempotent*, since for any point x in X, properties (a) and (b) imply $\gamma(\gamma(x)) = \gamma(A^*x) = \gamma(x)$, and furthermore γ must be norm preserving, since $\|\gamma(x)\| = \|A^*x\| = \|x\|$. Our first result, which is somewhat less trivial, has the following important corollary.

- The condition for equality in property (c) is the existence of an operator A in $\mathcal{G}$ with $x = A\gamma(x)$ and $w = A\gamma(w)$.

THEOREM 2.2. *A subset K of X has the property that $\langle x, w\rangle = \langle \gamma(x), \gamma(w)\rangle$ for every pair of elements x and w of K if and only if there is an operator A in $\mathcal{G}$ satisfying $x = A\gamma(x)$ for all x in K.*

Proof. The "if" direction is easy, so consider the "only if" direction. Without loss of generality, K is nonempty, so choose a point z in $\operatorname{ri}(\operatorname{conv} K)$ and an A in $\mathcal{G}$ for which $z = A\gamma(z)$. If there is a point x in K with $x \neq A\gamma(x)$ then the Cauchy–Schwartz inequality implies that $\langle x, A\gamma(x)\rangle < \|x\|^2$. It is easy to write z as a convex combination $z = \alpha_0 x + \sum_{i>0} \alpha_i x^i$ for strictly positive α_i's with sum 1 and points x^i in K. But now we have

$$\begin{aligned}
\langle \gamma(x), \gamma(z)\rangle &= \langle A\gamma(x), A\gamma(z)\rangle = \langle A\gamma(x), z\rangle \\
&< \alpha_0 \|x\|^2 + \langle A\gamma(x), \textstyle\sum_{i>0} \alpha_i x^i\rangle \\
&\leq \alpha_0 \|x\|^2 + \sum_{i>0} \alpha_i \langle \gamma(x), \gamma(x^i)\rangle \\
&= \alpha_0 \|x\|^2 + \sum_{i>0} \alpha_i \langle x, x^i\rangle \\
&= \langle x, z\rangle \leq \langle \gamma(x), \gamma(z)\rangle,
\end{aligned}$$

which is a contradiction. □

We defer a systematic discussion of examples until the end of the paper. However, for the sake of concreteness the reader may wish to keep in mind the following extremely simple example: $X = \mathbf{R}$ (with $\langle x, w\rangle = xw$), $\mathcal{G} = \{\pm \mathrm{id}\}$, and $\gamma(x) = |x|$. The properties are easily verified.

We will only use the closedness of $\mathcal{G}$ very rarely (specifically, in Theorem 3.3), but it does not rule out much of interest. We think of the formula $x = A\gamma(x)$ in property (b) as being a "normal form decomposition" of x. Property (c) expresses the fact that γ contracts the angle between the vectors x and w unless they have a simultaneous normal form decomposition (in which case the angle remains constant). If we write

$$\mathcal{G}^x = \{A \in \mathcal{G} \mid x = A\gamma(x)\}, \tag{2.1}$$

then (b) says that $\mathcal{G}^x$ is nonempty, while the condition for equality in (c) is that $\mathcal{G}^x \cap \mathcal{G}^w$ be nonempty.

PROPOSITION 2.3. *For points x and w in X,*

$$\max_{A \in \mathcal{G}} \langle x, Aw \rangle = \langle \gamma(x), \gamma(w) \rangle.$$

Proof. Note that $\langle x, Aw \rangle \leq \langle \gamma(x), \gamma(Aw) \rangle = \langle \gamma(x), \gamma(w) \rangle$ for any operator A in $\mathcal{G}$. On the other hand, since there exist operators B and C in $\mathcal{G}$ with $x = B\gamma(x)$ and $w = C\gamma(w)$, we have that $\langle \gamma(x), \gamma(w) \rangle = \langle B^*x, C^*w \rangle = \langle x, BC^*w \rangle$, so the maximum is attained by $A = BC^*$. □

Given a subset K of X, the *dual cone* of K is defined to be the closed, convex cone

$$K^+ = \{w \in X \mid \langle x, w \rangle \geq 0 \text{ for all } x \text{ in } K\}.$$

The set K is a closed, convex cone if and only if $K = K^{++}$ [32, Thm. 14.1]. The function γ is K^+-*convex* if the real function $\langle \gamma(\cdot), w \rangle$ is convex for all vectors w in K, and a function $f : K \to [-\infty, +\infty]$ is K^+-*isotone* if $f(x) \geq f(w)$ for any x and w in K satisfying $x - w \in K^+$.

It transpires that Definition 2.1 has strong implications for possible maps γ.

THEOREM 2.4. *The range $R(\gamma)$ of the map γ is a closed, convex cone. Furthermore, γ is norm preserving, positively homogeneous, and $R(\gamma)^+$-convex with global Lipschitz constant* 1.

Proof. For any point x in X it follows from Definition 2.1 that $\langle x, w \rangle \leq \langle \gamma(x), w \rangle$ for all points w in $R(\gamma)$, and hence $\gamma(x) - x \in R(\gamma)^+$. If in particular x lies in $R(\gamma)^{++}$ then

$$0 \leq \langle x, \gamma(x) - x \rangle = \langle x, \gamma(x) \rangle - \|x\|^2 \leq 0$$

since, as we have seen, $\|\gamma(x)\| = \|x\|$. It follows that $x = \gamma(x) \in R(\gamma)$, so $R(\gamma)^{++} \subset R(\gamma)$, and hence $R(\gamma)$ is a closed, convex cone.

Supposing once more that x lies in X and that the scalar λ is nonnegative, we have

$$\begin{aligned} \|\gamma(\lambda x) - \lambda\gamma(x)\|^2 &= \|\gamma(\lambda x)\|^2 + \lambda^2\|\gamma(x)\|^2 - 2\lambda\langle \gamma(\lambda x), \gamma(x) \rangle \\ &\leq \|\lambda x\|^2 + \lambda^2\|x\|^2 - 2\lambda\langle \lambda x, x \rangle \\ &= 0; \end{aligned}$$

whence $\gamma(\lambda x) = \lambda\gamma(x)$. Thus γ is positively homogeneous.

By Proposition 2.3, for any w in $R(\gamma)$ we have

$$\langle \gamma(x), w \rangle = \max_{A \in \mathcal{G}} \langle x, Aw \rangle,$$

and hence $\langle \gamma(\cdot), w \rangle$ is convex, being a pointwise maximum of linear functions. Thus γ is $R(\gamma)^+$-convex.

Finally, for any x and w in X,

$$\begin{aligned} \|\gamma(x) - \gamma(w)\|^2 &= \langle \gamma(x), \gamma(x) \rangle + \langle \gamma(w), \gamma(w) \rangle - 2\langle \gamma(x), \gamma(w) \rangle \\ &\leq \|x\|^2 + \|w\|^2 - 2\langle x, w \rangle \\ &= \|x - w\|^2, \end{aligned}$$

whence the Lipschitz constant 1. □

Various algebraic ideas can be applied naturally to the concept of a normal decomposition system. For example, we say that two normal decomposition systems $(X_1, \mathcal{G}_1, \gamma_1)$ and $(X_2, \mathcal{G}_2, \gamma_2)$ are *isomorphic* if there is an inner product space isomorphism $\alpha : X_1 \to X_2$ and a group isomorphism $\beta : \mathcal{G}_1 \to \mathcal{G}_2$ such that for all points x in X_1 and operators A in $\mathcal{G}_1$ we have $\gamma_2(\alpha(x)) = \alpha(\gamma_1(x))$ and $(\beta(A))(\alpha(x)) = \alpha(Ax)$. There is also a natural notion of the Cartesian product of two normal decomposition systems. Observe finally that, given any inner product space X and subgroup $\mathcal{G}$ of $O(X)$, easy examples show that there may be no map γ with $(X, \mathcal{G}, \gamma)$ a normal decomposition system.

3. $\mathcal{G}$-invariant functions and sets. The main aim of this paper is to study functions $f : X \to [-\infty, +\infty]$ on the inner product space X which are $\mathcal{G}$-invariant: $f(Ax) = f(x)$ for all points x in X and operators A in the group $\mathcal{G}$. As usual, we assume that the map γ induces a $\mathcal{G}$-invariant normal form on X in the sense of Definition 2.1.

We will be particularly interested in *convex* functions f, which we define by requiring that the *epigraph*

$$\operatorname{epi} f = \{(x, \alpha) \in X \times \mathbf{R} \mid \alpha \geq f(x)\}$$

be a convex set. The function f is *closed* if its epigraph is closed and is *proper* if it never takes the value $-\infty$ and has nonempty *domain*,

$$\operatorname{dom} f = \{x \in X \mid f(x) < +\infty\}.$$

The *(Fenchel) conjugate* of f is the closed, convex function $f^* : X \to [-\infty, +\infty]$ defined by

$$f^*(w) = \sup\{\langle x, w\rangle - f(x) \mid x \in X\}.$$

For proper, convex f, the conjugate f^* is also proper with $f^{**} = f$ providing that f is also closed. For proper f we can define the *(convex) subdifferential* at a point x in $\operatorname{dom} f$ by

$$\partial f(x) = \{w \in X \mid f(x) + f^*(w) = \langle x, w\rangle\}.$$

Elements of the subdifferential are called *subgradients*. For all of these ideas the standard reference is [32].

The following result is rather reminiscent of the discussion in [32, pp. 110–111]. It shows that conjugacy preserves $\mathcal{G}$-invariance.

PROPOSITION 3.1. *If the function $f : X \to [-\infty, +\infty]$ is $\mathcal{G}$-invariant then so is the conjugate function f^*, and*

$$f^*(w) = \sup\{\langle x, \gamma(w)\rangle - f(x) \mid x \in R(\gamma)\}$$

for any point x in X.

Proof. For any operator A in $\mathcal{G}^w$ (whence $w = A\gamma(w)$),

$$\begin{aligned} f^*(w) &= \sup\{\langle z, w\rangle - f(z) \mid z \in X\} \\ &= \sup\{\langle Bx, A\gamma(w)\rangle - f(Bx) \mid x \in R(\gamma),\ B \in \mathcal{G}\} \\ &= \sup\{\sup\{\langle x, B^*A\gamma(w)\rangle \mid B \in \mathcal{G}\} - f(x) \mid x \in R(\gamma)\} \\ &= \sup\{\langle x, \gamma(w)\rangle - f(x) \mid x \in R(\gamma)\} \end{aligned}$$

by Proposition 2.3. Since γ is $\mathcal{G}$-invariant, so is f^*. □

LEMMA 3.2. *A $\mathcal{G}$-invariant function $f : X \to [-\infty, +\infty]$ is (Frechet) differentiable at the point x in X if and only if it is differentiable at $\gamma(x)$.*

Proof. For any operator B in $\mathcal{G}$, we know that $f(Bw) = f(w)$ for all points w in X, and hence by the chain rule, if f is differentiable at Bw then it is differentiable at w. Choosing an operator A in $\mathcal{G}^x$ (so that $x = A\gamma(x)$), the result follows by setting $w = x$, $B = A^*$ and $w = \gamma(x)$, $B = A$ in turn. □

The next result is our first rather nontrivial observation. A consequence, for example, is that symmetric, convex functions on $\mathbf{R}^n$ are "Schur convex" (see Example 7.1).

THEOREM 3.3. *If the $\mathcal{G}$-invariant function $f : X \to [-\infty, +\infty]$ is convex then it is $R(\gamma)^+$-isotone on $R(\gamma)$: if points x and w lie in $R(\gamma)$ with $x - w$ in $R(\gamma)^+$ then $f(x) \geq f(w)$.*

Proof. The coset $\mathcal{G}x$ is compact (since $\mathcal{G}$ is compact). If w lay outside its convex hull then there would exist a separating hyperplane defined by a vector v in X with

$$\langle \gamma(v), w \rangle \geq \langle v, w \rangle > \max_{A \in \mathcal{G}} \langle v, Ax \rangle = \langle \gamma(v), x \rangle,$$

by Proposition 2.3, and then $\langle \gamma(v), x - w \rangle < 0$, contradicting the assumption that $x - w$ lies in $R(\gamma)^+$. Hence there exist positive scalars $\lambda_1, \lambda_2, \ldots, \lambda_r$ with sum 1 and operators $A_1, A_2, \ldots, A_r$ in $\mathcal{G}$ with $w = \sum_1^r \lambda_i A_i x$.

Suppose that $f(x) < f(w)$. Then we can choose a real number α in the interval $(f(x), f(w))$, and since f is $\mathcal{G}$-invariant, $f(A_i x) = f(x) < \alpha$ for each $i = 1, 2, \ldots, r$. Now since f is convex,

$$f(w) = f\left(\sum_1^r \lambda_i A_i x\right) < \sum_1^r \lambda_i \alpha = \alpha$$

(see [32, Thm. 4.2]), which is a contradiction. □

We will also be interested in $\mathcal{G}$-invariant subsets of X, so we will conclude this section with some simple observations illustrating how various algebraic and topological constructions preserve $\mathcal{G}$-invariance. Notice first that the class of $\mathcal{G}$-invariant sets is easily seen to be closed under arbitrary unions, intersections, and complements. Suppose that the subset D of X is $\mathcal{G}$-invariant (that is, $x \in D$, $A \in \mathcal{G}$ implies $Ax \in D$). Then the interior of D is quickly seen to be $\mathcal{G}$-invariant; whence the closure and boundary of D are also $\mathcal{G}$-invariant.

For each $r = 1, 2, \ldots$, suppose that Λ_r is a subset of $\mathbf{R}^r$ and define a subset of X,

$$\left\{ \sum_{i=1}^r \lambda_i x_i \;\middle|\; r \in \mathbf{N},\ \lambda \in \Lambda_r,\ x_1, x_2, \ldots, x_r \in D \right\}.$$

It is immediate that this set is $\mathcal{G}$-invariant. By taking Λ_r to be $\mathbf{R}^r$, $\mathbf{R}^r_+$, $\{\lambda \in \mathbf{R}^r \mid \sum \lambda_i = 1\}$, and $\{\lambda \in \mathbf{R}^r_+ \mid \sum \lambda_i = 1\}$ in turn we see that the *linear hull*, the *conical hull*, the *affine hull*, and the *convex hull* of D are all $\mathcal{G}$-invariant.

We say that a point x lies in the *intrinsic core* of D, written $\operatorname{icr} D$, if for any point w in the affine hull $\operatorname{aff} D$, $x + \delta(w - x)$ lies in D for all real δ sufficiently small. When D is convex its intrinsic core coincides with its *relative interior* $\operatorname{ri} D$, the interior of D relative to its affine hull (see [32]), and in this case the *relative boundary* $\operatorname{rb} D$ is just $\operatorname{cl} D \setminus \operatorname{ri} D$. Since it is easy to check that $\operatorname{icr} D$ is $\mathcal{G}$-invariant, it follows that for

convex, $\mathcal{G}$-invariant D, the relative interior and boundary of D are also $\mathcal{G}$-invariant. Finally, for any $\mathcal{G}$-invariant set D the dual cone D^+ and the orthogonal complement $D^\perp$ are both $\mathcal{G}$-invariant.

4. Reduction. Let us assume once more that $(X, \mathcal{G}, \gamma)$ is a normal decomposition system in the sense of Definition 2.1. If the function $f : X \to [-\infty, +\infty]$ is $\mathcal{G}$-invariant then since $f(x) = f(\gamma(x))$ for all points x in X, the behaviour of f is determined by its behaviour on $R(\gamma)$, the range of γ. The key idea of this paper is then rather simple—we reduce questions about f to corresponding questions about the restriction of f to a subspace Y containing $R(\gamma)$: typically, $Y = R(\gamma) - R(\gamma)$.

Given a subspace Y of X, we denote the *stabilizer* of Y in $\mathcal{G}$ by

$$\mathcal{G}_Y = \{A \in \mathcal{G} \mid AY = Y\}.$$

We will frequently abuse notation and write $\mathcal{G}_Y$ for the group of restricted operators

$$\mathcal{G}_Y|_Y = \{A|_Y \mid A \in \mathcal{G}_Y\}.$$

In other words, we think of operators in $\mathcal{G}_Y$ as orthogonal transformations on Y (as well as on X). When Y contains $R(\gamma)$ we can consider the restricted map $\gamma|_Y : Y \to Y$: we will frequently write γ in place of $\gamma|_Y$.

The following central assumption will remain in force throughout the remainder of the paper.

ASSUMPTION 4.1. *In the sense of Definition* 2.1, $(X, \mathcal{G}, \gamma)$ *is a normal decomposition system. The inner product space* Y *is a subspace of* X *(with the inherited inner product) and contains the range of* γ. *Furthermore,* $(Y, \mathcal{G}_Y, \gamma)$ *is also a normal decomposition system.*

This amounts to the additional assumption that if, in Definition 2.1, the point x in fact lies in Y then the operator A in property (b) can actually be chosen to leave Y invariant. Of course a trivial example is $Y = X$. Once again, since we are deferring examples until the end of the paper, it may be helpful to keep a simple (although nontrivial) example in mind. We take X to be $\mathbf{R}^n$ with the standard inner product, $\mathcal{G} = \mathcal{O}_n$, the orthogonal group on $\mathbf{R}^n$, and let e^1 be the vector $(1, 0, 0, \ldots, 0)$. Then it is easily verified that if we define $\gamma(x) = \|x\|e^1$ for all x in $\mathbf{R}^n$ then we obtain a normal decomposition system, and that if $Y = \operatorname{span}\{e^1\}$ then Assumption 4.1 holds.

An interesting general framework in which Assumption 4.1 holds is developed in [22]. In summary, suppose that G is a real, semisimple Lie group with a maximal compact subgroup K, and that $\mathbf{g} = \mathbf{k} \oplus \mathbf{t}$ is the corresponding Cartan decomposition (where $\mathbf{g}$ and $\mathbf{k}$ are the tangent algebras of G and K, respectively). Now let $X = \mathbf{t}$, let the group $\mathcal{G}$ consist of the adjoint actions of elements of K on $\mathbf{t}$, and let Y be a maximal $\mathbf{R}$-diagonalizable subspace of $\mathbf{t}$. Then $\mathcal{G}_Y$ is (essentially) the associated Weyl group. Finally, fix a closed Weyl chamber $D \subset Y$ and for any point x in $\mathbf{t}$ define $\gamma(x)$ to be the (singleton) intersection of the $\mathcal{G}$-orbit of x with the chamber D. Then Assumption 4.1 holds; see [22] for details. In fact, all of the concrete examples which we develop later fall into this framework.

In what follows, $\circ$ denotes composition. Thus $(h \circ \gamma)(x) = h(\gamma(x))$.

PROPOSITION 4.2. *A function* f *is* $\mathcal{G}$*-invariant on* X *if and only if it can be written in the form* $f = h \circ \gamma$ *for some* $\mathcal{G}_Y$*-invariant function* h *on* Y.

Proof. Any function of the form $h \circ \gamma$ is $\mathcal{G}$-invariant since γ is. If, on the other hand, f is $\mathcal{G}$-invariant then it is immediate that $f = f|_Y \circ \gamma$, and clearly $f|_Y$ is $\mathcal{G}_Y$-invariant. □

Thus henceforth we will restrict attention to $\mathcal{G}$-invariant functions $h \circ \gamma$, where the function h is $\mathcal{G}_Y$-invariant. We now follow two distinct approaches to the elegant fact that such an extended real-valued function $h \circ \gamma$ is convex on X if and only if h is convex on Y. The first approach is direct, using Theorem 3.3, and hence relies on the underlying assumption that the group $\mathcal{G}$ is closed. The second approach does not require this assumption, but instead assumes that the function h is closed and employs an attractive Fenchel conjugacy argument.

THEOREM 4.3 (convex and closed functions). *Suppose that the function $h : Y \to [-\infty, +\infty]$ is $\mathcal{G}_Y$-invariant. Then the function $h \circ \gamma$ is convex (respectively, closed) on X if and only if h is convex (respectively, closed) on Y. Hence, a $\mathcal{G}$-invariant function on X is convex (respectively, closed) if and only if its restriction to Y is convex (respectively, closed).*

Proof. Since $h = (h \circ \gamma)|_Y$, one direction is clear. Conversely, suppose that h is convex. For any points x and w in X and real λ in $(0,1)$, we know by Theorem 2.4 that $\lambda\gamma(x) + (1-\lambda)\gamma(w)$ and $\gamma(\lambda x + (1-\lambda)w)$ both lie in $R(\gamma)$, and that $(\lambda\gamma(x) + (1-\lambda)\gamma(w)) - \gamma(\lambda x + (1-\lambda)w)$ lies in $R(\gamma)^+$. Hence by Theorem 3.3 (applied to the system $(Y, \mathcal{G}_Y, \gamma)$), we have

$$h(\lambda\gamma(x) + (1-\lambda)\gamma(w)) \geq h(\gamma(\lambda x + (1-\lambda)w)).$$

Now for any real numbers $\alpha > h(\gamma(x))$ and $\beta > h(\gamma(w))$, since h is convex we deduce that $h(\gamma(\lambda x + (1-\lambda)w)) < \lambda\alpha + (1-\lambda)\beta$; whence $h \circ \gamma$ is convex [32, Thm. 4.2].

Turning now to the closed case, since $h = (h \circ \gamma)|_Y$ we have that

$$\operatorname{epi} h = \operatorname{epi}(h \circ \gamma) \cap (Y \times \mathbf{R}), \tag{4.1}$$

so that if $h \circ \gamma$ is closed, then so is h. Suppose on the other hand that h is closed. If $\{(x_i, r_i)\}$ is a sequence of points in $\operatorname{epi}(h \circ \gamma)$ approaching the point (x, r), then since the sequence $((\gamma(x_i), r_i))$ lies in the closed set $\operatorname{epi} h$ and approaches $(\gamma(x), r)$ (as γ is continuous by Theorem 2.4), it follows that $(\gamma(x), r) \in \operatorname{epi} h$, and so $(x, r) \in \operatorname{epi}(h \circ \gamma)$. □

The second approach to convexity is rather more transparent once we have derived the following elegant formula.

THEOREM 4.4 (conjugacy). *Suppose that the function $h : Y \to [-\infty, +\infty]$ is $\mathcal{G}_Y$-invariant. Then on the space X,*

$$(h \circ \gamma)^* = h^* \circ \gamma.$$

Proof. By Proposition 3.1 applied in turn to the systems $(X, \mathcal{G}, \gamma)$ and $(Y, \mathcal{G}|_Y, \gamma)$, we see that for any point w in X,

$$\begin{aligned}(h \circ \gamma)^*(w) &= \sup\{\langle x, \gamma(w)\rangle - h(\gamma(x)) \mid x \in R(\gamma)\} \\ &= \sup\{\langle x, \gamma(w)\rangle - h(x) \mid x \in R(\gamma)\} \\ &= h^*(\gamma(w)). \qquad \square\end{aligned}$$

It is an immediate consequence of this conjugacy formula that a $\mathcal{G}_Y$-invariant function $h : Y \to (-\infty, +\infty]$ (note that we exclude $-\infty$) is closed and convex exactly when the function $h \circ \gamma$ is closed and convex on X. One direction is clear from equation (4.1). On the other hand, if h is closed and convex then $h = h^{**}$, so that $h \circ \gamma = (h \circ \gamma)^{**}$ by Theorem 4.4, and hence $h \circ \gamma$ is also closed and convex.

Proposition 4.2 shows that the restriction operation which maps an extended real-valued function h on X to its restriction $h|_Y$ gives a one-to-one correspondence between $\mathcal{G}$-invariant functions on X and $\mathcal{G}_Y$-invariant functions on Y. Theorem 4.3 (convex and closed functions) shows that this correspondence preserves convexity and closedness, and Theorem 4.4 (conjugacy) shows that it also preserves the conjugacy operation. We shall see in §6 that restriction also preserves essential strict convexity and smoothness (Corollary 6.2).

The next result provides perhaps a more compelling motivation for the conjugacy approach. Recall that for a point x in X, the set $\mathcal{G}^x$ describes the possible decompositions of x: $\mathcal{G}^x = \{A \in \mathcal{G} \mid x = A\gamma(x)\}$.

THEOREM 4.5 (subdifferentials). *Given a function $h : Y \to (-\infty, +\infty]$ which is $\mathcal{G}_Y$-invariant, suppose that the point x in X satisfies $\gamma(x) \in \operatorname{dom}(h)$. Then the element w of X is a subgradient of the function $h \circ \gamma$ at x if and only if $\gamma(w)$ is a subgradient of h at $\gamma(x)$ with x and w having simultaneous decompositions: $\mathcal{G}^x \cap \mathcal{G}^w \neq \emptyset$. In fact, the following "chain rule" holds:*

$$\partial(h \circ \gamma)(x) = \mathcal{G}^x \partial h(\gamma(x)). \tag{4.2}$$

Proof. By definition, $w \in \partial(h \circ \gamma)(x)$ if and only if

$$\langle x, w \rangle = (h \circ \gamma)(x) + (h \circ \gamma)^*(w) = h(\gamma(x)) + h^*(\gamma(w)),$$

using Theorem 4.4 (conjugacy). But then, since

$$h(\gamma(x)) + h^*(\gamma(w)) \geq \langle \gamma(x), \gamma(w) \rangle \geq \langle x, w \rangle,$$

equality holds throughout, and the first part of the result follows using Theorem 2.2.

Suppose that $w \in \partial(h \circ \gamma)(x)$. Then by the above, $\gamma(w) \in \partial h(\gamma(x))$ and we can choose an operator A in $\mathcal{G}^x \cap \mathcal{G}^w$. Then

$$w = A\gamma(w) \in A\partial h(\gamma(x)) \subset \mathcal{G}^x \partial h(\gamma(x)).$$

Conversely, suppose that $y \in \partial h(\gamma(x))$ and that $A \in \mathcal{G}^x$. Then

$$\begin{aligned}(h \circ \gamma)(x) + (h \circ \gamma)^*(Ay) &= h(\gamma(x)) + h^*(\gamma(Ay)) \\ &= h(\gamma(x)) + h^*(\gamma(y)) \\ &= h(\gamma(x)) + h^*(y) \\ &= \langle \gamma(x), y \rangle = \langle x, Ay \rangle,\end{aligned}$$

using Theorem 4.4 (conjugacy) and the $\mathcal{G}_Y$-invariance of h^*. Thus Ay lies in $\partial(h \circ \gamma)(x)$, as required. □

Notice that this result is the first point at which we have used the condition for equality in property (c) of Definition 2.1.

COROLLARY 4.6. *Suppose that the function $f : X \to (-\infty, +\infty]$ is $\mathcal{G}$-invariant and that the point x lies in $\operatorname{dom} f$. Then the element w of X is a subgradient of f at x if and only if $\gamma(w)$ is a subgradient of f at $\gamma(x)$ with x and w having simultaneous decompositions: $\mathcal{G}^x \cap \mathcal{G}^w \neq \emptyset$. In fact, $\partial f(x) = \mathcal{G}^x \partial f(\gamma(x))$.*

Proof. Take $Y = X$ in Theorem 4.5 (subdifferentials). □

COROLLARY 4.7. *Suppose that the function $f : X \to (-\infty, +\infty]$ is $\mathcal{G}$-invariant and convex. If f is differentiable at the point x then $\nabla f(\gamma(x)) = \gamma(\nabla f(x))$.*

Proof. By Lemma 3.2, f is differentiable at $\gamma(x)$. By Corollary 4.6, since $\nabla f(x) \in \partial f(x)$ it follows that $\gamma(\nabla f(x)) \in \partial f(\gamma(x)) = \{\nabla f(\gamma(x))\}$. □

Given functions $h, p : Y \to (-\infty, +\infty]$, we define the *infimal convolution* $h \square p : Y \to [-\infty, +\infty]$ by

$$(h \square p)(y) = \inf_{w \in Y} \{h(w) + p(y - w)\}.$$

An analogous definition holds on X.

THEOREM 4.8 (infimal convolution). *Suppose that the functions $h, p : Y \to (-\infty, +\infty]$ are $\mathcal{G}_Y$-invariant and convex. Then*

$$(h \square p) \circ \gamma = (h \circ \gamma) \square (p \circ \gamma).$$

Proof. Given any two points x and z in X, define two compact convex subsets of Y by

$$C = \operatorname{conv} \mathcal{G}_Y \gamma(z), \quad \text{and}$$
$$D = \gamma(x) - \operatorname{conv} \mathcal{G}_Y \gamma(x - z).$$

These two sets are not disjoint, since a separating hyperplane would give an element u of Y and a scalar β with

$$\begin{aligned} \langle \gamma(u), \gamma(z) \rangle &= \max_{A \in \mathcal{G}_Y} \langle u, A\gamma(z) \rangle \\ &\leq \beta < \min_{A \in \mathcal{G}_Y} \langle u, \gamma(x) - A\gamma(x - z) \rangle \\ &= \langle u, \gamma(x) \rangle - \langle \gamma(u), \gamma(x - z) \rangle \\ &\leq \langle \gamma(u), \gamma(x) - \gamma(x - z) \rangle, \end{aligned}$$

which contradicts the convexity and positive homogeneity of $\langle \gamma(u), \gamma(\cdot) \rangle$ (Theorem 2.4). Thus there is a point w in $C \cap D$, and this point must satisfy $h(w) \leq h(\gamma(z))$ and $p(\gamma(x) - w) \leq p(\gamma(x - z))$.

Now consider any fixed point x in X. By the above argument we see that

$$\begin{aligned} (h \square p)(\gamma(x)) &= \inf_{w \in Y} \{h(w) + p(\gamma(x) - w)\} \\ &\leq \inf_{z \in X} \{h(\gamma(z)) + p(\gamma(x - z))\} \\ &= ((h \circ \gamma) \square (p \circ \gamma))(x). \end{aligned}$$

On the other hand, we can choose an operator A in $\mathcal{G}^x$, and then

$$\begin{aligned} ((h \circ \gamma) \square (p \circ \gamma))(x) &= \inf_{z \in X} \{h(\gamma(z)) + p(\gamma(x - z))\} \\ &\leq \inf_{w \in Y} \{h(\gamma(Aw)) + p(\gamma(A(\gamma(x) - w)))\} \\ &= \inf_{w \in Y} \{h(w) + p(\gamma(x) - w)\} \\ &= (h \square p)(\gamma(x)). \end{aligned}$$

The result follows. □

This result strengthens and generalizes those in [33, 22].

5. Invariant sets. In the last section we studied $\mathcal{G}$-invariant functions on the space X. In this section we consider analogous questions for $\mathcal{G}$-invariant subsets of X. As usual, $(X, \mathcal{G}, \gamma)$ is a normal decomposition system with a subsystem $(Y, \mathcal{G}_Y, \gamma)$ (where Y contains the range of γ). In other words, Assumption 4.1 holds.

PROPOSITION 5.1. *A subset D of X is $\mathcal{G}$-invariant if and only if it has the form $D = \gamma^{-1}(C)$ for some $\mathcal{G}_Y$-invariant subset C of Y.*

Proof. Clearly any set of the form $\gamma^{-1}(C) = \{x \in X \mid \gamma(x) \in C\}$ is $\mathcal{G}$-invariant because γ is. On the other hand, if D is $\mathcal{G}$-invariant then it is easily checked that we can write $D = \gamma^{-1}(D \cap Y)$, which has the required form. □

Thus henceforth we will restrict our attention to $\mathcal{G}$-invariant sets $\gamma^{-1}(C)$ (or, equivalently, $\mathcal{G}C$), where the set $C \subset Y$ is $\mathcal{G}_Y$-invariant. Sets can be effectively studied via their indicator functions,

$$\delta_C(y) = \begin{cases} 0, & y \in C, \\ +\infty, & y \notin C. \end{cases}$$

Notice, for example, that $\delta_{\gamma^{-1}(C)} = \delta_C \circ \gamma$ for any subset C of Y.

COROLLARY 5.2 (closed and convex sets). *Suppose that the subset C of Y is $\mathcal{G}_Y$-invariant. Then the set $\gamma^{-1}(C)$ is closed (respectively, convex) if and only if C is closed (respectively, convex).*

Proof. Apply Theorem 4.3 to the function δ_C. □

A fundamental idea in optimization is the *(convex) normal cone* to a subset C of Y at a point y in C, defined by

$$N(y|C) = \{w \in Y \mid \langle w, z - y \rangle \leq 0 \text{ for all } z \in C\}.$$

It is easily checked that $N(y|C) = \partial\delta_C(y)$, whence the following useful formula.

COROLLARY 5.3 (normal cones). *Suppose that the subset C of Y is $\mathcal{G}_Y$-invariant and that the point x in X satisfies $\gamma(x) \in C$. Then the element w of X lies in the normal cone $N(x|\gamma^{-1}(C))$ if and only if $\gamma(w)$ lies in $N(\gamma(x)|C)$ with x and w having simultaneous decompositions ($\mathcal{G}^x \cap \mathcal{G}^w \neq \emptyset$). In fact,*

$$N(x|\gamma^{-1}(C)) = \mathcal{G}^x N(\gamma(x)|C).$$

Proof. Apply Theorem 4.5 (subdifferentials) to the function $h = \delta_C$. □

Other convex-analytic formulae follow easily from Theorem 4.4 (conjugacy). For convenience, we collect some similar-looking results in a single theorem. The *polar set* of a subset C of Y is defined by

$$C^\circ = \{z \in Y \mid \langle z, y \rangle \leq 1 \text{ for all } y \in C\},$$

while the *polar cone* is

$$C^- = \{z \in Y \mid \langle z, y \rangle \leq 0 \text{ for all } y \in C\}.$$

THEOREM 5.4. *Suppose that the subset C of Y is $\mathcal{G}_Y$-invariant. Then*

(i) $(\gamma^{-1}(C))^- = \gamma^{-1}(C^-)$,

(ii) $(\gamma^{-1}(C))^\circ = \gamma^{-1}(C^\circ)$, *and*

(iii) $\operatorname{int}_X \gamma^{-1}(C) = \gamma^{-1}(\operatorname{int}_Y C)$.

Furthermore, $C = \gamma^{-1}(C) \cap Y$, and if C is also convex then

(iv) $\operatorname{ri} \gamma^{-1}(C) = \gamma^{-1}(\operatorname{ri} C)$, *and*

(v) $\operatorname{aff}\gamma^{-1}(C) = \gamma^{-1}(\operatorname{aff} C)$.

Proof. An element w of X lies in $(\gamma^{-1}(C))^-$ if and only if

$$0 \geq \delta^*_{\gamma^{-1}(C)}(w) = (\delta_C \circ \gamma)^*(w) = \delta^*_C(\gamma(w)),$$

whence (i), and (ii) is similar.

To see (iii), note that since γ may be regarded as a map from X to Y and is continuous by Theorem 2.4, $\gamma^{-1}(\operatorname{int}_Y C)$ is an open subset of $\gamma^{-1}(C)$, and hence $\gamma^{-1}(\operatorname{int}_Y C) \subset \operatorname{int}_X \gamma^{-1}(C)$. Conversely, suppose that the point x lies in $\operatorname{int}_X \gamma^{-1}(C)$, and yet $\gamma(x) \notin \operatorname{int}_Y(C)$. Then there is a sequence of points (y_n) in $Y \setminus C$ approaching $\gamma(x)$. Each point has a decomposition $y_n = A_n \gamma(y_n)$ for some operator A_n in $\mathcal{G}_Y$, and since $\mathcal{G}_Y$ is compact there is a convergent subsequence $A_{n'} \to A \in \mathcal{G}_Y$. Now notice that the sequence $\gamma(y_{n'}) = A^*_{n'} y_{n'}$ approaches $A^*\gamma(x)$. Since $\gamma^{-1}(C)$ is $\mathcal{G}$-invariant, so is $\operatorname{int}_X \gamma^{-1}(C)$, and hence since x lies in $\operatorname{int}_X \gamma^{-1}(C)$, so does $A^*\gamma(x)$. Thus for sufficiently large n' we have $\gamma(y_{n'}) \in \gamma^{-1}(C)$; whence $\gamma(y_{n'}) \in C$. Now since C is $\mathcal{G}_Y$-invariant, $y_{n'} \in C$, which is a contradiction.

For any point y in C, y lies in Y and there is an operator A in $\mathcal{G}_Y$ with $y = A\gamma(y)$. Since C is $\mathcal{G}_Y$-invariant it follows that $\gamma(y) \in C$, so that $y \in \gamma^{-1}(C) \cap Y$. Conversely, if $y \in \gamma^{-1}(C) \cap Y$ then again there exists A in $\mathcal{G}_Y$ with $y = A\gamma(y)$; whence $y \in C$ since C is $\mathcal{G}_Y$-invariant. Thus $C = \gamma^{-1}(C) \cap Y$.

Now suppose that C is convex and, without loss of generality, nonempty. By Corollary 5.2, $\gamma^{-1}(C)$ is a nonempty, $\mathcal{G}$-invariant, convex set, so there exists a point x in $\operatorname{ri}\gamma^{-1}(C)$. Since $\operatorname{ri}\gamma^{-1}(C)$ is $\mathcal{G}$-invariant, $\gamma(x)$ lies in $Y \cap \operatorname{ri}\gamma^{-1}(C)$. Hence the relative interiors of the convex sets $\gamma^{-1}(C)$ and Y intersect, so $\operatorname{ri} C = \operatorname{ri}(Y \cap \gamma^{-1}(C)) = Y \cap \operatorname{ri}\gamma^{-1}(C)$, by [32, Cor. 6.5.1], and it is elementary to check that $\operatorname{aff}(Y \cap \gamma^{-1}(C)) = Y \cap \operatorname{aff}\gamma^{-1}(C)$. Now, a point z belongs to $\operatorname{ri}\gamma^{-1}(C)$ if and only if $\gamma(z) \in Y \cap \operatorname{ri}\gamma^{-1}(C) = \operatorname{ri} C$, by the $\mathcal{G}$-invariance of $\gamma^{-1}(C)$, and (iv) follows. Equation (v) is similar. □

The pattern of these results is clear. *If the convex subset C of Y is $\mathcal{G}_Y$-invariant then for many set operations '#' the following metaformula holds*:

$$\boxed{\#(\gamma^{-1}(C)) = \gamma^{-1}(\#(C)).} \tag{5.1}$$

The utility of this formula lies in expressing the result of an operation in the larger space X on a complicated set, $\gamma^{-1}(C)$, in terms of the result of the same operation in a smaller space Y on the simpler set C. A straightforward deduction (in light of Theorem 5.4) is that if the convex subset D of X is $\mathcal{G}$-invariant then for many set operations '#' the following metaformula holds:

$$\boxed{Y \cap \#D = \#(Y \cap D).} \tag{5.2}$$

We will see another example of this pattern in the next section—we will show that $\exp(\gamma^{-1}(C)) = \gamma^{-1}(\exp C)$ for a closed, convex set C, where $\exp C$ denotes the set of exposed points of C. To end this section we prove the analogous result for the set of *extreme points* of C, denoted $\operatorname{ext} C$, which are those points x in C for which $C \setminus \{x\}$ is convex.

THEOREM 5.5 (extreme points). *If the subset C of Y is convex and $\mathcal{G}_Y$-invariant then*

$$\operatorname{ext}(\gamma^{-1}(C)) = \gamma^{-1}(\operatorname{ext} C).$$

Proof. Suppose first that the point x in X does not belong to $\gamma^{-1}(\operatorname{ext} C)$, so that $\gamma(x) \notin \operatorname{ext} C$. If $\gamma(x) \notin C$ then clearly $x \notin \operatorname{ext}(\gamma^{-1}(C))$, so suppose that for some points u and v in C distinct from $\gamma(x)$ and some scalar α in $(0,1)$ we have $\gamma(x) = \alpha u + (1-\alpha)v$. For any operator A in $\mathcal{G}^x$ we now have $x = A\gamma(x) = \alpha Au + (1-\alpha)Av$, and since the points Au and Av are distinct from x in the set $\gamma^{-1}(C)$, it follows that x is not extreme in this set.

On the other hand, suppose that $\gamma(x)$ is extreme in C and yet x is not extreme in $\gamma^{-1}(C)$—we will derive a contradiction. Pick points u_1 and v_1 distinct from x in $\gamma^{-1}(C)$ and a scalar α_1 in $(0,1)$ with $x = \alpha_1 u_1 + (1-\alpha_1)v_1$. Now define a new point

$$u = \begin{cases} \frac{1}{2}(x+u_1) & \text{if } \gamma(x) = \gamma(u_1), \\ u_1 & \text{otherwise.} \end{cases}$$

Since $\gamma^{-1}(C)$ is convex, u lies in $\gamma^{-1}(C)$, and if $\gamma(x) = \gamma(u_1)$ then $\|x\| = \|u_1\|$ by Theorem 2.4 with

$$\|\gamma(u)\| = \|u\| < \frac{\|x\| + \|u_1\|}{2} = \|x\| = \|\gamma(x)\|.$$

Hence in either case $\gamma(u) \neq \gamma(x)$. By defining a point v in an analogous fashion we arrive at a representation $x = \alpha u + (1-\alpha)v$ for a scalar α in $(0,1)$ with $\gamma(u)$ and $\gamma(v)$ distinct from $\gamma(x)$ in C.

Now certainly $\gamma(x)$ does not belong to either of the cosets $\mathcal{G}_Y\gamma(u)$ or $\mathcal{G}_Y\gamma(v)$. For example, if $\gamma(x) = A\gamma(u)$ for some operator A in $\mathcal{G}_Y$ then applying γ gives a contradiction. Thus, since $\gamma(x)$ is extreme,

$$\gamma(x) \notin \operatorname{conv}(\mathcal{G}_Y\gamma(u) \cup \mathcal{G}_Y\gamma(v)).$$

Since the set on the right-hand side is compact, we can choose an element y of Y (defining a separating hyperplane) so that

$$\begin{aligned} \langle \gamma(y), \gamma(x)\rangle \geq \langle y, \gamma(x)\rangle &> \max\langle y, \mathcal{G}_Y\gamma(u) \cup \mathcal{G}_Y\gamma(v)\rangle \\ &= \max\{\langle \gamma(y), \gamma(u)\rangle, \langle \gamma(y), \gamma(v)\rangle\}, \end{aligned}$$

using Proposition 2.3. But this contradicts the fact that $x = \alpha u + (1-\alpha)v$ and the function $\langle \gamma(y), \gamma(\cdot)\rangle$ is convex (Theorem 2.4). □

6. Smoothness, strict convexity, and invariant norms. Our aim in this section is to investigate the dual concepts of smoothness and strict convexity for $\mathcal{G}$-invariant convex functions. Once again, we assume throughout that $(X, \mathcal{G}, \gamma)$ is a normal decomposition system, and that Assumption 4.1 holds, which is to say that $(Y, \mathcal{G}_Y, \gamma)$ is a subsystem where the space Y contains the range of γ.

The first result shows that a $\mathcal{G}$-invariant convex function $h \circ \gamma$ (where the function h is $\mathcal{G}_Y$-invariant—see Proposition 4.2) is differentiable at a point x in X if and only if h is differentiable at $\gamma(x)$.

THEOREM 6.1 (differentiability). *Let the function $h : Y \to (-\infty, +\infty]$ be $\mathcal{G}_Y$-invariant. If $h \circ \gamma$ is differentiable at a point x in X then h is differentiable at $\gamma(x)$, and the following chain rule holds:*

$$\nabla(h \circ \gamma)(x) = A\nabla h(\gamma(x)) \quad \textit{for any operator } A \in \mathcal{G}^x. \tag{6.1}$$

Conversely, if h is in addition convex, and differentiable at $\gamma(x)$, then $h \circ \gamma$ is differentiable at x and, furthermore, $\gamma(\nabla(h \circ \gamma)(x)) = \nabla h(\gamma(x))$.

Proof. For any operator A in $\mathcal{G}^x$ we have $x = A\gamma(x)$, and for all points y in Y

$$(h \circ \gamma)(Ay) = h(\gamma(Ay)) = h(\gamma(y)) = h(y),$$

since h is $\mathcal{G}_Y$-invariant. The left-hand side is differentiable at $y = \gamma(x)$, by the chain rule, hence so is the right-hand side with $\nabla h(\gamma(x)) = A^* \nabla (h \circ \gamma)(x)$. The first part of the result follows.

On the other hand, if h is also convex, and differentiable at $\gamma(x)$, then $\partial h(\gamma(x)) = \{\nabla h(\gamma(x))\}$ by [32, Thm. 25.1]. Now by Theorem 4.5 (subdifferentials), if an element w of X belongs to $\partial(h \circ \gamma)(x)$ then $\gamma(w) \in \partial h(\gamma(x))$, and so $\gamma(w) = \nabla h(\gamma(x))$. In particular, since γ is norm preserving (Theorem 2.4), any such subgradient has norm $\|\nabla h(\gamma(x))\|$. Since $\partial(h \circ \gamma)(x)$ is a convex set and $\|\cdot\|$ is a strict norm, $\partial(h \circ \gamma)(x)$ has at most one element. However, it is nonempty by the chain rule (4.2). Thus it is a singleton, whence $h \circ \gamma$ is differentiable at x by [32, Thm. 25.1], and the result follows. □

We say that a proper, closed, convex function $h : Y \to (-\infty, +\infty]$ is *essentially smooth* if it is differentiable at any point where it has a subgradient, and is *essentially strictly convex* if it is strictly convex on any convex set on which the subdifferential is everywhere nonempty. These two concepts are dual to each other: h is essentially smooth if and only if its conjugate is essentially strictly convex and vice versa [32, Thm. 26.3].

COROLLARY 6.2 (essential smoothness and strict convexity). *Suppose that the function $h : Y \to (-\infty, +\infty]$ is $\mathcal{G}_Y$-invariant, closed, proper, and convex. Then the function $h \circ \gamma$ is essentially smooth (respectively, essentially strictly convex) if and only if h is essentially smooth (respectively, essentially strictly convex).*

Proof. Suppose first that $h \circ \gamma$ is essentially smooth. If h has a subgradient $v \in Y$ at the point $y \in Y$ then by Corollary 4.6 we have $\gamma(v) \in \partial h(\gamma(y))$. Since the identity operator lies in $\mathcal{G}^{\gamma(y)}$ it follows from the subdifferential formula (4.2) that $\gamma(v) \in \partial(h \circ \gamma)(\gamma(y))$. Thus because $h \circ \gamma$ is essentially smooth, it is differentiable at $\gamma(y)$, and hence by Theorem 6.1 (differentiability), h is differentiable at $\gamma(y)$, and therefore also at y by Lemma 3.2. Thus h must be essentially smooth.

Conversely, suppose that h is essentially smooth. If $h \circ \gamma$ has a subgradient at a point x in X then the subdifferential formula (4.2) implies that $\partial h(\gamma(x))$ is nonempty. Hence h is differentiable at $\gamma(x)$, and therefore $h \circ \gamma$ is differentiable at x by Theorem 6.1 (differentiability). Thus $h \circ \gamma$ is essentially smooth.

The essentially strictly convex case follows by taking conjugates. □

The following result is another example of the pattern (5.1) that we observed in the last section: $\#(\gamma^{-1}(C)) = \gamma^{-1}(\#(C))$. If the subset C of Y is closed and convex then we say that a point y in C is *exposed* if there is an element z of Y with $\langle z, y \rangle > \langle z, u \rangle$ for all points u in $C \setminus \{y\}$. Equivalently, a point y in C is exposed if and only if it lies in the range of $\nabla \delta_C^*$ [32, Cor. 25.1.3]. We denote the set of exposed points by $\exp(C)$. A generalization of this result to exposed *faces* appears in [21].

COROLLARY 6.3 (exposed points). *Suppose that the subset C of Y is $\mathcal{G}_Y$-invariant, closed, and convex. Then*

$$\exp(\gamma^{-1}(C)) = \gamma^{-1}(\exp(C)).$$

Proof. If the point x in X satisfies $\gamma(x) \in \exp(C)$ then for some element v of Y we have $\gamma(x) = \nabla \delta_C^*(v)$, and by Corollary 4.7 it follows that $\gamma(x) = \nabla \delta_C^*(\gamma(v))$. Notice

that $\delta_{\gamma^{-1}(C)} = \delta_C \circ \gamma$, and hence by Theorem 4.4 (conjugacy) we have $\delta^*_{\gamma^{-1}(C)} = \delta^*_C \circ \gamma$. Choose an operator A in $\mathcal{G}^x$, so that $x = A\gamma(x)$, and observe that $A \in \mathcal{G}^{A\gamma(v)}$. Thus, applying the chain rule (6.1),

$$\nabla \delta^*_{\gamma^{-1}(C)}(A\gamma(v)) = \nabla(\delta^*_C \circ \gamma)(A\gamma(v)) = A\nabla\delta^*_C(\gamma(v)) = A\gamma(x) = x,$$

so that $x \in \exp(\gamma^{-1}(C))$.

Conversely, if $x \in \exp(\gamma^{-1}(C))$ then for some element w of X we have $x = \nabla\delta^*_{\gamma^{-1}(C)}(w) = \nabla(\delta^*_C \circ \gamma)(w)$. It follows by Theorem 6.1 (differentiability) that $\gamma(x) = \nabla\delta^*_C(\gamma(w))$, whence $\gamma(x) \in \exp(C)$. □

To end this section we examine our results for the special case of invariant norms. If p is a norm on Y then we denote the *dual norm* on Y by p^D, where for an element z of Y,

$$p^D(z) = \max\{\langle y, z\rangle \mid y \in Y,\ p(y) = 1\}.$$

We relate the dualizing operation for norms with conjugacy by the following standard and straightforward trick.

LEMMA 6.4. *If p is a norm on Y then $(\frac{1}{2}p(\cdot)^2)^* = \frac{1}{2}(p^D(\cdot))^2$.*

A norm p on Y is *smooth* if it is differentiable except at the origin. Equivalently, the proper, closed, convex function $p^2/2$ is essentially smooth. Furthermore, p is *strict* if $p(u+v) < 2$ for all distinct points u and v in the unit ball for p, namely, $\{y \in Y \mid p(y) \le 1\}$. Equivalently, $p^2/2$ is essentially strictly convex. A point y in Y is a *smooth point* of the unit ball if $p(y) = 1$ and p is differentiable at y.

THEOREM 6.5 (norms). *The $\mathcal{G}$-invariant norms on X are those functions of the form $p \circ \gamma$, where p is a $\mathcal{G}_Y$-invariant norm on Y. The dual of such a norm is $p^D \circ \gamma$. The norm $p \circ \gamma$ is smooth (respectively, strict) if and only if p is smooth (respectively, strict). A point x in X is an extreme (respectively, exposed, smooth) point of the unit ball for $p \circ \gamma$ if and only if $\gamma(x)$ is an extreme (respectively, exposed, smooth) point of the unit ball for p.*

Proof. By Proposition 4.2, the $\mathcal{G}$-invariant functions on X are those of the form $p \circ \gamma$ with p a $\mathcal{G}_Y$-invariant function on Y. If $p \circ \gamma$ is actually a norm on X then p is a norm on Y, since by $\mathcal{G}_Y$-invariance, p agrees with $p \circ \gamma$ on Y. Conversely, suppose that p is a $\mathcal{G}_Y$-invariant norm. Then certainly $(p \circ \gamma)(x) = p(\gamma(x)) \ge 0$ for all points x in X with equality if and only if $\gamma(x) = 0$ or, equivalently, $x = 0$. Positive homogeneity of $p \circ \gamma$ follows from that of γ (Theorem 2.4). Finally, $p \circ \gamma$ is convex by Theorem 4.3, and hence is a norm.

By Lemma 6.4 we have

$$\begin{aligned}((p \circ \gamma)^D)^2/2 = ((p \circ \gamma)^2/2)^* &= (p^2/2 \circ \gamma)^* \\ &= (p^2/2)^* \circ \gamma = (p^D)^2/2 \circ \gamma = (p^D \circ \gamma)^2/2,\end{aligned}$$

using Theorem 4.4 (conjugacy). Hence $(p \circ \gamma)^D = p^D \circ \gamma$. The norm $p \circ \gamma$ is smooth if and only if $(p \circ \gamma)^2/2 = p^2/2 \circ \gamma$ is essentially smooth, which by Corollary 6.2 is equivalent to the essential smoothness of $p^2/2$, and hence to the smoothness of p. The strict case is analogous.

The last statement follows by applying Theorem 5.5 (extreme points) and Corollary 6.3 (exposed points) to the unit ball for p, and by applying Theorem 6.1 (differentiability) to p. □

7. Examples. The idea of a normal decomposition system that we introduced in Definition 2.1 works well as an abstract mechanism. Its real significance, however, is in the variety of examples that it models. In this section we discuss these examples. They fall into two distinct categories: "discrete" examples, where the group $\mathcal{G}$ is a reflection group (in fact a "Weyl group") and the range of the map γ has full dimension in the underlying inner product space X, and "continuous" examples, where γ maps X into a strictly smaller space Y. Both categories are important for our purposes. Further discussion of the role of Weyl groups in this construction may be found in [22].

First we explain some notation for various sets of matrices. The trace of a matrix w is denoted by $\operatorname{tr}(w)$ and the Hermitian conjugate by w^*.

$\mathcal{O}_n$:	The (multiplicative) group of $n \times n$ real orthogonal matrices.
$\mathcal{U}_n$:	The (multiplicative) group of $n \times n$ complex unitary matrices.
$\mathcal{P}_n$:	The (multiplicative) group of $n \times n$ permutation matrices.
$\mathcal{P}_n^{\pm}$:	The (multiplicative) group of $n \times n$ "signed" permutation matrices (having exactly one nonzero entry, ± 1, in each row and each column.
S_n:	The inner product space of $n \times n$ real symmetric matrices with $\langle w, v \rangle = \operatorname{tr}(wv)$.
H_n:	The (*real*) inner product space of $n \times n$ complex Hermitian matrices with $\langle w, v \rangle = \operatorname{tr}(wv)$.
$M_{m,n}(\mathbf{R})$:	The inner product space of $m \times n$ real matrices with $\langle w, v \rangle = \operatorname{tr}(w^T v)$.
$M_{m,n}(\mathbf{C})$:	The (*real*) inner product space of $m \times n$ complex matrices with $\langle w, v \rangle = \operatorname{Re} \operatorname{tr}(w^* v)$.

For a matrix w in S_n or H_n, the vector $\lambda(w) \in \mathbf{R}^n$ has components the eigenvalues of w, arranged in nonincreasing order. For a matrix w in $M_{m,n}(\mathbf{R})$ or $M_{m,n}(\mathbf{C})$, the vector $\sigma(w) \in \mathbf{R}_+^l$ (where $l = \min\{m, n\}$) has components the singular values of w, arranged in nonincreasing order.

Recall that a normal decomposition system consists of a real inner product space X, a subgroup $\mathcal{G}$ of the orthogonal group on X, $O(X)$, and a map $\gamma : X \to X$ satisfying Definition 2.1.

EXAMPLE 7.1 (reordering on $\mathbf{R}^n$). We take $X = \mathbf{R}^n$ (with the standard inner product), $\mathcal{G} = \mathcal{P}_n$ (considered as a subgroup of $O(\mathbf{R}^n) = \mathcal{O}_n$ in the natural way), and $\gamma(x) = \bar{x}$, where the vector $\bar{x} \in \mathbf{R}^n$ has components $\{x_1, x_2, \ldots, x_n\}$ arranged in nonincreasing order. The conditions in Definition 2.1 are immediate except for (c), which states that

$$\langle x, z \rangle \leq \langle \bar{x}, \bar{z} \rangle \quad \textit{for all } x \textit{ and } z \textit{ in } \mathbf{R}^n \tag{7.1}$$

(with equality if and only if $x = A\bar{x}$ and $z = A\bar{z}$ for some permutation matrix A). Inequality (7.1) is classical; see, for example, [14, Thm. 368] and [18, Lem. 2.1].

The range of γ is $\mathbf{R}^n_{\geq} = \{x \in \mathbf{R}^n \mid (x_i) \text{ nonincreasing}\}$. The dual cone is straightforward to compute. In fact, a vector z lies in $(\mathbf{R}^n_{\geq})^+$ if and only if

$$\sum_1^j z_i \geq 0 \quad \text{for } j = 1, 2, \ldots, n$$

with equality for $j = n$. We say that a real function f on $\mathbf{R}^n_{\geq}$ is *Schur convex* if $f(x) \geq f(w)$ whenever x and w lie in $\mathbf{R}^n_{\geq}$ with $x - w$ in $(\mathbf{R}^n_{\geq})^+$. Theorem 3.3 now shows that any symmetric, convex function is Schur convex [23, Prop. 3.C.2].

EXAMPLE 7.2 (absolute reordering on $\mathbf{R}^n$). We take $X = \mathbf{R}^n$, $\mathcal{G} = \mathcal{P}_n^{\pm}$, and $\gamma(x) = \overline{|x|}$ (where $|x| = (|x_1|, |x_2|, \ldots, |x_n|)$. Thus

$$(\overline{|x|})_1 = \max\{|x_1|, |x_2|, \ldots, |x_n|\},$$

and so forth. The conditions in Definition 2.1 are easy to check: (c) follows from inequality (7.1).

Diagonal matrices will play an important role in our continuous examples. We denote the smaller of the two dimensions m and n by $l = \min\{m, n\}$, and then we define a map $\operatorname{Diag} : \mathbf{R}^l \to M_{m,n}(\mathbf{C})$ by

$$(\operatorname{Diag}\alpha)_{ij} = \begin{cases} \alpha_i & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

EXAMPLE 7.3 (symmetric matrices). We take $X = S_n$ and $\mathcal{G}$ to be the group of orthogonal similarity transformations $x \mapsto u^T x u$ for symmetric matrices x and orthogonal matrices u. Finally, we define $\gamma(x) = \operatorname{Diag}\lambda(x)$.

More formally, define the *adjoint representation* of $\mathcal{O}_n$ on S_n, which we write $\operatorname{Ad} : \mathcal{O}_n \to O(S_n)$, by $(\operatorname{Ad}(u))x = u^T x u$ for orthogonal u and symmetric x. Then $\mathcal{G}$ is just the range of this representation, which has kernel $\{\pm\mathrm{id}\}$, and so $\mathcal{G}$ is isomorphic to $\mathcal{O}_n/\{\pm\mathrm{id}\}$.

Let us check the conditions of Definition 2.1. Condition (a), the $\mathcal{G}$-invariance of γ, amounts to the invariance of the set of eigenvalues under orthogonal similarity. Condition (b), the decomposition axiom, follows from the spectral decomposition. Condition (c), the angle contraction axiom, becomes the following inequality:

$$\operatorname{tr}(wx) \leq \langle \lambda(w), \lambda(x) \rangle \quad \text{for all } w, x \in S_n$$

(with equality if and only if there exists an orthogonal matrix u satisfying $x = u^T(\operatorname{Diag}\lambda(x))u$ and $w = u^T(\operatorname{Diag}\lambda(w))u$). The inequality appears in [25], for example, and the conditions for equality may be found in [35], using algebraic techniques. A variational proof is given in [18]. The result is closely connected with earlier work of von Neumann—see Example 7.5.

The natural choice for the subspace Y is the space of diagonal matrices $\operatorname{Diag}\mathbf{R}^n$. A standard calculation shows that an orthogonal u has $\operatorname{Ad}(u)$ in the stabilizer $\mathcal{G}_Y$ if and only if $u \in \mathcal{P}_n^{\pm}$. However, since

$$(\operatorname{Ad}(\operatorname{Diag}(\pm 1, \pm 1, \ldots, \pm 1)))|_Y = \mathrm{id},$$

we see that the group $\mathcal{G}_Y$ acting on the space Y of diagonal matrices is simply the permutation group $\mathcal{P}_n$ acting on the diagonal entries.

Notice also that for any vector α in $\mathbf{R}^n$ we have $\gamma(\operatorname{Diag}\alpha) = \operatorname{Diag}\bar{\alpha}$. Hence the subsystem $(Y, \mathcal{G}_Y, \gamma)$ is a normal decomposition system isomorphic to the "reordering" system described in Example 7.1. In particular, Assumption 4.1 holds, so that all of the machinery that we have developed can be applied. We list some consequences in the final section.

EXAMPLE 7.4 (Hermitian matrices). The complex analogue of the previous example is very similar (and there is a quaternionic analogue). We take $X = H_n$, which

we consider a *real* inner product space (since we are primarily concerned with properties of real vector spaces, such as convexity). The group $\mathcal{G}$ now consists of unitary similarity transformations $x \mapsto u^*xu$ for Hermitian x and unitary u and, as before $\gamma(x) = \operatorname{Diag}\lambda(x)$.

Formally, we define the adjoint representation of $\mathcal{U}_n$ on H_n, written $\operatorname{Ad} : \mathcal{U}_n \to O(H_n)$, by $(\operatorname{Ad}(u))(x) = u^*xu$ for unitary u and Hermitian x. Then $\mathcal{G}$ is just the range of this representation, which has kernel $\mathbf{T}\mathrm{id}$, where $\mathbf{T}$ is the circle group $\{\tau \in \mathbf{C} \mid |\tau| = 1\}$. Thus $\mathcal{G}$ is isomorphic to $\mathcal{U}_n/\mathbf{T}\mathrm{id}$. Checking Definition 2.1 is entirely analogous to the previous example.

An aside is illustrative at this point. If we choose the subspace Y as S_n then the stabilizer $\mathcal{G}_Y$ acts on Y exactly as $\operatorname{Ad}\mathcal{O}_n$. Thus in this case the subsystem $(Y, \mathcal{G}_Y, \gamma)$ is a normal decomposition system isomorphic to the previous "symmetric matrix" Example 7.3. In particular, Assumption 4.1 holds.

The natural choice, however, is again to choose Y as the subspace of diagonal matrices $\operatorname{Diag}\mathbf{R}^n$. Then it is once again straightforward to identify the action of the stabilizer $\mathcal{G}_Y$ on this subspace with the permutation group $\mathcal{P}_n$ acting on the diagonal entries. Thus the subsystem $(Y, \mathcal{G}_Y, \gamma)$ is a normal decomposition system isomorphic to the "reordering" system, Example 7.1. Again Assumption 4.1 holds, so our machinery applies.

EXAMPLE 7.5 (real matrices). We take $X = M_{m,n}(\mathbf{R})$ and $\mathcal{G}$ to be the group of transformations $x \mapsto u^Txv$ for orthogonal matrices u in $\mathcal{O}_m$ and v in $\mathcal{O}_n$. Then we define $\gamma(x) = \operatorname{Diag}\sigma(x)$.

Formally, we define a representation of $\mathcal{O}_m \times \mathcal{O}_n$ on $M_{m,n}(\mathbf{R})$, written $\operatorname{Ac} : \mathcal{O}_m \times \mathcal{O}_n \to O(M_{m,n}(\mathbf{R}))$, by $(\operatorname{Ac}(u,v))x = u^Txv$. Then $\mathcal{G}$ is the range of this representation: since the kernel of Ac is just $\{\pm(\mathrm{id}, \mathrm{id})\}$, the group $\mathcal{G}$ is isomorphic to $(\mathcal{O}_m \times \mathcal{O}_n)/\{\pm(\mathrm{id}, \mathrm{id})\}$.

Checking Definition 2.1, $\mathcal{G}$-invariance amounts to the invariance of the set of singular values under the transformations we consider. Condition (b), the decomposition axiom, follows from the singular value decomposition, and condition (c), the angle contraction axiom, becomes "von Neumann's lemma" [27]:

$$\operatorname{tr}(w^Tx) \leq \langle \sigma(w), \sigma(x) \rangle, \quad \text{for all } w, x \in M_{m,n}(\mathbf{R}) \tag{7.2}$$

(with equality if and only if w and x have simultaneous singular value decompositions $w = u^T(\operatorname{Diag}\sigma(w))v$ and $x = u^T(\operatorname{Diag}\sigma(x))v$ for some u in $\mathcal{O}_m$ and v in $\mathcal{O}_n$)—see the discussion in [7].

The natural choice for Y is the space of diagonal matrices $\operatorname{Diag}\mathbf{R}^l$ (where $l = \min\{m,n\}$). A little thought then identifies the action of the stabilizer $\mathcal{G}_Y$ on the space Y with the group of transformations $\operatorname{Diag}(\alpha) \mapsto \operatorname{Diag}(p\alpha)$ for a vector α in $\mathbf{R}^l$ and a matrix p in $\mathcal{P}_l^{\pm}$. To see this, note that any such transformation clearly belongs to $\mathcal{G}_Y$, whereas on the other hand a transformation in $\mathcal{G}_Y$ must preserve diagonality and the singular values.

Notice also that $\gamma(\operatorname{Diag}\alpha) = \overline{|\alpha|}$ for any vector α in $\mathbf{R}^l$. Thus the subsystem $(Y, \mathcal{G}_Y, \gamma)$ is a normal decomposition system isomorphic to the "absolute reordering" system described in Example 7.2. Since Assumption 4.1 holds, our machinery applies. Some consequences appear in the final section.

Two special cases deserve mention. The case $m = 1$ gives exactly the example discussed after Assumption 4.1. The even more special case $m = n = 1$ gives our very first example of a normal decomposition system, discussed after Definition 2.1.

EXAMPLE 7.6 (complex matrices). The complex analogue of the previous example is very similar (and again there is a quaternionic analogue). We take $X = M_{m,n}(\mathbf{C})$

and $\mathcal{G}$ to be the group of transformations $x \mapsto u^* x v$ for a matrix x in $M_{m,n}(\mathbf{C})$ and unitary matrices u in $\mathcal{U}_m$ and v in $\mathcal{U}_n$. Once again we define $\gamma(x) = \operatorname{Diag}\sigma(x)$.

Formally, we define a representation of $\mathcal{U}_m \times \mathcal{U}_n$ on $M_{m,n}(\mathbf{C})$, written $\mathrm{Ac} : \mathcal{U}_m \times \mathcal{U}_n \to O(M_{m,n}(\mathbf{C}))$, by $(\mathrm{Ac}(u,v))x = u^* x v$. Then $\mathcal{G}$ is the range of this representation. Since the kernel of Ac is easily checked to be $\mathbf{T}(\mathrm{id}, \mathrm{id})$ (where $\mathbf{T}$ is once again the circle group), we see that the group $\mathcal{G}$ is isomorphic to $(\mathcal{U}_m \times \mathcal{U}_n)/\mathbf{T}\mathrm{id}$. Checking Definition 2.1 is analogous to the previous example. In fact, if we choose $Y = M_{m,n}(\mathbf{R})$ then the subsystem $(Y, \mathcal{G}_Y, \gamma)$ is isomorphic to the previous example.

If we make the natural choice for Y, namely, the space of real diagonal matrices $\operatorname{Diag}\mathbf{R}^l$, then a similar argument to the previous example identifies the action of the stabilizer $\mathcal{G}_Y$ on the space Y as the group of transformations $\operatorname{Diag}(\alpha) \mapsto \operatorname{Diag}(p\alpha)$ for a vector α in $\mathbf{R}^l$ and a matrix p in $\mathcal{P}_l^{\pm}$. Thus just as in the previous example, the subsystem $(Y, \mathcal{G}_Y, \gamma)$ is isomorphic to the "absolute reordering" system, Example 7.2. Again, all our machinery applies.

8. Consequences for matrix functions. In this concluding section we consider how our results can be applied to the examples in the previous section to derive a variety of interesting results in the literature. We begin with the case of symmetric matrices, Example 7.3. The complex analogue is entirely similar, and we do not pursue it.

Symmetric matrices. A function $h : \mathbf{R}^n \to [-\infty, +\infty]$ is *symmetric* if for any vector α in $\mathbf{R}^n$ the value $h(\alpha)$ is unchanged by permuting the components of α—using the notation of Example 7.1, $h(\alpha) = h(\bar{\alpha})$. Similarly, a subset C of $\mathbf{R}^n$ is *symmetric* when $\alpha \in C$ if and only if $\bar{\alpha} \in C$. A function on the space of symmetric matrices $f : S_n \to [-\infty, +\infty]$ is *weakly orthogonally invariant* if $f(u^T x u) = f(x)$ for any matrices x in S_n and u in $\mathcal{O}_n$. Such functions have also been called *spectral* [13]. Analogously, a subset D of S_n is *weakly orthogonally invariant* if $u^T x u \in D$ whenever $x \in D$ (for orthogonal u).

The following result follows immediately by applying our machinery to Example 7.3. We make no attempt to be exhaustive.

THEOREM 8.1 (convex spectral functions). *Weakly orthogonally invariant extended real-valued functions on S_n are exactly those functions of the form $h \circ \lambda$ for a symmetric function $h : \mathbf{R}^n \to (-\infty, +\infty]$. Such a function on S_n is convex (respectively, closed, essentially strictly convex, essentially smooth) if and only if h is convex (respectively, closed, essentially strictly convex, essentially smooth). For any such symmetric function h we have*

$$(h \circ \lambda)^* = h^* \circ \lambda. \tag{8.1}$$

Suppose further that some symmetric matrix x satisfies $\lambda(x) \in \operatorname{dom} h$. Then the symmetric matrix w is a subgradient of $h \circ \lambda$ at x if and only if $\lambda(w)$ is a subgradient of h at $\lambda(x)$ and x and w have simultaneous spectral decompositions $x = u^T(\operatorname{Diag}\lambda(x))u$ and $w = u^T(\operatorname{Diag}\lambda(w))u$ for some orthogonal matrix u. In fact, the following "chain rule" holds:

$$\partial(h \circ \lambda)(x) = \{u^T(\operatorname{Diag}\mu)u \mid u \in \mathcal{O}_n,\ u^T(\operatorname{Diag}\lambda(x))u = x,\ \mu \in \partial h(\lambda(x))\}.$$

If h is convex then $h \circ \lambda$ is differentiable at x if and only if h is differentiable at $\lambda(x)$.

EXAMPLE 8.2 (the log barrier). Let us define a symmetric function $h : \mathbf{R}^n \to (-\infty, +\infty]$ by

$$h(\alpha) = \begin{cases} -\sum_{i=1}^n \log \alpha_i & \text{if } \alpha > 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Then h is a closed, convex function, essentially smooth, and essentially strictly convex, with conjugate

$$h^*(\mu) = \begin{cases} -n - \sum_{i=1}^n \log(-\mu_i) & \text{if } \mu < 0, \\ +\infty & \text{otherwise.} \end{cases}$$

It follows that the matrix function $h \circ \lambda : S_n \to (-\infty, +\infty]$ defined by

$$(h \circ \lambda)(x) = \begin{cases} -\log(\det x) & \text{if } x \text{ is positive definite,} \\ +\infty & \text{otherwise} \end{cases}$$

is also closed, convex, essentially smooth, and essentially strictly convex with conjugate

$$(h \circ \lambda)^*(w) = \begin{cases} -n - \log(\det(-w)) & \text{if } w \text{ is negative definite,} \\ +\infty & \text{otherwise.} \end{cases}$$

It is easy to check, using the chain rule, that for a positive-definite symmetric matrix x,

$$\nabla(h \circ \lambda)(x) = -x^{-1}.$$

The convexity part of Theorem 8.1 was essentially first proved in [6]. It was rediscovered in [3]. A characterization of convexity in the differentiable case was proved in [13] via Schur convexity, and the closed case was proved via the conjugacy formula (8.1) in [18]. The latter paper also contains the remainder of Theorem 8.1. A proof appears in [36] that $h \circ \lambda$ is analytic at x if and only if h is analytic at $\lambda(x)$. Somewhat related results appear in [20]. Numerous formulae for subgradients of specific matrix functions appear, for example, in [29, 30, 15, 16]. The chain rule in Theorem 8.1 provides a simple unified approach to these.

THEOREM 8.3 (spectral convex sets). *Weakly orthogonally invariant subsets of S_n are exactly those sets of the form $\lambda^{-1}(C)$ for symmetric subsets C of $\mathbf{R}^n$. If the symmetric matrix x has $\lambda(x)$ in the symmetric set C then a symmetric matrix w lies in the normal cone $N(x|\lambda^{-1}(C))$ if and only if $\lambda(w)$ lies in $N(\lambda(x)|C)$ with x and w having simultaneous spectral decompositions, $x = u^T(\operatorname{Diag}\lambda(x))u$ and $w = u^T(\operatorname{Diag}\lambda(w))u$ for some orthogonal matrix u. In fact,*

$$\begin{aligned} &N(x|\lambda^{-1}(C)) \\ &= \{u^T(\operatorname{Diag}\mu)u \mid u \in \mathcal{O}_n,\ u^T(\operatorname{Diag}\lambda(x))u = x,\ \mu \in N(\lambda(x)|C)\}. \end{aligned}$$

Furthermore,

$$\begin{aligned} (\lambda^{-1}(C))^- &= \lambda^{-1}(C^-), \\ (\lambda^{-1}(C))^\circ &= \lambda^{-1}(C^\circ), \\ \operatorname{int}(\lambda^{-1}(C)) &= \lambda^{-1}(\operatorname{int}(C)), \quad \textit{and} \\ \operatorname{Diag} C &= \lambda^{-1}(C) \cap \operatorname{Diag}\mathbf{R}^n. \end{aligned}$$

The set $\lambda^{-1}(C)$ is convex (respectively, closed) if and only if C is convex (respectively, closed). If C is convex then

$$\begin{aligned} \operatorname{ri}(\lambda^{-1}(C)) &= \lambda^{-1}(\operatorname{ri}(C)), \\ \operatorname{aff}(\lambda^{-1}(C)) &= \lambda^{-1}(\operatorname{aff}(C)), \quad \textit{and} \\ \operatorname{ext}(\lambda^{-1}(C)) &= \lambda^{-1}(\operatorname{ext}(C)), \end{aligned}$$

and if C is in addition closed then

$$\exp(\lambda^{-1}(C)) = \lambda^{-1}(\exp(C)).$$

EXAMPLE 8.4 (the simplex). Let us define a symmetric subset of $\mathbf{R}^n$ by

$$C = \left\{ \alpha \in \mathbf{R}^n \;\middle|\; \alpha \geq 0,\ \sum_{i=1}^{n} \alpha_i = 1 \right\}.$$

Then C is a closed, convex set with the standard unit vectors as extreme (in fact, exposed) points. We deduce that the set of symmetric matrices

$$\lambda^{-1}(C) = \{x \in S_n \mid x \text{ positive semidefinite, } \operatorname{tr}(x) = 1\}$$

is closed and convex with extreme (exposed) points yy^T for unit column vectors y in $\mathbf{R}^n$.

The fact that the set $\lambda^{-1}(C)$ is convex if and only if C is convex, for a symmetric closed set C, was proved in [11].

Unitarily invariant norms. A function $h : \mathbf{R}^l \to [-\infty, +\infty]$ is *absolutely symmetric* if the value $h(\alpha)$ at a vector α in $\mathbf{R}^l$ is independent of the order and signs of the components α_i: in the notation of Example 7.2, $h(\alpha) = h(|\alpha|)$ for all α. In particular, if such a function is also a norm then it is called a *symmetric gauge function.* A matrix function $f : M_{m,n}(\mathbf{C}) \to [-\infty, +\infty]$ is *(strongly) unitarily invariant* if $f(u^*xv) = f(x)$ for any matrix x in $M_{m,n}(\mathbf{C})$ and unitary matrices u and v.

The following result is a consequence of applying our machinery to Example 7.6. The real analogue is entirely similar. For brevity, we restrict ourselves to the norm case.

THEOREM 8.5 (unitarily invariant norms). *Unitarily invariant norms on $M_{m,n}(\mathbf{C})$ are exactly those functions of the form $h \circ \sigma$ for symmetric gauge functions h on $\mathbf{R}^l$ (where $l = \min\{m, n\}$). In this case the dual norm is given by*

$$(p \circ \sigma)^D = p^D \circ \sigma, \tag{8.2}$$

$p \circ \sigma$ is smooth (respectively, strict) if and only if p is smooth (respectively, strict), and a matrix x is an extreme (respectively, exposed, smooth) point of the unit ball for $p \circ \sigma$ if and only if $\sigma(x)$ is an extreme (respectively, exposed, smooth) point of the unit ball for p. Furthermore, a matrix w is a subgradient of $p \circ \sigma$ at x if and only if $\sigma(w)$ is a subgradient of p at $\sigma(x)$ with x and w having simultaneous singular value decompositions $x = u^*(\operatorname{Diag}\sigma(x))v$ and $w = u^*(\operatorname{Diag}\sigma(w))v$ for unitary matrices u and v. In fact,

$$\begin{aligned}&\partial(p \circ \sigma)(x)\\ &= \{u^*(\operatorname{Diag}\mu)v \mid u \in \mathcal{U}_m,\ v \in \mathcal{U}_n,\ u^*(\operatorname{Diag}\sigma(x))v = x,\ \mu \in \partial p(\sigma(x))\}.\end{aligned}$$

The classical examples are the symmetric gauge function $\|\cdot\|_p$ (for $1 \leq p \leq \infty$), which gives the "Schatten p-norm," and the functions

$$p(\alpha) = \sum_{i=1}^{k} (|\alpha|)_i \quad (\text{for } k = 1, 2, \ldots, l),$$

which give the "Ky Fan k-norms."

The fundamental characterization of unitarily invariant norms is due to von Neumann [27]. He proved the result in an analogous fashion to our conjugacy argument following Theorem 4.4 by proving the duality formula (8.2) via his lemma (7.2). Some interesting analogous results appear in [4]. The characterization of extreme, exposed, and smooth points was proved in [2]; see also [40, 41, 8, 7, 9]. Versions of the subdifferential formula appear in [38, 39].

Acknowledgments. Many thanks to Jean-Pierre Haeberly for simplifying the presentation of Definition 2.1, and for a number of other helpful comments. Many thanks also to Juan-Enrique Martinez-Legaz for helpful discussions concerning the presentation of Definition 2.1.

REFERENCES

[1] F. ALIZADEH, *Optimization over the positive definite cone: interior point methods and combinatorial applications*, in Advances in Optimization and Parallel Computing, P. Pardolos, ed., North–Holland, Amsterdam, 1992, pp. 1–25.

[2] J. ARAZY, *On the geometry of the unit ball of unitary matrix spaces*, Integral Equations Operator Theory, 4 (1981), pp. 151–171.

[3] J. M. BALL, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Rational Mech. Anal., 63 (1976), pp. 337–403.

[4] A. BARBARA AND J.-P. CROUZEIX, *Concave gauge functions and applications*, ZOR—Math. Meth. Oper. Res., 40 (1994), pp. 43–74.

[5] J. CULLUM, W. E. DONATH, AND P. WOLFE, *The minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices*, Math. Programming Study, 3 (1975), pp. 35–55.

[6] C. DAVIS, *All convex invariant functions of hermitian matrices*, Archiv der Mathematik, 8 (1957), pp. 276–278.

[7] E. M. DE SÁ, *Exposed faces and duality for symmetric and unitarily invariant norms*, Linear Algebra Appl., 197, 198 (1994), pp. 429–450.

[8] ———, *Faces and traces of the unit ball of a symmetric gauge function*, Linear Algebra Appl., 197, 198 (1994), pp. 349–395.

[9] ———, *Faces of the unit ball of a unitarily invariant norm*, Linear Algebra Appl., 197, 198 (1994), pp. 451–493.

[10] K. FAN, *On a theorem of Weyl concerning eigenvalues of linear transformations*, Proc. Nat. Acad. Sci. U.S.A., 35 (1949), pp. 652–655.

[11] P. A. FILLMORE AND J. P. WILLIAMS, *Some convexity theorems for matrices*, Glasgow Math. J., 12 (1971), pp. 110–117.

[12] R. FLETCHER, *Semi-definite matrix constraints in optimization*, SIAM J. Control Optim., 23 (1985), pp. 493–513.

[13] S. FRIEDLAND, *Convex spectral functions*, Linear and multilinear algebra, 9 (1981), pp. 299–316.

[14] G. H. HARDY, J. E. LITTLEWOOD, AND G. PÓLYA, *Inequalities*, Cambridge University Press, Cambridge, UK, 1952.

[15] J.-B. HIRIART-URRUTY, A. SEEGER, AND D. YE, *Sensitivity Analysis for a Class of Convex Functions Defined Over a Space of Symmetric Matrices*, Lecture Notes in Economics and Mathematical Systems, Vol. 382, Springer, 1992.

[16] J.-B. HIRIART-URRUTY AND D. YE, *Sensitivity analysis of all eigenvalues of a symmetric matrix*, Numer. Math., 70 (1995), pp. 45–72.

[17] F. JARRE, *An interior-point method for minimizing the maximum eigenvalue of a linear combination of matrices*, SIAM J. Control Optim., 31 (1993), pp. 1360–1377.

[18] A. S. LEWIS, *Convex analysis on the Hermitian matrices*, SIAM J. Optim., 6 (1996), pp. 164–177.

[19] ———, *The convex analysis of unitarily invariant functions*, J. Convex Anal., 2 (1995), pp. 173–183.

[20] ———, *Derivatives of spectral functions*, Math. Oper. Res., 1996, to appear.

[21] ———, *Eigenvalue-constrained faces*, Linear Algebra Appl., submitted.

[22] ———, *Von Neumann's lemma and a Chevalley-type theorem for convex functions on Cartan subspaces*, Trans. Amer. Math. Soc., submitted.

[23] A. W. MARSHALL AND I. OLKIN, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.
[24] J.-E MARTÍNEZ-LEGAZ, *On Convex and Quasi-Convex Spectral Functions*, Tech. report, Universitat Autònoma de Barcelona, Barcelona, Spain, 1995.
[25] L. MIRSKY, *On the trace of matrix products*, Mathematische Nachrichten, 20 (1959), pp. 171–174.
[26] Y. E. NESTEROV AND A. S. NEMIROVSKII, *Interior Point Polynomial Methods in Convex Programming*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1994.
[27] J. VON NEUMANN, *Some matrix inequalities and metrization of matric-space*, Tomsk University Rev., 1 (1937), pp. 286–300. In Collected Works, Vol. IV, Pergamon, Oxford, 1962, pp. 205–218.
[28] M. L. OVERTON, *On minimizing the maximum eigenvalue of a symmetric matrix*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 256–268.
[29] ———, *Large-scale optimization of eigenvalues*, SIAM J. Optim., 2 (1992), pp. 88–120.
[30] M. L. OVERTON AND R. S. WOMERSLEY, *Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices*, Math. Programming, Series B, 62 (1993), pp. 321–357.
[31] E. POLAK AND Y. WARDI, *A nondifferentiable optimization algorithm for structural problems with eigenvalue inequality constraints*, J. Structural Mech., 11 (1983), pp. 561–577.
[32] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[33] A. SEEGER, *Convex analysis of spectrally-defined matrix functions*, SIAM J. Optim., to appear.
[34] A. SHAPIRO, *Extremal problems on the set of nonnegative definite matrices*, Linear Algebra Appl., 67 (1985), pp. 7–18.
[35] C. M. THEOBALD, *An inequality for the trace of the product of two symmetric matrices*, Math. Proc. Cambridge Philosophical Soc., 77 (1975), pp. 265–266.
[36] N.-K. TSING, M. K. H. FAN, AND E. I. VERRIEST, *On analyticity of functions involving eigenvalues*, Linear Algebra Appl., 207 (1994), pp. 159–180.
[37] G. A. WATSON, *An algorithm for optimal l_2 scaling of matrices*, IMA J. Numer. Anal., 11 (1991), pp. 481–492.
[38] ———, *Characterization of the subdifferential of some matrix norms*, Linear Algebra Appl., 170 (1992), pp. 33–45.
[39] ———, *On matrix approximation problems with Ky Fan k norms*, Numer. Algorithms, 5 (1993), pp. 263–272.
[40] K. ZIĘTAK, *On the characterization of the extremal points of the unit sphere of matrices*, Linear Algebra Appl., 106 (1988), pp. 57–75.
[41] ———, *Subdifferentials, faces and dual matrices*, Linear Algebra Appl., 185 (1993), pp. 125–141.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 4, pp. 950–983, October 1996

014

STABILIZING THE GENERALIZED SCHUR ALGORITHM*

S. CHANDRASEKARAN† AND ALI H. SAYED†

Abstract. This paper provides a detailed analysis that shows how to stabilize the *generalized* Schur algorithm, which is a fast procedure for the Cholesky factorization of positive-definite structured matrices R that satisfy displacement equations of the form $R - FRF^T = GJG^T$, where J is a 2×2 signature matrix, F is a stable lower-triangular matrix, and G is a generator matrix. In particular, two new schemes for carrying out the required hyperbolic rotations are introduced and special care is taken to ensure that the entries of a Blaschke matrix are computed to high relative accuracy. Also, a condition on the smallest eigenvalue of the matrix, along with several computational enhancements, is introduced in order to avoid possible breakdowns of the algorithm by assuring the positive-definiteness of the successive Schur complements. We use a perturbation analysis to indicate the best accuracy that can be expected from *any* finite-precision algorithm that uses the generator matrix as the input data. We then show that the modified Schur algorithm proposed in this work essentially achieves this bound when coupled with a scheme to control the generator growth. The analysis further clarifies when pivoting strategies may be helpful and includes illustrative numerical examples. For all practical purposes, the major conclusion of the analysis is that the modified Schur algorithm is backward stable for a large class of structured matrices.

Key words. displacement structure, generalized Schur algorithm, Cholesky factorization, hyperbolic rotations, generator matrices, pivoting, Schur functions, error analysis

AMS subject classifications. 65F05, 65G05, 65F30, 15A23

1. Introduction.

We show how to stabilize the generalized Schur algorithm and give a finite-precision error analysis to support our conclusions. The notion of structured matrices, along with the algorithm itself, is reviewed in the next two sections. Here we proceed with a general overview of earlier relevant work in the literature.

One of the most frequent structures, at least in signal processing applications, is the Toeplitz structure, with constant entries along the diagonals of the matrix. A classical algorithm for the Cholesky factorization of the *inverses* of such matrices is the so-called Levinson–Durbin algorithm [14, 8], an error analysis of which was provided by Cybenko [7]. He showed that, in the case of positive reflection coefficients, the residual error produced by the Levinson–Durbin procedure is comparable to the error produced by the Cholesky factorization [8, p. 191].

A related analysis was carried out by Sweet [22] for the Bareiss algorithm [2], which is also closely related to an algorithm of Schur [20, 12]. These are fast procedures for the Cholesky factorization of the Toeplitz matrix itself rather than its inverse. Sweet concluded that the Bareiss algorithm is asymptotically stable.

In recent work, Bojanczyk et al. [3] further extended and strengthened the conclusions of Sweet [22] by employing elementary downdating techniques [1, 4, 5] that are also characteristic of array formulations of the Schur algorithm [13, 17]. They considered the larger class of quasi-Toeplitz matrices [13], which includes the Toeplitz matrix as a special case, and provided an error analysis that establishes that the Schur algorithm for this class of matrices is asymptotically stable.

The interesting formulation of Bojanczyk et al. [3] motivated us to take a closer look at the numerical stability of a generalized Schur algorithm [13, 15, 18] that applies

* Received by the editors June 9, 1995; accepted for publication (in revised form) by N. J. Higham November 28, 1995. This work was supported in part by National Science Foundation award MIP-9409319.

† The Center for Control Engineering and Computation, Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 (shiv@ece.ucsb.edu, sayed@ece.ucsb.edu).

to a wider class of positive-definite structured matrices R that satisfy displacement equations of the form $R - FRF^T = GJG^T$, where J is a signature matrix, F is a stable lower-triangular matrix, and G is a generator matrix. This class is briefly introduced in the next section, where the lower-triangular matrix F is shown to be pivotal in characterizing the structure of the matrix. For example, in the Toeplitz or quasi-Toeplitz case, the matrix F is equal to the shift matrix Z (i.e., a Jordan block with zero eigenvalue and ones on the first subdiagonal). Multiplying a column vector u by Z simply corresponds to shifting down the entries of u by one position. In general, however, the matrix F can be any lower-triangular matrix (for example, diagonal, bidiagonal, strictly lower triangular, etc.). This creates several complications that we address closely in order to guarantee a reliable algorithm.

For this purpose, we propose several modifications to the generalized Schur algorithm (Matlab codes for the new modified algorithm are provided at the end of this paper). In particular, two new schemes for carrying out the required hyperbolic rotations are introduced and special care is taken to ensure that the entries of the Blaschke matrix are computed to high relative accuracy. Also, a condition on the smallest eigenvalue of the matrix, along with several computational enhancements, is introduced in order to avoid possible breakdowns of the algorithm by assuring the positive-definiteness of the successive Schur complements.

We further use a perturbation analysis to indicate the best accuracy that can be expected from *any* finite-precision algorithm (slow or fast) that uses the generator matrix as the input data. We then show that the modified Schur algorithm proposed in this work essentially achieves this bound when coupled with a scheme to control the generator growth.

Another interesting idea that was recently suggested by Heinig [10] is the introduction of pivoting into algorithms for structured matrices when F is diagonal. In this paper, we have tried to clarify when pivoting may be helpful for positive-definite matrices. Numerical examples are included to support our observations. In particular, we emphasize that, in the diagonal F case, pivoting becomes necessary only when $\|F\|$ is very close to one. Furthermore, we note the following.

- If F is positive (or negative), a good strategy is shown to be the reordering of the entries of F in increasing order of magnitude.
- If F has both positive and negative entries, then numerical examples indicate that pivoting may not help in controlling the growth of the generators.

In our opinion, for positive-definite structured matrices, with diagonal or strictly lower-triangular F, the stabilization of the generalized Schur algorithm is critically dependent on the following:

- proper implementations of the hyperbolic rotations,
- proper evaluation of the Blaschke matrix–vector product,
- enforcement of positive-definiteness to avoid early breakdowns,
- control of the generator growth.

1.1. Notation. In the discussion that follows we use $\|\cdot\|$ to denote the 2-norm of its argument. We further assume, without loss of generality, that F is represented exactly in the computer. Also, the $\hat{}$ notation denotes computed quantities, while the $\bar{}$ notation denotes intermediate exact quantities. We further let ϵ denote the machine precision and n the matrix size. We also use subscripted δ's to denote quantities bounded by machine precision in magnitude, and subscripted c's to denote low-order polynomials in n.

We assume that in our floating point model, additions, subtractions, multiplica-

tions, divisions, and square roots are done to high relative accuracy; i.e.,

$$fl(x \circ y) = (x \circ y)(1 + \delta),$$

where $\circ$ denotes $+, -, \times, \div$ and $|\delta| \leq \epsilon$. The same holds for the square root operation. This is true for floating point processors that adhere to the IEEE standards.

2. Displacement structure. Consider an $n \times n$ symmetric positive-definite matrix R and an $n \times n$ lower-triangular real-valued matrix F. The displacement of R with respect to F is denoted by ∇_F and defined as

$$\nabla_F = R - FRF^T. \tag{1}$$

The matrix R is said to have low displacement rank with respect to F if the rank of ∇_F is considerably lower than n. In this case, R is said to have displacement structure with respect to F [13].

Let $r \ll n$ denote the rank of ∇_F. It follows that we can factor ∇_F as

$$\nabla_F = GJG^T, \tag{2}$$

where G is an $n \times r$ matrix and J is a signature matrix of the form

$$J = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix}, \quad p + q = r. \tag{3}$$

The integer p denotes the number of positive eigenvalues of ∇_F, while the integer q denotes the number of its negative eigenvalues. Factorization (2) is highly nonunique. If G satisfies (2) then $G\Theta$ also satisfies (2) for any J-unitary matrix Θ, i.e., for any Θ such that $\Theta J \Theta^T = J$. This follows from the trivial identity

$$(G\Theta)J(G\Theta)^T = G(\Theta J \Theta^T)G^T = GJG^T.$$

Combining (1) and (2), a matrix R is said to be structured with respect to the displacement operation defined by (1) if it satisfies a displacement equation of the form

$$R - FRF^T = GJG^T, \tag{4}$$

with a "low" rank matrix G. Equation (4) uniquely defines R (i.e., it has a unique solution R) if and only if the diagonal entries of the lower-triangular matrix F satisfy the condition

$$1 - f_i f_j \neq 0 \text{ for all } i, j.$$

This uniqueness condition will be assumed throughout the paper, although it can be relaxed in some instances [13].

The pair (G, J) is said to be a generator pair for R since, along with F, it completely identifies R. Note, however, that while R has n^2 entries, the matrix G has nr entries and r is usually much smaller than n. Therefore, algorithms that operate on the entries of G, with the purpose of obtaining a triangular factorization for R, will generally be an order of magnitude faster than algorithms that operate on the entries of R itself. The generalized Schur algorithm is one such fast $O(rn^2)$ procedure, which receives as input data the matrices (F, G, J) and provides as output data the Cholesky factor of R. A recent survey on various other forms of displacement structure and on the associated forms of Schur algorithms is [13].

2.1. Illustrative examples. The concept of displacement structure is perhaps best introduced by considering the much-studied special case of a symmetric Toeplitz matrix $T = [t_{|i-j|}]_{i,j=1}^{n}$, $t_0 = 1$.

Let Z denote the $n \times n$ lower-triangular shift matrix with ones on the first subdiagonal and zeros elsewhere (i.e., a lower-triangular Jordan block with eigenvalue 0):

$$Z = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}. \tag{5}$$

It can be easily checked that the difference $T - ZTZ^T$ has displacement rank 2 (except when all $t_i, i \neq 0$, are zero), and a generator for T is $\{G, (1 \oplus -1)\}$, where

$$T - ZTZ^T = \begin{bmatrix} 1 & 0 \\ t_1 & t_1 \\ \vdots & \vdots \\ t_{n-1} & t_{n-1} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ t_1 & t_1 \\ \vdots & \vdots \\ t_{n-1} & t_{n-1} \end{bmatrix}^T = GJG^T. \tag{6}$$

Another example is the so-called Pick matrix, which arises in the study of interpolation problems in the unit disc [19]:

$$R = \left[\frac{1 - \beta_i \beta_j}{1 - f_i f_j} \right]_{i,j=1}^{n},$$

where the β_i are real scalars and the f_i are distinct real points in the interval $(-1, 1)$. Let F denote the diagonal matrix $F = \operatorname{diag}[f_1, f_2, \ldots, f_n]$; then it can be verified that the above Pick matrix R has displacement rank 2 with respect to F since

$$R - FRF^T = \begin{bmatrix} 1 & \beta_1 \\ 1 & \beta_2 \\ \vdots & \vdots \\ 1 & \beta_n \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & \beta_1 \\ 1 & \beta_2 \\ \vdots & \vdots \\ 1 & \beta_n \end{bmatrix}^T. \tag{7}$$

More generally, one can allow for complex-valued quantities and define the Pick matrix [19] as

$$R = \left[\frac{x_i x_j^H - y_i y_j^H}{1 - f_i f_j^H} \right]_{i,j=1}^{n},$$

where ^{H} denotes Hermitian conjugation (complex conjugation for scalars), x_i and y_i are $1 \times p$ and $1 \times q$ row vectors, and f_i are complex points inside the open unit disc ($|f_i| < 1$). For the same diagonal matrix $F = \operatorname{diag}[f_1, f_2, \ldots, f_n]$, the above Pick matrix has displacement rank $r = (p + q)$, since

$$R - FRF^H = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix} \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix} \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix}^H. \tag{8}$$

Without loss of generality, the analysis provided in this paper focuses on real-valued data, i.e., on displacement equations of form (4) and on the important special case of matrices with displacement rank 2 (i.e., G has two columns and $J = (1 \oplus -1)$). The results can be extended to *higher displacement ranks* and to the complex case.

The displacement structure implied by (4) applies to symmetric matrices R. The case of nonsymmetric matrices will be pursued elsewhere since it also includes important matrices as special cases such as the Vandermonde matrix

$$V = \begin{bmatrix} 1 & \alpha_1 & \alpha_1^2 & \dots & \alpha_1^n \\ 1 & \alpha_2 & \alpha_2^2 & \dots & \alpha_2^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \alpha_n & \alpha_n^2 & \dots & \alpha_n^n \end{bmatrix}.$$

It is immediate to verify that the matrix V has displacement rank 1 since

$$V - FVZ^T = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}, \tag{9}$$

where F is now the diagonal matrix

$$F = \text{diag}\,[\alpha_1, \dots, \alpha_n].$$

3. The generalized Schur algorithm. The discussion in what follows focuses on symmetric positive-definite matrices R with displacement rank 2 with respect to a lower-triangular matrix F, viz., matrices R that satisfy displacement equations of the form

$$R - FRF^T = \begin{bmatrix} u_1 & v_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} u_1 & v_1 \end{bmatrix}^T, \tag{10}$$

where u_1 and v_1 denote the $n \times 1$ column vectors of G. The diagonal entries of F are further assumed to be strictly inside the open unit disc ($|f_i| < 1$). In this case, the matrix F is said to be stable. This condition is clearly satisfied for the Toeplitz case (5), where $f_i = 0$, and for the Pick matrix (7). In applications, the following forms are the most frequent occurrences for F: $F = Z$, $F =$ a diagonal matrix with distinct entries, $F =$ a Jordan block,

$$F = \begin{bmatrix} f_1 & & & \\ 1 & f_1 & & \\ & \ddots & \ddots & \\ & & 1 & f_1 \end{bmatrix},$$

F in bidiagonal form,

$$F = \begin{bmatrix} f_1 & & & \\ 1 & f_2 & & \\ & \ddots & \ddots & \\ & & 1 & f_n \end{bmatrix},$$

or F strictly lower triangular such as Z, Z^2, $(Z \oplus Z)$, etc.

Also, since a generator matrix G is highly nonunique, it can always be chosen to be of the form

$$G = \begin{bmatrix} x & 0 \\ x & x \\ x & x \\ \vdots & \vdots \\ x & x \end{bmatrix}. \tag{11}$$

That is, the top entry of v_1, v_{11}, can always be chosen to be zero. Indeed, assume that a generator G for R is found that does not satisfy this requirement, say

$$G = \begin{bmatrix} u_{11} & v_{11} \\ x & x \\ \vdots & \vdots \\ x & x \end{bmatrix}.$$

It then follows from (10) that the $(1,1)$ entry of R, which is positive, is given by

$$R_{11} = \frac{|u_{11}|^2 - |v_{11}|^2}{1 - |f_1|^2} > 0.$$

Consequently, $|u_{11}| > |v_{11}|$ and a hyperbolic rotation Θ can always be found in order to reduce the row $\begin{bmatrix} u_{11} & v_{11} \end{bmatrix}$ to the form $\begin{bmatrix} \sqrt{|u_{11}|^2 - |v_{11}|^2} & 0 \end{bmatrix}$. The matrix $G\Theta$ can then be used instead of G as a generator for R.

A generator matrix of form (11) is said to be in proper form. Note that in the Toeplitz case (6), the generator G is already in proper form.

The following algorithm is known as the generalized Schur algorithm: it operates on the entries of (F, G, J) and provides the Cholesky factor of R. (We remark that the algorithm can be extended to more general scenarios, e.g., an unstable F, nonsymmetric matrices R, etc.—see [13, 18, 15].)

ALGORITHM 3.1 (the generalized Schur algorithm).

- *Input data: A stable lower-triangular matrix F, a generator $G_1 = G$ in proper form, with columns denoted by u_1 and v_1, and $J = (1 \oplus -1)$.*
- *Output data: The lower-triangular Cholesky factor L of the unique matrix R that satisfies* (10), *$R = LL^T$.*

The algorithm operates as follows: start with (u_1, v_1) and repeat for $i = 1, 2, \ldots, n$:

1. *Compute the $n \times n$ matrix $\Phi_i = (F - f_i I)(I - f_i F)^{-1}$. Note that the (i,i) diagonal entry of Φ_i is zero.*
2. *Form the prearray of numbers $\begin{bmatrix} \Phi_i u_i & v_i \end{bmatrix}$. At step i, the top i entries of $\Phi_i u_i$ and v_i will be zero.*
3. *Apply a hyperbolic rotation Θ_i in order to annihilate the $(i+1)$ entry of v_i. Denote the resulting column vectors by (u_{i+1}, v_{i+1}):*

$$\begin{bmatrix} u_{i+1} & v_{i+1} \end{bmatrix} = \begin{bmatrix} \Phi_i u_i & v_i \end{bmatrix} \Theta_i. \tag{12}$$

The matrix $G_{i+1} = \begin{bmatrix} u_{i+1} & v_{i+1} \end{bmatrix}$ *will also be in proper form, with the top* i *rows equal to zero:*

$$G_{i+1} = \begin{bmatrix} 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ x & 0 \\ \vdots & \vdots \\ x & x \end{bmatrix}.$$

4. *The* i*th column of the Cholesky factor* L *is given by*

$$l_i = \sqrt{1 - |f_i|^2}(I - f_i F)^{-1} u_i. \tag{13}$$

The top $(i-1)$ *entries of* l_i *are zero.*

After n steps, the algorithm provides the Cholesky decomposition

$$R = \sum_{i=1}^{n} l_i l_i^T, \tag{14}$$

as shown below. Moreover, the successive matrices G_i that are obtained via the recursion have an interesting interpretation. Let R_i denote the Schur complement of R with respect to its leading $(i-1) \times (i-1)$ submatrix. That is, $R_1 = R$, R_2 is the Schur complement with respect to the $(1,1)$ top left entry of R, R_3 is the Schur complement with respect to the 2×2 top left submatrix of R, and so on. The matrix R_i is therefore $(n-i+1) \times (n-i+1)$. Define the $n \times n$ embedding

$$\tilde{R}_i = \begin{bmatrix} 0 & 0 \\ 0 & R_i \end{bmatrix}.$$

Then it can be shown that [13]

$$\tilde{R}_i - F\tilde{R}_i F^T = G_i J G_i^T. \tag{15}$$

In other words, G_i is a generator matrix for the ith Schur complement, which is also structured.

THEOREM 3.2. *The generalized Schur algorithm provides the Cholesky decomposition of* R*, viz.,*

$$R = \sum_{i=1}^{n} l_i l_i^T. \tag{16}$$

Proof. It follows from relation (12) that

$$u_{i+1}u_{i+1}^T - v_{i+1}v_{i+1}^T = \Phi_i u_i u_i^T \Phi_i^T - v_i v_i^T, \tag{17}$$

where, in view of (13),

$$u_i = \frac{1}{\sqrt{1 - |f_i|^2}}(I - f_i F)l_i, \quad \Phi_i u_i = \frac{1}{\sqrt{1 - |f_i|^2}}(F - f_i I)l_i.$$

Summing (17) over i, up to $n-1$, we obtain

$$\sum_{i=1}^{n-1} u_{i+1}u_{i+1}^T - \sum_{i=1}^{n-1} \Phi_i u_i u_i^T \Phi_i^T = -v_1 v_1^T \text{ since } v_n = 0,$$

which is equivalent to

$$\sum_{i=1}^{n} u_i u_i^T - \sum_{i=1}^{n} \Phi_i u_i u_i^T \Phi_i^T = u_1 u_1^T - v_1 v_1^T \text{ since } \Phi_n u_n = 0.$$

Using the above expressions for u_i and $\Phi_i u_i$ in terms of l_i we obtain

$$\sum_{i=1}^{n} \left[\frac{(I - f_i F) l_i l_i^T (I - f_i F)^T}{1 - |f_i|^2} - \frac{(F - f_i I) l_i l_i^T (F - f_i I)^T}{1 - |f_i|^2} \right] = u_1 u_1^T - v_1 v_1^T.$$

Expanding and simplifying the ith term of the sum on the left-hand side we get

$$\sum_{i=1}^{n} \frac{(1 - |f_i|^2) l_i l_i^T - (1 - |f_i|^2) F l_i l_i^T F^T}{1 - |f_i|^2} = u_1 u_1^T - v_1 v_1^T.$$

Therefore,

$$\sum_{i=1}^{n} l_i l_i^T - F \left(\sum_{i=1}^{n} l_i l_i^T \right) F^T = u_1 u_1^T - v_1 v_1^T = GJG^T.$$

This shows that $\sum_{i=1}^{n} l_i l_i^T$ satisfies the displacement equation (10). Hence, by uniqueness,

$$R = \sum_{i=1}^{n} l_i l_i^T. \qquad \square$$

4. Limits to numerical accuracy. Given a symmetric positive-definite matrix R (not necessarily structured), if its Cholesky factor is evaluated by any standard backward stable method that operates on the entries of R, e.g., Gaussian elimination [8, Chap. 4], the corresponding error bound is given by

$$\|R - \hat{L}\hat{L}^T\| \leq c_1 \epsilon \|R\|,$$

where ϵ is the machine precision and c_1 is a low-order polynomial in n, the matrix size.

A fundamental question that needs to be answered then is the following: Given (F, G, J), but not R, how accurately can we expect to be able to compute the Cholesky factorization of R *irrespective* of the algorithm used (*slow or fast*)?

To address this issue we note that just representing (F, G) in finite precision already induces round-off errors. This fact in turn imposes limits on how accurate an algorithm that employs (F, G) can be. We demonstrate this point by the following example.

Let F be a stable diagonal matrix with distinct entries $\{f_i\}$ and assume f_1 is the largest in magnitude. Let the entries of the column vectors u_1 and v_1 be given by

$$u_{i1} = \left(\frac{1}{2}\right)^{i-1}, \quad v_{i1} = \gamma f_i u_{i1}, \quad i \geq 1,$$

where γ is chosen such that $0 < \gamma < 1$.

The unique matrix R that solves (10) for the given (F, u_1, v_1) is symmetric positive definite. This can be verified by invoking Theorem A.1 and by noting that

$$v_{i1} = u_{i1}\, s(f_i),$$

where $s(z)$ is the Schur function (i.e., analytic and strictly bounded by one in $|z| < 1$)

$$s(z) = \gamma z.$$

For this example, we have

$$\frac{|u_{11}|^2}{\|u_1\|^2} \geq \frac{1}{1/(1-\frac{1}{4})} = \frac{3}{4}. \tag{18}$$

Now define the perturbed vectors $\hat{u}_1$ and $\hat{v}_1$ with

$$\hat{u}_{11} = u_{11}(1+\delta), \quad \hat{u}_{i1} = u_{i1}, \quad i \geq 2, \quad \hat{v}_1 = v_1.$$

That is, we make only a relative perturbation in the first entry of u_1 and keep all other entries of u_1 and v_1 unchanged. Here, δ is a small number (for example, for round-off errors, $|\delta|$ is smaller than machine precision).

Let $\hat{R}$ be the unique solution of the displacement equation with the perturbed generator matrix,

$$\hat{R} - F\hat{R}F^T = \hat{G}J\hat{G}^T, \quad \hat{G} = \begin{bmatrix} \hat{u}_1 & \hat{v}_1 \end{bmatrix},$$

and introduce the error matrix $E = R - \hat{R}$. Then E is the unique solution of

$$E - FEF^T = GJG^T - \hat{G}J\hat{G}^T = u_1u_1^T - \hat{u}_1\hat{u}_1^T,$$

from which we find

$$|E_{11}| = \left| \frac{u_{11}^2(2\delta + \delta^2)}{1 - f_1^2} \right|.$$

Therefore,

$$|E_{11}| = \frac{u_{11}^2}{\|u_1\|^2} \, \frac{(2|\delta| + \delta^2)}{1 - f_1^2} \, \|u_1\|^2. \tag{19}$$

But since F is diagonal and $|f_1| < 1$ is its largest entry, we have

$$(1 - f_1^2)^{-1} = \|(I - F \otimes F)^{-1}\|, \tag{20}$$

where $\otimes$ denotes the Kronecker product of two matrices.

Using (18), we conclude that

$$|E_{11}| \geq \frac{3}{4} 2|\delta| \, \|(I - F \otimes F)^{-1}\| \, \|u_1\|^2,$$

from which we get a *lower bound* on the norm of the error matrix

$$\|E\| \geq |E_{11}| \geq \frac{3}{2}|\delta| \, \|(I - F \otimes F)^{-1}\| \, \|u_1\|^2.$$

We now show that by suitably choosing γ, the norm of R can be much smaller than the above bound. Indeed,

$$\begin{aligned}\|R\| &\le n \max_i R_{ii} \\ &= n \max_i \frac{u_{ii}^2 - v_{ii}^2}{1 - f_i^2} \\ &= n \max_i \left[\frac{1 - \gamma^2 f_i^2}{1 - f_i^2} \; u_{ii}^2 \right],\end{aligned}$$

from which we see that as $\gamma \to 1$, the norm of R can be bounded by $n\|u_1\|^2$. Therefore, in this example, $\|R\|$ can be made much smaller than $\|(I - F \otimes F)^{-1}\| \; \|u_1\|^2$ by choosing f_1 and γ close to one.

In summary, we have shown that, at the same time,

- $\|R - \hat{R}\|$ can be larger than $|\delta| \; \|(I - F \otimes F)^{-1}\| \; \|u_1\|^2$ and
- $\|R\|$ can be much smaller than $\|(I - F \otimes F)^{-1}\| \; \|u_1\|^2$.

Hence, in general, we cannot expect the error norm, $\|R - \hat{L}\hat{L}^T\|$, for any algorithm (slow or fast) that uses (F, G, J) as input data (but not R), to be as small as $c_1|\delta|\|R\|$ for some constant c_1.

Therefore, we conclude that *irrespective* of the algorithm we use (*slow or fast*), if the input data is (F, G, J), for a general lower-triangular F, we cannot expect a better bound than

$$\|R - \hat{L}\hat{L}^T\| \le c_2 \epsilon \; \|(I - F \otimes F)^{-1}\| \; \|u_1\|^2. \tag{21}$$

5. Hyperbolic rotation. Each step (12) of the generalized Schur algorithm requires the application of a hyperbolic rotation Θ_i. The purpose of the rotation is to rotate the $(i+1)$th row of the prearray $\begin{bmatrix} \Phi_i u_i & v_i \end{bmatrix}$ to proper form (recall that the top i rows of $\begin{bmatrix} \Phi_i u_i & v_i \end{bmatrix}$ are zero by construction). If we denote the top nonzero row of the prearray by

$$\begin{bmatrix} (\Phi_i u_i)_{i+1} & (v_i)_{i+1} \end{bmatrix} = \begin{bmatrix} \alpha_i & \beta_i \end{bmatrix},$$

then the expression for a hyperbolic rotation that transforms it to the form

$$\begin{bmatrix} \sqrt{|\alpha_i|^2 - |\beta_i|^2} & 0 \end{bmatrix}$$

is given by

$$\Theta_i = \frac{1}{\sqrt{1 - \rho_i^2}} \begin{bmatrix} 1 & -\rho_i \\ -\rho_i & 1 \end{bmatrix}, \quad \text{where } \rho_i = \frac{\beta_i}{\alpha_i}. \tag{22}$$

The positive-definiteness of R guarantees $|\rho_i| < 1$.

For notational convenience, we rewrite equation (12) in the compact form

$$G_{i+1} = \bar{G}_{i+1} \Theta_i, \tag{23}$$

where we have denoted the prearray $\begin{bmatrix} \Phi_i u_i & v_i \end{bmatrix}$ by $\bar{G}_{i+1}$. Note that both G_{i+1} and $\bar{G}_{i+1}$ can be regarded as generator matrices for the $(i+1)$th Schur complement.

Expression (23) shows that in infinite precision, the generator matrices G_{i+1} and $\bar{G}_{i+1}$ must satisfy the fundamental requirement

$$G_{i+1} J G_{i+1}^T = \bar{G}_{i+1} J \bar{G}_{i+1}^T. \tag{24}$$

Obviously, this condition cannot be guaranteed in finite precision. But it turns out that with an appropriate implementation of transformation (23), equality (24) can be guaranteed to within a "small" error. (The need to enforce the condition in finite precision was first observed for the $F = Z$ case by Bojanczyk et al. [3].) To see how, we consider the case when $\bar{G}_{i+1}$ is available exactly in the following subsections.

5.1. Direct implementation. A naive implementation of the hyperbolic transformation (23) can lead to large errors. Indeed, in finite precision, if we apply Θ_i directly to $\bar{G}_{i+1}$ we obtain a computed matrix $\hat{G}_{i+1}$ such that

$$\hat{G}_{i+1} = \bar{G}_{i+1}\Theta_i \; + \; E_{i+1},$$

where the norm of the error matrix E_{i+1} satisfies [8, p. 66]

$$\|E_{i+1}\| \; \leq \; c_3\epsilon \, \|\bar{G}_{i+1}\| \, \|\Theta_i\|.$$

The constant c_3 is a low-order polynomial in the size of the matrices and ϵ is the machine precision. Consequently,

$$\hat{G}_{i+1}J\hat{G}_{i+1}^T = \bar{G}_{i+1}J\bar{G}_{i+1}^T + E_{i+1}JE_{i+1}^T + \bar{G}_{i+1}\Theta_i E_{i+1}^T + E_{i+1}\Theta_i\bar{G}_{i+1}^T,$$

which shows that

$$\|\hat{G}_{i+1}J\hat{G}_{i+1}^T - \bar{G}_{i+1}J\bar{G}_{i+1}^T\| \leq c_4\epsilon \, \|\bar{G}_{i+1}\|^2 \, \|\Theta_i\|^2. \tag{25}$$

But since $\|\Theta_i\|$ can be large, the computed quantities are not guaranteed to satisfy relation (24) to sufficient accuracy.[1] This possibly explains the disrepute to which fast algorithms have fallen.

5.2. Mixed downdating. One possible way to ameliorate the above problem is to employ the mixed-downdating procedure as suggested by Bojanczyk et al. [3, 4]. This scheme guarantees that

$$\|\hat{G}_{i+1}J\hat{G}_{i+1}^T - \bar{G}_{i+1}J\bar{G}_{i+1}^T\| \leq c_5\epsilon \, \left(\|\bar{G}_{i+1}\|^2 + \|\hat{G}_{i+1}\|^2\right).$$

This bound is sufficient, when combined with other modifications suggested in §§6 and 8.4, to make the algorithm numerically reliable (§7).

5.3. A new method: The OD procedure. An alternate scheme is now proposed which is based on using the SVD of the hyperbolic rotation Θ_i (it is a modification of a scheme in [6]). Its good numerical properties come from the fact that the hyperbolic rotation is applied as a sequence of orthogonal and diagonal matrices, which we shall refer to as the OD (orthogonal–diagonal) procedure. Its other advantage is that it is a general technique that can be applied in other situations, such as the Schur algorithms for nonsymmetric matrices. It can be implemented with the same operation count as the mixed-downdating algorithm of [3].

[1] Interestingly, though, Bojanczyk et al. [3] showed that for the special case $F = Z$ and displacement rank $r = 2$, the direct implementation of the hyperbolic rotation still leads to an asymptotically backward stable algorithm. This conclusion, however, does not hold for higher displacement ranks. Stewart and Van Dooren [21] showed that for $F = Z$ and $r > 2$, the direct implementation of the hyperbolic rotation can be unstable.

It is straightforward to verify that any hyperbolic rotation of form (22) admits the following eigen(svd-)decomposition:

$$\Theta_i = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{\frac{\alpha_i+\beta_i}{\alpha_i-\beta_i}} & 0 \\ 0 & \sqrt{\frac{\alpha_i-\beta_i}{\alpha_i+\beta_i}} \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \frac{1}{\sqrt{2}} = Q_i D_i Q_i^T, \tag{26}$$

where the matrix

$$Q_i = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

is orthogonal ($Q_i Q_i^T = I$).

If the eigendecomposition $Q_i D_i Q_i^T$ is now applied to the prearray $\bar{G}_{i+1}$ in (23), then it can be shown that the computed generator matrix $\hat{G}_{i+1}$ satisfies (see Appendix C)

$$(\hat{G}_{i+1} + E_{2,i+1}) = (\bar{G}_{i+1} + E_{1,i+1})\Theta_i, \tag{27}$$

with

$$\|E_{1,i+1}\| \le c_6 \epsilon \, \|\bar{G}_{i+1}\|, \quad \|E_{2,i+1}\| \le c_7 \epsilon \, \|\hat{G}_{i+1}\|.$$

It further follows from (27) that $\hat{G}_{i+1}$ satisfies

$$(\hat{G}_{i+1} + E_{2,i+1}) J (\hat{G}_{i+1} + E_{2,i+1})^T = (\bar{G}_{i+1} + E_{1,i+1}) J (\bar{G}_{i+1} + E_{1,i+1})^T, \tag{28}$$

which shows that

$$\|\hat{G}_{i+1} J \hat{G}_{i+1}^T - \bar{G}_{i+1} J \bar{G}_{i+1}^T\| \le c_8 \epsilon \left(\|\bar{G}_{i+1}\|^2 + \|\hat{G}_{i+1}\|^2 \right). \tag{29}$$

ALGORITHM 5.1 (the OD procedure). *Given a hyperbolic rotation Θ with reflection coefficient $\rho = \beta/\alpha$, $|\rho| < 1$, and a prearray row vector $\begin{bmatrix} x & y \end{bmatrix}$, the postarray row vector $\begin{bmatrix} x_1 & y_1 \end{bmatrix}$ can be computed as follows:*

$$\begin{bmatrix} x' & y' \end{bmatrix} \leftarrow \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

$$\begin{bmatrix} x'' & y'' \end{bmatrix} \leftarrow \begin{bmatrix} x' & y' \end{bmatrix} \begin{bmatrix} \frac{1}{2}\sqrt{\frac{\alpha+\beta}{\alpha-\beta}} & 0 \\ 0 & \frac{1}{2}\sqrt{\frac{\alpha-\beta}{\alpha+\beta}} \end{bmatrix}$$

$$\begin{bmatrix} x_1 & y_1 \end{bmatrix} \leftarrow \begin{bmatrix} x'' & y'' \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

The algorithm guarantees (cf. (27)–(29)) the following error bounds:

$$\begin{bmatrix} \hat{x}_1 + e_1 & \hat{y}_1 + e_2 \end{bmatrix} = \begin{bmatrix} x + e_3 & y + e_4 \end{bmatrix} \Theta,$$

with

$$\left\| \begin{bmatrix} e_1 & e_2 \end{bmatrix} \right\| \le c_9 \epsilon \left\| \begin{bmatrix} \hat{x}_1 & \hat{y}_1 \end{bmatrix} \right\|, \quad \left\| \begin{bmatrix} e_3 & e_4 \end{bmatrix} \right\| \le c_{10} \epsilon \left\| \begin{bmatrix} x & y \end{bmatrix} \right\|. \tag{30}$$

5.4. Another new method: The H procedure. Let $\rho = \beta/\alpha$ be the reflection coefficient of a hyperbolic rotation Θ,

$$\Theta = \frac{1}{\sqrt{1-\rho^2}} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix},$$

with $|\rho| < 1$. Let $\begin{bmatrix} x_1 & y_1 \end{bmatrix}$ and $\begin{bmatrix} x & y \end{bmatrix}$ be the postarray and prearray rows, respectively,

$$\begin{bmatrix} x_1 & y_1 \end{bmatrix} = \begin{bmatrix} x & y \end{bmatrix} \Theta, \text{ with } |x| > |y|.$$

The advantage of the method to be described in this section is that the computed quantities $\hat{x}_1$ and $\hat{y}_1$ satisfy the equation

$$\begin{bmatrix} \hat{x}_1 + e_1' & \hat{y}_1 + e_2' \end{bmatrix} = \begin{bmatrix} x & y \end{bmatrix} \Theta, \tag{31}$$

where the error terms satisfy

$$|e_1'| \leq c_{11}\epsilon|\hat{x}_1|, \quad |e_2'| \leq c_{12}\epsilon(|\hat{x}_1| + |\hat{y}_1|). \tag{32}$$

Compare this with (27), where the prearray is also perturbed. Moreover, we shall show in §10.2 that by a slight modification we can further enforce that $|\hat{x}_1| > |\hat{y}_1|$, which is needed to prevent breakdown in the algorithm. (If $|x| < |y|$, then it can be seen that $\begin{bmatrix} y & x \end{bmatrix} \Theta = \begin{bmatrix} y_1 & x_1 \end{bmatrix}$. Therefore, without loss of generality, we shall only consider the case $|x| > |y|$.)

The expression for x_1 can be written in the form

$$x_1 = \frac{|\alpha| x}{\sqrt{(\alpha-\beta)(\alpha+\beta)}} \left[1 - \frac{\beta}{\alpha}\frac{y}{x}\right].$$

The term $\xi = 1 - \frac{\beta}{\alpha}\frac{y}{x}$ can be evaluated to high relative accuracy as follows:

If $\frac{\beta}{\alpha}\frac{y}{x} < 1/2$
  then $\xi = 1 - \frac{\beta}{\alpha}\frac{y}{x}$
else
  $d_1 = \frac{|\alpha|-|\beta|}{|\alpha|}$, $d_2 = \frac{|x|-|y|}{|x|}$
  $\xi = d_1 + d_2 - d_1 d_2$

The argument employed in §6.1 establishes that

$$\hat{\xi} = \xi(1 + 33\delta_1),$$

where δ_1 denotes a quantity that is smaller than the machine precision in magnitude. Therefore, x_1 can be computed to high relative accuracy from the expression

$$x_1 = \frac{|\alpha| x \xi}{\sqrt{(\alpha-\beta)(\alpha+\beta)}};$$

i.e.,

$$\hat{x}_1 = x_1(1 + c_{13}\delta_2)$$

for some constant c_{13}.

To compute y_1 we use the expression

$$y_1 = x_1 - \sqrt{\frac{\alpha+\beta}{\alpha-\beta}}\,(x-y).$$

Then the computed y_1 satisfies

$$\hat{y}_1 = \left(x_1(1+c_{13}\delta_2) - \sqrt{\frac{\alpha+\beta}{\alpha-\beta}}\,(x-y)(1+c_{14}\delta_3) \right)(1+\delta_4),$$

from which we get

$$\hat{y}_1 = y_1 + x_1 c_{15}\delta_5 + \hat{y}_1 c_{16}\delta_6.$$

Therefore,

$$|\hat{y}_1 - y_1| \le c_{17}\epsilon\,[|\hat{x}_1| + |\hat{y}_1|]\,.$$

In summary, the H procedure is the following.

ALGORITHM 5.2 (the H procedure). *Given a hyperbolic rotation Θ with reflection coefficient $\rho = \beta/\alpha$, $|\rho| < 1$, and a prearray $\begin{bmatrix} x & y \end{bmatrix}$ with $|x| > |y|$, the postarray $\begin{bmatrix} x_1 & y_1 \end{bmatrix}$ can be computed as follows:*

If $\frac{\beta}{\alpha}\frac{y}{x} < 1/2$
  then $\xi \leftarrow 1 - \frac{\beta}{\alpha}\frac{y}{x}$
else
  $d_1 \leftarrow \frac{|\alpha|-|\beta|}{|\alpha|}$, $d_2 \leftarrow \frac{|x|-|y|}{|x|}$
  $\xi \leftarrow d_1 + d_2 - d_1 d_2$
endif
$x_1 \leftarrow \frac{|\alpha| x \xi}{\sqrt{(\alpha-\beta)(\alpha+\beta)}}$
$y_1 \leftarrow x_1 - \sqrt{\frac{\alpha+\beta}{\alpha-\beta}}\,(x-y).$

This algorithm guarantees (31) and (32). We remark that the H procedure requires $5n$ to $7n$ multiplications and $3n$ to $5n$ additions. It is therefore costlier than the OD procedure, which requires $2n$ multiplications and $4n$ additions. But the H procedure is forward stable (cf. (31)), whereas the OD method is only stable (cf. (27)).

From now on we shall denote by $\hat{u}_{i+1}$ and $\hat{v}_{i+1}$ the computed generator columns at step i; i.e.,

$$\hat{G}_{i+1} = \begin{bmatrix} \hat{u}_{i+1} & \hat{v}_{i+1} \end{bmatrix},$$

starting with

$$\bar{G}_i = \begin{bmatrix} \Phi_i \hat{u}_i & \hat{v}_i \end{bmatrix}.$$

6. Blaschke matrix. Each step of the algorithm also requires multiplying the Blaschke matrix Φ_i by u_i. (Note that the top i rows of $\Phi_i u_i$ are zero and, hence, can be ignored in the computation.) In this section, we consider the following two cases.

- F is stable and diagonal, in which case Φ_i itself is diagonal and given by

$$(\Phi_i)_{jj} = \frac{f_j - f_i}{1 - f_i f_j}.$$

- F is strictly lower triangular; e.g., $F = Z$, $F = (Z \oplus Z)$, or other more involved choices. In these situations, the matrix Φ_i is equal to F since the f_i are all zero,

$$\Phi_i = F.$$

6.1. The case of diagonal F. The goal of this section is to show how to compute $\Phi_i \hat{u}_i$ to high componentwise relative accuracy (i.e., high relative accuracy for each component of the computed vector). Here, $\hat{u}_i$ denotes the computed value of u_i.

The numerator of $(\Phi_i)_{jj}$ can be computed to high relative accuracy as

$$fl(f_j - f_i) = (f_j - f_i)(1 + \delta_1).$$

Computing the denominator $x_{ij} = (1 - f_i f_j)$ to high relative accuracy is a bit trickier, as the following example shows.

Let $f_1 = f_2 = 0.998842$. Then in 6-digit arithmetic $1 - f_1 f_2 \approx 2.31500 \times 10^{-3}$, whereas the actual answer is $2.31465903600 \times 10^{-3}$. Therefore, the relative error is approximately 1.5×10^{-4}. Using the scheme given below, we find $1 - f_1 f_2 \approx 2.31466 \times 10^{-3}$. The relative error is now approximately 4.2×10^{-7}.

The scheme we use to compute x_{ij} is as follows:

If $f_i f_j < 1/2$
  then $x_{ij} = 1 - f_i f_j$
else
  $d_j = 1 - |f_j|$, $d_i = 1 - |f_i|$
  $x_{ij} = d_i + d_j - d_i d_j$

We now show that this scheme ensures that x_{ij} is computed to high relative accuracy. Indeed, when $f_i f_j < 1/2$, we have $|1 - f_i f_j| > 1/2$. Moreover,

$$\begin{aligned} \hat{x}_{ij} &= (1 - f_i f_j (1 + \delta_2))(1 + \delta_3) \\ &= x_{ij} \left(1 - \frac{f_i f_j}{x_{ij}} \delta_2\right)(1 + \delta_3). \end{aligned}$$

Since $|f_i f_j / x_{ij}| < 1$, we have $\hat{x}_{ij} = x_{ij}(1 + 3\delta_4)$. On the other hand, for $f_i f_j \geq 1/2$, we note that

$$\hat{d}_j = d_j(1 + \delta_5), \quad \hat{d}_i = d_i(1 + \delta_6),$$

and

$$\hat{x}_{ij} = \left[(\hat{d}_i + \hat{d}_j)(1 + \delta_7) - \hat{d}_i \hat{d}_j (1 + \delta_8)\right](1 + \delta_9),$$

which, when simplified, gives

$$\hat{x}_{ij} = x_{ij} \left(1 + 11 \frac{(d_i + d_j + d_i d_j)}{x_{ij}} \delta_{10}\right).$$

We shall now show that $(d_i + d_j + d_i d_j)/x_{ij} < 3$. Indeed, first note that d_i and d_j are positive numbers smaller than $1/2$ since $|f_i| > 1/(2|f_j|) > 1/2$. It then follows from

$$d_i < \frac{1}{2} < \frac{1}{2} + \frac{d_i}{2d_j}$$

that $d_i + d_j > 2d_i d_j$. Therefore,

$$\frac{d_i + d_j + d_i d_j}{d_i + d_j - d_i d_j} < \frac{\frac{3}{2}(d_i + d_j)}{\frac{1}{2}(d_i + d_j)} = 3,$$

which shows that

$$\hat{x}_{ij} = x_{ij}\,(1 + 33\delta_{11})\,.$$

In summary, we have shown how to compute Φ_i to componentwise accuracy. Therefore, since Φ_i is diagonal, $\Phi_i \hat{u}_i$ can be computed to componentwise high relative accuracy. More specifically,

$$fl\,(\Phi_i \hat{u}_i)_j = (\Phi_i \hat{u}_i)_j (1 + 72\delta_{12}). \tag{33}$$

We should remark that if the denominator entries $(1 - f_i f_j)$ were instead computed directly, the error in computing $(\Phi_i \hat{u}_i)_j$ would also depend on the norm of $(I - f_i F)^{-1}$, which can be large. For this reason, we have introduced the above computational scheme for evaluating $(1 - f_i f_j)$. This scheme, however, is not totally successful when F is a general triangular matrix (for example, when F is bidiagonal). A way around this difficulty will be addressed elsewhere. But for a strictly lower-triangular F, the situation is far simpler, as shown in the next subsection.

But for now, let us consider how to compute the l_i's. Define

$$\bar{l}_i = \sqrt{1 - f_i^2}\;(I - f_i F)^{-1} \hat{u}_i. \tag{34}$$

We use the expression

$$(\bar{l}_i)_j = \frac{\sqrt{(1 - f_i)(1 + f_i)}}{1 - f_i f_j} (\hat{u}_i)_j$$

to compute $\bar{l}_i$ with the technique explained above for the denominator $(1 - f_i f_j)$. Then we can show that

$$(\hat{l}_i)_j = (\bar{l}_i)_j (1 + c_{18}\delta_{13}). \tag{35}$$

6.2. The case of strictly lower-triangular F. For a strictly lower-triangular F, we use the standard matrix–vector multiplication. In this case, the computed quantities satisfy the relation [8, p. 66]

$$\|fl(F\hat{u}_i) - F\hat{u}_i\| \leq c_{19}\epsilon \|F\|\,\|\hat{u}_i\|.$$

Also, since $f_i = 0$,

$$\bar{l}_i = \hat{u}_i = \hat{l}_i. \tag{36}$$

6.3. Enforcing positive-definiteness. Condition (44) on the matrix R in §7.1 guarantees the positive-definiteness of the successive Schur complements and, therefore, that $|\Phi_i \hat{u}_i|_{i+1} > |\hat{v}_i|_{i+1}$. This assures that the reflection coefficients will be smaller than one in magnitude, a condition that we now enforce in finite precision as follows:

If $|fl(\Phi_i \hat{u}_i)|_{i+1} < |\hat{v}_{i+1,i}|$ then
$fl(\Phi_i \hat{u}_i)_{i+1} \leftarrow |\hat{v}_{i+1,i}|(1+3\epsilon)\text{sign}(fl(\Phi_i \hat{u}_i)_{i+1})$.

This enhancement, along with condition (44), will be shown in §7.1 to guarantee that the algorithm will complete without any breakdowns.

7. Error analysis of the algorithm. The argument in this section is motivated by the analysis in Bojanczyk et al. [3].

Note that for diagonal and for strictly lower-triangular F we can write

$$\|fl(\Phi_i \hat{u}_i) - \Phi_i \hat{u}_i\| \leq c_{20}\epsilon \|\Phi_i\| \|\hat{u}_i\|.$$

Therefore, from the error analysis (27) of the hyperbolic rotation we obtain

$$(\hat{G}_{i+1} + E_{2,i+1}) = (\begin{bmatrix} \Phi_i \hat{u}_i & \hat{v}_i \end{bmatrix} + E_{3,i+1})\Theta_i, \tag{37}$$

where

$$\|E_{3,i+1}\| \leq c_{21}\epsilon(\|\Phi_i\|\|\hat{u}_i\| + \|\hat{v}_i\|).$$

It then follows that

$$\hat{u}_{i+1}\hat{u}_{i+1}^T - \hat{v}_{i+1}\hat{v}_{i+1}^T = \Phi_i \hat{u}_i \hat{u}_i^T \Phi_i^T - \hat{v}_i \hat{v}_i^T - M_{i+1}, \tag{38}$$

where

$$\|M_{i+1}\| \leq c_{22}\epsilon(\|\hat{u}_{i+1}\|^2 + \|\Phi_i\|^2\|\hat{u}_i\|^2 + \|\hat{v}_{i+1}\|^2 + \|\hat{v}_i\|^2). \tag{39}$$

Since

$$\bar{l}_i = \sqrt{1-f_i^2}\,(I - f_i F)^{-1}\hat{u}_i,$$

the following two equations hold:

$$\hat{u}_i = \frac{1}{\sqrt{1-f_i^2}}(I - f_i F)\bar{l}_i, \quad \Phi_i \hat{u}_i = \frac{1}{\sqrt{1-f_i^2}}(F - f_i I)\bar{l}_i.$$

Hence, following the proof of Theorem 3.2, we can establish that

$$\sum_{i=1}^{n} \bar{l}_i \bar{l}_i^T - F\left(\sum_{i=1}^{n} \bar{l}_i \bar{l}_i^T\right) F^T = \hat{u}_1 \hat{u}_1^T - \hat{v}_1 \hat{v}_1^T - \sum_{i=1}^{n} M_i. \tag{40}$$

Define

$$\bar{R} = \sum_{i=1}^{n} \bar{l}_i \bar{l}_i^T$$

and

$$\bar{E} = R - \bar{R}.$$

Then note that $\bar{E}$ satisfies

$$\bar{E} - F\bar{E}F^T = \sum_{i=1}^{n} M_i \tag{41}$$

since, we assume, $\hat{u}_1 = u_1$ and $\hat{v}_1 = v_1$.

Now if

$$E = R - \sum_{i=1}^{n} \hat{l}_i \hat{l}_i^T = (R - \bar{R}) + \left(\bar{R} - \sum_{i=1}^{n} \hat{l}_i \hat{l}_i^T \right),$$

then

$$\|E\| \le \|\bar{E}\| + \left\| \sum_{i=1}^{n} \bar{l}_i \bar{l}_i^T - \sum_{i=1}^{n} \hat{l}_i \hat{l}_i^T \right\|.$$

Using (35) and (36) we can establish that, for diagonal or strictly lower-triangular F,

$$\left\| \sum_{i=1}^{n} \bar{l}_i \bar{l}_i^T - \sum_{i=1}^{n} \hat{l}_i \hat{l}_i^T \right\| \le c_{23}\epsilon \|\bar{R}\|.$$

Therefore,

$$\begin{aligned} \|E\| &\le c_{24}\epsilon \|\bar{R}\| + \|(I - F \otimes F)^{-1}\| \sum_{i=1}^{n} \|M_i\| \\ &\le c_{25}\epsilon \left[\|\bar{R}\| + \|(I - F \otimes F)^{-1}\| \sum_{i=1}^{n} \left\{ (1 + \|\Phi_i\|^2) \, \|\hat{u}_i\|^2 + \|\hat{v}_i\|^2 \right\} \right]. \end{aligned} \tag{42}$$

7.1. Avoiding breakdown. The above error analysis assumes that the algorithm does not break down. That is, at every iteration the reflection coefficients of the hyperbolic rotations are assumed to be strictly less than one in magnitude. In this section we show that this can be guaranteed by imposing condition (44) on R *and* by employing the enhancement suggested in §6.3.

The argument is inductive. We know that $|\Phi_1 u_1|_2 > |v_1|_2$. Now assume that the algorithm has successfully gone through the first i steps and define the $\bar{l}_i$ as in (34). Also define the matrix S_i that solves the displacement equation

$$S_i - FS_iF^T = \hat{u}_i \hat{u}_i^T - \hat{v}_i \hat{v}_i^T, \tag{43}$$

as well as the matrix

$$\bar{R}_i = \sum_{j=1}^{i-1} \bar{l}_j \bar{l}_j^T + S_i.$$

Following the proof of Theorem 3.2, we can establish that

$$\bar{R}_i - F\bar{R}_iF^T = \hat{u}_1 \hat{u}_1^T - \hat{v}_1 \hat{v}_1^T - \sum_{j=1}^{i-1} M_j.$$

If we further introduce the error matrix

$$\bar{E}_i = R - \bar{R}_i,$$

we then note that it satisfies

$$\bar{E}_i - F\bar{E}_i F^T = \sum_{j=1}^{i-1} M_j$$

since, we assume, $\hat{u}_1 = u_1$ and $\hat{v}_1 = v_1$.

Then, as before, we can establish that

$$\begin{aligned}\|R - \bar{R}_i\| &\leq c_{26}\epsilon\|\bar{R}_i\| + \|(I - F \otimes F)^{-1}\| \sum_{j=1}^{i-1} \|M_j\| \\ &\leq c_{27}\epsilon \left[\|\bar{R}_i\| + \|(I - F \otimes F)^{-1}\| \sum_{j=1}^{i-1} \left\{ (1 + \|\Phi_j\|^2)\, \|\hat{u}_j\|^2 \;+\; \|\bar{R}_i\|(1 + \|F\|^2) \right\} \right],\end{aligned}$$

where in the second step we use Lemma B.1. Now note that for diagonal and stable F, $\|\Phi_i\| \leq 1$, and if F is strictly lower triangular then $\|\Phi_i\| \leq \|F\|$. Therefore, combining both cases, we get $1 + \|\Phi_i\|^2 \leq 2 + \|F\|^2$, which leads to

$$\|R - \bar{R}_i\| \leq c_{28}\epsilon\|(I - F \otimes F)^{-1}\|\, (2 + \|F\|^2) \left[\|\bar{R}_i\| + \sum_{j=1}^{i-1} \|\hat{u}_j\|^2 \right].$$

It now follows that if the minimum eigenvalue of R meets the lower bound

$$\lambda_{\min}(R) \;>\; c_{28}\epsilon\|(I - F \otimes F)^{-1}\|\, (2 + \|F\|^2) \left[\|\bar{R}_i\| + \sum_{j=1}^{i-1} \|\hat{u}_j\|^2 \right],$$

then $\bar{R}_i$ will be guaranteed to be positive definite and, consequently, S_i will be positive definite. Then, by positive-definiteness,

$$|\Phi_i \hat{u}_i|_{i+1} \;>\; |\hat{v}_i|_{i+1}.$$

But since we enforce $fl(|\Phi_i \hat{u}_i|_{i+1}) > |\hat{v}_i|_{i+1}$, the algorithm can continue to the next iteration.

This suggests the following lower bound on the smallest eigenvalue of R in order to avoid breakdown of the algorithm (i.e., in order to ensure the positive-definiteness of the computed Schur complements):

$$\lambda_{\min}(R) \;>\; c_{28}\epsilon\|(I - F \otimes F)^{-1}\|\, (2 + \|F\|^2) \left[\|\bar{R}\| + \sum_{j=1}^{n} \|\hat{u}_j\|^2 \right]. \tag{44}$$

7.2. Error bound. From the discussion in the previous section, we can conclude the following.

THEOREM 7.1 (error bound). *The generalized Schur algorithm for a structured positive-definite matrix R satisfying* (44), *with a stable diagonal or strictly lower-triangular F, implemented as detailed in this paper (see listing in Appendix* D*), guarantees the following error bound:*

$$\|E\| \leq c_{29}\epsilon\|(I - F \otimes F)^{-1}\| \, (2 + \|F\|^2) \left[\|\bar{R}\| + \sum_{j=1}^{n} \|\hat{u}_j\|^2 \right]. \tag{45}$$

The term $\|(I - F \otimes F)^{-1}\|$ in the error bound is expected from the perturbation analysis of §4. However, the presence of the norms of the successive generators makes the error bound larger than the bound suggested by the perturbation analysis, which depends only on the norm of the first generator matrix.

7.3. Growth of generators. The natural question then is as follows: How big can the norm of the generators be? An analysis based on the norm of the hyperbolic rotations used in the algorithm gives the following bound:

$$\|\hat{u}_i\| \leq c_{30}\|u_1\| \prod_{k=1}^{i} \sqrt{\frac{1 + |\rho_k|}{1 - |\rho_k|}}, \tag{46}$$

which is reminiscent of the error bounds of Cybenko for the Levinson algorithm [7]. A tighter bound can be obtained by using the fact that R is positive definite to get (recall (13))

$$\|u_i\|^2 \leq \|(I - F \otimes F)^{-1}\| \, (1 + \|F\|^2)^2 \, \|R\|. \tag{47}$$

This shows that the growth of the generators depends on $\|(I - F \otimes F)^{-1}\|$. But for a strictly lower-triangular F a better bound is $\|u_i\|^2 \leq \|R\|$. Therefore, for strictly lower-triangular F the error bound is as good as can be expected from the perturbation analysis of §4.

What does this say about the stability of the generalized Schur algorithm for a diagonal and stable F? Clearly, when the eigenvalues of F are sufficiently far from 1 the method has excellent numerical stability. The algorithm degrades as the eigenvalues of F get closer to 1. This is to be expected from the perturbation analysis (whether we use a slow or a fast algorithm). However, if the generators grow rapidly (i.e., as fast as (47)) then the algorithm degrades faster than the rate predicted by the perturbation analysis.

Is there anything further we can do to ameliorate this problem? One thing we have not considered yet is *pivoting*, which is possible only when F is diagonal. We discuss this in §8.

7.4. F strictly lower triangular. When F is strictly lower triangular the error bound can be written in the alternate form

$$\|E\| \leq c_{31}\epsilon(2 + \|F\|^2) \left(\sum_{i=1}^{n} \|F^i\|^2 \right) \left[\|\bar{R}\| + \left(\sum_{i=1}^{n} \|\hat{u}_i\|^2 \right) \right].$$

This shows that when F is contractive ($\|F\| \leq 1$), the error bound is as good as can be expected from the perturbation analysis; i.e.,

$$\|E\| \leq c_{32}\epsilon \left[\|\bar{R}\| + \left(\sum_{i=1}^{n} \|\hat{u}_i\|^2 \right) \right].$$

We observed in §7.3 that $\|u_i\|^2 \leq \|R\|$. Therefore, if F is strictly lower triangular and contractive, then the algorithm is backward stable.

This includes the important class of positive-definite quasi-Toeplitz matrices, which correspond to $F = Z$. In this case, we strengthen the result of Bojanczyk et al. [3], which states that for quasi-Toeplitz symmetric positive-definite matrices, the Schur algorithm is asymptotically backward stable. Our analysis shows that the modified algorithm proposed here is backward stable provided the smallest eigenvalue of the quasi-Toeplitz matrix satisfies

$$\lambda_{\min}(R) \; > \; c_{32}\epsilon \left[\|\bar{R}\| + \sum_{j=1}^{n} \|\hat{u}_j\|^2 \right].$$

If F is strictly lower triangular but noncontractive then the error norm can possibly depend on $\|(I - F \otimes F)^{-1}\|$.

7.5. F diagonal. For the special case of a stable diagonal F, the bound in (45) may suggest that the norm of the error can become very large when the magnitude of the diagonal entries of F become close to one. But this is not necessarily the case (see also the numerical example in the next section).

First note that the bound in (39) can be strengthened since the hyperbolic rotations are applied to each row independently. By assumption (44), the Schur complement generated by $(\hat{u}_i, \hat{v}_i)$ is positive definite. Hence, by Theorem A.1,

$$|\hat{u}_i| \; > \; |\Phi_i \hat{u}_i| \; > \; |\hat{v}_i|,$$

and we conclude that

$$|M_{i+1}| \; \leq c_{33}\epsilon(|\hat{u}_{i+1}||\hat{u}_{i+1}|^T + |\hat{u}_i||\hat{u}_i|^T).$$

Define

$$\rho_{j,i} = \frac{\hat{v}_{j,i}}{\hat{u}_{j,i}}.$$

It then follows from (43) that

$$(S_i)_{j,k} = \frac{\hat{u}_{j,i}\hat{u}_{k,i}}{1 - f_j f_k} \left(1 - \rho_{j,i}\rho_{k,i}\right).$$

Therefore,

$$\frac{|\hat{u}_{j,i}\hat{u}_{k,i}|}{1 - f_j f_k} \; \leq \; \frac{|(S_i)_{j,k}|}{1 - \rho_{j,i}\rho_{k,i}} \leq \frac{\|\bar{R}_i\|}{1 - \max_j \left(\rho_{j,i}^2\right)}.$$

Now using expression (41) we obtain

$$|\bar{E}|_{j,k} = \frac{|\sum_{i=1}^{n} M_i|_{j,k}}{1 - f_j f_k} \leq c_{34}\epsilon \frac{\sum_{i=1}^{n} \|\bar{R}_i\|}{1 - \max_{j,i} \left(\rho_{j,i}^2\right)}.$$

This establishes the following alternative bound for the error matrix E:

$$\|E\| \leq c_{35}\epsilon \, \frac{\sum_{i=1}^{n} \|\bar{R}_i\|}{1 - \max_{j,i} \left(\rho_{j,i}^2\right)},$$

which is independent of the $\{f_i\}$. In other words, if the coefficients $\rho_{j,i}$ are sufficiently smaller than one, then the algorithm will be backward stable irrespective of how close the $\{|f_i|\}$ are to one.

8. Pivoting with diagonal F. When F is diagonal it is possible to accommodate pivoting into the algorithm, as suggested by Heinig [10]; it corresponds to reordering the f_j's, $u_{j,i}$'s, and $v_{j,i}$'s identically at the ith iteration of the algorithm. This has the effect of computing the Cholesky factorization of PRP^T, where P is the product of all the permutations that were carried out during the algorithm.

In finite precision, pivoting strategies are employed in classical Cholesky factorization algorithms when the positive-definite matrix is numerically singular. In the context of the generalized Schur algorithm of this paper, the main motivation for pivoting should be to keep the norm of the generator matrices as small as possible! This is suggested by the expression for the error bound in (45), which depends on the norm of the generators. Note that this motivation has little to do with the size of the smallest eigenvalue of the matrix.

We would like to emphasize that pivoting is necessary only when the norm of F is very close to one since otherwise the generators do not grow appreciably (47).

8.1. A numerical example. The first question that arises then is whether there exists a pivoting strategy that guarantees a small growth in the norm of the generators. Unfortunately, we have numerical examples that show that irrespective of what pivoting strategy is employed, the norms of the generators may not exhibit significant reduction.

Consider the matrix R the satisfies the displacement equation $R - FRF^T = GJG^T$ with

$$G = \begin{bmatrix} 0.26782811166721 & 0.26782805810159 \\ 0.65586390188981 & -0.65586311485320 \\ 0.65268528182561 & 0.65268365011256 \\ 0.26853783287812 & -0.26853149538590 \end{bmatrix}, \quad J = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

and

$$F = \text{diagonal}\{0.9999999, -0.9999989, 0.9999976, -0.9999765\}.$$

The matrix R is positive definite since the entries of the column vector v_1 were generated from the relation $v_{i,1} = u_{i,1}s(f_i)$, where $s(z)$ is the Schur function $s(z) = 0.9999999z$. Table 8.1 lists the values of

$$\sum_{i=1}^{n} \|\hat{u}_i\|^2$$

for all 24 possible pivoting options of the rows of the generator matrix G. The results indicate that none of the pivoting options significantly reduces the size of the generators. Indeed, note that the norm of u_1 is approximately one, while the best growth rate we achieve with pivoting is approximately 10^4. This best case is achieved when the diagonal entries of F are approximately in increasing order of magnitude.

8.2. The case of positive F. This raises the next question: Is pivoting useful at all? It *is* useful when the the entries of the F matrix are strictly positive (or negative). In this case, we permute the entries of F (and, correspondingly, the entries of u_1 and v_1) such that the diagonal of F is in increasing order of magnitude. Then it is shown below that

$$\|\hat{u}_i\| \leq \|\bar{R}\|^{1/2},$$

TABLE 8.1

$10^{-6}\sum_{i=1}^{n}\|\hat{u}_i\|^2$	
5.30	0.41
5.30	0.41
5.21	0.40
5.21	0.40
5.03	0.40
5.03	0.40
0.83	0.04
0.83	0.04
0.83	0.04
0.83	0.04
0.83	0.04
0.83	0.04

which makes the first-order term of the upper bound on E depend only on the first power of $\|(I - F \otimes F)^{-1}\|$. Indeed, we know that

$$\hat{u}_i = \frac{1}{\sqrt{1-|f_i|^2}}\,(I - f_i F)\bar{l}_i,$$

where the top $(i-1)$ entries of u_i and l_i are zero. For $j \geq i$,

$$\hat{u}_{j,i} = \frac{1 - f_i f_j}{\sqrt{1-|f_i|^2}}\,\bar{l}_{j,i},$$

and due to the ordering of the entries of F, and since $|f_i| < 1$, we have

$$1 - f_i f_j \;\leq\; 1 - f_i^2 \;\leq\; \sqrt{1-|f_i|^2}.$$

Therefore,

$$\frac{1 - f_i f_j}{\sqrt{1-|f_i|^2}} \leq 1,$$

and we conclude that

$$\|\hat{u}_i\| \leq \|\bar{l}_i\| \;\leq\; \|\bar{R}\|^{1/2},$$

as desired.

8.3. The nonpositive case. When F is not positive, the example in §8.1 suggests that pivoting may not help in general. However, it may still be beneficial to try a heuristic pivoting strategy to control the growth of the generators. Ideally, at the ith iteration we should pick the row of the prearray $\bar{G}_{i+1}$ which would lead to the smallest (in norm) postarray G_{i+1}. Since there seems to be no efficient way to do this we suggest picking the row that leads to the smallest reflection coefficient (in magnitude) for the hyperbolic rotation Θ_i. As suggested by the example of §8.1, an alternate strategy would be to order the f_i's in increasing order of magnitude.

We stress that pivoting is relevant *only* when the norm of F is very close to one, as indicated by the error bound (45) and (47).

8.4. Controlling the generator growth. We have shown in §§7.4 and 8.2 that the generators do not grow (i) if F is strictly lower triangular and contractive or (ii) if F is a positive diagonal matrix with increasing diagonal entries. We now show how to control the generator growth in general using an idea suggested by Gu [9].

It follows from (43) that

$$\|\hat{G}_i J \hat{G}_i^T\| = \|S_i - F S_i F^T\|.$$

Let $W_i \Lambda_i W_i^T$ denote the eigendecomposition of $\hat{G}_i J \hat{G}_i^T$, where Λ_i is a 2×2 real diagonal matrix with $(\Lambda_i)_{11} > 0$ and $(\Lambda_i)_{22} < 0$. Then $W_i \sqrt{|\Lambda_i|}$ can be taken as a generator for S_i with the desirable property that

$$\left\| W_i \sqrt{|\Lambda_i|} \right\|^2 = \|\Lambda_i\| = \|S_i - F S_i F^T\| \leq \|S_i\|(1 + \|F\|^2) \leq \|\bar{R}_i\|(1 + \|F\|^2),$$

where $\|\bar{R}_i\| \approx \|R\|$ to first order in ϵ.

Therefore, whenever the generator grows, i.e., $\|\hat{G}_i\|^2$ becomes larger than a given threshold (say, $2\|R\|(1 + \|F\|^2)$), we can replace it by $W_i \sqrt{|\Lambda_i|}$. This computation can be done in $O((n-i)r^2 + r^3)$ flops ($r = 2$ in the case under consideration) by first computing the QR factorization of $\hat{G}_i$, say

$$\hat{G}_i = Q_i P_i, \quad Q_i Q_i^T = I,$$

and then computing the eigendecomposition of the 2×2 matrix $P_i J P_i^T$. We can then get W_i by multiplying Q_i by the orthogonal eigenvector matrix of $P_i J P_i^T$.

9. Solution of linear systems of equations. The analysis in the earlier sections suggests that for $\|F\|$ sufficiently close to one, the error norm can become large. However, if our original motivation is the solution of the linear system of equations

$$Rx = b,$$

then the error can be improved by resorting to iterative refinement if either the matrix R is given or if it can be computed accurately from (F, G). In what follows we show that for a diagonal F, the matrix R can be evaluated to high relative accuracy if u_1 and v_1 are exact.

9.1. Computing the matrix R. Given a positive-definite structured matrix R that satisfies

$$R - FRF^T = u_1 u_1^T - v_1 v_1^T,$$

with F diagonal and stable, its entries can be computed to high relative accuracy, as we explain below.

It follows from the displacement equation that

$$r_{ij} = \frac{u_{i,1}u_{j,1} - v_{i,1}v_{j,1}}{1 - f_i f_j} = \frac{u_{i,1}u_{j,1}\left(1 - \frac{v_{i,1}}{u_{i,1}}\frac{v_{j,1}}{u_{j,1}}\right)}{1 - f_i f_j},$$

where, by positive-definiteness, the ratios

$$\frac{v_{i,1}}{u_{i,1}} \quad \text{and} \quad \frac{v_{j,1}}{u_{j,1}}$$

are strictly less than one.

The term $\xi = 1 - \frac{v_{i,1}}{u_{i,1}}\frac{v_{j,1}}{u_{j,1}}$ can be evaluated to high relative accuracy, as explained earlier in the paper in §5.4, viz.,

If $\frac{v_{i,1}}{u_{i,1}}\frac{v_{j,1}}{u_{j,1}} < 1/2$
  then $\xi = 1 - \frac{v_{i,1}}{u_{i,1}}\frac{v_{j,1}}{u_{j,1}}$
else
  $d_1 = \frac{|u_{i,1}|-|v_{i,1}|}{|u_{i,1}|}$, $d_2 = \frac{|u_{j,1}|-|v_{j,1}|}{|u_{j,1}|}$
  $\xi = d_1 + d_2 - d_1 d_2$

Likewise, we evaluate $\mu = (1 - f_i f_j)$ and then

$$r_{ij} = \frac{u_{i,1}u_{j,1}\xi}{\mu}.$$

This guarantees that

$$\hat{r}_{i,j} = r_{i,j}(1 + c_{36}\delta_7).$$

9.2. Iterative refinement. If the factorization $\hat{L}\hat{L}^T$ is not too inaccurate and if R is not too ill conditioned, then it follows from the analysis in [11] that the solution $\hat{x}$ of $Rx = b$ can be made backward stable by iterative refinement.

ALGORITHM 9.1 (iterative refinement).

Set $\hat{x}_0 = \hat{x}$, $r = b - R\hat{x}_0$
repeat until $\|r\| \leq c_{37}\epsilon\|R\|\,\|\hat{x}\|$
  solve $\hat{L}\hat{L}^T\delta x = r$
  set $\hat{x}_i = \hat{x}_{i-1} + \delta x$
  $r = b - R\hat{x}_i$
endrepeat

10. Enhancing the robustness of the algorithm. We now suggest enhancements to further improve the robustness of the algorithm.

To begin with, carrying out the hyperbolic rotation as in (26) enforces the relation (28),

$$\hat{u}_{i+1}\hat{u}_{i+1}^T - \hat{v}_{i+1}\hat{v}_{i+1}^T = \Phi_i\hat{u}_i\hat{u}_i^T\Phi_i^T - \hat{v}_i\hat{v}_i^T - N_{i+1}, \tag{48}$$

where

$$\|N_{i+1}\| \leq c_{38}\epsilon(\|\hat{u}_{i+1}\|^2 + \|\Phi_i\|^2\|\hat{u}_i\|^2 + \|\hat{v}_{i+1}\|^2 + \|\hat{v}_i\|^2). \tag{49}$$

But the positive-definiteness of R further imposes conditions on the columns of the generator matrix. Indeed,

- for a diagonal and stable F, by Theorem A.1, a necessary condition for the positive-definiteness of the matrix is that we must have

  $$|\hat{u}_{i+1}| > |\hat{v}_{i+1}|, \tag{50}$$

  where the inequality holds componentwise;
- for a lower-triangular contractive F, Lemma B.2 shows that a necessary condition for positive-definiteness is

  $$\|u_i\| \geq \|v_i\|;$$

- in all cases, the condition $|\Phi_i u_i|_{i+1} > |v_i|_{i+1}$ is required to ensure that the reflection coefficient of the hyperbolic rotation Θ_i is less than 1.

We have found that if all these necessary conditions are enforced explicitly the algorithm is more reliable numerically. An example of this can be found in §10.3.

We now show how the OD and H methods can be modified to preserve the sign of the J-norm of each row of the prearray.

10.1. Enhancing the OD method. The OD method can be enhanced to preserve the sign of the J-norm of the row it is being applied to. For this purpose, assume that

$$|\Phi_i \hat{u}_i|_j > |\hat{v}_i|_j.$$

Then from (27) we see that if the jth row of the perturbed prearray has a positive J-norm then by adding a small perturbation to the jth row of the computed postarray we can guarantee a positive J-norm. If the jth row of the perturbed prearray does not have a positive J-norm, then in general there does not exist a small perturbation for the jth row of the postarray that will guarantee a positive J-norm. For such a row, the prearray must be perturbed to make its J-norm sufficiently positive and then the hyperbolic rotation must be reapplied by the OD method to that row. The new jth row of the postarray can now be made to have a positive J-norm by a small perturbation. The details are given in the algorithm below. For the case of a diagonal and stable F, all the rows of the prearray should have a positive J-norm. The algorithm should enforce this property.

In the statement of the algorithm, $\begin{bmatrix} x & y \end{bmatrix}$ stands for a particular row of the prearray $\bar{G}_{i+1}$, $\begin{bmatrix} \hat{x}_1 & \hat{y}_1 \end{bmatrix}$ stands for the corresponding row of the postarray $\hat{G}_{i+1}$, and Θ stands for the hyperbolic rotation. Here we are explicitly assuming that $|x| > |y|$, which is automatically the case when F is diagonal and stable. Otherwise, since $\begin{bmatrix} y & x \end{bmatrix}\Theta = \begin{bmatrix} y_1 & x_1 \end{bmatrix}$, the technique must be used with the elements of the input row interchanged.

ALGORITHM 10.1 (enhanced OD method).

Assumption: $|x| > |y|$.
if $|\hat{x}_1| < |\hat{y}_1|$
  $\gamma_1 \leftarrow c_7\epsilon(|\hat{x}_1| + |\hat{y}_1|)\text{sign}(\hat{x}_1)$
  $\gamma_2 \leftarrow c_7\epsilon(|\hat{x}_1| + |\hat{y}_1|)\text{sign}(\hat{y}_1)$
  if $|\hat{x}_1 + \gamma_1| > |\hat{y}_1 - \gamma_2|$ then
    $\hat{x}_1 \leftarrow \hat{x}_1 + \gamma_1$
    $\hat{y}_1 \leftarrow \hat{y}_1 - \gamma_2$
  else
    $\eta_1 \leftarrow c_6\epsilon(|x| + |y|)\text{sign}(x)$
    $\eta_2 \leftarrow c_6\epsilon(|x| + |y|)\text{sign}(y)$
    $\begin{bmatrix} \hat{x}_1 & \hat{y}_1 \end{bmatrix} \leftarrow \begin{bmatrix} x + \eta_1 & y - \eta_2 \end{bmatrix}\Theta$ (via the OD method)
    if $|\hat{x}_1| > |\hat{y}_1|$ then $\hat{x}_1 \leftarrow \hat{x}_1$ and $\hat{y}_1 \leftarrow \hat{y}_1$
    else
      $\gamma_1 \leftarrow c_7\epsilon(|\hat{x}_1| + |\hat{y}_1|)\text{sign}(\hat{x}_1)$
      $\gamma_2 \leftarrow c_7\epsilon(|\hat{x}_1| + |\hat{y}_1|)\text{sign}(\hat{y}_1)$
      $\hat{x}_1 \leftarrow \hat{x}_1 + \gamma_1$
      $\hat{y}_1 \leftarrow \hat{y}_1 - \gamma_2$
    endif

endif
endif

The computed columns $\hat{u}_{i+1}$ and $\hat{v}_{i+1}$ continue to satisfy a relation of form (27).

10.2. Enhancing the H procedure. Here again, $\begin{bmatrix} x & y \end{bmatrix}$ stands for a particular row of the prearray $\bar{G}_{i+1}$ and $\begin{bmatrix} \hat{x}_1 & \hat{y}_1 \end{bmatrix}$ stands for the corresponding row of the postarray. We shall again assume that $|x| > |y|$. If that is not the case then the procedure must be applied to $\begin{bmatrix} y & x \end{bmatrix}$, since $\begin{bmatrix} y & x \end{bmatrix} \Theta = \begin{bmatrix} y_1 & x_1 \end{bmatrix}$.

It follows from $|x| > |y|$ and relation (31) that

$$|\hat{x}_1 + e_1|^2 - |\hat{y}_1 + e_2|^2 > 0$$

for the H procedure. Therefore, by adding small numbers to $\hat{x}_1$ and $\hat{y}_1$ we can guarantee $|\hat{x}_1| > |\hat{y}_1|$.

ALGORITHM 10.2 (enhanced H method).

Assumption: $|x| > |y|$.
Apply the hyperbolic rotation Θ to $\begin{bmatrix} x & y \end{bmatrix}$ using the H procedure.
If $|\hat{x}_1| < |\hat{y}_1|$ then
$\hat{y}_1 \leftarrow |\hat{x}_1|(1 - 3\epsilon)\text{sign}(\hat{y}_1)$

10.3. A numerical example. The following example exhibits a positive-definite matrix R for which a direct implementation of the Schur algorithm, without the enhancements and modifications proposed herein, breaks down. On the other hand, the modified Schur algorithm enforces positive-definiteness and avoids breakdown, as the example shows. The data is given in Appendix E.

A straightforward implementation of the generalized Schur algorithm (i.e., with a naive implementation of the hyperbolic rotation and the Blaschke matrix–vector multiply) breaks down at the 8th step and declares the matrix indefinite.

On the other hand, our implementation, using the enhanced H procedure (§10.2) and the enhanced Blaschke matrix–vector multiply (§6.3), successfully completes the matrix factorization and yields a relative error

$$\frac{\|R - \hat{L}\hat{L}^T\|}{\epsilon(1 - \|F\|^2)^{-2}\|R\|} \approx 0.15.$$

Furthermore, the relative backward error $\|R - \hat{L}\hat{L}^T\|/\|R\|$ is approximately 10^{-11} (using a machine precision of approximately 10^{-16}).

11. Summary of results. The general conclusion is the following.

The modified Schur algorithm is backward stable for a large class of structured matrices. Generally, it is as stable as can be expected from the analysis in §4.

More specifically, we have the following.

- If F is strictly lower triangular and contractive (e.g., $F = Z$), then the modified algorithm is backward stable with no generator growth.
- If F is stable, diagonal, and positive, then by reordering the entries of F in increasing order, there will be no generator growth and the algorithm will be as stable as can be expected from §4. In particular, it will be backward stable if $\|F\|$ is not too close to one (e.g., $\|F\|^2 \leq 1 - \frac{1}{n^2}$).

- In all other cases, we can use the technique outlined in §8.4 to control the generator growth and make the algorithm as stable as can be expected from §4. In particular, it is backward stable if $\|F\|$ is not too close to one (e.g., $\|F\|^2 \leq 1 - \frac{1}{n^2}$).
- If R is given or can be computed accurately (e.g., when F is diagonal), iterative refinement can be used to make the algorithm backward stable for the solution of linear equations.

As far as pivoting is concerned, in the diagonal F case, we emphasize that it is necessary only when $\|F\|$ is very close to one.

- If F is positive (or negative), a good strategy is to reorder the entries of F in increasing order of magnitude.
- If F has both positive and negative entries, then our numerical example of §8.1 indicates that pivoting may not help control the growth of the generators.

In our opinion, for positive-definite structured matrices, with diagonal or strictly lower-triangular F, the stabilization of the generalized Schur algorithm is critically dependent on the following:

- proper implementations of the hyperbolic rotations (using the OD or H procedures),
- proper evaluation of the Blaschke matrix–vector product,
- enforcement of positive-definiteness to avoid early breakdowns,
- control of the generator growth.

12. Concluding remarks. The analysis and results of this paper can be extended to positive-definite structured matrices with displacement rank larger than 2, as well as to other forms of displacement structure, say the Hankel-like case

$$FR + RF^T = GJG^T.$$

While the current analysis can also be extended to the bidiagonal F case, the error bound will further depend on the norm of the Blaschke matrices, which need not be smaller than one. Further improvements seem possible and will be discussed elsewhere.

We are currently pursuing the extension of our results to general nonsymmetric structured matrices. In these cases, the hyperbolic rotations are replaced by coupled rotations [16] and the OD and H procedures can be generalized to implement these rotations accurately.

Moreover, the results of this work suggest improvements to certain fast algorithms in adaptive filtering and state–space estimation in view of the connections of these algorithms to the Schur algorithm [15]. This is a subject of ongoing investigation.

The H procedure can also be extended to other elementary transformations like Gauss transforms and Givens rotations. The implications of this fact will be addressed elsewhere.

Appendix A. Schur functions. A function $s(z)$ that is analytic and strictly bounded by one (in magnitude) in the closed unit disc ($|z| \leq 1$) will be referred to as a Schur function. Such functions arise naturally in the study of symmetric positive-definite matrices R that satisfy (10) with a stable diagonal matrix F with distinct entries. Indeed, let $\{u_{i1}\}$ and $\{v_{i1}\}$ denote the entries of the generator column vectors

u_1 and v_1:

$$u_1 = \begin{bmatrix} u_{11} \\ u_{21} \\ \vdots \\ u_{n1} \end{bmatrix}, \quad v_1 = \begin{bmatrix} v_{11} \\ v_{21} \\ \vdots \\ v_{n1} \end{bmatrix}.$$

The following theorem guarantees the existence of a Schur function $s(z)$ that maps the u_{i1} to the v_{i1} [19, Thm. 2.1], [15, 17].

THEOREM A.1. *The matrix R that solves* (10) *with a stable diagonal matrix F with distinct entries is positive definite if and only if there exists a Schur function $s(z)$ such that*

$$v_{i1} = u_{i1} s(f_i). \tag{A.51}$$

It follows from (A.51) that $|v_{i1}| < |u_{i1}|$ and, consequently, that $\|v_1\| < \|u_1\|$.

Appendix B. Useful lemmas. The following two results are used in the body of the paper.

LEMMA B.1. *If $R - FRF^T = uu^T - vv^T$ then*

$$\|v\|^2 \leq \|u\|^2 + \|R\| \, (1 + \|F\|^2).$$

The following is an extension of a result in [3].

LEMMA B.2. *If $R - FRF^T = uu^T - vv^T$, R is a positive-definite matrix, and F is a contractive matrix, then*

$$\|v\|^2 \leq \|u\|^2.$$

Proof. Taking the trace of both sides of the displacement equation we get

$$\operatorname{tr}(R) - \operatorname{tr}(FRF^T) = \|u\|^2 - \|v\|^2.$$

Introduce the SVD of F: $F = U\Sigma V^T$ with $\Sigma_{ii} \leq 1$. Then $\operatorname{tr}(FRF^T) = \operatorname{tr}(\Sigma V^T R V \Sigma) \leq \operatorname{tr}(V^T R V) = \operatorname{tr}(R)$, from which the result follows. □

Appendix C. Error analysis of the OD method. In this section we analyze the application of a hyperbolic rotation Θ using its eigendecomposition $\Theta = QDQ^T$, where

$$Q = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix},$$

and

$$D = \begin{bmatrix} \sqrt{\frac{\alpha+\beta}{\alpha-\beta}} & 0 \\ 0 & \sqrt{\frac{\alpha-\beta}{\alpha+\beta}} \end{bmatrix}$$

for given number α and β such that $|\alpha| > |\beta|$.

Let $B = \Theta A$. We shall now show that

$$(\hat{B} + E_2) = \Theta(A + E_1),$$

where $\|E_2\| \leq c_{39}\epsilon\|\hat{B}\|$ and $\|E_1\| \leq c_{40}\epsilon\|A\|$.

First note that using the above expressions for D and Q, their entries can be computed to high componentwise relative accuracy (assuming that the square roots can be computed to high relative accuracy).

Next observe that, for any vector x,

$$fl((\hat{D}x)_j) = (Dx)_j(1 + c_{41}\delta_8).$$

Therefore, $\|\hat{D}x - Dx\| \leq c_{42}\epsilon\|Dx\|$. Also, note that

$$fl(\hat{Q}_{ij}x_j) = Q_{ij}x_j(1 + c_{43}\delta_9).$$

Hence, we can show that

$$fl(\hat{Q}x) = Q(x + e) = Qx + Qe = Qx + e_1,$$

where $\|e_1\| = \|e\| \leq c_{44}\epsilon\|x\| = c_{45}\epsilon\|fl(Qx)\|$.

Now, let $y = QDQ^Tx$. Then in finite precision we have

$$\begin{aligned} fl(\hat{Q}^Tx) &= Q^T(x + e_3), \\ fl(\hat{D}Q^T(x + e_3)) &= DQ^T(x + e_3) + e_4 \end{aligned}$$

and

$$\begin{aligned} fl(\hat{Q}(DQ^T(x + e_3) + e_4)) &= Q(DQ^T(x + e_3) + e_4) + e_5 \\ &= QDQ^T(x + e_1) - e_6 = \hat{y}. \end{aligned}$$

Thus,

$$\hat{y} + e_6 = QDQ^T(x + e_3) = \Theta(x + e_3),$$

where, by the bounds above,

$$\|e_3\| \leq c_{46}\epsilon\|x\|, \quad \|e_6\| \leq c_{47}\epsilon\|\hat{y}\|.$$

Appendix D. Matlab programs. We include here a Matlab listing of the stabilized Schur algorithm suggested in this paper. The program assumes that the input matrix is positive definite and tries to enforce it. The algorithm listed here can be easily modified to test if a structured matrix is positive definite.

D.1. The H procedure. Input data: The ratio *beta/alpha* represents the reflection coefficient, which is smaller than one in magnitude. Also, y/x is assumed smaller than one in magnitude.

Output data: The entries $\begin{bmatrix} x_1 & y_1 \end{bmatrix}$ that result by applying a hyperbolic rotation to $\begin{bmatrix} x & y \end{bmatrix}$ with $|x_1| > |y_1|$.

function $\begin{bmatrix} x1, & y1 \end{bmatrix} = h\_procedure(x, y, beta, alpha)$

```
c = (beta * y)/(alpha * x);
if c < 0.5
  xi = 1 - c;
else
```

```
    d1 = (abs(alpha) − abs(beta))/abs(alpha);
    d2 = (abs(x) − abs(y))/abs(x);
    xi = d1 + d2 − d1 * d2;
end
x1 = (abs(alpha) * x * xi)/sqrt((alpha − beta) * (alpha + beta));
y1 = x1 − sqrt((alpha + beta)/(alpha − beta)) * (x − y);
if abs(x1) < abs(y1)
  y1 = abs(x1) * (1 − 3 * eps) * sign(y1)
end
```

D.2. The Blaschke matrix–vector product. We now list the program that computes $\Phi_i u_i$ for both a diagonal F and a strictly lower-triangular F.

Input data: An $n \times n$ stable and diagonal matrix F, a vector u, a vector v (such that $|v| < |u|$), and an index i $(1 \leq i \leq n)$.

Output data: The matrix–vector product $z = \Phi_i u$, where $\Phi_i = (I - f_i F)^{-1}(F - f_i I)$, and the vector $ub = (I - f_i F)^{-1} u$.

function $[\ z,\ \ ub\ ] = blaschke\_1(F, u, v, i, n)$

```
ub = u;
z = u;
for j = i : n
    if F(i,i) * F(j,j) < 0.5
      xi = 1/(1 − F(j,j) * F(i,i));
    else
        d1 = 1 − abs(F(i,i));
        d2 = 1 − abs(F(j,j));
        xi = 1/(d1 + d2 − d1 * d2);
    end
    ub(j) = xi * z(j);
    z(j) = (F(j,j) − F(i,i)) * ub(j);
    if abs(z(j)) < abs(v(j))
      z(j) = abs(v(j)) * (1 + 3 * eps) * sign(z(j));
    end
end
```

For a strictly lower-triangular F we use the following.

Input data: An $n \times n$ strictly lower-triangular matrix F, a vector u, a vector v (such that $|v| < |u|$), and an index i $(1 \leq i \leq n)$.

Output data: The matrix–vector product $z = Fu$ and $ub = u$.

function $[\ z,\ \ ub\ ] = blaschke\_2(F, u, v, i, n)$

```
ub = u;
z = F * u;
z(i) = 0;
```

```
if abs(z(i + 1)) < abs(v(i + 1))
  z(i + 1) = abs(v(i + 1)) * (1 + 3 * eps) * sign(z(i + 1));
end
```

D.3. The stable Schur algorithm. We now list two versions of the stable modified Schur algorithm—one for diagonal stable F and the other for strictly lower-triangular F.

Input data: An $n \times n$ diagonal and stable matrix F, a generator $G = \begin{bmatrix} u & v \end{bmatrix}$ in proper form (i.e., $v_1 = 0$) with column vectors u, v.

Output data: A lower-triangular Cholesky factor L such that $\|R - LL^T\|$ satisfies (45).

function $L = stable\_schur\_1(u, v, F)$

```
n = size(F, 1);
for i = 1 : n − 1
    [ u,   ub ] = blaschke_1(F, u, v, i, n);
    L(:, i) = sqrt((1 − F(i, i)) * (1 + F(i, i))) * ub;
    a = v(i + 1);
    b = u(i + 1);
    for j = i + 1 : n
        [ u(j),   v(j) ] = h_procedure(u(j), v(j), a, b);
    end
    v(i + 1) = 0;
end
L(n, n) = (1/sqrt((1 − F(n, n)) * (1 + F(n, n)))) * u(n);
L(1 : n − 1, n) = zeros(n − 1, 1);
```

Input data: An $n \times n$ strictly lower-triangular matrix F, a generator $G = \begin{bmatrix} u & v \end{bmatrix}$ in proper form (i.e., $v_1 = 0$) with column vectors u, v.

Output data: A lower-triangular Cholesky factor L such that $\|R - LL^T\|$ satisfies (45).

function $L = stable\_schur\_2(u, v, F)$

```
n = size(F, 1);
for i = 1 : n − 1
    [ u,   ub ] = blaschke_2(F, u, v, i, n);
    L(:, i) = sqrt((1 − F(i, i)) * (1 + F(i, i))) * ub;
    a = v(i + 1);
    b = u(i + 1);
    for j = i + 1 : n
        if abs(u(j)) > abs(v(j))
            [ u(j),   v(j) ] = h_procedure(u(j), v(j), a, b);
        else
            [ temp_v,   temp_u ] = h_procedure(v(j), u(j), a, b);
            v(j) = temp_v; u(j) = temp_u;
```

```
        endif
    end
    v(i+1) = 0;
end
L(n,n) = (1/sqrt((1 - F(n,n)) * (1 + F(n,n)))) * u(n);
L(1 : n-1, n) = zeros(n-1, 1);
```

Appendix E. Example of breakdown.

$$G = \begin{bmatrix} 0.29256168393970 & 0 \\ 0.28263551029525 & -0.10728616660709 \\ 0.09633626413940 & 0.01541380240248 \\ 0.06797943459994 & -0.02572176567354 \\ 0.55275012712414 & 0.22069874528633 \\ 0.42631253478657 & 0.06821000412583 \\ 0.50468895704517 & 0.20125628531328 \\ 0.23936358366577 & -0.09527653751206 \\ 0.14608901804405 & 0.02337424345679 \end{bmatrix}, \quad F = \operatorname{diag} \begin{bmatrix} 0.40000000000000 \\ 0.97781078411630 \\ -0.00000000433051 \\ 0.97646762001746 \\ -0.99577002371173 \\ 0.00000001005313 \\ -0.99285659894698 \\ 0.99789820799463 \\ -0.00000001100000 \end{bmatrix}.$$

The matrix R that solves $R - FRF^T = GJG^T$ is positive definite since the entries of v were computed from

$$v_i = u_i s(f_i),$$

where $s(z)$ is the Schur function

$$s(z) = 0.4\,\frac{0.4 - z}{1 - 0.4z}.$$

Acknowledgments. The authors would like to thank Dr. Ming Gu for useful discussions and the anonymous referees and the Editor for their helpful comments.

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] A. W. Bojanczyk, R. P. Brent, F. R. de Hoog, and D. R. Sweet, *On the stability of the Bareiss and related Toeplitz factorization algorithms*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 40–57.

[4] A. W. Bojanczyk, R. P. Brent, P. Van Dooren, and F. R. de Hoog, *A note on downdating the Cholesky factorization*, SIAM J. Sci. Stat. Comput., 8 (1987), pp. 210–221.

[5] A. W. Bojanczyk and A. O. Steinhardt, *Matrix downdating techniques for signal processing*, in Proc. SPIE Conference on Advanced Algorithms and Architectures for Signal Processing, San Diego, CA, Society of Photo-optical Instrumentation Engineers, 1988, pp. 68–75.

[6] S. Chandrasekaran and M. Gu, *Notes on the PSVD and QSVD*, in preparation.

[7] G. Cybenko, *The numerical stability of the Levinson–Durbin algorithm for Toeplitz systems of equations*, SIAM J. Sci. Stat. Comput., 1 (1980), pp. 303–319.

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

[9] M. Gu, *Stable and Efficient Algorithms for Structured Systems of Linear Equations*, Tech. report LBL-37690, Lawrence Berkeley Laboratory, University of California, Berkeley, CA, 1995.

[10] G. Heinig, *Inversion of generalized Cauchy matrices and other classes of structured matrices*, in Linear Algebra for Signal Processing, Vol. 69, A. Bojanczyk and G. Cybenko, eds., Springer-Verlag, New York, Berlin, 1995, pp. 63–81.

[11] M. JANKOWSKI AND M. WOZNIAKOWSKI, *Iterative refinement implies numerical stability*, BIT, 17 (1977), pp. 303–311.

[12] T. KAILATH, *A theorem of I. Schur and its impact on modern signal processing*, in Operator Theory: Advances and Applications, Vol. 18, I. Gohberg, ed., Birkhäuser, Boston, 1986, pp. 9–30.

[13] T. KAILATH AND A. H. SAYED, *Displacement structure: Theory and applications*, SIAM Rev., 37 (1995), pp. 297–386.

[14] N. LEVINSON, *The Wiener r.m.s. (root-mean-square) error criterion in filter design and prediction*, J. Math. Phys., 25 (1947), pp. 261–278.

[15] A. H. SAYED, *Displacement Structure in Signal Processing and Mathematics*, Ph.D. thesis, Stanford University, Stanford, CA, 1992.

[16] ———, *Time-variant structured matrices: An application to instrumental variable methods*, in Proc. SPIE Conference on Advanced Signal Processing: Algorithms, Architectures, and Implementations, San Diego, CA, 1994, Society of Photo-optical Instrumentation Engineers, pp. 516–527.

[17] A. H. SAYED, T. CONSTANTINESCU, AND T. KAILATH, *Square-Root Algorithms for Structured Matrices, Interpolation, and Completion Problems*, IMA Volumes in Mathematics and Its Applications, Vol. 69, Springer-Verlag, 1995, pp. 153–184. Linear Algebra for Signal Processing, A. Bojanczyk and G. Cybenko, eds.

[18] A. H. SAYED AND T. KAILATH, *Fast algorithms for generalized displacement structures and lossless systems*, Linear Algebra Appl., 219 (1995), pp. 49–78.

[19] A. H. SAYED, T. KAILATH, H. LEV-ARI, AND T. CONSTANTINESCU, *Recursive solutions of rational interpolation problems via fast matrix factorization*, Integral Equations Operator Theory, 20 (1994), pp. 84–118.

[20] 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*, Vol. 18, I. Gohberg, ed., Birkhäuser, Boston, 1986, pp. 31–88.)

[21] M. STEWART AND P. VAN DOOREN, *Stability Issues in the Factorization of Structured Matrices*, SIAM J. Matrix Anal. Appl., to appear.

[22] D. R. SWEET, *Numerical Methods for Toeplitz Matrices*, Ph.D. thesis, University of Adelaide, Adelaide, Australia, 1982.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 4, pp. 984–991, October 1996

015

BEST AVAILABLE BOUNDS FOR DEPARTURE FROM NORMALITY*

STEVEN L. LEE†

Abstract. The best available bounds for the departure from normality of a matrix are given. The significant properties of these lower and upper bounds are also described. For example, one of the upper bounds is a practical estimate that costs (at most) $2m$ multiplications, where m is the number of nonzeros in the matrix. In terms of applications, the results can be used to bound from above the sensitivity of eigenvalues to matrix perturbations or to bound from below the distance to the closest normal matrix.

Key words. nonnormal matrix, departure from normality, condition numbers, eigenvalues

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 various matrix bounds. If A is an $n \times n$ matrix, its departure from normality in the Frobenius norm is defined to be [8]

$$\operatorname{dep}(A) := \left(\|A\|^2 - \|\Lambda\|^2\right)^{1/2}, \tag{1.1}$$

where Λ is a diagonal matrix whose entries are the eigenvalues, λ_k, of A. This measure of matrix nonnormality can be used to bound the spectral norm of matrix functions [2], [5]; the sensitivity of eigenvalues to matrix perturbations [8], [15]; and the distance to the closest normal matrix [10], [16], for example. It is impractical to compute $\operatorname{dep}(A)$ if A is large and its eigenvalues are unknown. This difficulty motivates us to seek lower and upper bounds for $\operatorname{dep}(A)$ that are practical to compute or optimal in some sense.

In terms of eigenvalues, bounds for $\operatorname{dep}(A)$ can be used to obtain lower and upper bounds for

$$\|\Lambda\|^2, \quad \|\operatorname{Re}(\Lambda)\|^2, \quad \text{and} \quad \|\operatorname{Im}(\Lambda)\|^2, \tag{1.2}$$

where $\operatorname{Re}(\Lambda)$ and $\operatorname{Im}(\Lambda)$ are the real and imaginary parts of Λ. In particular, such results can be obtained by substituting lower and upper bounds for $\operatorname{dep}(A)$ into [12, Lem. 3.1]

$$\|\Lambda\|^2 = \|A\|^2 - \operatorname{dep}^2(A), \tag{1.3}$$

$$\|\operatorname{Re}(\Lambda)\|^2 = \|M\|^2 - \frac{1}{2}\operatorname{dep}^2(A), \tag{1.4}$$

$$\|\operatorname{Im}(\Lambda)\|^2 = \|N\|^2 - \frac{1}{2}\operatorname{dep}^2(A), \tag{1.5}$$

where

$$M = \frac{1}{2}\left(A + A^H\right) \quad \text{and} \quad N = \frac{1}{2}\left(A - A^H\right)$$

* Received by the editors April 28, 1995; accepted for publication (in revised form) by N. J. Higham November 28, 1995. This research was supported by Applied Mathematical Sciences Research Program, Office of Energy Research, and U. S. Department of Energy contract DE-AC05-96OR22464 with Lockheed Martin Energy Research Corporation.

† Mathematical Sciences Section, Oak Ridge National Laboratory, P. O. Box 2008, Building 6012, Oak Ridge, TN 37831-6367 (na.slee@na-net.ornl.gov).

are the Hermitian and skew-Hermitian parts of A, respectively. Upper bounds for $\|\Lambda\|^2$ can be used to bound the spectral radius [14] and the spread of a matrix [1]. As we show later, the lower and upper bounds for $\|\Lambda\|^2$ in [5], [9] are especially useful because they can be substituted into (1.1) to help obtain better bounds for $\mathrm{dep}(A)$.

The outline of this paper is as follows. In §2, we give the notation, definitions, and observations that will be needed in later sections. In §3, we present various bounds for $\|\Lambda\|^2$ that are then used to obtain better bounds for $\mathrm{dep}^2(A)$. In §4, we describe the significant properties of the newly improved bounds. In §5, we group the currently known a priori bounds for $\mathrm{dep}^2(A)$ into two main categories and then show that the new bounds are among the best available.

2. Preliminaries. Let $A = (a_{ij})$ be an $n \times n$ matrix with conjugate transpose $A^H = (\bar{a}_{ji})$ and Frobenius norm

$$\|A\|^2 := \sum_{i,j} |a_{ij}|^2.$$

Also, recall that A is normal if and only if (iff), for example, the following are true [7]:

(2.1a) A has a complete, orthogonal set of eigenvectors,

(2.1b) $$\|A\| = \|\Lambda\| = \left(\sum |\lambda_k|^2\right)^{1/2}, \text{ or}$$

(2.1c) $$A^H A - AA^H = 0.$$

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. It is easily seen that $\mathrm{dep}(A)$ is invariant with respect to shifts and rotations. That is,

$$\mathrm{dep}(A) = \mathrm{dep}(e^{-i\theta}(A - \alpha I)) \tag{2.2}$$

for any complex scalar α and $0 \leq \theta < 2\pi$. The commutator in (2.1c) is another measure of nonnormality [4] that is invariant to shifts and rotations of A; for example,

$$\|(A - \alpha I)^H (A - \alpha I) - (A - \alpha I)(A - \alpha I)^H\| = \|A^H A - AA^H\|. \tag{2.3}$$

It is easy to show that the quadratic function $\|A - \alpha I\|^2$ is minimized for $\alpha = \frac{\mathrm{tr}(A)}{n}$, where $\mathrm{tr}(A)$ is the trace of A. If $\mathrm{tr}(A) = 0$, we shall say A is a *centered* matrix. Centered matrices such as

$$\widetilde{A} = A - \frac{\mathrm{tr}(A)}{n} I$$

will be denoted with a tilde accent. Finally, we give a lemma that relates the norm of the shifted matrix $A - \alpha I$ to the norm of the centered matrix $\widetilde{A}$.

LEMMA 2.1. *For any $n \times n$ matrix A and complex scalar α,*

$$\|A - \alpha I\|^2 = \|\widetilde{A}\|^2 + \frac{|\mathrm{tr}(A - \alpha I)|^2}{n}. \tag{2.4}$$

Proof. First, we relate the norm of A to the norm of $\widetilde{A}$. For $\sigma = \frac{\mathrm{tr}(A)}{n}$, we have

$$\|A\|^2 - \|\widetilde{A}\|^2 = \|A\|^2 - \left\|A - \frac{\mathrm{tr}(A)}{n} I\right\|^2 \tag{2.5}$$

$$
\begin{aligned}
&= \sum(|a_{ii}|^2) - \sum(|a_{ii} - \sigma|^2) \\
&= \sum(a_{ii}^H a_{ii}) - \sum\left[a_{ii}^H a_{ii} - a_{ii}^H \sigma - \sigma^H a_{ii} + \sigma^H \sigma\right] \\
&= \sigma \sum(a_{ii}^H) + \sigma^H \sum(a_{ii}) - n\sigma^H \sigma \\
&= \frac{\operatorname{tr}(A)}{n} \operatorname{tr}^H(A) + \frac{\operatorname{tr}^H(A)}{n} \operatorname{tr}(A) - n\left|\frac{\operatorname{tr}(A)}{n}\right|^2 \\
&= \frac{2\,|\operatorname{tr}(A)|^2}{n} - \frac{|\operatorname{tr}(A)|^2}{n} \\
&= \frac{|\operatorname{tr}(A)|^2}{n}. \qquad (2.6)
\end{aligned}
$$

If we replace A with $A - \alpha I$ on the right-hand side of (2.5) and (2.6), we obtain

$$\|A - \alpha I\|^2 - \left\|A - \alpha I - \frac{\operatorname{tr}(A - \alpha I)}{n} I\right\|^2 = \frac{|\operatorname{tr}(A - \alpha I)|^2}{n},$$

and the second term simplifies to

$$\left\|A - \alpha I - \frac{\operatorname{tr}(A - \alpha I)}{n} I\right\|^2 = \|\widetilde{A}\|^2$$

to obtain (2.4). □

3. Bounds for eigenvalues and departure from normality. We now present several bounds for $\|\Lambda\|^2$ and $\operatorname{dep}^2(A)$, along with their important properties. An upper bound for $\|\Lambda\|^2$ is given by Kress, de Vries, and Wegmann [9]. Moreover, the authors exhibit nonnormal matrices for which the bound is sharp and prove that the upper bound is the best possible in terms of $\|A\|$ and $\|A^H A - AA^H\|$.

THEOREM 3.1 (see [9, Thm. 1]). *For nonnormal A there holds*

$$\|\Lambda\|^2 \leq \left(\|A\|^4 - \frac{1}{2}\|A^H A - AA^H\|^2\right)^{1/2} \qquad (3.1)$$

with equality iff

$$A = \gamma(vw^H + rwv^H) \qquad (3.2)$$

where γ is a nonzero complex scalar; $0 \leq r < 1$ is a real scalar; and v, w are orthonormal vectors.

The practical lower bound for $\|\Lambda\|^2$ in [5],

$$|\operatorname{tr}(A^2)| \leq \|\Lambda\|^2, \qquad (3.3)$$

comes from the triangle inequality applied to the eigenvalues of A^2:

$$
\begin{aligned}
|\operatorname{tr}(A^2)| = |\operatorname{tr}(\Lambda^2)| &= \left|\lambda_1^2 + \cdots + \lambda_n^2\right| \qquad (3.4) \\
&\leq \left|\lambda_1^2\right| + \cdots + \left|\lambda_n^2\right| = |\lambda_1|^2 + \cdots + |\lambda_n|^2 = \|\Lambda\|^2.
\end{aligned}
$$

The lower bound is sharp iff zero and the eigenvalues of A are collinear. Moreover, the bound is cheap to compute since only the diagonal of A^2 is needed. This diagonal can

be computed with (at most) m multiplications, where m is the number of nonzeros in A.

The lower bound [17, p. 161] and the upper bound [5] for $\mathrm{dep}^2(A)$,

$$\|A\|^2 - \left(\|A\|^4 - \frac{1}{2}\|A^HA - AA^H\|^2\right)^{1/2} \leq \mathrm{dep}^2(A) \leq \|A\|^2 - |\mathrm{tr}(A^2)|, \tag{3.5}$$

can be obtained by substituting (3.1) and (3.3) into (1.1). The upper bound in (3.5) is sharp iff zero and the eigenvalues of A are collinear, and it can be computed with (at most) $2m$ multiplications. The lower bound is an $O(n^3)$ computation that is sharp iff A is normal or satisfies condition (3.2). This lower bound inherits the properties of the upper bound (3.1) via (1.1); thus, it is the best possible in terms of $\|A\|$ and $\|A^HA - AA^H\|$.

It is straightforward to improve the $\mathrm{dep}^2(A)$ lower bound in (3.5) if we recall that $\mathrm{dep}(A)$ is invariant with respect to the shift parameter α; see (2.2). For normal matrices, the bound

$$\|A-\alpha I\|^2 - \left(\|A-\alpha I\|^4 - \frac{1}{2}\|A^HA - AA^H\|^2\right)^{1/2} \leq \mathrm{dep}^2(A-\alpha I) = \mathrm{dep}^2(A) \tag{3.6}$$

is zero for any choice of α. For nonnormal matrices, however, there is a unique value of α that maximizes (3.6). In particular, by substituting (2.4) into (3.6), we seek to maximize the function

$$f(z(\alpha)) = (\beta^2 + z^2(\alpha)) - \left[(\beta^2 + z^2(\alpha))^2 - \frac{1}{2}K^2\right]^{1/2} \tag{3.7}$$

where

$$z^2(\alpha) = \frac{|\mathrm{tr}(A-\alpha I)|^2}{n}$$

and

$$\beta^2 = \|\widetilde{A}\|^2, \quad K^2 = \|A^HA - AA^H\|^2 > 0.$$

By solving $\frac{df}{dz} = 0$, we find that the unique solution $z = 0$ is a global maximum since $\frac{d^2f}{dz^2}(0) < 0$. By solving $z(\alpha) = 0$, we find that the lower bound is maximized for $\alpha = \frac{\mathrm{tr}(A)}{n}$.

In §4, we prove that $\alpha = \frac{\mathrm{tr}(A)}{n}$ also optimizes the upper bound

$$\mathrm{dep}^2(A) = \mathrm{dep}^2(A-\alpha I) = \|\widetilde{A}\|^2 - \|\widetilde{\Lambda}\|^2 \leq \|\widetilde{A}\|^2 - |\mathrm{tr}(\widetilde{A}^2)|. \tag{3.8}$$

4. Main results. We now establish the significant properties of the new bounds given in §3. To begin, recall that the lower bound in (3.5) is sharp for any nonnormal matrix that satisfies condition (3.2). The improved lower bound (3.6), with $\alpha = \frac{\mathrm{tr}(A)}{n}$, is unaffected by complex shifts; thus, it is sharp for

$$A = \gamma(vw^H + rwv^H) - \sigma I$$

for any choice of the scalar σ. Note that we have

$$\alpha = \frac{\mathrm{tr}(\gamma(vw^H + rwv^H) - \sigma I)}{n} = -\sigma$$

so that $\alpha = \frac{\mathrm{tr}(A)}{n}$ cancels the arbitrary shift σ. The improved bound is also unaffected by rotations. We summarize the above results as follows.

THEOREM 4.1. *For any $n \times n$ matrix A,*

$$\mathrm{dep}^2(A) \geq \|\widetilde{A}\|^2 - \left(\|\widetilde{A}\|^4 - \frac{1}{2}\|A^HA - AA^H\|^2\right)^{1/2} \tag{4.1}$$

where $\widetilde{A} = A - \frac{\mathrm{tr}(A)}{n}I$. The bound is sharp, with equality iff A is normal or

$$A = e^{-i\theta}(\gamma(vw^H + rwv^H) - \sigma I) \tag{4.2}$$

where γ, r, and σ are complex scalars; $0 \leq \theta < 2\pi$; and v, w are orthonormal vectors.

We will now prove that the upper bound (3.8) is sharp iff the eigenvalues of A are collinear in the complex plane. Before doing so, we must establish a natural measure of the noncollinearity of matrix eigenvalues. One approach is to define "departure from collinearity" as

$$\mathrm{depcol}(A) := \sum |d_k|^2 \tag{4.3}$$

where $|d_k|$ is the perpendicular distance from λ_k to the straight line, total least squares (TLS) fit of the eigenvalues of A. Recall that a TLS fit minimizes the sum of the squares of the perpendicular distances from the points to the fitted line and that $\sum |d_k|^2$ is the TLS error [6]. Given the definition (4.3), we find $\mathrm{depcol}(A)$ to be a sensible metric for quantifying departure from collinearity, especially since $\mathrm{depcol}(A) = 0$ iff A has collinear eigenvalues.

A useful result concerning departure from collinearity follows from [11, Thm. 2.2].

THEOREM 4.2. *Given the complex numbers z_k, $k = 1, \ldots, n$, let $\bar{z} = \frac{1}{n}\sum z_k$ so that*

$$\tilde{z}_k = z_k - \bar{z}.$$

The error for the TLS fit is

$$\sum |d_k|^2 = \frac{1}{2}\left(\sum |\tilde{z}_k|^2 - \left|\sum \tilde{z}_k^2\right|\right) \tag{4.4}$$

where $|d_k|$ is the perpendicular distance from z_k to the fit.

In the context of matrix eigenvalues, (4.4) yields

$$\mathrm{depcol}(A) = \frac{1}{2}\left(\|\widetilde{\Lambda}\|^2 - |\mathrm{tr}(\widetilde{A}^2)|\right). \tag{4.5}$$

If we arrange (4.5) as

$$\|\widetilde{\Lambda}\|^2 = |\mathrm{tr}(\widetilde{A}^2)| + 2\ \mathrm{depcol}(A),$$

we can substitute to obtain

$$\mathrm{dep}^2(A) = \|\widetilde{A}\|^2 - \|\widetilde{\Lambda}\|^2 = \|\widetilde{A}\|^2 - \left(|\mathrm{tr}(\widetilde{A}^2)| + 2\ \mathrm{depcol}(A)\right). \tag{4.6}$$

The upper bound in (3.8) is a special case of the equality in (4.6).

THEOREM 4.3. *For any $n \times n$ matrix A,*

$$\mathrm{dep}^2(A) \leq \|\widetilde{A}\|^2 - |\mathrm{tr}(\widetilde{A}^2)| \tag{4.7}$$

where $\widetilde{A} = A - \frac{\mathrm{tr}(A)}{n}I$. The bound is sharp, with equality iff the eigenvalues of A are collinear in the complex plane.

Proof. The bound (4.7) is obtained from (4.6) by dropping the term $2\ \mathrm{depcol}(A)$. The bound is sharp iff $\mathrm{depcol}(A) = 0$, that is, iff the eigenvalues of A are collinear. □

5. Discussion and summary. To the best of our knowledge, a priori bounds for $\operatorname{dep}^2(A)$ fall into one of two distinct categories. The bounds in the first category are based on computing the Frobenius norm of the commutator $A^HA - AA^H$ [3], [4], [8], [13], [17]. The bounds in the second category are based on inequalities that are sharp iff the eigenvalues of A have a certain alignment in the complex plane [5], [12]. For each of these categories, we give the best bounds known to us at this time.

Bounds based on $\|A^HA - AA^H\|$.

$$\operatorname{dep}^2(A) \leq \left(\frac{n^3 - n}{12}\right)^{1/2} \|A^HA - AA^H\|, \tag{5.1}$$

$$\operatorname{dep}^2(A) \geq \|\widetilde{A}\|^2 - \left(\|\widetilde{A}\|^4 - \frac{1}{2}\|A^HA - AA^H\|^2\right)^{1/2}. \tag{5.2}$$

Remarks. The upper bound (5.1) is due to Henrici [8, Thm. 1]. The lower bound (5.2) is given in Theorem 4.1. Sun's lower bound (3.5) is the best possible in terms of $\|A^HA - AA^H\|$ and $\|A\|$; thus it is stronger than the bounds in [3], [4]. The bound (5.2) improves upon Sun's lower bound, and it is also stronger than the one in [13].

Bounds based on eigenvalue alignment.

$$\operatorname{dep}^2(A) = \|\widetilde{A}\|^2 - \left(|\operatorname{tr}(\widetilde{A}^2)| + 2\ \operatorname{depcol}(A)\right) \leq \|\widetilde{A}\|^2 - |\operatorname{tr}(\widetilde{A}^2)|. \tag{5.3}$$

Remarks. The upper bound (5.3) is sharp iff the eigenvalues of A are collinear. In contrast, note that for $\alpha = \frac{\operatorname{tr}(A)}{n}$, we have [12, Thm. 3.3]

$$\operatorname{dep}^2(A) \leq 2\ \min\left\{\|M - \operatorname{Re}(\alpha)I\|^2, \|N - i\operatorname{Im}(\alpha)I\|^2\right\}. \tag{5.4}$$

This bound is sharp only when the eigenvalues are horizontally or vertically aligned in the complex plane. Furthermore, the bound (5.3) is about half as expensive to compute as (5.4). Despite these shortcomings, the latter bound is useful and has some noteworthy properties. In particular, (5.3) and (5.4) yield the same value if A is a real matrix. We also remark that (5.4) explicitly bounds matrix nonnormality in terms of the nonsymmetry of A.

Besides its practicality, the bound (5.3) is also appealing because it sometimes enables us to compute $\operatorname{dep}(A)$ for matrices with extremely sensitive eigenvalues. For example, consider the $n \times n$ matrix

$$\widehat{W}_n = U^H W_n U$$

where $\widehat{W}_n$ is dense and unitarily similar to the Wilkinson matrix [18, p. 90]

$$W_n = \begin{bmatrix} n & n & & & \\ & (n-1) & n & & \\ & & \ddots & \ddots & \\ & & & 2 & n \\ & & & & 1 \end{bmatrix}.$$

The eigenvalues of $\widehat{W}_n$ are real, and the interior eigenvalues are notoriously difficult to compute for $n \gg 20$. Thus, we cannot determine the departure from normality

for $\widehat{W}_{50}$ via (1.1) because $\|\Lambda\|^2$ cannot be computed accurately. However, we can precisely determine that $\operatorname{dep}(\widehat{W}_{50}) = 350$ via (5.3) since the sharpness of the formula (modulo rounding errors) only depends upon eigenvalue collinearity—not eigenvalue sensitivity.

The $\operatorname{dep}^2(A)$ bounds (5.1)–(5.3), together with equations (1.3)–(1.5), lead to better bounds for $\|\Lambda\|^2$, $\|\operatorname{Re}(\Lambda)\|^2$, and $\|\operatorname{Im}(\Lambda)\|^2$. For example, substituting (5.2) into (1.3) yields

$$\|\Lambda\|^2 \leq \left(\|\widetilde{A}\|^4 - \frac{1}{2}\|A^HA - AA^H\|^2\right)^{1/2} + \frac{|\operatorname{tr}(A)|^2}{n} \tag{5.5}$$

which improves upon the original result in (3.1). This bound is sharp, with equality iff A is normal or satisfies condition (4.2). For additional results concerning bounds for eigenvalues, see [19] and the references therein.

To summarize, we have developed new bounds for $\operatorname{dep}^2(A)$ and described their significant properties. We have also grouped these and the other known a priori bounds into two categories. Within each category, we have given the best available bounds. The bounds based on $\|A^HA - AA^H\|$ have an important property: they reduce to zero if A is normal. Unfortunately, such bounds are often weak and impractical to compute if A is large. On the other hand, the bounds based on eigenvalue alignment are often good estimates (e.g., [12, Table 1]), and they are practical to compute if A is large and sparse. A minor drawback is that these bounds only reduce to zero for normal matrices with collinear eigenvalues (e.g., Hermitian and skew-Hermitian matrices). Theorem 4.1, Theorem 4.3, and [8, Thm. 1] describe nonnormal matrices for which the best available bounds in (5.1)–(5.3) are sharp. The significance of our results is described in §1.

REFERENCES

[1] N. A. Derzko and A. M. Pfeffer, *Bounds for the spectral radius of a matrix*, Math. Comp., 19 (1965), pp. 62–67.

[2] J. Descloux, *Bounds for the spectral norm of functions of matrices*, Numer. Math., 5 (1963), pp. 185–190.

[3] P. J. Eberlein, *On measures of non-normality for matrices*, Amer. Math. Monthly, 72 (1965), pp. 995–996.

[4] L. Elsner and M. H. C. Paardekooper, *On measures of nonnormality of matrices*, Linear Algebra Appl., 92 (1987), pp. 107–124.

[5] M. Gil', *Estimate for the norm of matrix-valued functions*, Linear and Multilinear Algebra, 35 (1993), pp. 65–73.

[6] G. H. Golub and C. F. Van Loan, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 883–893.

[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] R. Kress, H. L. de Vries, and R. Wegmann, *On nonnormal matrices*, Linear Algebra Appl., 8 (1974), pp. 109–120.

[10] L. László, *An attainable lower bound for the best normal approximation*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1035–1043.

[11] S. L. Lee, *A Note on the Total Least Squares Problem for Coplanar Points*, Technical Report TM-12852, Oak Ridge National Laboratory, Oak Ridge, TN, 1994.

[12] ———, *A practical upper bound for departure from normality*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 462–468.

[13] G. Loizou, *Nonnormality and Jordan condition numbers of matrices*, J. Assoc. Comput. Mach., 16 (1969), pp. 580–584.

[14] L. MIRSKY, *The spread of a matrix*, Mathematika, 3 (1956), pp. 127–130.
[15] A. RUHE, *On the closeness of eigenvalues and singular values for almost normal matrices*, Linear Algebra Appl., 11 (1975), pp. 87–94.
[16] ———, *Closest normal matrix finally found!* BIT, 27 (1987), pp. 585–598.
[17] J.-G. SUN, *Matrix Perturbation Analysis*, Academic Press, Beijing, China, 1987. (In Chinese.)
[18] J. WILKINSON, *The Algebraic Eigenvalue Problem*, Claredon Press, Oxford, England, 1965.
[19] H. WOLKOWICZ AND G. P. H. STYAN, *More bounds for eigenvalues using traces*, Linear Algebra Appl., 31 (1980), pp. 1–17.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 4, pp. 992–997, October 1996

016

GENERALIZED MONOTONE AFFINE MAPS*

JEAN-PIERRE CROUZEIX† AND SIEGFRIED SCHAIBLE‡

Dedicated to Professor Bela Martos on the occasion of his 75th birthday.

Abstract. In this paper we derive new necessary and sufficient conditions for an affine map to be quasimonotone on a convex set.

Key words. affine map, quasimonotonicity, pseudomonotonicity

AMS subject classifications. Primary, 90C30; Secondary, 26B25

1. Introduction. Recently, quasimonotone and pseudomonotone maps have been introduced and studied in the context of complementarity problems and variational inequalities problems [10, 11, 12, 17]. Several existence results have been obtained and algorithmic implications are being studied. For a recent survey see [17].

In the particular case where a map F is the gradient of a function f, quasiconvexity (pseudoconvexity) of f is equivalent to the quasimonotonicity (pseudomonotonicity) of F. In this paper, we consider affine maps $F(x) = Ax + q$; here A is an $n \times n$ matrix and q is a vector of $\Re^n$.

If A is symmetric, then F is the gradient of a quadratic function. Quasiconvexity and pseudoconvexity of quadratic functions have been studied extensively [1, 4, 6, 13, 15, 16], and therefore characterizations of generalized monotone affine maps are well known in this case. The purpose of this paper is to treat the case where A is not necessarily symmetric. For some initial results see [7, 8, 9, 12, 14].

Generalized monotone affine maps arise in linear complementarity problems and linear variational inequality problems; see, for example, [7]–[10], [17].

2. Definitions and notation. Let U be a convex subset of $\Re^n$ and $F : U \to \Re^n$. The map F is *quasimonotone* on U if for every $x, y \in U$

$$\langle y - x, F(x)\rangle > 0 \Rightarrow \langle y - x, F(y)\rangle \geq 0, \tag{1}$$

and F is *pseudomonotone* on U if for every $x, y \in U$

$$\langle y - x, F(x)\rangle > 0 \Rightarrow \langle y - x, F(y)\rangle > 0. \tag{2}$$

For various definitions of generalized monotonicity see for instance [11]. For any subset U of $\Re^n$, its (positive) polar cone is given by

$$U^+ = \{y \in \Re^n : \langle y, x\rangle \geq 0 \ \forall x \in U\}.$$

Given a square $n \times n$ matrix D, its (Moore–Penrose) pseudoinverse [2] is the uniquely defined $n \times n$ matrix $D^\dagger$ which satisfies the following conditions:

$$DD^\dagger D = D,\ D^\dagger DD^\dagger = D^\dagger,\ (DD^\dagger)^t = DD^\dagger, \text{ and } (D^\dagger D)^t = D^\dagger D.$$

* Received by the editors August 21, 1995; accepted for publication (in revised form) by R. Cottle November 30, 1995.

† Applied Mathematics, Université Blaise Pascal, 63177 Aubière Cedex, France (crouzeix@ucfma.univ-bpclermont.fr).

‡ Graduate School of Management, University of California, Riverside, CA 92521 (schaible@ucrac1.ucr.edu).

Given a symmetric $n \times n$ matrix B, its inertia is the triple

$$In(B) = (n_+, n_-, n_0)$$

where n_+, n_- and n_0 denote the number of positive, negative, and zero eigenvalues of B, respectively; then $n_+ + n_- + n_0 = n$. Lagrange–Sylvester's law on inertia says that $In(B) = In(P^t BP)$ for any nonsingular $n \times n$ matrix P.

Given a partitioned symmetric matrix

$$M = \begin{pmatrix} A & B \\ B^t & C \end{pmatrix}$$

with A nonsingular, the Schur complement of A is the matrix $M/A = C - B^t A^{-1} B$ [3]. The inertias of M, A, and M/A are related to each other as follows:

$$In(M) = In(A) + In(M/A).$$

In what follows,
$F(x) = Ax + q$ where A is an $n \times n$ matrix and q is a vector of $\Re^n$,
$B = \frac{1}{2}(A + A^t)$,
$C = A^t B^\dagger A$,
$In(B) = (n_+, n_-, n_0)$ and $r = \dim(Kern(A))$,
$f(x) = \langle Ax + q, B^\dagger(Ax + q)\rangle$,
$S = \{x : f(x) \leq 0\}$,
$T = \{x : \langle Cx, x\rangle \leq 0\}$,
and U is a convex set of $\Re^n$ with nonempty interior.

3. The main results. It is known [12] that an affine map is quasimonotone on an open convex set if and only if it is pseudomonotone on that set. This result is not true for a convex set in general [12]. For a continuous (not necessarily affine) map G, we have the following result which uses the nonemptiness of $\text{int}(U)$.

LEMMA 1. *Let $G : U \to \Re^n$. Assume that G is continuous on U and quasimonotone on* $\text{int}(U)$; *then it is also quasimonotone on U.*

Proof. Assume that G is not quasimonotone on U; then $x, y \in U$ exist such that

$$\langle y - x, G(x)\rangle > 0 \text{ and } \langle y - x, G(y)\rangle < 0.$$

Then there exist $x', y' \in \text{int}(U)$ sufficiently close to x, y, respectively, such that

$$\langle y' - x', G(x')\rangle > 0 \text{ and } \langle y' - x', G(y')\rangle < 0,$$

contradicting quasimonotonicity of G on $\text{int}(U)$. □

On the other hand, we have the following characterization of pseudomonotone affine maps on open convex sets.

PROPOSITION 1 (see [12, Thm. 5.2]). *The map F is pseudomonotone on* $\text{int}(U)$ *if and only if*

$$x \in \text{int}(U), h \in \Re^n, \text{ and } \langle Ax + q, h\rangle = 0 \text{ imply } \langle Bh, h\rangle \geq 0. \tag{3}$$

Combining Lemma 1 and Proposition 1, we obtain the following characterization.

PROPOSITION 2. *F is quasimonotone on U (and pseudomonotone on* $\text{int}(U)$*) if and only if one of the following conditions holds:*

(i) $n_- = 0$; *i.e., B is positive semidefinite and F is monotone on $\Re^n$;*

(ii) $n_- = 1$, $-q \notin A(\mathrm{int}(U))$, $q \in B(\Re^n) \supseteq A(\Re^n)$, *and* $U \subseteq S$.

Proof. It is known [5] that for a vector $b \in \Re^n$ the condition

$$h \in \Re^n, \langle b, h\rangle = 0 \text{ imply } \langle Bh, h\rangle \geq 0 \tag{4}$$

holds if and only if either

(i) $n_- = 0$; i.e., B is positive semidefinite; or

(ii) $n_- = 1$, $b \neq 0$, and there exists $v \in \Re^n$ such that $Bv = b$ and $\langle b, v\rangle \leq 0$.

We note that (ii) is equivalent to

$$n_- = 1, 0 \neq b \in B(\Re^n), \text{ and } \langle b, B^\dagger b\rangle \leq 0.$$

Assume that $n_- = 1$. Then condition (3) is equivalent to

$$0 \notin A(\mathrm{int}(U)) + q \subseteq B(\Re^n), \text{ and } \mathrm{int}(U) \subseteq S. \tag{5}$$

Assume that (5) holds. Fix some $x \in \mathrm{int}(U)$. Then for any $h \in \Re^n$, $t > 0$ exists such that $x + th \in \mathrm{int}(U)$. Since $Ax + q$ and $Ax + tAh + q$ belong to the same linear space $B(\Re^n)$, Ah also belongs to this space. Hence $A(\Re^n) \subseteq B(\Re^n)$. Now, $Ax + q \in B(\Re^n)$ and $Ax \in B(\Re^n)$ together imply $q \in B(\Re^n)$. It is then easy to complete the proof. □

Since $B = \frac{1}{2}(A + A^t)$, we have the following relationship between the inertias of B and C.

LEMMA 2. *If* $A(\Re^n) \subseteq B(\Re^n)$, *then*

$$In(C) = (n_+ + n_0 - r, n_- + n_0 - r, 2r - n_0). \tag{6}$$

Hence it follows that

$$0 \leq r - n_0 \leq n_-. \tag{7}$$

Proof. (a) First we show that

$$In\begin{pmatrix} 0 & A \\ A^t & 0 \end{pmatrix} = (n - r, n - r, 2r). \tag{8}$$

To see this, let us consider the equation

$$\begin{pmatrix} 0 & A \\ A^t & 0 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \lambda\begin{pmatrix} x \\ y \end{pmatrix}.$$

Then $Ay = \lambda x$, $A^t x = \lambda y$, and $A^t A y = \lambda^2 y$. It is easy to see that if the vector $\binom{x}{y}$ is an eigenvector for the eigenvalue λ , then the vector $\binom{x}{-y}$ is an eigenvector for the eigenvalue $-\lambda$.

(b) Next we prove that

$$In\begin{pmatrix} B & A \\ A^t & 0 \end{pmatrix} = In\begin{pmatrix} 0 & A \\ A^t & 0 \end{pmatrix}.$$

This is a consequence of Lagrange–Sylvester's law on inertia and the identity

$$\begin{pmatrix} B & A \\ A^t & 0 \end{pmatrix} = \begin{pmatrix} I & \frac{1}{2}I \\ 0 & I \end{pmatrix}\begin{pmatrix} 0 & A \\ A^t & 0 \end{pmatrix}\begin{pmatrix} I & 0 \\ \frac{1}{2}I & I \end{pmatrix}.$$

(c) Since B is symmetric, there is an $n \times n$ nonsingular matrix P and an $(n - n_0) \times (n - n_0)$ nonsingular diagonal matrix D such that

$$P^tP = I \text{ and } PBP^t = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}.$$

Since $A(\Re^n) \subseteq B(\Re^n)$, $n_0 \leq r$ and there exists an $(n - n_0) \times n$ matrix R such that $PAP^t = \binom{R}{0}$.

Then

$$In \begin{pmatrix} B & A \\ A^t & 0 \end{pmatrix} = In \begin{pmatrix} D & 0 & R \\ 0 & 0 & 0 \\ R^t & 0 & 0 \end{pmatrix} = (0, 0, n_0) + In \begin{pmatrix} D & R \\ R^t & 0 \end{pmatrix}.$$

Hence by the inertia formula for the Schur complement

$$In \begin{pmatrix} B & A \\ A^t & 0 \end{pmatrix} = (0, 0, n_0) + In(D) + In(-R^tD^{-1}R).$$

Now $In(D) = (n_+, n_-, 0)$, and

$$PCP^t = PA^tP^tPB^\dagger P^tPAP^t = R^tD^{-1}R,$$

from which the result follows in view of (8). □

Accordingly, (ii) in Proposition 2 reduces to two cases: $r = n_0 + 1$ and $r = n_0$. We now state the main theorem.

THEOREM 1. *F is quasimonotone on U (and pseudomonotone on $\text{int}(U)$) if and only if one of the following conditions holds:*

(i) $n_- = 0$; *i.e., B is positive semidefinite and F is monotone on $\Re^n$;*

(ii1) $n_- = 1$, $r = n_0 + 1$, $-q \notin A(\text{int}(U))$, $q \in B(\Re^n) \supseteq A(\Re^n)$, *$C$ is positive semidefinite, S is a closed convex set, and $U \subseteq S$;*

(ii2) $n_- = 1$, $r = n_0$, $-q \notin A(\text{int}(U))$, $q \in B(\Re^n) = A(\Re^n)$, *and $T = T_+ \cup -T_+$ where T_+ is a closed convex cone with nonempty interior; and if $\bar{x}$ is such that $A\bar{x} = q$, then either $U + \bar{x} \subseteq T_+$ or $U + \bar{x} \subseteq -T_+$.*

Proof. Let us consider case (ii) of Proposition 2. If $r = n_0 + 1$, then by (6) C is positive semidefinite. Hence f is convex and S is a convex set.

Assume now that $r = n_0$. Then $A(\Re^n) = B(\Re^n)$, so that $\bar{x}$ exists such that $A\bar{x} = q$. Hence $f(x) = \langle C(x + \bar{x}), x + \bar{x} \rangle$.

The case where $n_+ = 0$ deserves special attention. Then A has rank 1. It follows that two nonzero vectors u and v exist such that $A = uv^t$. Since B also has rank 1, u and v are collinear and $k < 0$ exists such that $A = B = kuu^t$. Take $T_+ = \{x : \langle u, x \rangle \geq 0\}$. Since $-q \notin A(\text{int}(U))$, $U + \bar{x}$ is contained in either T_+ or $-T_+$.

Assume now that $n_+ \neq 0$. Then the discussion follows the lines developed for generalized convex quadratic functions by Ferland [6] and Schaible [15, 16]. Since C is symmetric, a diagonal matrix Δ (with entries δ_i) and a nonsingular matrix Q exist such that $Q^tQ = I$, $Q^tCQ = \Delta$, and $\delta_1 < 0 \leq \delta_2 \leq \cdots \leq \delta_n > 0$.

Then T has the following form:

$$T = \{x = Qy : \delta_1 y_1^2 \leq -(\delta_2 y_2^2 + \cdots + \delta_n y_n^2)\} = T_+ \cup -T_+,$$

where

$$T_+ = \left\{ x = Qy : y_1 \geq \sqrt{\frac{-(\delta_2 y_2^2 + \cdots + \delta_n y_n^2)}{\delta_1}} \right\}.$$

It is easy to see that T_+ is a closed convex cone and that $\operatorname{int}(T) = \operatorname{int}(T_+) \cup \operatorname{int}(-T_+)$ where

$$\operatorname{int}(T_+) = \left\{ x = Qy : y_1 > \sqrt{\frac{-(\delta_2 y_2^2 + \cdots + \delta_n y_n^2)}{\delta_1}} \right\}.$$

Since the open convex set $\operatorname{int}(U + \bar{x})$ is contained in $\operatorname{int}(T)$, it is contained in either $\operatorname{int}(T_+)$ or $\operatorname{int}(-T_+)$. Hence the result follows. □

This theorem gives the maximal domain of quasimonotonicity of F: $\Re^n$ in case (i), S in case (ii1), and $T_+ - \bar{x}$ or $-T_+ - \bar{x}$ in case (ii2). It is worth noticing that quasimonotonicity of the map F on U implies quasiconvexity of the associated function f on this set.

Now we apply this theorem to the case where U is a cone.

THEOREM 2. *Assume that U is a convex cone with nonempty interior. Then F is quasimonotone on U (and pseudomonotone on $\operatorname{int}(U)$) if and only if one of the following conditions holds:*

(i) $n_- = 0$; *i.e., B is positive semidefinite and F is monotone on $\Re^n$;*

(ii1) $n_- = n_+ = 1$, *A has rank 1, $q \in B(\Re^n) \supseteq A(\Re^n)$, $\langle q, B^\dagger q \rangle \leq 0$, and $-A^t B^\dagger q \in U^+$;*

(ii2) $n_- = 1$, $r = n_0$, *$q \in B(\Re^n) = A(\Re^n)$, and $T = T_+ \cup -T_+$ where T_+ is a closed convex cone with nonempty interior; and if $\bar{x}$ is such that $A\bar{x} = q$, then either ($U \subseteq T_+$ and $\bar{x} \in T_+$) or ($U \subseteq -T_+$ and $\bar{x} \in -T_+$).*

Proof. Let us first consider case (ii1). Let $x \in U$. Then for all $t > 0$

$$0 \geq f(tx) = t^2 \langle Cx, x \rangle + 2t \langle x, A^t B^\dagger q \rangle + \langle q, B^\dagger q \rangle.$$

It follows that $\langle Cx, x \rangle \leq 0$ and, since C is positive semidefinite, that $\langle Cx, x \rangle = 0$ and therefore $Cx = 0$ for all $x \in U$. Then $C = 0$ since U has a nonempty interior.

Hence, by (6), $n_+ + n_0 - r = 0$; thus $n_+ = 1$, $n_0 = n - 2$, and $r = n - 1$. Conversely, if $n_+ = n_- = 1$ and $r = n - 1$, then there are u and v noncollinear such that $A = uv^t$. It is easy to see that then $C = 0$.

The other results are immediate.

Now, case (ii2) is a direct consequence of Theorem 1 and the assumption that U is a convex cone. □

An important case is $U = \Re^n_+$, the nonnegative orthant. In this setting, the last theorem becomes Corollary 1.

COROLLARY 1. *Assume that $U = \Re^n_+$. Then F is quasimonotone on $\Re^n_+$ (and pseudomonotone on $\operatorname{int}(\Re^n_+)$) if and only if one of the following conditions holds:*

(i) $n_- = 0$; *i.e., B is positive semidefinite, and F is monotone on $\Re^n$;*

(ii1) $n_- = n_+ = 1$, *A has rank 1, $q \in B(\Re^n) \supseteq A(\Re^n)$, $\langle q, B^\dagger q \rangle \leq 0$, and $-A^t B^\dagger q \in \Re^n_+$;*

(ii2) $n_- = 1$, $r = n_0$, *$-C$ is copositive, $q \in B(\Re^n) = A(\Re^n)$, and $T = T_+ \cup -T_+$ where T_+ is a closed convex cone with nonempty interior; and if $\bar{x}$ is such that $A\bar{x} = q$, then either ($\Re^n_+ \subseteq T_+$ and $\bar{x} \in T_+$) or ($\Re^n_+ \subseteq -T_+$ and $\bar{x} \in -T_+$).*

We now consider some special cases.

If A is symmetric, then $C = B = A$, case (ii1) cannot occur since $r = n_0$, and F is the gradient of $\frac{1}{2}f$. Theorems 1 and 2 and their corollary recover the characterizations of generalized convex quadratic functions [1, 4, 6, 13, 15, 16].

If A is invertible, then case (ii1) in Theorem 1 (and a fortiori in Theorem 2 and its corollary) cannot occur since $n_0 = r - 1 = -1$, which is not possible.

REFERENCES

[1] M. AVRIEL, W. E. DIEWERT, S. SCHAIBLE, AND I. ZANG, *Generalized Concavity*, Plenum Publishing Corporation, New York, 1988.

[2] A. BEN-ISRAEL AND T. N. E. GREVILLE, *Generalized Inverses: Theory and Applications*, John Wiley and Sons, New York, 1974.

[3] R. W. COTTLE, *Manifestations of the Schur complement*, Linear Algebra Appl., 8 (1974), pp. 189–211.

[4] R. W. COTTLE AND J. A. FERLAND, *Matrix-theoretic criteria for the quasiconvexity and pseudoconvexity of quadratic functions*, Linear Algebra Appl., 5 (1972), pp. 123–136.

[5] J.-P. CROUZEIX AND J. A. FERLAND, *Criteria for quasiconvexity and pseudoconvexity: Relationships and comparisons*, Math. Programming, 23 (1982), pp. 193–205.

[6] J. A. FERLAND, *Maximal domains of quasiconvexity and pseudoconvexity for quadratic functions*, Math. Programming, 3 (1972), pp. 178–192.

[7] M. S. GOWDA, *Pseudomonotone and copositive star matrices*, Linear Algebra Appl., 113 (1989), pp. 107–118.

[8] ———, *On the transpose of a pseudomonotone matrix and the linear complementarity problem*, Linear Algebra Appl., 120 (1990), pp. 129–137.

[9] ———, *Affine pseudomonotone mappings and the linear complementarity problem*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 373–380.

[10] S. KARAMARDIAN, *Complementarity over cones with monotone and pseudomonotone maps*, J. Optim. Theory Appl., 18 (1976), pp. 445–454.

[11] S. KARAMARDIAN AND S. SCHAIBLE, *Seven kinds of monotone maps*, J. Optim. Theory Appl., 66 (1990), pp. 37–46.

[12] S. KARAMARDIAN, S. SCHAIBLE, AND J.-P. CROUZEIX, *Characterizations of generalized monotone maps*, J. Optim. Theory Appl., 76 (1993), pp. 399–413.

[13] B. MARTOS, *Subdefinite matrices and quadratic forms*, SIAM J. Appl. Math., 17 (1969), pp. 1215–1223.

[14] R. PINI AND S. SCHAIBLE, *Invariance properties of generalized monotonicity*, Optimization, 28 (1994), pp. 211–222.

[15] S. SCHAIBLE, *Beiträge zur quasikonvexen Programmierung*, Doctoral Dissertation, Köln, Germany, 1971.

[16] ———, *Second order characterizations of pseudoconvex quadratic functions*, J. Optim. Theory Appl., 21 (1977), pp. 15–26.

[17] ———, *Generalized monotonicity-concepts and uses*, in Variational Inequalities and Network Equilibrium Problems, F. Giannessi and A. Maugeri, eds., Plenum Publishing Corporation, New York, 1995, pp. 289–299.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 4, pp. 998–1006, October 1996

017

EVERY NORMAL TOEPLITZ MATRIX IS EITHER OF TYPE I OR OF TYPE II*

TAKASHI ITO†

Abstract. In their 1995 manuscript, Farenick and Lee proved that every normal Toeplitz matrix of order less than or equal to 5 is either of type I or of type II and had left open the case of higher order as a conjecture.

The author of this paper settled the conjecture affirmatively. Almost simultaneously, Farenick and Lee themselves proved the conjecture in a work with Krupnik and Krupnik [*SIAM J. Matrix Anal. Appl.*, 17 (1996), pp. 1037–1043]. Since the idea of proof in the work just mentioned is somewhat different from the one of this paper, the editor has recommended that the author publish in a separate paper.

Key words. normal matrices, Toeplitz matrices

AMS subject classifications. 15A57, 47B15, 47B35

1. Introduction: Normal Toeplitz matrices. Toeplitz operators have been studied in connection with many branches of mathematics. Our concern here is finite dimensional Toeplitz operators, especially the *normality* of them. For the background of the problem and for previously known results related to this problem, see [2].

Denote by $[a_1, a_2, \dots, a_N; a_{-1}, a_{-2}, \dots, a_{-N}]$ the Toeplitz matrix T_N of order $N+1$ generated by $\{a_j\ ;\ -N \le j \le N\}$:

$$T = \begin{bmatrix} a_0 & a_{-1} & \dots & a_{-N} \\ a_1 & a_0 & \dots & a_{-(N-1)} \\ \dots & \dots & \dots & \dots \\ a_{N-1} & a_{N-2} & \dots & a_{-1} \\ a_N & a_{N-1} & \dots & a_0 \end{bmatrix}.$$

Here we always assume $a_0 = 0$.

Consider the self-commutator $[T_N, T_N^*] \stackrel{def}{=} T_N T_N^* - T_N^* T_N \equiv [\alpha_{m,n}]_{m,n=1}^{N+1}$. Then we have

$$\alpha_{m,n} = \sum_{k=1}^{N+1} a_{m-k}\overline{a}_{n-k} - \sum_{k=1}^{N+1} \overline{a}_{-(m-k)} a_{-(n-k)}.$$

It is easy to see from this relation the skew-symmetry of the self-commutator with respect to the second diagonal:

$$\alpha_{m,n} = -\alpha_{N+2-n,N+2-m} \quad (1 \le m, n \le N+1).$$

Therefore a necessary and sufficient condition for T_N to be normal is that

$$\alpha_{m+1,n+1} = \alpha_{m,n} \quad (1 \le m, n \le N).$$

When expressed in terms of a_n's this shows that T_N is normal if and only if

$$(1)\quad a_m\overline{a}_n - \overline{a}_{-m}a_{-n} + \overline{a}_{N+1-m}a_{N+1-n} - a_{-(N+1-m)}\overline{a}_{-(N+1-n)} = 0 \quad (1 \le m, n \le N).$$

* Received by the editors October 12, 1995; accepted for publication by T. Ando December 1, 1995.

† Department of Mathematics, Musashi Institute of Technology, Tamazutsumi 1, Setagaya-ku, Tokyo 158, Japan (uafuruta@ipc.musashi-tech.ac.jp).

Suppose that T_N is normal; that is, (1) holds. If a subset $\{j_1, j_2, \dots, j_M\}$ of $\{1, 2, \dots, N\}$ with $j_1 < j_2 < \cdots < j_M$ is closed under the mapping $j \longmapsto N+1-j$, that is,

$$j_{M+1-m} = N + 1 - j_m \quad (1 \le m \le M),$$

it is seen from (1) that the Toeplitz submatrix $[a_{j_1}, a_{j_2}, \dots, a_{j_M}; a_{-j_1}, a_{-j_2}, \dots, a_{-j_M}]$ becomes normal. Let us denote this Toeplitz matrix by $[j_1, j_2, \dots, j_M]$:

$$[j_1, j_2, \dots, j_M] \stackrel{def}{=} [a_{j_1}, a_{j_2}, \dots, a_{j_M}; a_{-j_1}, a_{-j_2}, \dots, a_{-j_M}].$$

In particular, the Toeplitz matrices of order 5

$$[m, n, N+1-n, N+1-m] \quad \left(1 \le m < n < \frac{N+1}{2}\right) \tag{2}$$

and those of order 4 or 3

$$\left[m, \frac{N+1}{2}, N+1-m\right] \quad \left(1 \le m < \frac{N+1}{2};\ N \text{ odd}\right), \tag{2'}$$

$$[m, N+1-m] \quad \left(1 \le m < \frac{N+1}{2}\right) \tag{2"}$$

are normal.

As an important consequence of (1), we summarize the above-mentioned facts in a proposition.

PROPOSITION 1. *A Toeplitz matrix T_N is normal if and only if all Toeplitz submatrices with order less than or equal to 5 of the form (2), (2'), and (2") are normal.*

Another consequence of (1) is the following proposition.

PROPOSITION 2. *If for some $1 \le m \le N$*

$$\begin{bmatrix} a_m & a_{-m} \\ a_{N+1-m} & a_{-(N+1-m)} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

then normality of T_N is equivalent to that of the Toeplitz submatrix generated $\{a_j\ ;\ j \ne m, -m, N+1-m, -(N+1-m)\}$ with a canonical indexing.

Therefore throughout this paper we shall assume

$$\begin{bmatrix} a_m & a_{-m} \\ a_{N+1-m} & a_{-(N+1-m)} \end{bmatrix} \ne \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \quad (1 \le m \le N). \tag{3}$$

2. Type I and type II. Recall the definitions of types of normal Toeplitz matrices in [1, 2]. A normal Toeplitz matrix $T_N = [a_1, \dots, a_n; a_{-1}, \dots, a_{-N}]$ is said to be of type I if it is of the form $T = \alpha I + \beta H$ for some scalars α and β and for Hermitian Toeplitz matrix H. It is said to be of type II if it is a *generalized circulant* in the sense that there is $0 \le \omega < 2\pi$ such that

$$a_{-j} = e^{i\omega} a_{N+1-j} \quad (j = 1, 2, \dots, N).$$

It is obvious that a normal Toeplitz matrix T_N (with $a_0 = 0$) is of type I if and only if

$$\begin{bmatrix} a_{-1} \\ a_{-2} \\ \vdots \\ a_{-N} \end{bmatrix} = e^{i\theta} \begin{bmatrix} \overline{a}_1 \\ \overline{a}_2 \\ \vdots \\ \overline{a}_N \end{bmatrix} \quad \text{for some } 0 \le \theta < 2\pi, \tag{4}$$

whereas T_N is of type II if and only if

$$\begin{bmatrix} a_{-1} \\ a_{-2} \\ \vdots \\ a_{-N} \end{bmatrix} = e^{i\omega} \begin{bmatrix} a_N \\ a_{N-1} \\ \vdots \\ a_1 \end{bmatrix} \quad \text{for some } 0 \le \omega < 2\pi. \tag{5}$$

The following property was observed originally by Farenick and Lee [1]. This property will be used later on.

PROPOSITION 3. *Suppose that a Toeplitz matrix T_N is normal. If*

$$\begin{bmatrix} a_{k_0} \\ a_{-k_0} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a_{N+1-k_0} \\ a_{-(N+1-k_0)} \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{for some } 1 \le k_0 \le N, \tag{6}$$

then T_N is of type I.

Proof. Since T_N is normal, by putting $m = k_0$ in (1) we have

$$\overline{a}_{N+1-k_0} a_{N+1-n} = a_{-(N+1-k_0)} \overline{a}_{-(N+1-n)} \quad (1 \le n \le N).$$

By putting $n = k_0$ it follows $|a_{N+1-k_0}| = |a_{-(N+1-k_0)}|$. Then by assumption (6) we have

$$|a_{N+1-k_0}| = |a_{-(N+1-k_0)}| > 0.$$

Thus

$$a_{N+1-n} = e^{i\theta} \overline{a}_{-(N+1-n)} \quad (1 \le n \le N) \quad \text{where} \quad e^{i\theta} = \frac{a_{-(N+1-k_0)}}{\overline{a}_{N+1-k_0}}. \qquad \square$$

Corresponding to Proposition 3 we have the following for type II.

PROPOSITION 4. *Suppose that a Toeplitz matrix T_N is normal. If*

$$\begin{bmatrix} a_{k_0} \\ a_{-(N+1-k_0)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a_{N+1-k_0} \\ a_{-k_0} \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{for some} \quad 1 \le k_0 \le N, \tag{7}$$

then T_N is of type II.

Proof. By putting $m = k_0$ in (1) we have

$$-\overline{a}_{-k_0} a_{-n} + \overline{a}_{N+1-k_0} a_{N+1-n} = 0 \quad (1 \le n \le N).$$

By putting $n = k_0$ in, it follows $|a_{k_0}| = |a_{N+1-k_0}|$. By (7) we have

$$|a_{k_0}| = |a_{N+1-k_0}| > 0.$$

Thus

$$a_{-n} = e^{i\omega} a_{N+1-n} \quad (1 \le n \le N) \quad \text{where} \quad e^{i\omega} = \frac{\overline{a}_{N+1-k_0}}{\overline{a}_{-k_0}}. \qquad \square$$

3. Proof of theorems. Farenick and Lee [1] proved first that every normal Toeplitz matrix of order less than or equal to 5 is either of type I or of type II. Based on this result, we will show that the same statement holds for general orders. Since Farenick et al. [2] does not contain explicitly the original argument of Farenick and Lee [1], we first show their result with a simplified proof.

THEOREM 1 (see Farenick and Lee [1]). *For $1 \le N \le 4$ every normal Toeplitz matrix $T_N = [a_1, \ldots, a_N; a_{-1}, \ldots, a_{-N}]$ is either of type* I *or of type* II.

Proof. 1. Case of $N = 1$. T_N is normal if and only if $|a_1| = |a_{-1}|$. Thus T_N is trivially of type I (and also type II).

2. Case of $N = 2$. Normality of T_N implies from (1) that

$$a_1\overline{a}_2 = a_{-2}\overline{a}_{-1} \tag{8}$$

and

$$|a_1|^2 + |a_2|^2 = |a_{-1}|^2 + |a_{-2}|^2. \tag{9}$$

By our assumption (3) and Proposition 3, if

$$\begin{bmatrix} a_1 \\ a_{-1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} a_2 \\ a_{-2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

then T_N is of type I. Similarly, by (3) and Proposition 4, if

$$\begin{bmatrix} a_1 \\ a_{-2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} a_2 \\ a_{-1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

then T_N is of type II. Thus we consider only the case in which none of a_1, a_2, a_{-1}, a_{-2} is zero. Then $|a_1\overline{a}_2| = |a_{-2}\overline{a}_{-1}|$ and (9) imply that

$$|a_1| = |a_{-1}| \text{ and } |a_2| = |a_{-2}| \quad \text{or} \quad |a_1| = |a_{-2}| \text{ and } |a_{-1}| = |a_2|.$$

When $|a_1| = |a_{-1}|$ and $|a_2| = |a_{-2}|$ holds, (8) shows

$$\frac{a_1}{\overline{a}_{-1}} = \frac{a_{-2}}{\overline{a}_2} = e^{i\theta} \quad \text{for some } 0 \leq \theta < 2\pi;$$

thus T_N is of type I. Similarly, if $|a_1| = |a_{-2}|$ and $|a_{-1}| = |a_2|$ holds, then T_N is of type II.

3. Case of $N = 3$. By Proposition 1, the Toeplitz submatrix of order 3 $[a_1, a_3; a_{-1}, a_{-3}]$ is normal. Thus, the previous case of $N = 2$ shows that

$$\begin{bmatrix} a_{-1} \\ a_{-3} \end{bmatrix} = e^{i\theta} \begin{bmatrix} \overline{a}_1 \\ \overline{a}_3 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} a_{-1} \\ a_{-3} \end{bmatrix} = e^{i\theta} \begin{bmatrix} a_3 \\ a_1 \end{bmatrix} \quad \text{for some } 0 \leq \theta < 2\pi. \tag{10}$$

Furthermore (1) implies

$$|a_2| = |a_{-2}| \tag{11}$$

and

$$a_1\overline{a}_2 + a_2\overline{a}_3 = \overline{a}_{-1}a_{-2} + \overline{a}_{-2}a_{-3}. \tag{12}$$

By our assumption (3) it follows that $|a_2| = |a_2| > 0$.

We have to consider two cases separately.

(i) Suppose

$$\begin{bmatrix} a_{-1} \\ a_{-3} \end{bmatrix} = e^{i\theta} \begin{bmatrix} \overline{a}_1 \\ \overline{a}_3 \end{bmatrix}$$

holds. Set $a_{-2} = e^{i\omega}\overline{a}_2$. If $\omega = \theta$, it is clear that T_N is of type I. Assume $\omega \neq \theta$. By substituting

$$a_{-1} = e^{i\theta}\overline{a}_1, a_{-3} = e^{i\theta}\overline{a}_3, \text{ and } a_{-2} = e^{i\omega}\overline{a}_2$$

into (12), we have

$$(1 - e^{i(\omega-\theta)})(a_1\overline{a}_2 - a_{-3}\overline{a}_{-2}) = 0.$$

Hence we have $a_1\overline{a}_2 = a_{-3}\overline{a}_{-2}$. Thus, by setting $a_{-2} = e^{i\gamma}a_2$, we have

$$a_{-3} = \frac{\overline{a}_2}{\overline{a}_{-2}}a_1 = e^{i\gamma}a_1$$

and

$$a_{-1} = e^{i\theta}\overline{a}_1 = e^{i\theta}e^{i\gamma}\overline{a}_{-3} = e^{i\theta}e^{i\gamma}e^{-i\theta}a_3 = e^{i\gamma}a_3.$$

Thus T_N is of type II.

With a similar argument to that in case (i) we can see the following.

(ii) Suppose

$$\begin{bmatrix} a_{-1} \\ a_{-3} \end{bmatrix} = e^{i\theta}\begin{bmatrix} a_3 \\ a_1 \end{bmatrix}$$

holds. Then T_N is either of type I or of type II.

Before going into the case of $N = 4$, we formulate as a lemma the argument which was used just above, because the same argument will be used later in different situations several times.

LEMMA 1.

(i) $\begin{bmatrix} a_{-1} \\ a_{-3} \end{bmatrix} = e^{i\theta}\begin{bmatrix} \overline{a}_1 \\ \overline{a}_3 \end{bmatrix}$, $a_1\overline{c} = a_{-3}\overline{d}$, *and* $|c| = |d| > 0$ *imply*

$$\begin{bmatrix} a_{-1} \\ d \\ a_{-3} \end{bmatrix} = e^{i\gamma}\begin{bmatrix} a_3 \\ c \\ a_1 \end{bmatrix} \quad \textit{where} \quad e^{i\gamma} = \frac{d}{c}.$$

(ii) $\begin{bmatrix} a_{-1} \\ a_{-3} \end{bmatrix} = e^{i\theta}\begin{bmatrix} a_3 \\ a_1 \end{bmatrix}$, $\overline{a}_1 d = a_{-1}\overline{c}$, *and* $|c| = |d| > 0$ *imply*

$$\begin{bmatrix} a_{-1} \\ d \\ a_{-3} \end{bmatrix} = e^{i\gamma}\begin{bmatrix} \overline{a}_1 \\ \overline{c} \\ \overline{a}_3 \end{bmatrix} \quad \textit{where} \quad e^{i\gamma} = \frac{d}{\overline{c}}.$$

4. Case of $N = 4$. By Proposition 1, the Toeplitz submatrices of order 3, $[a_1, a_4; a_{-1}, a_{-4}]$, and $[a_2, a_3; a_{-2}, a_{-3}]$ are normal. The case of $N = 3$ shows that for some $0 \leq \theta, \omega < 2\pi$,

$$\begin{bmatrix} a_{-1} \\ a_{-4} \end{bmatrix} = e^{i\theta}\begin{bmatrix} \overline{a}_1 \\ \overline{a}_4 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} a_{-1} \\ a_{-4} \end{bmatrix} = e^{i\theta}\begin{bmatrix} a_4 \\ a_1 \end{bmatrix} \tag{13}$$

and

$$\begin{bmatrix} a_{-2} \\ a_{-3} \end{bmatrix} = e^{i\omega}\begin{bmatrix} \overline{a}_2 \\ \overline{a}_3 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} a_{-2} \\ a_{-3} \end{bmatrix} = e^{i\omega}\begin{bmatrix} a_3 \\ a_2 \end{bmatrix}. \tag{14}$$

Furthermore, (1) implies

$$a_1\overline{a}_2 + a_3\overline{a}_4 = a_{-2}\overline{a}_{-1} + a_{-4}\overline{a}_{-3} \tag{15}$$

and

$$a_1\overline{a}_3 + a_2\overline{a}_4 = a_{-3}\overline{a}_{-1} + a_{-4}\overline{a}_{-2}. \tag{16}$$

Here notice one remark. Under conditions (13) and (14), Propositions 3 and 4 show that if one of the eight elements a_i $(-4 \leq i \leq 4)$ is zero, then T_N is either of type I or of type II. Therefore we assume that none of a_i is zero. There are four (essentially two) cases to be considered.

(i) Suppose

$$\begin{bmatrix} a_{-1} \\ a_{-4} \end{bmatrix} = e^{i\theta} \begin{bmatrix} \overline{a}_1 \\ \overline{a}_4 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a_{-2} \\ a_{-3} \end{bmatrix} = e^{i\omega} \begin{bmatrix} \overline{a}_2 \\ \overline{a}_3 \end{bmatrix}. \tag{17}$$

If $\theta = \omega$, then T_N is of type I. Assume $\theta \neq \omega$. By substituting (17) into (15) and (16) we have

$$(1 - e^{i(\omega-\theta)})(a_1\overline{a}_2 - a_{-4}\overline{a}_{-3}) = 0$$

and

$$(1 - e^{i(\omega-\theta)})(a_1\overline{a}_3 - a_{-4}\overline{a}_{-2}) = 0.$$

Since $\theta \neq \omega$, we have $a_1\overline{a}_2 = a_{-4}\overline{a}_{-3}$ and $a_1\overline{a}_3 = a_{-4}\overline{a}_{-2}$. By multiplying these two equalities, we have $a_1a_{-4}\overline{a}_2\overline{a}_{-2} = a_1a_{-4}\overline{a}_3\overline{a}_{-3}$. Thus $a_2a_{-2} = a_3a_{-3}$; hence we have $|a_{-2}| = |a_2| = |a_3| = |a_{-3}|$. By applying Lemma 1, we can see that

$$\begin{bmatrix} a_{-1} \\ a_{-4} \end{bmatrix} = e^{i\theta} \begin{bmatrix} \overline{a}_1 \\ \overline{a}_2 \end{bmatrix},$$

$a_1\overline{a}_3 = a_{-4}\overline{a}_{-2}$, and $|a_2| = |a_{-3}| > 0$ imply

$$\begin{bmatrix} a_{-1} \\ a_{-3} \\ a_{-4} \end{bmatrix} = e^{i\gamma} \begin{bmatrix} a_4 \\ a_2 \\ a_1 \end{bmatrix} \quad \text{where} \quad e^{i\gamma} = \frac{a_{-3}}{a_2}.$$

Similarly by applying Lemma 1 again, we have

$$\begin{bmatrix} a_{-1} \\ a_{-2} \\ a_{-4} \end{bmatrix} = e^{i\delta} \begin{bmatrix} a_4 \\ a_3 \\ a_1 \end{bmatrix} \quad \text{where} \quad e^{i\delta} = \frac{a_{-2}}{a_3}.$$

However it is clear from these two equations that

$$e^{i\gamma} = \frac{a_{-1}}{a_4} = e^{i\delta}.$$

Thus T_N is of type II.

Using arguments similar to those employed in case (i), we can deduce the following.

(ii) Suppose

$$\begin{bmatrix} a_{-1} \\ a_{-4} \end{bmatrix} = e^{i\theta} \begin{bmatrix} a_4 \\ a_1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a_{-2} \\ a_{-3} \end{bmatrix} = e^{i\omega} \begin{bmatrix} a_3 \\ a_2 \end{bmatrix}.$$

Then T_N is either of type I or of type II.

(iii) Suppose

$$\begin{bmatrix} a_{-1} \\ a_{-4} \end{bmatrix} = e^{i\theta} \begin{bmatrix} \overline{a}_1 \\ \overline{a}_4 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a_{-2} \\ a_{-3} \end{bmatrix} = e^{i\omega} \begin{bmatrix} a_3 \\ a_2 \end{bmatrix}. \tag{18}$$

By substituting the first equality of (18) into (15) and (16), we have

$$a_1(\overline{a}_2 - e^{-i\theta}a_{-2}) + \overline{a}_4(a_3 - e^{i\theta}\overline{a}_{-3}) = 0 \tag{19}$$

and

$$a_1(\overline{a}_3 - e^{-i\theta}a_{-3}) + \overline{a}_4(a_2 - e^{i\theta}\overline{a}_{-2}) = 0. \tag{20}$$

By multiplying these two, we have

$$a_1\overline{a}_4|a_2 - e^{i\theta}\overline{a}_{-2}|^2 = a_1\overline{a}_4|a_3 - e^{i\theta}\overline{a}_{-3}|^2.$$

Thus we have

$$|a_2 - e^{i\theta}\overline{a}_{-2}| = |a_3 - e^{i\theta}\overline{a}_{-3}|.$$

If

$$|a_2 - e^{i\theta}\overline{a}_{-2}| = |a_3 - e^{i\theta}\overline{a}_{-3}| = 0,$$

it is clear that T_N is of type I. Suppose

$$|a_2 - e^{i\theta}\overline{a}_{-2}| = |a_3 - e^{i\theta}\overline{a}_{-3}| > 0;$$

then we see $|a_1| = |a_4|$ from (19) or (20). Set $\overline{a}_4 = e^{i\alpha}a_1$, and substitute this into (19) and (20). Then we have

$$\overline{a}_2 - e^{-i\theta}a_{-2} + e^{i\alpha}(a_3 - e^{i\theta}\overline{a}_{-3}) = 0$$

and

$$a_3 - e^{i\theta}\overline{a}_{-3} + e^{i\alpha}(\overline{a}_2 - e^{-i\theta}a_{-2}) = 0.$$

By using the second equality of (18), furthermore, we have

$$(1 - e^{i(\alpha+\theta-\omega)})(\overline{a}_2 - e^{i(\omega-\theta)}a_3) = 0.$$

Thus we have two possibilities: $\alpha + \theta - \omega = 0 \pmod{2\pi}$ or $\overline{a}_2 = e^{i(\omega-\theta)}a_3$.

If $\alpha + \theta - \omega = 0 \pmod{2\pi}$, then

$$\overline{a}_4 = e^{i\alpha}a_1 = e^{i(\omega-\theta)}a_1 = e^{i(\omega-\theta)}e^{i\theta}\overline{a}_{-1} = e^{i\omega}\overline{a}_{-1};$$

hence

$$a_{-1} = e^{i\omega}a_4 \text{ and } a_{-4} = e^{i\theta}\overline{a}_4 = e^{i\theta}e^{i(\omega-\theta)}a_1 = e^{i\omega}a_1.$$

Thus T_N is of type II.

If $\overline{a}_2 = e^{i(\omega-\theta)}a_3$, then

$$\overline{a}_2 = e^{i(\omega-\theta)}e^{-i\omega}a_{-2} = e^{-i\theta}a_{-2};$$

hence

$$a_{-2} = e^{i\theta}\overline{a}_2 \text{ and } a_{-3} = e^{i\omega}a_2 = e^{i\omega}e^{-i(\omega-\theta)}\overline{a}_3 = e^{i\theta}\overline{a}_3.$$

Thus T_N is of type I.

(iv) Suppose

$$\begin{bmatrix} a_{-1} \\ a_{-4} \end{bmatrix} = e^{i\theta}\begin{bmatrix} a_4 \\ a_1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a_{-2} \\ a_{-3} \end{bmatrix} = e^{i\omega}\begin{bmatrix} \overline{a}_2 \\ \overline{a}_3 \end{bmatrix}.$$

In this case, by the same argument as in the previous case (iii) we can show that T_N is either of type I or of type II. □

Before going into the general case, let us observe (Toeplitz) submatrices of T_N of order 5.

LEMMA 2. *Suppose that* $1 \le m < n < \frac{N+1}{2}, 1 \le k < l < \frac{N+1}{2}$, *and*

$$\{m, n\} \cap \{k, l\} = \{r\}, \{m, n\} \cup \{k, l\} \setminus \{r\} = \{p, q\}.$$

If both $[m, n, N+1-n, N+1-m]$ *and* $[k, l, N+1-l, N+1-k]$ *are simultaneously of type* I *(or type* II*), so is* $[p, q, N+1-p, N+1-q]$.

Proof. Let us consider only the type II case. Supposing $m = p, q = l, n = k = r$, for instance, we see that both $[p, r, N+1-r, N+1-p]$ and $[r, q, N+1-q, N+1-r]$ are of type II. Then there are $0 \le \theta, \omega < 2\pi$ such that

$$\begin{bmatrix} a_{-p} \\ a_{-r} \\ a_{-(N+1-r)} \\ a_{-(N+1-p)} \end{bmatrix} = e^{i\theta} \begin{bmatrix} a_{N+1-p} \\ a_{N+1-r} \\ a_r \\ a_p \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a_{-r} \\ a_{-q} \\ a_{-(N+1-q)} \\ a_{-(N+1-r)} \end{bmatrix} = e^{i\omega} \begin{bmatrix} a_{N+1-r} \\ a_{N+1-q} \\ a_q \\ a_r \end{bmatrix},$$

so that

$$a_{-r} = e^{i\theta} a_{N+1-r} \quad \text{and} \quad a_{-(N+1-r)} = e^{i\theta} a_r$$

and

$$a_{-r} = e^{i\omega} a_{N+1-r} \quad \text{and} \quad a_{-(N+1-r)} = e^{i\omega} a_r.$$

Since

$$\begin{bmatrix} a_r & a_{-r} \\ a_{N+1-r} & a_{-(N+1-r)} \end{bmatrix} \neq \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

by assumption (3), we can conclude $\theta = \omega$. Therefore we have

$$\begin{bmatrix} a_{-p} \\ a_{-q} \\ a_{-(N+1-q)} \\ a_{-(N+1-p)} \end{bmatrix} = e^{i\theta} \begin{bmatrix} a_{N+1-p} \\ a_{N+1-q} \\ a_q \\ a_p \end{bmatrix},$$

and consequently $[p, q, N+1-q, N+1-p]$ is of type II. □

THEOREM 2. *Every normal Toeplitz matrix* $T_N = [a_1, \ldots, a_N; a_{-1}, \ldots, a_{-N}]$ *with* $N \ge 5$ *is either of type* I *or of type* II.

Proof. (a) Case of even N. Since all submatrices of order 5 $[1, k, N+1-k, N]$ $(1 < k < \frac{N+1}{2})$ are normal, by Theorem 1 it is of type I or type II. We claim that they are of the same type. To prove this by contradiction, without loss of generality we may assume that there are k_0, k_1 such that $1 < k_0 < k_1 < \frac{N+1}{2}$ and $[1, k_0, N+1-k_0, N]$ is of type I but not of type II while $[1, k_1, N+1-k_1, N]$ is of type II but not of type I.

Again by Theorem 1 $[k_0, k_1, N+1-k_1, N+1-k_0]$ is either of type I or of type II. If it is of type I, by applying Lemma 2 to $[1, k_0, N+1-k_0, N]$ and $[k_0, k_1, N+1-k_1, N+1-k_0]$, we see that $[1, k_1, N+1-k_1, N]$ is of type I, which is a contradiction. In the same way, assuming type II of $[k_0, k_1, N+1-k_1, N+1-k_0]$ leads to a contradiction again. Thus we have proved that all $[1, k, N+1-k, N]$ $(1 < k < \frac{N+1}{2})$ are of the same type. Suppose, for instance, they are of type I. Then for $1 < k < m < \frac{N+1}{2}$

$$\begin{bmatrix} a_{-1} \\ a_{-k} \\ a_{-(N+1-k)} \\ a_{-N} \end{bmatrix} = e^{i\theta} \begin{bmatrix} \overline{a}_1 \\ \overline{a}_k \\ \overline{a}_{N+1-k} \\ \overline{a}_N \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a_{-1} \\ a_{-m} \\ a_{-(N+1-m)} \\ a_{-N} \end{bmatrix} = e^{i\omega} \begin{bmatrix} \overline{a}_1 \\ \overline{a}_m \\ \overline{a}_{N+1-m} \\ \overline{a}_N \end{bmatrix}.$$

Since

$$\begin{bmatrix} a_1 & a_{-1} \\ a_N & a_{-N} \end{bmatrix} \neq \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

by assumption (3), this implies $\theta = \omega$, and consequently

$$a_{-k} = e^{i\theta} \overline{a}_k \quad (1 \le k \le N).$$

This proves that T_N is of type I.

In a similar way we can treat the case of type II.

(b) Case of odd N. Since all submatrices of order 5 $[1, k, N+1-k, N]$ $(1 < k < \frac{N+1}{2})$ and the submatrix of order 4 $[1, \frac{N+1}{2}, N]$ are normal, by Theorem 1 each of them is either of type I or of type II. It is shown in case (a) that all such submatrices of order 5 are of the same type. Suppose that they are of type I. In this case we claim that the submatrix $[1, \frac{N+1}{2}, N]$ is of type I too. Suppose, by contradiction, that $[1, \frac{N+1}{2}, N]$ is of type II but not of type I, while for some $1 < k_1 < N$ the submatrix $[1, k_1, N+1-k_1, N]$ is of type I but not of type II. Consider $[k_1, \frac{N+1}{2}, N+1-k_1]$. This is either of type I or of type II by Theorem 1. Suppose that this is of type I. By applying the same idea as in Lemma 2 to the two submatrices $[k_1, \frac{N+1}{2}, N+1-k_1]$ and $[1, k_1, N+1-k_1, N]$, we can see that $[1, \frac{N+1}{2}, N]$ is of type I, which is a contradiction. Suppose that $[k_1, \frac{N+1}{2}, N+1-k_1]$ is of type II. Then, by applying the same idea as in Lemma 2 again to $[k_1, \frac{N+1}{2}, N+1-k_1]$ and $[1, \frac{N+1}{2}, N]$, we see that $[1, k_1, N+1-k_1, N]$ is of type II, which is also a contradiction. Thus we have that all the submatrices of order 5 $[1, k, N+1-k, N]$ $(1 < k < \frac{N+1}{2})$ and the submatrix of order 4 $[1, \frac{N+1}{2}, N]$ are of the same type, say type I. Then we can see, as in case (a), that T_N itself is of type I. □

REFERENCES

[1] D. R. FARENICK AND W. Y. LEE, *Normal Toeplitz matrices*, 1995, manuscript.

[2] D. R. FARENICK, M. KRUPNIK, N. KRUPNIK, AND W.Y. LEE, *Normal Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 1037–1043.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 4, pp. 1007–1019, October 1996

018

PRECONDITIONING STRATEGIES FOR HERMITIAN TOEPLITZ SYSTEMS WITH NONDEFINITE GENERATING FUNCTIONS*

STEFANO SERRA†

Abstract. In this paper we present new preconditioning techniques for the solution by the preconditioned conjugate gradient (PCG) method of Hermitian Toeplitz systems with real and nondefinite generating functions: actually we extend some results of Chan [*IMA J. Numer. Anal.*, 11 (1991), pp. 333–345] and Di Benedetto, Fiorentino, and Serra [*Comput. Math. Appl.*, 25 (1993), pp. 33–45] proved for positive definite Toeplitz systems.

Moreover we demonstrate some density properties of the spectra of the preconditioned matrices. Finally, we show that the convergence speed of this PCG method is independent of the dimension of the involved matrices.

Key words. linear system, Toeplitz matrix, conjugate gradient, preconditioner

AMS subject classifications. 65F10, 65F15

1. Introduction.

We consider the problem of solving a linear system $A_n(f)\mathbf{x} = \mathbf{b}$ where $A_n(f)$ is an $n \times n$ Hermitian Toeplitz [5] matrix; i.e., $a_{i,j} = a_{i-j}$, $a_k = \bar{a}_{-k}$ for any nonnegative integer k and the values a_k are the Fourier coefficients of an assigned continuous function $f : I \to \mathbf{R}$, $I = [-\pi, \pi]$, that is

$$a_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-\mathbf{i}kx} dx, \quad \mathbf{i}^2 = -1.$$

This kind of matrix arises in many applicative fields [5, 16], such as Markov chains, differential and integral equations, time series analysis, etc., where the efficient solution of very large Toeplitz systems of equations is frequently required.

In some of these applications the Toeplitz matrices are guaranteed to be positive definite, but in other applications, such as eigenfilter problems, harmonic retrieval, and linear prediction, the matrices may also be indefinite [10].

Extensive literature has been devoted to the study of resolution methods with low arithmetic cost which exploit the specific structure of a Toeplitz matrix. So-called "superfast" direct methods [1, 17, 20] compute the solution of a Toeplitz system in $O(n \log^2(n))$ operations (ops), but they are inherently sequential. In the parallel random access machines (PRAM) model of parallel computation, where at each step each processor can perform an arithmetic operation, $O(n)$ steps with n processors are required by these algorithms.

In the latest literature a certain attention has been devoted to the solution of Toeplitz linear systems by means of iterative methods and, in particular, by means of PCG methods and multigrid techniques [13, 14]. Many contributions have concerned devising good preconditioners that allow the solution of the system in a number of iterations independent of n [4], [6]–[9], [11, 12]. The case of $f(x) > 0$ is dealt with in [4, 6, 8, 9]; the case where the function $f(x)$ is nonnegative is analysed in [7, 11, 12, 21].

The algorithms devised in this way are very effective since they yield the solution, within a given precision, by performing just a few fast Fourier transforms (FFTs) [23]. In fact, the cost of a single iteration of the PCG method is determined by the computation of a product of a Toeplitz matrix and a vector and by the inversion

* Received by the editors July 20, 1994; accepted for publication (in revised form) by G. A. Watson December 1, 1995.

† Dipartimento di Informatica, Corso Italia 40, 56100 Pisa, Italy (serra@morse.dm.unipi.it).

of the preconditioner. The first computation can be performed by means of FFT in $O(n \log n)$ ops and a parallel cost of $O(\log n)$ steps with $O(n)$ processors in the parallel PRAM model of computation. The second one has the same cost, provided the preconditioner is chosen in some matrix algebra (such as the circulant, the Hartley, or the tau class) or possesses a band Toeplitz structure.

In this paper we consider the important case where $f(x)$ has a nondefinite sign. It is worth pointing out that Ku and Kuo treated the non-Hermitian case [18, 19] and also, consequently, the nondefinite case, obtaining superlinear methods in the class of PCG methods. While they use the assumption that $|f| \geq m > 0$ and $A_n(f)$ belongs to the Wiener class, however, in this paper we consider the more ill-conditioned case of $f \in \mathbf{C}[I]$ having zeros, and we prove good clustering properties comparable with those of Ku and Kuo. The main idea is the following: by extending the techniques of [7, 12], we transform the original nondefinite system (potentially singular) into an equivalent positive definite (at least nonnegative definite) system whose matrix is somehow related to the Toeplitz structure. We show that, for a wide class of associated functions $f(x)$, our PCG method performs the computation in a number of iterations independent of n, and we prove that each iteration is reduced to perform few FFTs for an overall cost of $O(n \log n)$ ops and $O(\log n)$ steps with $O(n)$ processors in the PRAM model of computation. Moreover, since the generating function $f(x)$ has nondefinite sign, it may occur that $A_n(f)$ is singular for particular values of the size n. Even in this case, however, by applying the CG method to the normal equations in the preconditioned system of **Stage 2** of our algorithm, we obtain an approximation, with a preassigned accuracy, of $A_n^+(f)\mathbf{b}$, i.e., the least squares solution of the proposed problem.

Let us assume that the entries of $A_n(f)$ are given and that the function $f(x)$ is known, say, by means of its formal expression. Here and hereafter we assume that $f(x)$ has zeros $x_1, \ldots, x_m \in [-\pi, \pi]$.

Our method is outlined by the following stages:

Stage 1: Find g such that $g(x_i) = 0$, g is positive elsewhere and the closed set

$$\overline{\mathcal{R}\left(\frac{f}{g}\right)} \stackrel{\text{def}}{=} \left\{ \overline{y : \exists x \neq x_i \text{ such that } y = \frac{f(x)}{g(x)}} \right\}$$

is contained in $[\alpha^-, \beta^-] \bigcup [\alpha^+, \beta^+]$ where $\alpha^- \leq \beta^- < 0 < \alpha^+ \leq \beta^+$; for instance, set $g = |f|$.

Stage 2: Compute the Toeplitz matrix $A_n(g)$, which is Hermitian and positive definite [7], and consider the equivalent non-Hermitian system

$$H_n \mathbf{x} = \hat{\mathbf{b}}$$

where $\hat{\mathbf{b}} = A_n^{-1}(g)\mathbf{b}$ and $H_n = A_n^{-1}(g)A_n(f)$.

Stage 3: Consider the new equivalent system

$$H_n^2 \mathbf{x} = \tilde{\mathbf{b}}$$

where the matrix H_n^2 is associated with a symmetrizable positive definite form and $\tilde{\mathbf{b}} = H_n \hat{\mathbf{b}}$. Solve it by means of the PCG method.

Actually, since $H_n^2 = A_n^{-1}(g)A_n(f)A_n^{-1}(g)A_n(f)$, we can look at the coefficient matrix H_n^2 as the product of $A_n^{-1}(g)$, which is a Hermitian positive definite (HPD)

matrix, and $T_n = A_n(f)A_n^{-1}(g)A_n(f)$, which is an HPD matrix if $A_n(f)$ is nonsingular and is nonnegative definite otherwise.

Therefore, in **Stage 3**, we may apply a PCG method in which T_n is the coefficient matrix of a new equivalent system and $A_n(g)$ is the preconditioner. Of course, the convergence features of this PCG method, that is the number k of iterations needed to reach the solution within a fixed accuracy, are determined by the spectral properties of the matrix H_n^2 or equivalently by the spectral behaviour of H_n.

The main goal of this paper is to prove that k is a constant independent of n. More precisely, we arrive at this result by proving the following facts: the closure of the union $S = \bigcup_{n=1}^{\infty} \sigma(H_n)$ of the spectra of the matrices H_n contains $\overline{\mathcal{R}(f/g)}$, and S is contained in (α^-, β^+). In the case where $f(x)$ and $g(x)$ are rational even functions, we demonstrate that there exists a constant q independent of n such that

$$\sigma(H_n^2) \subset \{\lambda_1^{(n)}, \ldots, \lambda_q^{(n)}\} \bigcup [c^-, c^+],$$

where $c^- = \min\{\alpha^+, |\beta^-|\}$, $c^+ = \max\{|\alpha^-|, \beta^+\}$, $\lambda_1^{(n)} \leq \cdots \leq \lambda_n^{(n)}$, and $\lambda_1^{(n)}, \ldots, \lambda_q^{(n)} \in [0, c^-)$. Therefore, by applying the result of [2], we expect that the conjugate gradient (CG) method, applied to the system $H_n^2 \mathbf{x} = \tilde{\mathbf{b}}$, converges to the solution with a preassigned accuracy ϵ in at most $k + q$ iterations where

$$k = \left\lceil \frac{\log 2/\epsilon + q \log c^+/\lambda_1^{(n)}}{\log 1/\delta} \right\rceil, \quad \delta = \frac{\sqrt{c^+} - \sqrt{c^-}}{\sqrt{c^+} + \sqrt{c^-}}. \tag{1}$$

If ϵ is fixed and $\lambda_1^{(n)} \geq \theta > 0$ or goes to zero *slowly*, then the desired precision is practically reached through a constant number of iterations. In the general case we have $q = o(n)$, but we conjecture that the relation $q = O(1)$ holds for sufficiently regular functions $f(x)$ and $g(x)$.

It is worth pointing out that the search of g in **Stage 1** is easy: if f has only zeros $x_1, \ldots, x_m$ of even orders $2k_1, \ldots, 2k_m$, then we can choose $g_{\min}(x) = \prod_{j=1}^{m} (1 - \cos(x - x_j))^{k_j}$ as the generating function of our preconditioner $A_n(g)$.

In this case g is a nonnegative trigonometric polynomial, $A_n(g)$ is an HPD band Toeplitz matrix, and the function f/g has a range contained in a set of the desired form $[\alpha^-, \beta^-] \bigcup [\alpha^+, \beta^+]$ with $\alpha^- \leq \beta^- < 0 < \alpha^+ \leq \beta^+$. Then we may use a band solver performing only $O(n)$ arithmetic ops [3, 15] and $O(\log(n))$ parallel steps [3].

In the general case it is always possible to choose $g(x) = |f(x)|$, but we lose the band structure of the preconditioner. Therefore, in **Stage 2**, if the function $g = |f|$ has only zeros of *even orders* then the vector $\hat{\mathbf{b}}$ can be calculated in $O(n \log n)$ arithmetic ops by using any PCG method of [7, 11, 12]. We emphasize that the calculation of the solution of a linear system whose coefficient matrix is $A_n(g)$ is a very important task of the proposed algorithm. Actually this kind of linear system occurs in four different computations:

- the calculation of $\hat{\mathbf{b}}$ from $\mathbf{b}$,
- the calculation of $\tilde{\mathbf{b}}$ from $\hat{\mathbf{b}}$,
- the preconditioning step in the PCG method used at **Stage 3**,
- the multiplication of T_n with the search vector at each PCG iteration.

However, in the recent work [22] a simple linear algebra trick is introduced to transform a system $A_n(f)\mathbf{x} = \mathbf{b}$, where f also has zeros of *odd orders*, into a new equivalent system $A_n(\hat{f})\mathbf{x} = \hat{\mathbf{b}}$ in which the new generating function $\hat{f}$ has only zeros of *even orders*. Consequently, our algorithm can be used to deal with the case of

functions f having also zeros of odd orders: we stress that this case may frequently occur when f has a nondefinite sign.

Finally, by using the fact that the product between a Toeplitz matrix and a vector costs $O(n \log n)$ ops and $O(\log n)$ parallel steps with n processors in the PRAM model of computation, it is easy to prove that **Stage 2** costs $O(n \log n)$ ops, provided that the entries of $A_n(g)$ can be computed within this time. Concerning **Stage 3** we recall that its cost is $O(kn \log n)$ ops where k is the number of iterations required by our PCG method to reach the solution with a preassigned accuracy.

The paper is organised as follows: in §2 we prove the theoretical results, and in §3 we perform some numerical experiments which confirm the analysis developed in the other section.

2. Main results. Let $A_n(f)$ be the Hermitian Toeplitz matrix associated with the function $f : I \to \mathbf{R}$ and denote $\|A\|_2$ the Euclidean norm of A. The following result holds as seen in Theorem 2.1.

THEOREM 2.1. *If $f(x)$ and $g(x)$ are continuous functions such that $f(x)$ has nonconstant sign and $g(x)$ is nonnegative, then we have the following:*

(a) *Zero is an accumulation point for $\bigcup_{n=1}^{\infty} \sigma(A_n(f))$.*
(b) *The condition number $\|A\|_2\|A^{-1}\|_2$ of $A = A_n(f)$ tends to ∞, as n tends to ∞, and $A_n(f)$ can be singular for some values of n.*
(c) *$A_n(g)$ is positive definite.*
(d) *$A_n^{-1}(g)A_n(f)$ has eigenvalues in the open set (α^-, β^+) where $\alpha^- = \inf_{x \in I} f/g$ and $\beta^+ = \sup_{x \in I} f/g$.*

Proof. Concerning parts (a) and (b) see the theorem on p. 64 in [16]. For the rest of the theorem, compare Lemma 1 in [7] and Lemma 3.1 and Theorem 3.1 in [12]. □

If $f(x)$ has zeros $x_1, \ldots, x_m$ in the fundamental interval $[-\pi, \pi]$ then we choose $g(x)$ such that $g(x_i) = 0$, g is positive elsewhere and

$$0 < \liminf_{x \to x_i} \left|\frac{f}{g}\right| \leq \limsup_{x \to x_i} \left|\frac{f}{g}\right| < \infty. \tag{2}$$

In this case we have

$$\overline{\mathcal{R}\left(\frac{f}{g}\right)} \subset [\alpha^-, \beta^-] \bigcup [\alpha^+, \beta^+], \tag{3}$$

where $\alpha^- \leq \beta^- = \sup_I \{f(x)/g(x) < 0\} < 0 < \alpha^+ = \inf_I \{f(x)/g(x) > 0\} \leq \beta^+$. Now we are ready for the main result.

THEOREM 2.2. *Let $f(x)$ and $g(x)$ be two continuous functions; suppose that $f(x), g(x)$ satisfy condition (2) and consequently relation (3). Then we have the following:*

(a) *The closure of the union $S = \bigcup_{n=1}^{\infty} \sigma(A_n^{-1}(g)A_n(f))$ of the spectra of $A_n^{-1}(g)A_n(f)$ contains $\overline{\mathcal{R}(f/g)}$.*
(b) *For any $\eta > 0$, $\#\{\sigma(A_n^{-1}(g)A_n(f)) \bigcap [\alpha^-, \beta^- + \eta] \bigcup [\alpha^+ - \eta, \beta^+]\} = n - o(n)$; i.e., most of the eigenvalues of $A_n^{-1}(g)A_n(f)$ belong to the image of f/g.*
(c) *If $f(x)$ and $g(x)$ are even rational functions with respect to e^{ix} we have that for any $\eta > 0$, $\#\{\sigma(A_n^{-1}(g)A_n(f)) \bigcap [\alpha^-, \beta^- + \eta] \bigcup [\alpha^+ - \eta, \beta^+]\} = n - O(1)$.* □

In order to prove this result we need some preliminary lemmas on the inertia of Toeplitz matrices and on the sign of the ratio of the generating functions.

LEMMA 2.3. *Let $f(x), g(x)$ be two continuous functions which satisfy condition* (2); *let α be a value in (β^-, α^+) and $f_t(x) \stackrel{\text{def}}{=} f(x) - tg(x)$ with t a real parameter. Then there exists $\epsilon = \min\{|\alpha^+ - \alpha|, |\beta^- - \alpha|\}$ such that $f_\alpha(x)$ and $f_{\tilde{\alpha}}(x)$ have the same sign for all $|\alpha - \tilde{\alpha}| < \epsilon$.*

Proof. Since α is in $(\beta^-, \alpha^+) \subset \mathbf{R}\backslash\overline{\mathcal{R}(f/g)}$, we find that $f_\alpha(x_i) = 0$ and $f_\alpha(x) \neq 0$ otherwise. Now let us consider $\tilde{\alpha}$ such that $|\alpha - \tilde{\alpha}| < \epsilon$; then we have

$$f_{\tilde{\alpha}}(x) = f_\alpha(x) + (\alpha - \tilde{\alpha})g(x).$$

Moreover $\operatorname{sign}(f_t(x)) = \operatorname{sign}(f_t(x)/g(x))$ for any $x \neq x_i$ and $t \in \mathbf{R}$. So we find

$$\frac{f_{\tilde{\alpha}}(x)}{g(x)} = \frac{f_\alpha(x)}{g(x)} + \alpha - \tilde{\alpha};$$

but the closure of the image of $f_{\tilde{\alpha}}(x)/g(x)$ is $[\alpha^- - \tilde{\alpha}, \beta^- - \tilde{\alpha}] \bigcup [\alpha^+ - \tilde{\alpha}, \beta^+ - \tilde{\alpha}]$ and therefore by choosing ϵ such that

$$\epsilon = \min\{|\alpha^+ - \alpha|, |\beta^- - \alpha|\} \quad \text{if } \alpha \in (\beta^-, \alpha^+),$$

it follows that $f_{\tilde{\alpha}}(x)/g(x)$ has the same sign of $f_\alpha(x)/g(x)$ for all $x \neq x_i$. Since $g(x)$ is positive for any $x \neq x_i$, the lemma is proved. □

Now we are interested in the inertia of a Toeplitz matrix generated by a continuous function $f(x)$. In the following we denote with $\lambda_j(X)$ an eigenvalue of the matrix X.

LEMMA 2.4. *Let $f(x)$ be such that $m\{x \in I : f(x) = 0\} = 0$ where $m(\cdot)$ is the Lebesgue measure in $\mathbf{R}$; then we have the following.*

(a) $$\lim_{n\to\infty} \frac{\#\{j : \lambda_j(A_n(f)) < 0\}}{n} = \frac{m\{x \in I : f(x) < 0\}}{2\pi}.$$

(b) $$\#\{j : \lambda_j(A_n(f)) < 0\} = \left\lceil n\frac{m\{x \in I : f(x) < 0\}}{2\pi} \right\rceil + o(n).$$

Proof. It follows from the theorem on page 64 of [16]. □

In the following we prove a refinement of the former result; that is,

$$\#\{j : \lambda_j(A_n(f)) < 0\} = \left\lceil n\frac{m\{x \in I : f(x) < 0\}}{2\pi} \right\rceil + O(1)$$

in the case where $f(x)$ is a symmetric trigonometric polynomial and subsequently in the case where $f(x)$ is a symmetric rational function.

For any nonnegative integer we define τ_n [3] as the matrix algebra generated by

$$H = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix}$$

whose eigenvalues $\mu_i^{(n)}$ are $2\cos(x_i^{(n)})$, $x_i^{(n)} = \frac{i\pi}{n+1}$. Let $A_n(f)$ be the band symmetric

Toeplitz matrix

$$A_n(f) = \begin{pmatrix} a_0 & a_1 & \cdots & a_k & & & \\ a_1 & \ddots & \ddots & & \ddots & & \\ \vdots & \ddots & & & & & \\ a_k & & & & & & a_k \\ & \ddots & & & & \ddots & \vdots \\ & & & & \ddots & \ddots & a_1 \\ & & & a_k & \cdots & a_1 & a_0 \end{pmatrix}$$

generated by the symmetric trigonometric polynomial $f(x) = a_0 + \sum_{j=1}^{k} a_j \cos(jx)$, $x \in [0, \pi]$, and let $\tilde{A}_n(f)$ be the τ_n matrix whose eigenvalues are $\tilde{\lambda}_i^{(n)} = f(x_i^{(n)})$, $i = 1, \ldots, n$. Then we obtain that [3]

$$A_n(f) = \tilde{A}_n(f) + \begin{pmatrix} b_0 & b_1 & \cdots & b_{k-2} & & & \\ b_1 & & \iddots & & & & \\ \vdots & \iddots & & & & & \\ b_{k-2} & & & & & & b_{k-2} \\ & & & & & \iddots & \vdots \\ & & & & \iddots & & b_1 \\ & & & b_{k-2} & \cdots & b_1 & b_0 \end{pmatrix}, \quad b_i = a_{i+2}.$$

Therefore $A_{n-2k}(f)$ is a principal submatrix of $\tilde{A}_n(f)$, and setting $\delta \in \{0, -1\}$ we have

$$\#\{j : \tilde{\lambda}_j^{(n)} < 0\} = \left\lceil n \frac{m\{x \in [0, \pi] : f(x) < 0\}}{\pi} \right\rceil + \delta.$$

Now we characterize the eigenvalues of Toeplitz matrices generated by symmetric trigonometric polynomials.

THEOREM 2.5. *Let $f(x)$ be a symmetric trigonometric polynomial and $A_n(f)$ the associated $n \times n$ Toeplitz matrix. If $f(x)$ has degree $\hat{k}$ and $\lambda_j^{(n)}$ are the eigenvalues of $A_n(f)$ sorted in nondecreasing order, then we have*

$$\#\{j : \lambda_j^{(n)} < 0\} = \left\lceil n \frac{m\{x \in [0, \pi] : f(x) < 0\}}{\pi} \right\rceil + O(1).$$

Proof. Let us consider $\tilde{A}_{n+2\hat{k}}(f) \in \tau_{n+2\hat{k}}$. By using the results in [3] it is known that $A_n(f)$ is a principal submatrix of $\tilde{A}_{n+2\hat{k}}(f)$ obtained by $\tilde{A}_{n+2\hat{k}}(f)$ deleting the first and the last $\hat{k}$ rows and columns. Now we call $\{\beta_j\}$ the values $\{\tilde{\lambda}_j^{(n+2\hat{k})}\}_{j=1}^{j=n+2\hat{k}}$ sorted in nondecreasing order; therefore, by using the Courant–Fischer minimax theorem [15] we have

$$\beta_j \le \lambda_j^{(n)} \le \beta_{j+2\hat{k}}, \quad j = 1, \ldots, n.$$

Consequently, since

$$\#\{j : \beta_j < 0\} = \left\lceil (n + 2\hat{k}) \frac{m\{x \in [0, \pi] : f(x) < 0\}}{\pi} \right\rceil + \delta,$$

the theorem is proved. □

Finally we generalize the preceding result to the case where $f(x)$ is an even trigonometric rational function.

THEOREM 2.6. *Let $A_n(f)$ be the $n \times n$ Toeplitz matrix generated by the continuous, even, and rational function $f(x)$. Then we have*

$$\#\{j : \lambda_j^{(n)} < 0\} = \left\lceil n \frac{m\{x \in [0,\pi] : f(x) < 0\}}{\pi} \right\rceil + O(1).$$

Proof. Let $f(x) = \hat{N}(x)/\hat{D}(x)$, where $\hat{N}(x)$ and $\hat{D}(x)$ are even trigonometric polynomials and have degrees k_1 and k_2. By the Dickinson characterization [18] there exist polynomials $A(z), B(z), C(z), D(z)$ such that

$$f(x) = \hat{f}(z) = \frac{A(z^{-1})}{B(z^{-1})} + \frac{C(z)}{D(z)},$$

where $z = e^{ix}$, $A(z) = C(z)$, $B(z) = D(z)$, and

$$A(z) = \sum_{j=0}^{q_1} \hat{a}_j z^j, \; B(z) = \sum_{j=0}^{q_2} \hat{b}_j z^j, \quad \hat{a}_j, \hat{b}_j \in \mathbf{R}.$$

Consequently we find that

$$f(x) = \frac{\hat{N}(x)}{\hat{D}(x)} = \frac{\sum_{j=0}^{k_1} \alpha_j \cos(jx)}{\sum_{j=0}^{k_2} \beta_j \cos(jx)},$$

with $k_1 = q_1$ and $k_2 = \max\{q_1, q_2\}$. Moreover it is well known that [18]

$$\Psi_n \stackrel{\text{def}}{=} A_n(B(z^{-1}))A_n(f)A_n(B(z)) = A_n(\hat{N}(x)) + L$$

where L is a matrix of rank r and where r is constant independent of the dimension n and with nonzero entries only in the northwest corner. Now we observe that Ψ_n has the same inertia as $A_n(f(x))$. Moreover, $A_n(\hat{N}(x))$ verifies the hypotheses of the preceding theorem, and therefore we find that

$$\#\{j : \lambda_j(A_n(\hat{N}(x))) < 0\} = \left\lceil n \frac{m\{x \in [0,\pi] : \hat{N}(x) < 0\}}{\pi} \right\rceil + O(1),$$

and the signs of $f(x)$ and $\hat{N}(x)$ are the same because $\hat{D}(x)$ has no zeros and we can assume that it is positive. If we call $\mu_j^{(n+r)}$ the eigenvalues of Ψ_{n+r} sorted in nondecreasing way and $\phi_j^{(n)}$ the eigenvalues of $A_n(\hat{N}(x))$, then we verify that $A_n(\hat{N}(x))$ is a principal submatrix of Ψ_{n+r} and by using the Courant–Fischer minimax theorem we have

$$\phi_j^{(n)} \le \mu_j^{(n+r)} \le \phi_{j+r}^{(n)}$$

and the theorem is proved. □

Now, before we prove Theorem 2.2 we need another subsidiary lemma.

LEMMA 2.7. *The set $\{t \in \mathbf{R}$ such that $m\{x \in I : f_t(x) = 0\} = 0\}$ is dense in $\mathbf{R}$, where the function $f_t(x) = f(x) - tg(x)$ is the function defined in Lemma 2.3.*

Proof. We define $Z_t = \{x \in I : f_t(x) = 0\} = \{x \in I : f/g = t\}$. Clearly $Z_t \subset I$ and $\forall t \neq s$, and we have that $Z_t \cap Z_s$ is an empty set. By contradiction we suppose that there exists an open set (a, b) for which $m(Z_t) > 0$ for any $t \in (a, b)$. Setting $Z = \bigcup_{t\in(a,b)} Z_t$ we find

$$m(Z) = \sum_{t\in(a,b)} m(Z_t) = \infty$$

as we perform a nondenumerable summation. But, since $Z_t \subset I$ for any value t, then we have $Z \subset I$ and, consequently, $m(Z) \leq 2\pi$, which is a contradiction. □

Proof of Theorem 2.2. We want to prove that any $\alpha \in \mathcal{R}(f/g)$ is the limit of a suitable sequence of eigenvalues of the matrices $A_n^{-1}(g)A_n(f)$. If ϵ is fixed, then we are looking for a size n and a real number $\hat{t}$ such that

$$\det(A_n(f) - \hat{t}A_n(g)) = 0 \quad \text{and} \quad |\hat{t} - \alpha| < \epsilon.$$

We want to study the spectra of $(A_n(f) - tA_n(g))_{n\in\mathbf{N}}$, when t lies in the interval $[\alpha - \epsilon, \alpha + \epsilon]$; observe that the associated function is $f_t(x) = f(x) - tg(x)$. For any $x \in I$, we have $f_{\alpha-\epsilon}(x) \geq f_{\alpha+\epsilon}(x)$. Moreover, because of relation (3), there exists $\bar{x} \neq x_i$ such that $g(\bar{x}) > 0$ and $f_\alpha(\bar{x}) = 0$, so that

$$f_{\alpha-\epsilon}(\bar{x}) = \epsilon g(\bar{x}) > 0 > -\epsilon g(\bar{x}) = f_{\alpha+\epsilon}(\bar{x}).$$

It follows that

$$m\{x : f_{\alpha-\epsilon}(x) < 0\} < m\{x : f_{\alpha+\epsilon}(x) < 0\} \tag{4}$$

where $m(\cdot)$ indicates the Lebesgue measure. Now we set

$$C_n^-(-\epsilon) = \#\{\lambda_i(A_n(f) - (\alpha - \epsilon)A_n(g)) < 0,\ i = 1, \ldots, n\}, \tag{5}$$

$$C_n^-(\epsilon) = \#\{\lambda_i(A_n(f) - (\alpha + \epsilon)A_n(g)) < 0,\ i = 1, \ldots, n\}, \tag{6}$$

and we consider the following classical result of Grenander and Szegö [16]: if $k(x)$ is a continuous function defined in I and $A_n(k)$ is the associated class of Hermitian Toeplitz matrices, then if $m\{x \in I : k(x) = 0\} = 0$,

$$\lim_{n\to\infty} \frac{\#\{\lambda_i(A_n(k)) < 0,\ i = 1, \ldots, n\}}{n} = m\{x \in I : k(x) < 0\}.$$

Therefore by recalling equation (4) we have

$$\lim_{n\to\infty} C_n^-(\epsilon) - C_n^-(-\epsilon) = \infty. \tag{7}$$

Notice that if $m\{x \in I : f_{\alpha-\epsilon}(x) = 0\} > 0$ or $m\{x \in I : f_{\alpha+\epsilon}(x) = 0\} > 0$ then the former equation can be false. However, by using the last lemma we can always find a value $0 < \epsilon^* < \epsilon$ such that $m\{x \in I : f_{\alpha-\epsilon^*}(x) = 0\} = m\{x \in I : f_{\alpha+\epsilon^*}(x) = 0\} = 0$ and consequently (7) holds with $\epsilon = \epsilon^*$.

In particular, by a continuity argument it follows that for a suitable size n there is an eigenvalue $\lambda(t)$ of $A_n(f) - tA_n(g)$ that is negative for $t = \alpha + \epsilon$ and positive for $t = \alpha - \epsilon$. Hence we find $\hat{t} \in (\alpha - \epsilon, \alpha + \epsilon)$ for which $A_n(f) - \hat{t}A_n(g)$ is singular. Therefore part (a) is proved.

To prove the second and the third parts of Theorem 2.2, we take $\alpha = (\beta^- + \alpha^+)/2$, so the function $f_t(x)$ vanishes if and only if $x \in \{x_1, \ldots, x_m\}$ and for all $t \in (\beta^-, \alpha^+)$.

Moreover, by taking $\epsilon = (\beta^- - \alpha^+)/2 - \eta$, $f_t(x)$ keeps the sign for all $t \in [\alpha - \epsilon, \alpha + \epsilon]$ (see Lemma 2.3), and it follows that

$$m(x \in I : f_{\alpha-\epsilon}(x) < 0) = m(x \in I : f_{\alpha+\epsilon}(x) < 0). \tag{8}$$

It is worth pointing out that we can choose $\eta = 0$ if $Z_{\beta+}$ and $Z_{\alpha-}$ are zero-measure sets; otherwise η has to be positive but it can be as small as we like.

Therefore, by means of the results of the former lemmas we determine that the inertia of $A_n(f) - (\alpha - \epsilon)A_n(g)$ is almost the same as $A_n(f) - (\alpha + \epsilon)A_n(g)$; i.e.,

$$\begin{aligned} &\#\{\lambda_i(A_n(f) - (\alpha - \epsilon)A_n(g)) < 0,\ i = 1, \ldots, n\} \\ &= \#\{\lambda_i(A_n(f) - (\alpha + \epsilon)A_n(g)) < 0,\ i = 1, \ldots, n\} + w, \end{aligned} \tag{9}$$

where $w = O(1)$ if $f(x)$ and $g(x)$ are symmetric rational functions (see Theorem 2.6) and $w = o(n)$ in the general case. But $A_n^{-1/2}(g)$ is positive definite and therefore the inertia of $A_n(f) - tA_n(g)$ coincides with that of $A_n^{-1/2}(g)A_n(f)A_n^{-1/2}(g) - tI$ for any $t \in \mathbf{R}$. The latter matrix is similar to $A_n^{-1}(g)A_n(f) - tI$, and so using (9) we have

$$\operatorname{sign}\{\lambda_i(A_n(f) - (\alpha - \epsilon)A_n(g))\} = \operatorname{sign}\{\lambda_i(A_n(f) - (\alpha + \epsilon)A_n(g))\}$$

for all $i \in J$ where $\#J = n - w$. Moreover, $\lambda_i(A_n(f) - (\alpha + \epsilon)A_n(g)) = \lambda_i(A_n(f) - (\alpha - \epsilon)A_n(g)) - 2\epsilon$. Consequently $A_n^{-1}(g)A_n(f) - \alpha I$ has at most only w eigenvalues in $(-\epsilon, \epsilon)$; that is, $A_n^{-1}(g)A_n(f)$ has at most only w eigenvalues in $(\alpha - \epsilon, \alpha + \epsilon)$. Since $(\alpha - \epsilon, \alpha + \epsilon)$ coincides with $(\beta^- + \eta, \alpha^+ - \eta)$ the theorem is proved with $\eta = 0$ if $Z_{\beta-}$ and $Z_{\alpha+}$ are zero-measure sets and with $\eta > 0$ but as small as we like otherwise. □

3. Numerical experiments. In this section we present five examples of functions $f(x)$ with the associated functions $g(x)$. In the first four cases we have chosen functions having $x_0 = 0$ as a unique zero in $(-\pi, \pi)$. When f has a unique zero in I this choice has no loss of generality since it is possible to prove (see [7, Lem. 2]) that for all $x_0 \in I, n \in \mathbf{N}$ we have that $A_n(f(x))$ is similar to $A_n(f(x + x_0))$. More precisely, we have

$$\begin{aligned} &A_n(f(x + x_0)) \\ &= \operatorname{Diag}(1, e^{\mathbf{i}x_0}, \ldots, e^{\mathbf{i}(n-1)x_0})A_n(f(x))\operatorname{Diag}(1, e^{-\mathbf{i}x_0}, \ldots, e^{-\mathbf{i}(n-1)x_0}). \end{aligned}$$

In this way it is possible to consider only continuous functions $f(x)$ such that $f(0) = 0$ and $f(-x)f(x) < 0$ for all $x \in (0, \pi]$. For the last example, we have chosen a function having two distinct zeros.

We considered matrices of size $n = 16$ and $n = 64$. From all the numerical experiments performed with MATLAB we observe that the distribution of the eigenvalues of the preconditioned matrices confirms the theoretical results of §2.

Example 1. Let

$$f(x) \equiv x = \sum_{k=1}^{\infty} \frac{\mathbf{i}(-1)^k}{k}(e^{\mathbf{i}kx} - e^{-\mathbf{i}kx}), \quad x \in I,$$

and choose $A_n(g)$ generated by

$$g(x) \equiv |x| = \frac{\pi}{2} - \sum_{k=1}^{\infty} \frac{1 - (-1)^k}{\pi k^2}(e^{\mathbf{i}kx} + e^{-\mathbf{i}kx}), \quad x \in I.$$

According to Theorem 2.2 we expect that the eigenvalues of $H_n = A_n^{-1}(g)A_n(f)$ form two clusters around -1 and 1 since $f/g = \text{sign}(x)$. For $n = 16$ we have

$$\sigma(H_n) = \{\pm 1.000 \text{ (4 times)}, \pm 0.9997, \pm 0.9946, \pm 0.9287, \pm 0.4773\}. \tag{10}$$

For $n = 64$ we have

$$\sigma(H_n) = \{\pm 1.000 \text{ (27 times)}, \pm 0.9995, \pm 0.9963, \pm 0.9737, \pm 0.8470, \pm 0.3830\}. \tag{11}$$

Example 2. Let

$$f(x) \equiv \text{sign}(x)x^2 = \sum_{k=1}^{\infty} \frac{\mathbf{i}}{\pi k^2}\left((-1)^k\pi^2 + \frac{2}{k^2}(1+(-1)^{(k+1)}\right) (e^{\mathbf{i}kx} - e^{-\mathbf{i}kx}), \quad x \in I.$$

We propose two different functions g_1 and g_2:

$$g_1(x) \equiv x^2 = \frac{\pi^2}{3} + 2\sum_{k=1}^{\infty} \frac{(-1)^k}{k^2}(e^{\mathbf{i}kx} + e^{-\mathbf{i}kx}), \quad x \in I,$$

$$g_2(x) = 2 - 2\cos(x), \quad x \in I.$$

According to the results of the previous section, we expect that $\sigma(H_n) = \sigma(A_n^{-1}(g_1)A_n(f))$ forms two clusters around -1 and 1 since $f/g_1 = \text{sign}(x)$. For $n = 16$ we have

$$\sigma(H_n) = \{\pm 1.000 \text{ (4 times)}, \pm 0.9998, \pm 0.9961, \pm 0.9412, \pm 0.500\}. \tag{12}$$

For $n = 64$ we have

$$\sigma(H_n) = \{\pm 1.000 \text{ (27 times)}, \pm 0.9997, \pm 0.9972, \pm 0.9787, \pm 0.8640, \pm 0.4002\}. \tag{13}$$

In the case of $H_n = A_n^{-1}(g_2)A_n(f)$ we expect (Theorem 2.2) that most of the eigenvalues belong to $\overline{\text{R}(f/g_2)} = [-\pi^2/4, -1] \bigcup [1, \pi^2/4]$. For $n = 16$ we obtain

14	eigenvalues in $[-\pi^2/4, -1] \bigcup [1, \pi^2/4]$,
2	eigenvalues $= \pm 0.7078$ in $(-1, 1)$.

For $n = 64$ we have

60	eigenvalues in $[-\pi^2/4, -1] \bigcup [1, \pi^2/4]$,
2	eigenvalues $= \pm 0.9938$ in a small neighbourhood of -1 and 1, respectively, and
2	eigenvalues $= \pm 0.5698$ in $(-1, 1)$.

Example 3. Let

$$f(x) \equiv e^x - 1 = \sum_{k=-\infty}^{\infty} \frac{(-1)^k(e^{\pi} - e^{-\pi})}{2\pi(1+k^2)} (1 + \mathbf{i}k)e^{\mathbf{i}kx} - 1, \quad x \in I,$$

and

$$g(x) \equiv |e^x - 1| = \sum_{k=-\infty}^{\infty} t_k + \frac{1+\mathbf{i}k}{2\pi(1+k^2)} ((e^{\pi} - e^{-\pi})(-1)^k - 2)e^{\mathbf{i}kx}, \quad x \in I,$$

where t_k is $2\mathbf{i}/\pi k$ if k is odd and 0 elsewhere.

According to Theorem 2.2 we expect two clusters around -1 and 1 for the spectrum of $H_n = A_n^{-1}(g)A_n(f)$. For $n = 16$ we have

$$\sigma(H_n) = \{\pm 1.000 \text{ (4 times)}, 0.9950, -0.9987, 0.9741, -0.9740, 0.6755, -0.6842, 0.3395, -0.3384\}. \tag{14}$$

For $n = 64$ we have

$$\sigma(H_n) = \{1.000 \text{ (27 times)}, -1 \text{ (26 times)}, 0.9999, -0.9998, 0.9993, -0.9978, 0.9950, -0.9824, 0.9710, -0.8764, 0.4212, -0.4076\}. \tag{15}$$

Example 4. Let

$$f(x) \equiv x^3 = \sum_{k=1}^{\infty} \frac{\mathbf{i}(-1)^k}{k} \left(\pi^2 - \frac{6}{k^2}\right)(e^{\mathbf{i}kx} - e^{-\mathbf{i}kx}),$$

and

$$g(x) \equiv |x|(2 - 2\cos(x)) = \sum_{k=-\infty}^{\infty} c_k e^{\mathbf{i}kx},$$

where $c_0 = \pi - 2a_1(|x|)$, $c_j = 2a_j(|x|) - a_{j-1}(|x|) - a_{j+1}(|x|) = c_{-j}$, and $a_j(|x|)$ are the Fourier coefficients of the function $|x|$ shown in the first example.

According to the theoretical results, we expect (Theorem 2.2) that most of the eigenvalues $H_n = A_n^{-1}(g)A_n(f)$ belong to $[-\pi^2/4, -1] \bigcup [1, \pi^2/4]$, which is the closure of the image of f/g. For $n = 16$ we obtain

14	eigenvalues in $[-\pi^2/4, -1] \bigcup [1, \pi^2/4]$,
2	eigenvalues $= \pm 0.7409$ in $(-1, 1)$.

For $n = 64$ we have

60	eigenvalues in $[-\pi^2/4, -1] \bigcup [1, \pi^2/4]$,
2	eigenvalues $= \pm 0.9980$ in a small neighbourhood of -1 and 1, respectively, and
2	eigenvalues $= \pm 0.5937$ in $(-1, 1)$.

It is very interesting to remark that f/g_2 in Example 2 coincides with f/g in this example, and, as a consequence, we have that the behaviours of the spectra of the related matrices H_n, $n = 16, 64$ are practically the same.

Example 5. Let

$$f(x) \equiv x\left(x - \frac{\pi}{2}\right) = \sum_{k=-\infty}^{\infty} a_k(f)e^{\mathbf{i}kx},$$

$$a_k(f) = a_k(x^2) - \frac{\pi}{2}a_k(x)$$

be a function having two zeros in the fundamental interval I.

Let

$$g(x) \equiv \left| x\left(x - \frac{\pi}{2}\right)\right| = \sum_{k=-\infty}^{\infty} c_k e^{\mathrm{i}kx},$$

where the values c_k are obtained by the Fourier coefficients of f in the following way:

$$\begin{aligned} c_0 &= \frac{17\pi^2}{48}, \\ r_k &= \frac{\pi}{2k}\mathrm{i}e^{-\mathrm{i}k\pi/2} + \frac{1}{k^2}\left(e^{-\mathrm{i}k\pi/2} - 1\right), \ k \neq 0, \\ s_k &= \frac{\pi^2}{4k}\mathrm{i}e^{-\mathrm{i}k\pi/2} - \frac{2}{k}\mathrm{i}r_k, \ k \neq 0, \\ c_k &= a_k(f) - \frac{1}{\pi}s_k + \frac{1}{2}r_k, \ k \neq 0. \end{aligned}$$

According to Theorem 2.2, we expect two clusters around -1 and 1 for the spectrum of $H_n = A_n^{-1}(g)A_n(f)$. For $n = 16$ we have

$$\begin{aligned} \sigma(H_n) = \{&1.000 \text{ (3 times)}, 0.9999 \text{ (6 times)}, 0.9993, 0.9863, \\ &0.7947, -0.9999, -0.9967, -0.9244, -0.2314\}. \end{aligned} \tag{16}$$

For $n = 64$ we have

$$\begin{aligned} \sigma(H_n) = \{&1.000 \text{ (35 times)}, -1 \text{ (4 times)}, \\ &0.9999 \text{ (9 times)}, -0.9999 \text{ (7 times)}, 0.9992, \\ &-0.9997, 0.9933, -0.9972, 0.9467, -0.9750, \\ &0.6624, -0.8199, -0.1723\}. \end{aligned} \tag{17}$$

We remark that the interval where $f/g = -1$ is small with respect to I, and, in fact, the number of the negative eigenvalues of H_n close to -1 is less than the number of those close to 1. Moreover, it is worth pointing out that the presence of two zeros causes a partial deterioration of the clustering property of the spectrum of H_n with respect to the case where the generating function f has a unique zero.

In cases (10)–(15), f/g is the function sign(x) (because $g(x) = |f(x)|$ and $f(0) = 0$), and it is interesting to compare these spectra with the spectrum of $A_n(\mathrm{sign}(x))$. The similarities are very deep: for $n = 16$ we have

$$\sigma(A_n(\mathrm{sign}(x))) = \{\pm 1.000 \text{ (4 times)}, \pm 0.9995, \pm 0.9913, \pm 0.9013, \pm 0.4294\}.$$

For $n = 64$ we have

$$\begin{aligned} \sigma(A_n(\mathrm{sign}(x))) = \{&\pm 1.000 \text{ (26 times)}, \pm 0.9999, \pm 0.9993, \pm 0.9945, \\ &\pm 0.9636, \pm 0.8122, \pm 0.3487\}. \end{aligned}$$

REFERENCES

[1] G. Ammar and W. Gragg, *Superfast solution of real positive definite Toeplitz systems*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 61–76.

[2] O. Axelsson and G. Lindskog, *On the rate of convergence of the preconditioned conjugate gradient method*, Numer. Math., 48 (1988), pp. 499–523.

[3] D. Bini and M. Capovani, *Spectral and computational properties of band symmetric Toeplitz matrices*, Linear Algebra Appl., 52 (1983), pp. 99–126.

[4] D. Bini and F. Di Benedetto, *A new preconditioner for the parallel solution of positive definite Toeplitz systems*, Proc. ACM Symposium on Parallel Algorithms and Architecture, Crete, Greece, 1990, pp. 220–223.

[5] J. R. Bunch, *Stability of methods for solving Toeplitz systems of equations*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 349–364.

[6] R. H. Chan, *Circulant preconditioners for Hermitian Toeplitz systems*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 542–550.

[7] ———, *Toeplitz preconditioners for Toeplitz systems with nonnegative generating functions*, IMA J. Numer. Anal., 11 (1991), pp. 333–345.

[8] R. H. Chan and G. Strang, *Toeplitz equations by conjugate gradients with circulant preconditioner*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 104–119.

[9] T. F. Chan, *An optimal circulant preconditioner for Toeplitz systems*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 766–771.

[10] T. F. Chan and P. C. Hansen, *A look-ahead Levinson algorithm for indefinite Toeplitz systems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 491–506.

[11] F. Di Benedetto, *Analysis of preconditioning techniques for ill-conditioned Toeplitz matrices*, SIAM J. Sci. Comput., 16 (1995), pp. 682–697.

[12] F. Di Benedetto, G. Fiorentino, and S. Serra, *C. G. preconditioning for Toeplitz matrices*, Comput. Math. Appl., 25 (1993), pp. 33–45.

[13] G. Fiorentino and S. Serra, *Multigrid methods for Toeplitz matrices*, Calcolo, 28 (1991), pp. 283–305.

[14] ———, *Multigrid methods for symmetric block Toeplitz matrices with nonnegative generating functions*, SIAM J. Sci. Comput., 17 (1996), to appear.

[15] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[16] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*, 2nd edition, Chelsea, New York, 1984.

[17] F. De Hoog, *On the solution of Toeplitz systems*, Linear Algebra Appl., 88/89 (1987), pp. 123–138.

[18] T. K. Ku and C. C. J. Kuo, *Mimimum phase LU factorization*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 1470–1487.

[19] ———, *Spectral properties of preconditioned rational Toeplitz matrices: The nonsymmetric case*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 521–542.

[20] B. Musicus, *Levinson and Fast Cholesky Algorithms for Toeplitz and Almost Toeplitz Matrices*, Tech. Rep., Research Laboratory of Electronics, MIT, Cambridge, MA, 1981.

[21] S. Serra, *Preconditioning techniques for ill-conditioned block Toeplitz systems with nonnegative generating functions*, BIT, 34-4 (1994), pp. 579–594.

[22] ———, *New PCG Based Algorithms for the Solution of Symmetric Toeplitz Systems with L^1 Generating Functions*, Tech. Rep. nr. 10, University of Calabria, Cosenza, Italy, 1995.

[23] 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. 17, No. 4, pp. 1020–1024, October 1996

019

INTERVAL *P*-MATRICES*

JIŘÍ ROHN† AND GEORG REX‡

Abstract. A characterization of interval P-matrices is given. The result implies that a symmetric interval matrix is a P-matrix if and only if it is positive definite (although nonsymmetric matrices may be involved). As a consequence it is proved that the problem of checking whether a symmetric interval matrix is a P-matrix is NP-hard.

Key words. interval matrix, P-matrix, positive definiteness, NP-hardness

AMS subject classifications. 15A48, 65G10, 68Q25

1. Introduction. As is well known, an $n \times n$ matrix A is called a P-matrix if all its principal minors are positive. P-matrices play an important role in several areas, e.g., in the linear complementarity theory, since they guarantee existence and uniqueness of the solution of a linear complementarity problem (see Murty [6]).

A basic characterization of P-matrices was given by Fiedler and Pták [3]: A is a P-matrix if and only if for each $x \in R^n, x \neq 0$ there exists an i such that $x_i(Ax)_i > 0$ holds. This result immediately implies that a *symmetric* matrix A is a P-matrix if and only if it is positive definite (Wilkinson [13]). In fact, if A is positive definite, then for each $x \neq 0$, from $\sum_i x_i(Ax)_i = x^T Ax > 0$ it follows that $x_i(Ax)_i > 0$ for some i; hence A is a P-matrix. Conversely, if A is a P-matrix, then all of its leading principal minors are positive; hence it is positive definite in view of the Sylvester determinant criterion [6].

In this paper we focus our attention on interval P-matrices. An interval matrix

$$A^I = [\underline{A}, \overline{A}] = \{A;\ \underline{A} \leq A \leq \overline{A}\},$$

where $\underline{A}$ and $\overline{A}$ are $n \times n$ matrices satisfying $\underline{A} \leq \overline{A}$ (componentwise), is said to be a P-matrix if each $A \in A^I$ is a P-matrix. In §2 we introduce a finite set of matrices A_z in A^I (whose cardinality is at most 2^{n-1}) such that A^I is a P-matrix if and only if all the matrices A_z are P-matrices (Theorem 2.3). In view of a similar characterization of positive definiteness of A^I via the matrices A_z (Theorem 2.4), it is then proved in §3 that a symmetric interval matrix A^I (i.e., with symmetric bounds $\underline{A}$, $\overline{A}$) is a P-matrix if and only if it is positive definite (Theorem 3.2). This is a generalization of the above result for real symmetric matrices, but it is not a simple consequence of it since here nonsymmetric matrices may be involved. As a consequence of this result we prove that the problem of checking whether a symmetric interval matrix is a P-matrix is NP-hard (Theorem 3.4). This result shows that the exponential number of test matrices A_z used in the necessary and sufficient condition of Theorem 2.3 is highly unlikely to be essentially reducible.

* Received by the editors February 6, 1995; accepted for publication (in revised form) by J. Kautsky September 16, 1995. This work was supported in part by the Charles University Grant Agency under grant GAUK 237.

† Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, Czech Republic (rohn@uivt.cas.cz) and Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (rohn@ms.mff.cuni.cz). Part of the work was done during the author's stay at the Center of Theoretical Sciences of the University of Leipzig.

‡ Institute of Mathematics, University of Leipzig, Augustusplatz 10-11, D-04109 Leipzig, Germany (rex@mathematik.uni-leipzig.d400.de).

2. Characterizations. Let us introduce an auxiliary set

$$Z = \{z \in R^n;\ z_j \in \{-1, 1\} \text{ for } j = 1, \dots, n\},$$

i.e., the set of all ± 1-vectors. The cardinality of Z is obviously 2^n. For an interval matrix

$$A^I = [\underline{A}, \overline{A}],$$

we define matrices $A_z, z \in Z$ by

$$(A_z)_{ij} = \tfrac{1}{2}(\underline{A}_{ij} + \overline{A}_{ij}) - \tfrac{1}{2}(\overline{A}_{ij} - \underline{A}_{ij})z_i z_j$$

$(i, j = 1, \dots, n)$. Clearly, $(A_z)_{ij} = \underline{A}_{ij}$ if $z_i z_j = 1$ and $(A_z)_{ij} = \overline{A}_{ij}$ if $z_i z_j = -1$. Hence $A_z \in A^I$ for each $z \in Z$, and the number of mutually different matrices A_z is at most 2^{n-1} (since $A_{-z} = A_z$ for each $z \in Z$) and equal to 2^{n-1} if $\underline{A} < \overline{A}$. The properties in question (P-property and positive definiteness) will be formulated below in terms of the finite set of matrices $A_z, z \in Z$. For a vector $x \in R^n$, let us define its sign vector

$$z = \operatorname{sgn} x$$

by

$$z_i = \begin{cases} 1 & \text{if } x_i \geq 0, \\ -1 & \text{if } x_i < 0 \end{cases}$$

$(i = 1, \dots, n)$ so that $\operatorname{sgn} x \in Z$. For a matrix $A = (A_{ij})$ we introduce its absolute value by $|A| = (|A_{ij}|)$; a similar notation also applies to vectors.

The basic property of the matrices $A_z, z \in Z$, is summed up in the following auxiliary result; notice that no assumptions on A^I are made.

THEOREM 2.1. *Let A^I be an $n \times n$ interval matrix, $x \in R^n$, and let $z = \operatorname{sgn} x$. Then for each $A \in A^I$ and each $i \in \{1, \dots, n\}$ we have*

$$x_i(Ax)_i \geq x_i(A_z x)_i. \tag{1}$$

Proof. Let $A \in A^I$ and $i \in \{1, \dots, n\}$. Then

$$|x_i(Ax)_i - x_i((\tfrac{1}{2}(\underline{A} + \overline{A}))x)_i| = |x_i((A - \tfrac{1}{2}(\underline{A} + \overline{A}))x)_i|$$

$$\leq |x_i|(|A - \tfrac{1}{2}(\underline{A} + \overline{A})| \cdot |x|)_i \leq |x_i|(\tfrac{1}{2}(\overline{A} - \underline{A})|x|)_i;$$

hence

$$x_i(Ax)_i \geq x_i((\tfrac{1}{2}(\underline{A} + \overline{A}))x)_i - |x_i|(\tfrac{1}{2}(\overline{A} - \underline{A})|x|)_i.$$

Since $z = \operatorname{sgn} x$, we have $|x_j| = z_j x_j$ for each j; hence

$$\begin{aligned} x_i(Ax)_i &\geq \sum_j (\tfrac{1}{2}(\underline{A}_{ij} + \overline{A}_{ij}) - \tfrac{1}{2}(\overline{A}_{ij} - \underline{A}_{ij})z_i z_j)x_i x_j \\ &= \sum_j (A_z)_{ij} x_i x_j = x_i(A_z x)_i, \end{aligned}$$

which concludes the proof. □

As the first consequence of this result, we prove a Fiedler–Pták type characterization of interval P-matrices. Notice that the inequality holds "uniformly" in Theorem 2.2.

THEOREM 2.2. *An interval matrix A^I is a P-matrix if and only if for each $x \in R^n$, $x \neq 0$ there exists an $i \in \{1, \dots, n\}$ such that*

$$x_i(Ax)_i > 0 \tag{2}$$

holds for each $A \in A^I$.

Proof. If (2) holds, then each $A \in A^I$ is a P-matrix by the Fiedler–Pták theorem. Conversely, let A^I be a P-matrix and let $x \neq 0$. Put $z = \operatorname{sgn} x$; then A_z is a P-matrix. Hence by the Fiedler–Pták theorem we have $x_i(A_z x)_i > 0$ for some i. Then (1) implies $x_i(Ax)_i \geq x_i(A_z x)_i > 0$ for each $A \in A^I$, and we are done. □

The characterization in Theorem 2.3, however, turns out to be much more useful.

THEOREM 2.3. *A^I is a P-matrix if and only if each $A_z, z \in Z$, is a P-matrix.*

Proof. If A^I is a P-matrix, then each $A_z \in A^I$ is obviously also a P-matrix. Conversely, let each $A_z, z \in Z$, be a P-matrix. Let $x \in R^n$, $x \neq 0$, and let $z = \operatorname{sgn} x$. Since A_z is a P-matrix, there exists an i with $x_i(A_z x)_i > 0$. Then from Theorem 2.1 we obtain $x_i(Ax)_i \geq x_i(A_z x)_i > 0$ for each $A \in A^I$; hence A^I is a P-matrix by Theorem 2.2. □

Another finite characterization of interval P-matrices, formulated in different terms, was proved by Białas and Garloff [1].

In keeping with the terminology introduced for P-matrices, an interval matrix A^I is said to be positive definite if each $A \in A^I$ is positive definite (i.e., satisfies $x^T Ax > 0$ for each $x \neq 0$). The following theorem was proved in [9, Thm. 2]. We give here another proof of this result to make the paper self-contained and to demonstrate that it is a simple consequence of Theorem 2.1.

THEOREM 2.4. *A^I is positive definite if and only if each $A_z, z \in Z$, is positive definite.*

Proof. The "only if" part is obvious since $A_z \in A^I$ for each $z \in Z$. To prove the "if" part, take $A \in A^I$ and $x \in R^n, x \neq 0$. For $z = \operatorname{sgn} x$ from Theorem 2.1 we have

$$x_i(Ax)_i \geq x_i(A_z x)_i$$

for each i; hence

$$x^T Ax = \sum_i x_i(Ax)_i \geq \sum_i x_i(A_z x)_i = x^T A_z x > 0$$

so that A is positive definite. Thus, by definition, A^I is positive definite. □

The last two theorems reveal that both the P-property and the positive definiteness of interval matrices are characterized by the same finite subset of matrices $A_z \in A^I$, $z \in Z$. This relationship will become even more apparent in the case of symmetric interval matrices, which we shall consider in the next section.

3. Symmetric interval matrices. For an interval matrix $A^I = [\underline{A}, \overline{A}]$, define an associated interval matrix A_s^I by

$$A_s^I = [\tfrac{1}{2}(\underline{A} + \underline{A}^T), \tfrac{1}{2}(\overline{A} + \overline{A}^T)].$$

A^I is called *symmetric* if $A^I = A^I_s$, which is clearly the case if and only if both $\underline{A}$ and $\overline{A}$ are symmetric. Hence, A^I_s is always a symmetric interval matrix. The relationship between positive definiteness and P-property is provided by Theorem 3.1.

THEOREM 3.1. *A^I is positive definite if and only if A^I_s is a P-matrix.*

Proof. For each $z \in Z$, let us denote by A^s_z the matrix A_z for A^I_s; i.e.,

$$(A^s_z)_{ij} = \tfrac{1}{4}(\underline{A}_{ij} + \underline{A}_{ji} + \overline{A}_{ij} + \overline{A}_{ji}) - \tfrac{1}{4}(\overline{A}_{ij} + \overline{A}_{ji} - \underline{A}_{ij} - \underline{A}_{ji})z_i z_j$$

$(i, j = 1, \ldots, n)$. Then A^s_z is symmetric and a direct computation shows that

$$x^T A^s_z x = x^T A_z x \tag{3}$$

holds for each $x \in R^n$. Now, if A^I is positive definite, then each $A_z, z \in Z$, is positive definite. Therefore, each A^s_z is positive definite due to (3), so that A^s_z is a P-matrix; hence A^I_s is a P-matrix by Theorem 2.3. Conversely, if A^I_s is a P-matrix, then each $A^s_z, z \in Z$, is a P-matrix. Therefore, it is positive definite due to its symmetry; hence each $A_z, z \in Z$, is positive definite by (3) and A^I is positive definite by Theorem 2.4. □

Our main result on symmetric interval matrices is now obtained as a simple consequence of Theorem 3.1.

THEOREM 3.2. *A symmetric interval matrix A^I is a P-matrix if and only if it is positive definite.*

Proof. The result follows immediately from Theorem 3.1 since a symmetric interval matrix A^I satisfies $A^I = A^I_s$ by definition. □

At the beginning of the Introduction we showed that a real symmetric matrix is a P-matrix if and only if it is positive definite. The result of Theorem 3.2 sounds alike, but it is not a simple consequence of the real case since here nonsymmetric matrices may be involved. In fact, it can be seen immediately that a symmetric interval matrix $A^I = [\underline{A}, \overline{A}]$ contains nonsymmetric matrices if and only if $\underline{A}_{ij} < \overline{A}_{ij}$ holds for some $i \neq j$.

An interval matrix A^I is called regular (cf. Neumaier [7]) if each $A \in A^I$ is nonsingular. The following result shows that for symmetric interval matrices the P-property is preserved by regularity. Several other results of this type are summed up in [10].

THEOREM 3.3. *A symmetric interval matrix A^I is a P-matrix if and only if it is regular and contains at least one symmetric P-matrix.*

Proof. A symmetric interval P-matrix A^I is regular (each $A \in A^I$ has a positive determinant) and contains a symmetric P-matrix $\underline{A}$. If A^I is regular and contains a symmetric P-matrix A_0, then A_0 is positive definite; hence A^I is positive definite by Theorem 3 in [9], which in light of Theorem 3.2 means that A^I is a P-matrix. □

Another relationship between regularity and the P-property of interval matrices was established in [8, Thm. 5.1, assert. (B1)]: an interval matrix $A^I = [\underline{A}, \overline{A}]$ is regular if and only if $(\underline{A} + \overline{A} - S(\overline{A} - \underline{A}))^{-1}(\underline{A} + \overline{A} + S(\overline{A} - \underline{A}))$ is a P-matrix for each signature matrix S (i.e., a diagonal matrix with ± 1 diagonal elements). This topic was recently studied by Johnson and Tsatsomeros [5].

The necessary and sufficient condition of Theorem 2.3 employs up to 2^{n-1} test matrices $A_z, z \in Z$. There is a natural question whether an essentially simpler criterion can be found. Theorem 3.4 gives an indirect answer to this question: it implies that an existence of a polynomial-time algorithm for checking the P-property of symmetric interval matrices would imply that the complexity classes P and NP are equal, thereby running contrary to the current (unproved) conjecture that P≠NP. We refer

the reader to the classic book by Garey and Johnson [4] for a detailed discussion of the problem "P = NP" and related issues.

THEOREM 3.4. *The following problem is NP-hard.*

Instance. *A symmetric interval matrix* $A^I = [\underline{A}, \overline{A}]$ *with rational bounds* $\underline{A}, \overline{A}$.

Question. *Is* A^I *a P-matrix?*

Proof. By Theorem 3.2, A^I is a P-matrix if and only if it is positive definite; checking positive definiteness of symmetric interval matrices was proved to be NP-hard in [11]. □

Coxson [2] proved that the P-matrix problem for real matrices is co-NP-complete. His result concerns nonsymmetric matrices, since the symmetric case can be solved by Sylvester determinant criterion, which can be performed in polynomial time (Schrijver [12]). Theorem 3.4 shows that for interval matrices even the symmetric case is NP-hard.

Acknowledgments. Correspondence with J. Garloff on the subject of this paper is highly appreciated.

REFERENCES

[1] S. BIAŁAS AND J. GARLOFF, *Intervals of P-matrices and related matrices*, Linear Algebra Appl., 58 (1984), pp. 33–41.

[2] G. E. COXSON, *The P-matrix problem is co-NP-complete*, Math. Programming, 64 (1994), pp. 173–178.

[3] M. FIEDLER AND V. PTÁK, *On matrices with non-positive off-diagonal elements and positive principal minors*, Czechoslovak Math. J., 12 (1962), pp. 382–400.

[4] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, 1979.

[5] C. R. JOHNSON AND M. J. TSATSOMEROS, *Convex sets of nonsingular and P-matrices*, Linear and Multilinear Algebra, 38 (1995), pp. 233–239.

[6] K. G. MURTY, *Linear Complementarity, Linear and Nonlinear Programming*, Helderman-Verlag, Berlin, 1988.

[7] A. NEUMAIER, *Interval Methods for Systems of Equations*, Cambridge University Press, Cambridge, 1990.

[8] J. ROHN, *Systems of linear interval equations*, Linear Algebra Appl., 126 (1989), pp. 39–78.

[9] ———, *Positive definiteness and stability of interval matrices*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 175–184.

[10] ———, *On some properties of interval matrices preserved by nonsingularity*, Z. Angew. Math. Mech., 74 (1994), p. T688.

[11] ———, *Checking positive definiteness or stability of symmetric interval matrices is NP-hard*, Comment. Math. Univ. Carolin., 35 (1994), pp. 795–797.

[12] A. SCHRIJVER, *Theory of Integer and Linear Programming*, Wiley, Chichester, 1986.

[13] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 4, pp. 1025–1036, October 1996

020

INVERSES OF UNIPATHIC M-MATRICES*

J. J. MCDONALD†, M. NEUMANN‡, H. SCHNEIDER§, AND M. J. TSATSOMEROS¶

Abstract. In this paper we characterize all nonnegative matrices whose inverses are M-matrices with unipathic digraphs. A digraph is called unipathic if there is at most one simple path from any vertex j to any other vertex k. The set of unipathic digraphs on n vertices includes the simple n-cycle and all digraphs whose underlying undirected graphs are trees (or forests). Our results facilitate the construction of nonnegative matrices whose inverses are M-matrices with unipathic digraphs. We highlight this procedure for inverses of tridiagonal M-matrices and of M-matrices whose digraphs are simple n-cycles with loops.

Key words. unipathic digraph, M-matrix, inverse, principal minor

AMS subject classifications. 15A09, 15A57

1. Introduction.

The inverse of an M-matrix is always a nonnegative matrix; however, characterizing the nonnegative matrices whose inverses are M-matrices is a long-standing open problem. In the present article we contribute to the solution of the inverse M-matrix problem by identifying a subclass of the inverse M-matrices. We provide necessary and sufficient conditions for a nonnegative matrix C to be the inverse of an M-matrix whose digraph is unipathic. A digraph is called unipathic if there is at most one simple path from any vertex j to any other vertex k.

Unipathic digraphs were introduced by Harary, Norman, and Cartwright [5], and they were proposed as a new direction of research in combinatorial matrix analysis by Maybee [11]. It is pointed out in [11] that unipathic digraphs can serve as a generalization and a theoretic unification of digraphs whose underlying undirected graphs are trees (or forests) and of directed simple cycles.

The conditions we obtain for a nonnegative matrix to be an inverse of an M-matrix whose digraph is unipathic (see Theorem 3.2) involve positivity of the diagonal entries and certain 2×2 principal minors as well as particular off-diagonal entries and 2×2 almost principal minors being zero. (An almost principal minor is the determinant of a submatrix whose row and column index sets differ by only one element.) Our proof is based on properties of unipathic digraphs (see Lemmas 2.1 and 2.2) and on a key observation in [12] that connects zero 2×2 almost principal minors of an inverse M-matrix to the digraph of the M-matrix (see Theorem 3.1).

Our results facilitate the construction of nonnegative matrices whose inverses are M-matrices with unipathic digraphs. We illustrate this procedure for inverses of tridiagonal M-matrices and of M-matrices whose digraphs are simple cycles with loops (see §4).

For definitions, references, and background on M-matrices and the inverse M-matrix problem the reader is referred to Berman and Plemmons [2] and Johnson [6].

* Received by the editors October 18, 1994; accepted for publication (in revised form) by R. Horn October 2, 1995.

† Department of Mathematics and Statistics, University of Regina, Regina, Saskatchewan S4S 0A2, Canada. This work was supported by NSERC grant 6-53120.

‡ Department of Mathematics, University of Connecticut, Storrs, CT 06269-3009. This work was supported by NSF grant DMS-8901860 and NSF grant DMS-9306357.

§ Department of Mathematics, University of Wisconsin, Madison, WI 53706. This work was supported by NSF grant DMS-9123318.

¶ Department of Mathematics and Statistics, University of Regina, Regina, Saskatchewan S4S 0A2, Canada (tsat@plato.math.uregina.ca). This work was supported by NSERC grant 6-53121.

In the following section we introduce the notation necessary to describe our results, summarize the properties of unipathic digraphs, and present some definitions and auxiliary results.

2. Notation and preliminaries. In this paper we let $\langle n\rangle = \{1, 2, \ldots, n\}$ and $\Gamma = (V, E)$ be a digraph with vertex set $V = \langle n\rangle$ and directed edge set $E = \{(i, j) \mid i, j \in V\}$. A *path* from j to k in Γ is a sequence of vertices $j = r_1, r_2, \ldots, r_t = k$, with $(r_i, r_{i+1}) \in E$ for $i = 1, \ldots, t-1$. A path is *simple* if $r_1, r_2, \ldots, r_t$ are distinct. A path $r_1, \ldots, r_t, r_1$ with $t > 1$ is called a *cycle*. It is called a *simple cycle* if the intermediate vertices are distinct.

A digraph is called *unipathic* if there is at most one simple path from any vertex j to any other vertex k.

We adopt the following notation to be used within proofs and in commentary:

$j \rightsquigarrow_\Gamma k$: if there is a path from j to k in Γ ("j has access to k in Γ").

$j \not\rightsquigarrow_\Gamma k$: if there is no path from j to k in Γ.

$j \rightarrow_\Gamma k$: if $(j, k) \in E$.

$j \not\rightarrow_\Gamma k$: if $(j, k) \notin E$.

We denote by Γ_i the digraph obtained from Γ when vertex i and any associated edges are removed (i.e., the subgraph of Γ induced by $\langle n\rangle \setminus i$). We denote by Γ^i the digraph obtained from Γ_i by adding an edge from a vertex j to a vertex k whenever $j \rightarrow_\Gamma i$ and $i \rightarrow_\Gamma k$. The *transitive closure* of Γ, denoted by $\overline{\Gamma}$, is obtained from Γ by adding an edge (i, j) whenever $i \rightsquigarrow_\Gamma j$. If $\Gamma = (V, E)$ and $\Gamma' = (V, E')$ are two digraphs, we let $\Gamma \cup \Gamma' = (V, E \cup E')$.

The *digraph of a matrix* $A = (a_{ij}) \in \mathbb{R}^{nn}$ is denoted by $\mathcal{D}(A) = (V, E)$, with $V = \langle n\rangle$ and $E = \{(i, j) \mid a_{ij} \neq 0\}$. We say that A is *irreducible* if $j \rightsquigarrow_{\mathcal{D}(A)} k$ for all $j, k \in V$. The matrix A is called *unipathic* if $\mathcal{D}(A)$ is a unipathic digraph.

The *underlying undirected graph* $\tilde{\Gamma} = (V, \tilde{E})$ of $\Gamma = (V, E)$ has a vertex set V and an edge set $\tilde{E} \subseteq \{\{i, j\} \mid i, j \in V\}$, where $\{i, j\} \in \tilde{E}$ if and only if $i \neq j$ and either $(i, j) \in E$ or $(j, i) \in E$. We define paths, cycles, and simple cycles of a graph to correspond to the definitions for a digraph. A graph with no cycles is called a *forest*. A connected forest is called a *tree*.

We continue with a summary of the characteristics of unipathic digraphs. Clearly in any digraph, if there is a path from j to k, then there is a simple path from j to k. By definition, if Γ is unipathic, then there may be several paths from j to k, but there can only be one simple path. More precisely we have the following lemma.

LEMMA 2.1. *Let Γ be a unipathic digraph. Let i, j, k be distinct vertices such that $j \rightsquigarrow_\Gamma k$. Then the following are equivalent:*

(i) *The vertex j does not have access to k in Γ_i.*

(ii) *Every path in Γ from j to k goes through i.*

(iii) *The simple path from j to k in Γ goes through i.*

Proof. (i) implies (ii): if $j \not\rightsquigarrow_{\Gamma_i} k$, then every path from j to k must go through i.

(ii) implies (iii): if every path in Γ from j to k goes through i, then the simple path from j to k in Γ goes through i.

(iii) implies (i): let i be any vertex in the simple path from j to k. Suppose $j \rightsquigarrow_{\Gamma_i} k$. Then there is a simple path from j to k in Γ_i. But then we would have two different simple paths from j to k in Γ. This is a contradiction. Hence $j \not\rightsquigarrow_{\Gamma_i} k$. □

A unipathic digraph may have loops on its vertices and, unlike a digraph whose underlying undirected graph is a tree, may have cycles of any length. However, no two cycles can have a common edge. As explained in [11], every strongly connected

unipathic digraph can be constructed from a tree (by adjoining chords and orienting the resulting cycles and by replacing edges with directed simple paths). Notice that if the digraph of a combinatorially symmetric matrix $A = (a_{ij})$ (i.e., $a_{ij} \neq 0$ implies $a_{ji} \neq 0$) is strongly connected and unipathic, then its underlying undirected graph must be a tree.

An important property of unipathic digraphs is given next. The *indegree* (resp., *outdegree*) of a vertex i of a digraph is the number of edges entering (resp., issuing from) vertex i.

LEMMA 2.2. *Let $\Gamma = (V, E)$ be a unipathic digraph. Then there is a vertex with indegree at most* 1 *and a vertex with outdegree at most* 1.

Proof. Let $r_1, \ldots, r_t \in V$ be a simple path in Γ of maximal length. Suppose r_1 has indegree greater than 1. Then there exist two distinct edges ending in r_1, and they must be of the form (r_i, r_1) and (r_j, r_1), with $1 < i < j$, by the maximality of the simple path $r_1, \ldots, r_t$. But then there are two simple paths from r_i to r_1, which is a contradiction. Similarly we can show that r_t has outdegree at most 1. □

We denote an entrywise nonnegative matrix C by $C \geq 0$. If all the entries of C are positive we write $C \gg 0$. Let $S, T \subseteq \langle n \rangle$ and $C \in \mathbb{R}^{nn}$. We write $C[S, T]$ for the submatrix of C whose rows and columns are indexed by S and T, respectively, in their natural order. If S or T is a singleton, e.g., $T = \{\ell\}$, we write $C[S, \ell]$ instead of $C[S, \{\ell\}]$. Let $R \subseteq \langle n \rangle$. The *Schur complement* of C with respect to an invertible principal submatrix $C[R, R]$ is

$$C/C[R, R] = C[Q, Q] - C[Q, R](C[R, R])^{-1}C[R, Q],$$

where $Q = \langle n \rangle \setminus R$.

The following lemma will be used in the proof of Theorem 3.2. It can be obtained by combining formulae from Brualdi and Schneider [3, (10), p. 773] and Watford [14, (4), p. 251].

LEMMA 2.3. *Let $C \in \mathbb{R}^{nn}$ and let $R \subseteq \langle n \rangle$, $Q = \langle n \rangle \setminus R$. If all the relevant inverses in the following block matrix exist, then C is invertible and its inverse is permutationally similar to*

$$\begin{bmatrix} (C/C[R, R])^{-1} & -(C[Q, Q])^{-1}C[Q, R](C/C[Q, Q])^{-1} \\ -(C[R, R])^{-1}C[R, Q](C/C[R, R])^{-1} & (C/C[Q, Q])^{-1} \end{bmatrix}.$$

We close this section with a characterization for a nonnegative matrix to be an inverse M-matrix. Owing to its generality, it gives less insight than one might wish. However, it can be used to obtain some additional characterizations for inverses of M-matrices.

LEMMA 2.4. *Let $A = sI - P$, $P \geq 0$, be a nonsingular M-matrix. Then for all $\beta \geq 0$, $A(A + \beta I)^{-1}$ is a nonsingular M-matrix which is given by*

$$A(A + \beta I)^{-1} = \frac{s}{s + \beta} I - \frac{\beta}{s + \beta} \sum_{j=1}^{\infty} \frac{P^j}{(s + \beta)^j}. \tag{1}$$

Proof. The proof is essentially to be found in the proof of Theorem 3 of Johnson [6]. □

Lemma 2.4 now yields the following characterization mentioned above.

THEOREM 2.5. *Let $C = (c_{ij}) \in \mathbb{R}^{nn}$ be a nonnegative matrix. Then C is the inverse of an M-matrix if and only if $c_{ii} > 0$ for all $1 \leq i \leq n$, $C + \alpha I$ is invertible*

for all $\alpha \geq 0$, $\mathcal{D}(C) = \overline{\mathcal{D}(C)}$, *and*

$$\mathcal{D}((C+\alpha I)^{-1}) = \mathcal{D}(C) \quad \text{for all} \quad \alpha > 0. \tag{2}$$

Proof. Assume first that C is the inverse of an M-matrix. Then it is well known that $c_{ii} > 0$, $i = 1, \ldots, n$, $C + \alpha I$ is invertible for all $\alpha \geq 0$, and $\mathcal{D}(C) = \overline{\mathcal{D}(C)}$. Now put $A = C^{-1}$ and observe that

$$(C+\alpha I)^{-1} = \beta A(A+\beta I)^{-1}, \tag{3}$$

where $\beta := 1/\alpha$. Since A is a nonsingular M-matrix, there exists $P \geq 0$ such that $A = sI - P$ and such that $s > \rho(P)$ (the spectral radius of P). But then, by the Neumann expansion,

$$C = \frac{1}{s}I + \frac{1}{s}\sum_{j=1}^{\infty} \frac{P^j}{s^j}. \tag{4}$$

The validity of (2) now follows by comparing, for $\alpha > 0$, the expansion for the matrix on the right-hand side of (3) which can be obtained via (1) and the expansion in (4).

Conversely, suppose that the equality in (2) holds for all $\alpha > 0$. If $((C+\alpha I)^{-1})_{ij} = 0$ for some $\alpha > 0$, then by (2) it must hold for all $\alpha > 0$ and hence, by continuity arguments, $(C^{-1})_{ij} = 0$. Similarly, if $((C+\alpha I)^{-1})_{ij} < 0$ then, again by (2), this entry must be negative for all $\alpha > 0$ so that $(C^{-1})_{ij} \leq 0$. Suppose now that $((C+\alpha I)^{-1})_{ij} > 0$, $i \neq j$, for some $\alpha > 0$ so that this entry is positive for all $\alpha > 0$. Then, for sufficiently large $\alpha > 0$, the Neumann expansion gives us that

$$((C+\alpha I)^{-1})_{ij} = \frac{1}{\alpha}\left(\left(I + \frac{C}{\alpha}\right)^{-1}\right)_{ij} = \frac{1}{\alpha}\left(I - \frac{C}{\alpha} + \frac{C^2}{\alpha^2} - \cdots\right)_{ij}.$$

In particular we see that as α increases, it will attain a value such that beyond this value the (i,j)th entry of $(C+\alpha I)^{-1}$ will become negative, contradicting the constancy of the sign implied by (2). Hence there cannot be a value of $\alpha > 0$ for which $((C+\alpha I)^{-1})_{ij} > 0$ and our proof is done. □

Our theorem has the following two corollaries.

COROLLARY 2.6. *Let* $C = (c_{ij}) \in \mathbb{R}^{nn}$ *be nonnegative. Then a necessary and sufficient condition for* C *to be the inverse of an M-matrix is that* $c_{ii} > 0$, *the matrix* $C + \alpha I$ *is invertible for all* $\alpha > 0$, $\mathcal{D}(C) = \overline{\mathcal{D}(C)}$, *and that for each pair* (i,j) *the minor of* $C + \alpha I$ *obtained after deleting the* i*th row and* j*th column has a constant sign as a function of* $\alpha > 0$.

COROLLARY 2.7. *If* $A \in \mathbb{R}^{nn}$ *is a nonsingular M-matrix, then for any* $\alpha > 0$, $\mathcal{D}(A^{-1} + \alpha I)^{-1} = \overline{\mathcal{D}(A)}$.

The last corollary has the following implication. Suppose that $A \in \mathbb{R}^{nn}$ is a nonsingular irreducible M-matrix that has some off-diagonal entries equal to zero. Invert A to obtain the positive matrix C. Now invert $C + \alpha I$. Then $(C+\alpha I)^{-1}$ is an M-matrix and, by the foregoing, all its entries are nonzero.

3. The inverse of a unipathic M-matrix. We begin with a theorem on nonsingular M-matrices proved in McDonald, Neumann, Schneider, and Tsatsomeros [12]. It represents a graph-theoretical refinement and generalization of a condition found in Willoughby [15] that is necessary for a matrix to be an inverse M-matrix.

THEOREM 3.1. *Let* $A \in \mathbb{R}^{nn}$ *be a nonsingular M-matrix and* $\Gamma = \mathcal{D}(A)$. *Let also* $C = A^{-1}$ *and* $\{i,j,k\} \subseteq \langle n \rangle$ *be distinct. Then*

(i) $c_{jk} = \frac{c_{ji}c_{ik}}{c_{ii}}$ *whenever* j *does not have access to* k *in* Γ_i,

(ii) $c_{jk} > \frac{c_{ji}c_{ik}}{c_{ii}}$ *whenever* j *has access to* k *in* Γ_i.

Notice that Theorem 3.1 refers to the value (zero or positive) of the almost principal minors of C formed from rows i, j and columns i, k.

In the next theorem, our main result, we provide necessary and sufficient conditions for $C \geq 0$ to be the inverse of a unipathic M-matrix. It is well known that if C is an inverse M-matrix then its diagonal entries and 2×2 principal minors are positive, the 2×2 almost principal minors satisfy Theorem 3.1, and $\mathcal{D}(C) = \overline{\mathcal{D}(C)} = \overline{\mathcal{D}(C^{-1})}$. These conditions are not in general sufficient for $C \geq 0$ to be an inverse M-matrix. However, as we will show in Theorem 3.2, a subset of these conditions, dictated by a unipathic digraph, is necessary and sufficient for C to be the inverse of a unipathic M-matrix.

THEOREM 3.2. *Let* Γ *be a unipathic digraph on* n *vertices and* $C \in \mathbb{R}^{nn}$. *Then the following are equivalent.*

(i) C *is nonsingular and* C^{-1} *is an M-matrix such that* $\mathcal{D}(C^{-1}) = \Gamma$.

(ii) $C \geq 0$ *and satisfies*

(a) $c_{ii} > 0$ *for all* $i \in \langle n \rangle$,

(b) $c_{jj}c_{kk} > c_{jk}c_{kj}$ *for all distinct* j *and* k *such that there is an edge from* j *to* k *in* Γ,

(c) $c_{jk} = 0$ *whenever there is no path from* j *to* k *in* Γ,

(d) $c_{jk} = \frac{c_{ji}c_{ik}}{c_{ii}}$ *for all distinct* i, j, k *such that there is a path from* j *to* k *in* Γ, *but there is no path from* j *to* k *in* Γ_i.

Proof. (i) implies (ii): since C^{-1} is an M-matrix, $C \geq 0$ and (ii)(a),(b) follow from well-known facts about M-matrices (see, e.g., [2]). Property (ii)(c) follows from the fact that the digraph of an inverse M-matrix is the transitive closure of the digraph of its inverse (see Lewin and Neumann [8] and Schneider [13]). Property (ii)(d) follows from Theorem 3.1.

(ii) implies (i): we proceed by induction on n. For $n = 1$, the result follows trivially. Assume $n \geq 2$ and that (ii) implies (i) for all $(n-1) \times (n-1)$ matrices.

Using the inductive hypothesis we will establish three claims which, combined with Lemma 2.3, will allow us to show that C is invertible and that its inverse is a Z-matrix (i.e., it has nonpositive off-diagonal entries) with digraph Γ.

CLAIM 1. $c_{jj}c_{kk} > c_{jk}c_{kj}$ *for all distinct* $j, k \in \langle n \rangle$.

Proof of Claim 1. If $j \not\rightsquigarrow_\Gamma k$ or $k \not\rightsquigarrow_\Gamma j$ then by (ii)(c) of Theorem 3.2, $c_{jk}c_{kj} = 0$, and the claim follows. Assume $j \rightsquigarrow_\Gamma k$ and $k \rightsquigarrow_\Gamma j$. We use induction on the length r of the simple path from j to k. If $r = 1$, the claim follows from (ii)(b). Assume $r > 1$ and that the claim holds for any two vertices connected by a simple path with length less than r. Either the simple path from j to k has no vertices, other than j and k, in common with the simple path from k to j, or there is an additional vertex i which is common to both paths. In the latter case, by (ii)(d) of Theorem 3.2, Lemma 2.1, and the induction hypothesis on the path length,

$$\frac{c_{jk}c_{kj}}{c_{jj}c_{kk}} = \frac{c_{ji}c_{ik}c_{ki}c_{ij}}{c_{jj}c_{ii}c_{kk}c_{ii}} = \left(\frac{c_{ji}c_{ij}}{c_{jj}c_{ii}}\right)\left(\frac{c_{ik}c_{ki}}{c_{kk}c_{ii}}\right) < 1.$$

In the former case, for any i in the simple path from j to k, there is a simple path from i to j through k. Hence by (ii)(d), Lemma 2.1, and the induction hypothesis on the path length,

$$\frac{c_{jk}c_{kj}}{c_{jj}c_{kk}} = \frac{c_{ji}c_{ik}c_{kj}}{c_{jj}c_{ii}c_{kk}} = \frac{c_{ji}c_{ij}}{c_{jj}c_{ii}} < 1.$$

This establishes Claim 1. □

CLAIM 2. *For any $\ell \in \langle n \rangle$, let $B = C/C[\ell, \ell]$. Then B is invertible and B^{-1} is an M-matrix with $\mathcal{D}(B^{-1}) = \Gamma_\ell$.*

Proof of Claim 2. By the induction hypothesis, it is enough to show B satisfies (ii) of Theorem 3.2 for the digraph Γ_ℓ. For ease of notation, we will assume that the indices of the entries of B correspond to those of C; i.e.,

$$b_{jk} = c_{jk} - \frac{c_{j\ell}c_{\ell k}}{c_{\ell\ell}}.$$

First we show that B is nonnegative. If $b_{jk} = 0$, we are done. So suppose that $b_{jk} \neq 0$. Then either $c_{jk} \neq 0$ or $c_{j\ell}c_{\ell k} \neq 0$. If $c_{j\ell}c_{\ell k} = 0$, then $b_{jk} = c_{jk} > 0$. If $c_{j\ell}c_{\ell k} \neq 0$, then by (ii)(c) $j \rightsquigarrow_\Gamma \ell$ and $\ell \rightsquigarrow_\Gamma k$ and the following cases have to be considered.

If $j \not\rightsquigarrow_{\Gamma_\ell} k$, then by (ii)(d)

$$b_{jk} = c_{jk} - \frac{c_{j\ell}c_{\ell k}}{c_{\ell\ell}} = c_{jk} - c_{jk} = 0.$$

If $j \rightsquigarrow_{\Gamma_\ell} k$, then joining the simple paths from j to ℓ and from ℓ to k forms a path from j to k through ℓ, which therefore cannot be simple. So let i be the first vertex in the simple path from j to ℓ which is also in the simple path from ℓ to k. Then the (sub)path from j to i and then from i to k forms a simple path from j to k. Hence $j \not\rightsquigarrow_{\Gamma_i} k$, $j \not\rightsquigarrow_{\Gamma_i} \ell$, and $\ell \not\rightsquigarrow_{\Gamma_i} k$. By Theorem 3.1,

$$b_{jk} = c_{jk} - \frac{c_{j\ell}c_{\ell k}}{c_{\ell\ell}} = c_{jk} - \frac{c_{ji}c_{i\ell}c_{\ell i}c_{ik}}{c_{ii}c_{ii}c_{\ell\ell}}$$

$$= c_{jk} - c_{jk}\frac{c_{i\ell}c_{\ell i}}{c_{ii}c_{\ell\ell}} = c_{jk}\left(1 - \frac{c_{i\ell}c_{\ell i}}{c_{ii}c_{\ell\ell}}\right) > 0 \quad \text{(by Claim 1)}.$$

Hence B is a nonnegative matrix.

We now show that conditions (ii)(a)–(d), labeled here as (a')–(d'), also hold for the matrix B with the digraph Γ_ℓ.

(a') By Claim 1,

$$b_{jj} = c_{jj} - \frac{c_{j\ell}c_{\ell j}}{c_{\ell\ell}} > 0.$$

(b') Suppose $j \rightarrow_{\Gamma_\ell} k$. If $k \not\rightsquigarrow_\Gamma \ell$, then either $k \not\rightsquigarrow_\Gamma j$ or $j \not\rightsquigarrow_\Gamma \ell$, so by (ii)(c),

$$c_{k\ell} = 0 = \frac{c_{kj}c_{j\ell}}{c_{jj}}. \tag{5}$$

If $k \rightsquigarrow_\Gamma \ell$, then $j \rightsquigarrow_\Gamma \ell$. It follows that either j is in the simple path from k to ℓ in Γ, in which case, by (ii)(d) and Lemma 2.1, we have that

$$c_{k\ell} = \frac{c_{kj}c_{j\ell}}{c_{jj}} \geq 0, \tag{6}$$

or the edge from j to k combined with the simple path from k to ℓ forms a simple path from j to ℓ which includes k. In this case we have, again by (ii)(d) and Lemma 2.1, that

$$c_{j\ell} = \frac{c_{jk}c_{k\ell}}{c_{kk}}. \tag{7}$$

If (5) or (6) holds, then by replacing $c_{k\ell}$ in the following expression we get

$$b_{jj}b_{kk} - b_{jk}b_{kj} = \left(c_{jj} - \frac{c_{j\ell}c_{\ell j}}{c_{\ell\ell}}\right)\left(c_{kk} - \frac{c_{k\ell}c_{\ell k}}{c_{\ell\ell}}\right) - \left(c_{jk} - \frac{c_{j\ell}c_{\ell k}}{c_{\ell\ell}}\right)\left(c_{kj} - \frac{c_{k\ell}c_{\ell j}}{c_{\ell\ell}}\right)$$

$$= (c_{jj}c_{kk} - c_{jk}c_{kj}) - \frac{c_{jj}c_{kk}c_{j\ell}c_{\ell j} - c_{jk}c_{kj}c_{j\ell}c_{\ell j}}{c_{jj}c_{\ell\ell}} - \frac{c_{kj}c_{j\ell}c_{\ell k} - c_{kj}c_{j\ell}c_{\ell k}}{c_{\ell\ell}}$$

$$= (c_{jj}c_{kk} - c_{jk}c_{kj})\left(1 - \frac{c_{j\ell}c_{\ell j}}{c_{jj}c_{\ell\ell}}\right) > 0 \quad \text{(by Claim 1).}$$

If equation (7) holds, then the above inequality follows in a similar manner by replacing c_{jl}.

(c') If $j \not\rightsquigarrow_{\Gamma_\ell} k$, then by (ii)(c) or (d),

$$b_{jk} = c_{jk} - \frac{c_{j\ell}c_{\ell k}}{c_{\ell\ell}} = c_{jk} - c_{jk} = 0.$$

(d') Let $i, j, k \in S$ be distinct. Assume $j \rightsquigarrow_{\Gamma_\ell} k$, but $j \not\rightsquigarrow_{(\Gamma_\ell)_i} k$. Then $j \not\rightsquigarrow_{\Gamma_i} k$. Hence by (ii)(d),

$$c_{jk} = \frac{c_{ji}c_{ik}}{c_{ii}}. \tag{8}$$

If $j \rightsquigarrow_\Gamma \ell$ and $\ell \rightsquigarrow_\Gamma k$, then by Lemma 2.1 i is in either the simple path from j to ℓ or the simple path from ℓ to k. Hence by (ii)(d) and Lemma 2.1,

$$c_{j\ell} = \frac{c_{ji}c_{i\ell}}{c_{ii}} \tag{9}$$

or

$$c_{\ell k} = \frac{c_{\ell i}c_{ik}}{c_{ii}}. \tag{10}$$

If $j \not\rightsquigarrow_\Gamma \ell$, then $i \not\rightsquigarrow_\Gamma \ell$, and hence by (ii)(c), equation (9) is satisfied.

If $\ell \not\rightsquigarrow_\Gamma k$, then $\ell \not\rightsquigarrow_\Gamma i$, and hence by (ii)(c), equation (10) is satisfied.

Hence either (9) holds or (10) holds. If (9) holds, then using (8) to replace c_{jk} and (9) to replace $c_{j\ell}$ in the following expression we get

$$b_{jk}b_{ii} - b_{ji}b_{ik} = \left(c_{jk} - \frac{c_{j\ell}c_{\ell k}}{c_{\ell\ell}}\right)\left(c_{ii} - \frac{c_{i\ell}c_{\ell i}}{c_{\ell\ell}}\right) - \left(c_{ji} - \frac{c_{j\ell}c_{\ell i}}{c_{\ell\ell}}\right)\left(c_{ik} - \frac{c_{i\ell}c_{\ell k}}{c_{\ell\ell}}\right)$$

$$= (c_{jk}c_{ii} - c_{ji}c_{ik}) - \left(\frac{c_{ji}c_{ik}c_{i\ell}c_{\ell i} - c_{ji}c_{i\ell}c_{\ell i}c_{ik}}{c_{ii}c_{\ell\ell}}\right)$$

$$- \left(\frac{c_{ii}c_{ji}c_{i\ell}c_{\ell k} - c_{ii}c_{i\ell}c_{\ell k}c_{ji}}{c_{ii}c_{\ell\ell}}\right) = 0.$$

If (10) holds, then the above equality follows in a similar manner.

This establishes Claim 2. □

By Lemma 2.2, there is a vertex with indegree at most 1. Without loss of generality, we can assume this vertex is labeled by n (otherwise we can work with a permutation similarity of C). Let $T = \langle n-1 \rangle$.

CLAIM 3. *The matrix $C[T,T]$ is invertible and its inverse is an M-matrix with digraph Γ^n. Moreover, $C/C[T,T] > 0$.*

Proof of Claim 3. Let $i, j \in T$. Since $i \rightsquigarrow_{\Gamma^n} j$ if and only if $i \rightsquigarrow_{\Gamma} j$, (a), (c), and (d) of (ii) must hold for $C[T,T]$. Also (ii)(b) follows from Claim 1. Hence, by the induction hypothesis, $C[T,T]$ is invertible and its inverse is an M-matrix with digraph Γ^n. To show that $C/C[T,T] > 0$, if the indegree of n is 1 choose m such that $m \rightarrow_{\Gamma} n$; otherwise choose any $m \in \langle n-1 \rangle$. Then by (c) and (d)

$$C[T,n] = \begin{bmatrix} c_{1n} \\ c_{2n} \\ \vdots \\ c_{n-1,n} \end{bmatrix} \tag{11}$$

$$= \begin{bmatrix} \frac{c_{1m}c_{mn}}{c_{mm}} \\ \frac{c_{2m}c_{mn}}{c_{mm}} \\ \vdots \\ \frac{c_{n-1,m}c_{mn}}{c_{mm}} \end{bmatrix} = \frac{c_{mn}}{c_{mm}} \begin{bmatrix} c_{1m} \\ c_{2m} \\ \vdots \\ c_{n-1,m} \end{bmatrix} = \frac{c_{mn}}{c_{mm}} C[T,m].$$

Thus, letting $e_m \in \mathbb{R}^{n-1}$ be the mth standard basis vector, by (11) we get

$$C/C[T,T] = c_{nn} - C[n,T](C[T,T])^{-1}C[T,n]$$

$$= c_{nn} - \frac{c_{mn}}{c_{mm}} C[n,T](C[T,T])^{-1}C[T,m] = c_{nn} - \frac{c_{mn}}{c_{mm}} C[n,T]e_m$$

$$= c_{nn} - \frac{c_{mn}c_{nm}}{c_{mm}} > 0 \quad \text{(by Claim 1).}$$

This establishes Claim 3. □

Recall now that since by Claims 2 and 3 $C/C[n,n]$, $C[T,T]$, and $C/C[T,T]$ are invertible, C has to be invertible and its inverse is, from Lemma 2.3 with $R = \{n\}$ and $Q = T$,

$$\begin{bmatrix} (C/C[n,n])^{-1} & -(C[T,T])^{-1}C[T,n](C/C[T,T])^{-1} \\ -(C[n,n])^{-1}C[n,T](C/C[n,n])^{-1} & (C/C[T,T])^{-1} \end{bmatrix}.$$

Moreover, by Claim 2, for all $\ell \in \langle n \rangle$ and for $S = \langle n \rangle \setminus \ell$, $C^{-1}[S,S] = (C/C[\ell,\ell])^{-1}$ is an M-matrix with digraph Γ_ℓ. It follows that C^{-1} is a Z-matrix with digraph Γ, whose inverse is a nonnegative matrix. This implies that C^{-1} is an M-matrix with digraph Γ. □

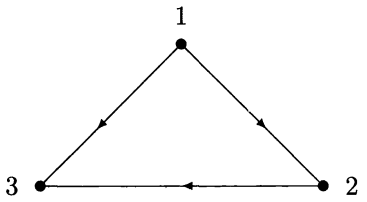
The contents of Claims 2 and 3 within the inductive proof of Theorem 3.2 can now be stated separately.

COROLLARY 3.3. *Let $C \in \mathbb{R}^{nn}$ be the inverse of a unipathic M-matrix whose digraph is Γ.*

(i) *For any $\ell \in \langle n \rangle$ and $S = \langle n \rangle \setminus \ell$, $C/C[\ell,\ell]$ is the inverse of a unipathic M-matrix whose digraph is Γ_ℓ.*

(ii) *For any* $\ell \in \langle n \rangle$ *of indegree at most* 1 *and* $T = \langle n \rangle \setminus \ell$, $C[T,T]$ *is the inverse of a unipathic M-matrix whose digraph is* Γ^{ℓ}. *Moreover,* $C/C[T,T] > 0$.

Example 3.4. The assumption in Theorem 3.2 that the digraph Γ is unipathic is critical for condition (ii) of the theorem to imply that the inverse of C is an M-matrix. For example, consider

$\Gamma =$ 

Notice that Γ is not unipathic (but if any edge is removed the digraph becomes unipathic). The matrix

$$C = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix}$$

satisfies condition (ii) of Theorem 3.2 for the digraph Γ, but its inverse

$$C^{-1} = \begin{bmatrix} 1 & -2 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix}$$

is not an M-matrix. In particular, this example shows that condition (ii) is not in general sufficient to imply that C^{-1} is an M-matrix when Γ belongs to the classes of digraphs discussed in [4] and [9].

Remark 3.5. It follows by Corollary 2.6 that if C is a nonnegative matrix that satisfies one and hence both of the equivalent conditions of Theorem 3.2, then for each pair $1 \leq i, j \leq n$, the minor of $C + \alpha I$ which is obtained by deleting the ith row and jth column has a constant sign as a function of $\alpha > 0$.

4. Construction of inverses of unipathic M-matrices. Theorem 3.2 enables us to construct a matrix C which is the inverse of an M-matrix with a given unipathic digraph. The process is as follows. Given a unipathic digraph Γ, first fill the diagonal entries with positive values. Then, for any simple cycle $r_1 \to \cdots \to r_t \to r_1$, choose $c_{r_i r_{i+1}}$ and $c_{r_t r_1}$ so that they are positive and

$$\frac{c_{r_1 r_1} c_{r_2 r_2} \cdots c_{r_t r_t}}{c_{r_1 r_2} c_{r_2 r_3} \cdots c_{r_t r_1}} > 1.$$

Since Γ is unipathic, no two simple cycles share a common edge, so making the above choices can proceed without overlap. If there is an edge from j to k in Γ with $j \neq k$, but this edge is not part of any simple cycle, assign any positive value to c_{jk}. If j does not have access to k in Γ, then set $c_{jk} = 0$. The remaining entries of C are uniquely determined by Theorem 3.2 (ii)(d).

We highlight this procedure for the inverse of a tridiagonal M-matrix and of an M-matrix whose digraph is the simple n-cycle with loops.

We begin with the inverse of a tridiagonal M-matrix. Conditions (ii)(c), (iii)(c), and (iii)(d) of the following theorem also appear in Barrett [1], who characterizes inverses of tridiagonal matrices in general.

THEOREM 4.1. *The following are equivalent for* $C \in \mathbb{R}^{nn}$.

(i) *C is nonsingular and C^{-1} is a tridiagonal M-matrix.*
(ii) *$C \geq 0$ and satisfies*
  (a) *$c_{ii} > 0$ for all $i \in \langle n \rangle$,*
  (b) *$c_{ii}c_{i+1,i+1} - c_{i+1,i}c_{i,i+1} > 0$ for all $i \in \langle n-1 \rangle$,*
  (c) *$c_{jk} = \frac{c_{ji}c_{ik}}{c_{ii}}$ for all $j > i > k$ and for all $k > i > j$.*
(iii) *$C \geq 0$ and satisfies*
  (a) *$c_{ii} > 0$ for all $i \in \langle n \rangle$,*
  (b) *$c_{ii}c_{i+1,i+1} - c_{i+1,i}c_{i,i+1} > 0$ for all $i \in \langle n-1 \rangle$,*
  (c) *$c_{ij} = \frac{c_{i,i+1}c_{i+1,i+2}\cdots c_{j-1,j}}{c_{i+1,i+1}c_{i+2,i+2}\cdots c_{j-1,j-1}}$ for all $j > i+1$,*
  (d) *$c_{ij} = \frac{c_{j,j+1}c_{j+1,j+2}\cdots c_{i-1,i}}{c_{j+1,j+1}c_{j+2,j+2}\cdots c_{i-1,i-1}}$ for all $i > j+1$.*
(iv) *C is a nonsingular matrix which is totally nonnegative (i.e., all the minors of C are nonnegative) and whose inverse is an M-matrix.*

Proof. The equivalence of (i) and (ii) follows directly from Theorem 3.2 applied to the digraph of a tridiagonal matrix.

The equivalence of (ii) and (iii) is also straightforward.

The equivalence of (i) and (iv) is a result due to Lewin [7]. □

We remark that in [10] Markham introduces a class of matrices which he calls type D. A real $n \times n$ matrix $C = (c_{ij})$ is of *type D* if

$$c_{ij} = \begin{cases} c_i & \text{if } i \leq j, \\ c_j & \text{if } j < i, \end{cases}$$

and if $c_n > c_{n-1} > \cdots > c_1$. Markham shows that if $c_1 > 0$, then C is nonsingular and C^{-1} is a tridiagonal M-matrix. It can be readily checked that if $c_{11} > 0$ and C is a matrix of type D, then its entries satisfy the conditions (ii)(a)–(c) stipulated in Theorem 4.1. Thus the class of matrices characterized in Theorem 4.1 contains the positive matrices of type D as a subclass.

Note that if the entries of the first superdiagonal, first subdiagonal, and the diagonal of $C = (c_{ij})$ satisfy (ii)(a) and (b) of Theorem 4.1, then (ii)(c) (or (iii)(c) and (iii)(d)) uniquely determines the remaining entries of C. This is illustrated in the following example.

Example 4.2. We construct inverses of tridiagonal M-matrices as follows. Begin by filling in the tridiagonal structure of $C = (c_{ij})$ so that (ii)(a) and (b) of Theorem 4.1 are satisfied. For example, let

$$C = \begin{bmatrix} 4 & 2 & * & * & * & * \\ 2 & 2 & 2 & * & * & * \\ * & 1 & 2 & 6 & * & * \\ * & * & 0 & 1 & 1 & * \\ * & * & * & 1 & 3 & 4 \\ * & * & * & * & 2 & 4 \end{bmatrix}.$$

Then use (ii)(c) (or (iii)(c) and (iii)(d)) to (uniquely) fill in the *'s, one (sub-) superdiagonal at a time:

$$C = \begin{bmatrix} 4 & 2 & 2 & 6 & 6 & 8 \\ 2 & 2 & 2 & 6 & 6 & 8 \\ 1 & 1 & 2 & 6 & 6 & 8 \\ 0 & 0 & 0 & 1 & 1 & 4/3 \\ 0 & 0 & 0 & 1 & 3 & 4 \\ 0 & 0 & 0 & 2/3 & 2 & 4 \end{bmatrix}.$$

Then

$$C^{-1} = \begin{bmatrix} 1/2 & -1/2 & 0 & 0 & 0 & 0 \\ -1/2 & 3/2 & -1 & 0 & 0 & 0 \\ 0 & -1/2 & 1 & -3 & 0 & 0 \\ 0 & 0 & 0 & 3/2 & -1/2 & 0 \\ 0 & 0 & 0 & -1/2 & 7/6 & -1 \\ 0 & 0 & 0 & 0 & -1/2 & 3/4 \end{bmatrix}.$$

Next we consider the case where the digraph is a simple cycle with loops.

THEOREM 4.3. *The following are equivalent for* $C \in \mathbb{R}^{nn}$.

(i) *The matrix* C *is nonsingular and* C^{-1} *is an M-matrix whose digraph is the simple* n*-cycle* $1 \to 2 \to \cdots \to n \to 1$ *with loops.*

(ii) $C \gg 0$ *and satisfies*

(a) $\frac{c_{11}c_{22}\cdots c_{nn}}{c_{12}c_{23}\cdots c_{n-1,n}c_{n1}} > 1$,

(b) $c_{jk} = \frac{c_{ji}c_{ik}}{c_{ii}}$ *for all* $i > j > k$, $k > i > j$, *and* $j > k > i$ *(i.e., for all distinct vertices* i, j, k *such that* i *lies on the path from* j *to* k*).*

Example 4.4. We construct an inverse of an M-matrix whose digraph is a simple cycle with loops as follows. Begin by filling in the cycle of $\mathcal{D}(C)$ so that Theorem 4.3 (ii)(a) is satisfied. For example, let

$$C = \begin{bmatrix} 4 & 2 & * & * & * & * \\ * & 2 & 2 & * & * & * \\ * & * & 2 & 6 & * & * \\ * & * & * & 1 & 1 & * \\ * & * & * & * & 3 & 4 \\ 1 & * & * & * & * & 4 \end{bmatrix}.$$

Then use (ii)(b) to (uniquely) fill in the *'s:

$$C = \begin{bmatrix} 4 & 2 & 2 & 6 & 6 & 8 \\ 2 & 2 & 2 & 6 & 6 & 8 \\ 2 & 1 & 2 & 6 & 6 & 8 \\ 1/3 & 1/6 & 1/6 & 1 & 1 & 4/3 \\ 1 & 1/2 & 1/2 & 3/2 & 3 & 4 \\ 1 & 1/2 & 1/2 & 3/2 & 3/2 & 4 \end{bmatrix}.$$

Then

$$C^{-1} = \begin{bmatrix} 1/2 & -1/2 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -6 & 0 & 0 \\ 0 & 0 & 0 & 2 & -2/3 & 0 \\ 0 & 0 & 0 & 0 & 2/3 & -2/3 \\ -1/8 & 0 & 0 & 0 & 0 & 1/2 \end{bmatrix}.$$

Let Z denote the $n \times n$ simple cycle permutation matrix. We can apply Theorem 4.3 to characterize all nonnegative matrices which are polynomials in Z and which are inverses of M-matrices whose digraph is a simple n-cycle with loops.

COROLLARY 4.5. *Let* Z *be the* $n \times n$ *simple cycle permutation matrix. Let* $k_1, \ldots, k_n$ *be nonnegative numbers and consider the matrix*

$$C = p(Z) = k_1 I + k_2 Z + \cdots k_n Z^{n-1}.$$

Then necessary and sufficient conditions for C to be the inverse of an M-matrix whose digraph is the simple cycle $1 \to 2 \to \cdots \to n \to 1$, with loops, are

(i) $k_1 > k_2 > 0$,

(ii) $k_j = (k_2)^{j-1} / (k_1)^{j-2}$, $j = 3, \ldots, n$.

Proof. Notice that

$$C = \begin{bmatrix} k_1 & k_2 & \cdots & k_{n-1} & k_n \\ k_n & k_1 & k_2 & \cdots & k_{n-1} \\ \vdots & k_n & k_1 & \ddots & \vdots \\ k_3 & \vdots & \ddots & k_1 & k_2 \\ k_2 & k_3 & \cdots & k_n & k_1 \end{bmatrix}.$$

The result now follows by applying Theorem 4.3 to C. □

Acknowledgments. We would like to thank Bjarne Toft for providing us with the simple proof of Lemma 2.2. We also thank Miroslav Fiedler for drawing our attention to [4].

REFERENCES

[1] W. W. BARRETT, *A theorem on inverses of tridiagonal matrices*, Linear Algebra Appl., 27 (1979), pp. 211–217.

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

[3] R. A. BRUALDI AND H. SCHNEIDER, *Determinantal identities: Gauss, Schur, Cauchy, Sylvester, Kronecker, Jacobi, Binet, Laplace, Muir, and Cayley*, Linear Algebra Appl., 52/53 (1983), pp. 769–791.

[4] M. FIEDLER, *On inverting partitioned matrices*, Czechoslovak Math. J., 13 (1963), pp. 574–586.

[5] F. HARARY, R. Z. NORMAN, AND D. CARTWRIGHT, *Structural Models*, Wiley, New York, 1965.

[6] C. R. JOHNSON, *Inverse M-matrices*, Linear Algebra Appl., 47 (1982), pp. 195–216.

[7] M. LEWIN, *Totally nonnegative, M–, and Jacobi matrices*, SIAM J. Alg. Disc. Meth., 1 (1980), pp. 419–421.

[8] M. LEWIN AND M. NEUMANN, *The inverse M-matrix problem for $(0,1)$–matrices*, Linear Algebra Appl., 30 (1980), pp. 41–50.

[9] T. J. LUNDY AND J. S. MAYBEE, *Uniformly One–Connected Matrices and Their Inverses*, preprint.

[10] T. L. MARKHAM, *Nonnegative matrices whose inverses are M-matrices*, Proc. Amer. Math Soc., 36 (1972), pp. 326–330.

[11] J. S. MAYBEE, *Some possible new directions for combinatorial matrix analysis*, Linear Algebra Appl., 107 (1988), pp. 23–40.

[12] J. J. MCDONALD, M. NEUMANN, H. SCHNEIDER, AND M. J. TSATSOMEROS, *Inverse M-matrix inequalities and generalized ultrametrix matrices*, Linear Algebra Appl., 220 (1995), pp. 321–341.

[13] H. SCHNEIDER, *Theorems on M–splittings of a singular M-matrix which depend on graph structure*, Linear Algebra Appl., 58 (1984), pp. 407–424.

[14] L. J. WATFORD, JR., *The Schur complement of generalized M-matrices*, Linear Algebra Appl., 5 (1972), pp. 247–255.

[15] R. A. WILLOUGHBY, *The inverse M-matrix problem*, Linear Algebra Appl., 18 (1977), pp. 75–94.

SIAM J. MATRIX ANAL. APPL.
Vol. 17, No. 4, pp. 1037–1043, October 1996

021

NORMAL TOEPLITZ MATRICES*

DOUGLAS R. FARENICK†, MARK KRUPNIK‡, NAUM KRUPNIK§, AND WOO YOUNG LEE¶

Abstract. It is well known from the work of Brown and Halmos [*J. Reine Angew. Math.*, 213 (1963/1964), pp. 89–102] that an infinite Toeplitz matrix is normal if and only if it is a rotation and translation of a Hermitian Toeplitz matrix. In the present article we prove that all finite normal Toeplitz matrices are either generalised circulants or are obtained from Hermitian Toeplitz matrices by rotation and translation.

Key words. normal matrix, circulant matrix, Toeplitz matrix

AMS subject classifications. Primary, 15A57; Secondary, 47B35

1. Introduction.

The purpose of the present article is to describe fully the structure of all finite normal Toeplitz matrices.

The algebraic theory of Toeplitz matrices and Toeplitz operators is now extensive, having been developed over many years. An overview of the theory for finite Toeplitz matrices is given in the monograph [3] of Iohvidov, whereas the classic paper of Brown and Halmos [1] contains many of the fundamental results on the algebraic properties of Toeplitz operators. A well-known theorem from that paper states that an infinite Toeplitz matrix (operator) is normal if and only if it is a rotation and translation of a Hermitian Toeplitz matrix. This theorem does not, however, apply to finite matrices: all circulant matrices, for example, are normal Toeplitz matrices. To date very little has been published about the general structure of a finite normal Toeplitz matrix. In fact it appears that the most informative work on this question is a recent article of Ikramov. In [2] Ikramov has shown that a normal Toeplitz matrix (of order at most 4) over the real field must be of one of four types: symmetric Toeplitz, skew-symmetric Toeplitz, circulant, or skew-circulant. A reading of Ikramov's paper suggests that it may be possible to characterise complex normal Toeplitz matrices of all orders, and we do so here. We first identify the two types of normal Toeplitz matrices that arise.

Type I *is a rotation and translation of a Hermitian Toeplitz matrix*, that is $T = \alpha I + \beta H$, for some complex α and β and for some Hermitian Toeplitz matrix H.

Type II *is a generalised circulant*, which is to mean a Toeplitz matrix of the form

$$T = \begin{pmatrix} a_0 & a_N e^{i\theta} & \ddots & a_1 e^{i\theta} \\ a_1 & a_0 & \ddots & \ddots \\ \ddots & \ddots & \ddots & a_N e^{i\theta} \\ a_N & \ddots & a_1 & a_0 \end{pmatrix},$$

* Received by the editors October 2, 1995; accepted for publication (in revised form) by T. Ando October 10, 1995.

† Department of Mathematics and Statistics, University of Regina, Regina, Saskatchewan S4S 0A2, Canada (farenick@abel.math.uregina.edu). Supported in part by an NSERC research grant.

‡ Department of Mathematics and Statistics, University of Regina, Regina, Saskatchewan S4S 0A2, Canada. Postdoctoral Fellow, supported in part by grants from NSERC.

§ Department of Mathematics and Computer Science, Bar-Ilan University, Ramat-Gan 52900, Israel.

¶ Department of Mathematics, Sung Kyun Kwan University, Suwon 440-746, Korea (wylee@yurim.skkv.ac.kr). Supported in part by KOSEF project 94-0701-02-01-3 and GARC-KOSEF (1995).

for some fixed real θ.

The main result of this paper is the following theorem.

THEOREM 1.1. *Every finite complex normal Toeplitz matrix is a generalised circulant or is a rotation and translation of a Hermitian Toeplitz matrix. In particular, every finite real normal Toeplitz matrix is symmetric Toeplitz, skew-symmetric Toeplitz, circulant, or skew-circulant.*

This paper consists of four sections. In §2 we give criteria for a Toeplitz matrix to be normal. The main result is proved in §3. Within the proof we use several technical lemmas, which are described in §4.

2. Key equalities. Let T be a Toeplitz $(N+1)\times(N+1)$ matrix

$$T=\begin{pmatrix} a_0 & a_{-1} & \ddots & a_{-N} \\ a_1 & a_0 & \ddots & \ddots \\ \ddots & \ddots & \ddots & a_{-1} \\ a_N & \ddots & a_1 & a_0 \end{pmatrix}.$$

Throughout the paper we use the following notation: $b_i := a_{-i}$ for $i = 1,\dots,N$. Note that neither the normality nor the form (I) or (II) of a Toeplitz matrix depends on the value of its diagonal entry; therefore we may assume that $a_0 = 0$.

THEOREM 2.1. *A Toeplitz matrix T of the form*

$$T=\begin{pmatrix} 0 & b_1 & \ddots & b_N \\ a_1 & 0 & \ddots & \ddots \\ \ddots & \ddots & \ddots & b_1 \\ a_N & \ddots & a_1 & 0 \end{pmatrix}$$

is normal if and only if for each p and q the following equalities hold:

$$\overline{a}_q a_p + a_{N-q+1}\overline{a}_{N-p+1} = b_q\overline{b}_p + \overline{b}_{N-q+1}b_{N-p+1}\,. \tag{1}$$

Proof. The matrix T is normal if and only if $R = TT^* - T^*T = 0$. Let $R = \{r_{i,j}\}_0^N$.

Let us write down the condition $r_{p,q} = 0$ for some $p \ge 0, q \ge 0$ in terms of the entries of the matrix T. Suppose first that $p \le q$. Then

$$\begin{aligned}[a_p\overline{a}_q + a_{p-1}\overline{a}_{q-1} + \cdots + a_1\overline{a}_{q-p+1}] + (b_1\overline{a}_{q-p-1} + \cdots + b_{q-p-1}\overline{a}_1) \\ +b_{q-p+1}\overline{b}_1 + \cdots + b_{N-p}\overline{b}_{N-q} = [\overline{b}_p b_q + \cdots + \overline{b}_1 b_{q-p+1}] \\ +(b_1\overline{a}_{q-p-1} + \cdots + b_{q-p-1}\overline{a}_1) + a_1\overline{a}_{q-p+1} + \cdots + a_{N-q}\overline{a}_{N-p}\,.\end{aligned} \tag{2}$$

(We suppose here that the expressions in ring brackets equal 0 for $p = q-1$ and $p = q$ and that the expressions in square brackets equal 0 when $p = 0$.)

Let $p \le N - q$; then after the simplification we obtain

$$a_{p+1}\overline{a}_{q+1} + \cdots + a_{N-q}\overline{a}_{N-p} = \overline{b}_{p+1}b_{q+1} + \cdots + b_{N-p}\overline{b}_{N-q}. \tag{3}$$

Now the condition $r_{p-1,q-1} - r_{p,q} = 0$ applied to the previous equalities with $1 \leq p \leq q$ gives

$$a_p \overline{a}_q + a_{N-q+1} \overline{a}_{N-p+1} = b_q \overline{b}_p + b_{N-p+1} \overline{b}_{N-q+1}. \tag{4}$$

It remains to show that these equalities hold for all p and q, without the restrictions $p \leq q$ and $p + q < N + 1$.

1. If $p > q$, it is enough to interchange in (4) p and q and consider conjugated equalities.

2. If $p+q > N+1$, then denote $s := N-q+1$, $r := N-p+1$, and we come to the same equalities with respect to r and s, with the conditions $s+r = 2N+2-p-q < 2N+2-N-1 = N+1$.

3. If, finally, $N - q = p - 1$, then from (2) it follows that $a_p \overline{a}_q = b_q \overline{b}_p$, which is a particular case of (1), corresponding to the choice $q + p = N + 1$.

The proof that equation (1) implies normality will not be given, for it is a consequence of our main theorem on the structure of normal Toeplitz matrices. All subsequent work requires only the implication that normal Toeplitz matrices satisfy equation (1). □

REMARK 2.2. *If we consider* (1) *with* $\tilde{p} = N-q+1, \tilde{q} = p, N-\tilde{p}+1 = q, N-\tilde{q}+1 = N-p+1$, *we obtain that for each* $\tilde{p}, \tilde{q}$ *the following equalities hold:*

$$\overline{a}_{N-\tilde{p}+1} a_{\tilde{q}} + a_{\tilde{p}} \overline{a}_{N-\tilde{q}+1} = b_{N-\tilde{p}+1} \overline{b}_{\tilde{q}} + \overline{b}_{\tilde{p}} b_{N-\tilde{q}+1}\,. \tag{5}$$

The following observation will be put to use in the proof of the main theorem.

REMARK 2.3. *If* $N = 2n-1$ *and* $a_n = b_n = 0$, *then using the notation* $\tilde{a}_p = a_p$, $\tilde{b}_p = b_p$ *for* $p < n$ *and* $\tilde{a}_p = a_{p+1}$, $\tilde{b}_p = b_{p+1}$ *for* $p > n$, *we again come to the equalities of the form* (1) *for* $4(n-1)$ *variables* $\tilde{a}_p$ *and* $\tilde{b}_p$.

3. Main result. Let

$$T = \begin{pmatrix} 0 & b_1 & \ddots & b_N \\ a_1 & 0 & \ddots & \ddots \\ \ddots & \ddots & \ddots & b_1 \\ a_N & \ddots & a_1 & 0 \end{pmatrix}$$

be a normal Toeplitz matrix and let $\{m_i\} \subseteq \{1, \ldots, N\}$ be a set of positive integers. We say that a set of pairs of diagonals of the matrix T with the indices m_i is *coconnected* (with argument θ) if $b_{m_i} = \overline{a}_{m_i} e^{i\theta}$ and *contraconnected* (with argument θ) if $b_{m_i} = a_{N-m_i+1} e^{i\theta}$. Now using these definitions, we cast the statement of the main theorem in the following equivalent form.

THEOREM 3.1. *If a finite complex Toeplitz matrix* T *of trace zero is normal, then there exists a single argument* $0 \leq \theta \leq 2\pi$ *such that with respect to* θ *either all pairs of diagonals of* T *are coconnected or all pairs of diagonals of* T *are contraconnected.*

The proof is based on the following principal idea: we prove first that each two pairs of diagonals with the indices p and $2n+1-p$, $1 \leq p \leq n$ (in case $N = 2n$) or each three pairs of diagonals with the indices $n-k, n, n+k$, $1 \leq k \leq n$ (in the case $N = 2n-1$) are either coconnected or contraconnected or are simultaneously co- and contraconnected. Then we show that for all sets of pairs of diagonals there is a

unique common argument θ, such that all pairs of diagonals are either coconnected or contraconnected or simultaneously co- and contraconnected with the same argument.

Proof. We split the proof of this theorem into three parts.

Part I. $N = 2n-1$ with $a_n \neq 0$. Take two natural numbers m and k such that $0 \leq m, k \leq n-1$ and apply Theorem 2.1 with $p = n-k$ and $q = n+m$. The equalities

$$a_{n-m}\overline{a}_{n+k} + a_{n-k}\overline{a}_{n+m} = b_{n+m}\overline{b}_{n-k} + b_{n+k}\overline{b}_{n-m} \tag{6}$$

hold for two arbitrary pairs of diagonals with indices $[n-m, n+m]$ and $[n+k, n-k]$. Consider then the following system of equalities:

$$a_n\overline{a}_{n+k} + a_{n-k}\overline{a}_n = b_n\overline{b}_{n-k} + b_{n+k}\overline{b}_n \quad (m=0), \tag{7}$$

$$a_{n-k}\overline{a}_{n+k} = b_{n+k}\overline{b}_{n-k} \quad (m = k \neq 0), \tag{8}$$

$$|a_n| = |b_n| \quad (m = k = 0) \tag{9}$$

for the three pairs of diagonals with the indices $n-k, n, n+k$.

Taking into account (9), one can suppose, without loss of generality, that $a_n = |a_n|, b_n = a_n e^{i\theta}$; then from (7) it follows that

$$\overline{a}_{n+k} + a_{n-k} = \overline{b}_{n-k}e^{i\theta} + b_{n+k}e^{-i\theta},$$

and from (8) it follows that

$$\overline{a}_{n+k}a_{n-k} = (\overline{b}_{n-k}e^{i\theta})(b_{n+k}e^{-i\theta}).$$

Therefore at least one of the following two pairs of equations holds for all k:

$$\overline{a}_{n+k} = \overline{b}_{n-k}e^{i\theta}, \quad a_{n-k} = b_{n+k}e^{-i\theta} \tag{10}$$

or

$$\overline{a}_{n+k} = b_{n+k}e^{-i\theta}, \quad a_{n-k} = \overline{b}_{n-k}e^{i\theta}. \tag{11}$$

Validity of (10) implies

$$b_{n-k} = a_{n+k}e^{i\theta},\ b_n = a_n e^{i\theta},\ b_{n+k} = a_{n-k}e^{i\theta}, \tag{12}$$

which means that pairs of diagonals $[n-k, n, n+k]$ are contraconnected. If (11) holds, then

$$b_{n-k} = \overline{a}_{n-k}e^{i\theta},\ b_n = \overline{a}_n e^{i\theta},\ b_{n+k} = \overline{a}_{n+k}e^{i\theta}, \tag{13}$$

and these diagonals are coconnected.

If it happens that both (12) and (13) hold for all $k = 1, \ldots, n-1$, then the proof of Part I of the theorem is complete. Otherwise let us suppose that for some k only one of (12) or (13) holds, say (12); i.e. $\overline{a}_{n-k} \neq a_{n+k}$. We will show that the equalities (12) hold for all m. Assume that for some m (13) holds; we show that (12) is valid too. Substitute (13) in (6); then

$$(\overline{a}_{n+m} - a_{n-m})(a_{n-k} - \overline{a}_{n+k}) = 0.$$

This means that $\overline{a}_{n+m} = a_{n-m}$, and therefore

$$b_{n-m} = a_{n+m}e^{i\theta},\ b_n = a_n e^{i\theta},\ b_{n+m} = a_{n-m}e^{i\theta}.$$

Part II. $N = 2n$. Set $N = 2n$ and $q = p$ in equation (1). We obtain

$$|a_p|^2 + |a_{2n-p+1}|^2 = |b_p|^2 + |b_{2n-p+1}|^2. \tag{14}$$

If we set $q = 2n + 1 - p$ in (1), we will have

$$a_p\overline{a}_{2n+1-p} = b_{2n+1-p}\overline{b}_p. \tag{15}$$

The system of equations (14) and (15), as in Part I, possesses two representations, namely

$$b_p = \overline{a}_p e^{i\theta_p}, \quad b_{2n+1-p} = \overline{a}_{2n+1-p}e^{i\theta_p}, \tag{16}$$

which means that the pairs of diagonals $[p, N-p+1]$ are coconnected, or

$$b_p = a_{2n+1-p}e^{i\gamma_p}, \quad b_{2n+1-p} = a_p e^{i\gamma_p}; \tag{17}$$

these pairs of diagonals are contraconnected. In other words, for each $p = 1, \ldots, n$ at least one of the two equations above holds.

Now we consider two possibilities.

1. $a_p \neq 0$ for each p. We have to consider three cases.

Case 1. All pairs of diagonals are coconnected; i.e., $b_p = \overline{a}_p e^{i\beta_p}$. If all β_p are equal, then we have finished the proof. Let $\beta_1 \neq \beta_2$. By Lemma 4.1 there exists α such that $b_1 = a_{2n}e^{i\alpha}$, $b_{2n} = a_1 e^{i\alpha}$, $b_2 = a_{2n-1}e^{i\alpha}$, $b_{2n-1} = a_2 e^{i\alpha}$.

Now take arbitrary $p > 2$. Hence $b_p = \overline{a}_p e^{i\beta_p}$, $b_{2n-p+1} = \overline{a}_{2n-p+1}e^{i\beta_p}$. It is important here that $\beta_p \neq \beta_k$ for either $k = 1$ or $k = 2$ (because $\beta_1 \neq \beta_2$). Applying Lemma 4.3 for $[b_p, b_{2n-p+1}, b_k, b_{2n-k+1}]$, we obtain $b_p = a_{2n-p+1}e^{i\alpha}$ and $b_{2n-p+1} = a_p e^{i\alpha}$.

Case 2. Assume now that all pairs of diagonals are contraconnected; i.e., $b_p = a_{2n-p+1}e^{i\gamma_p}$ for all p. Using the same arguments as in Case 1, we conclude that either all γ_p are equal to each other or for some γ, $b_p = \overline{a}_p e^{i\gamma}$ for all p.

Case 3. Now let some pairs of diagonals be coconnected and some pairs of diagonals be contraconnected but not all of them be coconnected. Without loss of generality, we can assume that

$$b_1 = a_{2n}e^{i\theta_1}, \quad b_{2n} = a_1 e^{i\theta_1}, \tag{18}$$

and $[b_1, b_{2n}, a_1, a_{2n}]$ are not coconnected. Take arbitrary p. Then

$$b_p = \overline{a}_p e^{i\theta_p}, \quad b_{2n-p+1} = a_{2n-p+1}e^{i\theta_p}. \tag{19}$$

According to Lemma 4.3 either all pairs of diagonals

$$[b_1, b_{2n}, a_1, a_{2n}, b_p, b_{2n-p+1}, a_p, a_{2n-p+1}]$$

are coconnected or contraconnected with the same argument. But because $[b_1, b_{2n}, a_1, a_{2n}]$ are not coconnected, then

$$[b_1, b_{2n}, a_1, a_{2n}, b_p, b_{2n-p+1}, a_p, a_{2n-p+1}]$$

are contraconnected. We have thus reduced this case to Case 2.

2. $a_p = 0$ for some p. Consider again three cases.

Case 1.

$$a_{2n-p+1} \neq 0, \quad b_{2n-p+1} = \overline{a}_{2n-p+1}e^{i\theta}, \quad b_p = \overline{a}_p e^{i\theta}. \tag{20}$$

Then $b_p = 0$ and for each q, substituting (20) in (5), we obtain

$$\overline{a}_{N-q+1}a_p + a_q\overline{a}_{N-p+1} = b_{N-q+1}\overline{b}_p + b_{N-p+1}\overline{b}_q,$$

but $a_p = b_p = 0$ and $a_{2n-p+1} \neq 0$, so $a_q = \overline{b}_q e^{i\theta}$ or $b_q = \overline{a}_q e^{i\theta}$ for each q.

Case 2.

$$a_{2n-p+1} \neq 0, \quad b_{2n-p+1} = a_p e^{i\theta}, \quad b_p = a_{2n-p+1}e^{i\theta}. \tag{21}$$

In this case $b_{2n-p+1} = 0$ and using the same arguments as in Case 1, we obtain $a_q = b_{2n-q+1}e^{-i\theta}$ or $b_{2n-q+1} = a_q e^{i\theta}$ for each q; i.e., $b_m = a_{2n-m+1}e^{i\theta}$ for each m.

Case 3. $a_p = a_{2n-p+1} = b_p = b_{2n-p+1} = 0$. In this case we reduce the order by 2 and consider the Toeplitz matrix of order $N-2$ without these four zero diagonals. To this normal Toeplitz matrix the results of the previous cases apply and yield the desired conclusion.

Part III. $N = 2n-1$, $a_n = 0$ (and so $b_n = 0$ by (9)). In light of Remark 2.3 of Theorem 2.1, equalities (1) hold for an even number of diagonals. This part of the proof, therefore, can be reduced to Part II. □

4. The technical lemmas. This section contains the technical lemmas used in the proof of Theorem 3.1. Throughout this section we assume that $N = 2n$ and that T_N is normal.

LEMMA 4.1. *If there exist* $p, q, \theta, \gamma, \theta \neq \gamma$ *such that*

$$b_p = \overline{a}_p e^{i\theta}, \quad b_{2n-p+1} = \overline{a}_{2n-p+1}e^{i\theta} \tag{22}$$

and

$$b_q = \overline{a}_q e^{i\gamma}, \quad b_{2n-q+1} = \overline{a}_{2n-q+1}e^{i\gamma}, \tag{23}$$

then there exists α *such that* $b_m = a_{2n-m+1}e^{i\alpha}$ *for* $m \in \{p, q, 2n-p+1, 2n-q+1\}$.

Proof. Substitute (22) and (23) into (1) to obtain

$$\overline{a}_q a_p + a_{2n-q+1}\overline{a}_{2n-p+1} = \overline{a}_q e^{i\gamma}a_p e^{-i\theta} + \overline{a}_{2n-p+1}e^{i\theta}a_{2n-q+1}e^{-i\gamma},$$

but $\theta \neq \gamma$, so

$$\overline{a}_q a_p = \overline{a}_{2n-p+1}a_{2n-q+1}e^{i(\theta-\gamma)}. \tag{24}$$

Analogously, substitute (22) and (23) into (5) to get

$$\overline{a}_{2n-p+1}a_q = a_p\overline{a}_{2n-q+1}e^{-i(\theta-\gamma)}. \tag{25}$$

Consider now a product of (24) and (25). We obtain

$$\overline{a}_{2n-p+1}|a_q|^2 a_p = \overline{a}_{2n-p+1}|a_{2n-q+1}|^2 a_p\,. \tag{26}$$

1. If $a_p = a_{2n-p+1} = 0$, then we take $\alpha = \gamma$.

2. If $a_p = 0, a_{2n-p+1} \neq 0$, then (25) implies $a_q = 0$ and (24) implies $a_{2n-q+1} = 0$. In this case we take $\alpha = \theta$.

3. The case $a_p \neq 0, a_{2n-p+1} = 0$ is the same as 2.

4. Let $a_p a_{2n-p+1} \neq 0$. Then (26) implies $|a_q| = |a_{2n-q+1}|$. Then if $a_q = 0$, let $\alpha = \theta$. If $a_q \neq 0$, there are $s > 0$ and $0 \leq \delta, \beta < 2\pi$ such that $a_q = se^{i\delta}$ and $a_{2n-q+1} = se^{i\beta}$. Now substituting these into (24) we have

$$a_{2n-p+1} = \overline{a}_p e^{i(\theta+\delta+\beta-\gamma)}$$

or

$$a_p = \overline{a}_{2n-p+1} e^{i(\theta+\delta+\beta-\gamma)}.$$

We come to such a system of equalities as follows:

$$b_p = a_{2n-p+1} e^{-i(\theta+\delta+\beta-\gamma)} e^{i\theta} = a_{2n-p+1} e^{-i(\delta+\beta-\gamma)},$$

$$b_{2n-p+1} = a_p e^{-i(\theta+\delta+\beta-\gamma)} e^{i\theta} = a_p e^{-i(\delta+\beta-\gamma)},$$

$$b_q = se^{-i\delta} e^{i\gamma} = a_{2n-q+1} e^{-i(\delta+\beta-\gamma)},$$

$$b_{2n-q+1} = se^{-i\beta} e^{i\gamma} = a_q e^{-i(\delta+\beta-\gamma)}.$$

Denoting $\delta + \beta - \gamma = -\alpha$, we obtain $b_m = a_{2n-m+1} e^{i\alpha}$ for $m \in \{p, 2n-p+1, q, 2n-q+1\}$. □

The proofs of the next two lemmas use the same ideas as Lemma 4.1.

LEMMA 4.2. *If there exist* $p, q, \theta, \gamma, \theta \neq \gamma$ *such that*

$$b_p = a_{2n-p+1} e^{i\theta}, \quad b_{2n-p+1} = a_p e^{i\theta}$$

and

$$b_q = a_{2n-q+1} e^{i\gamma}, \quad b_{2n-q+1} = a_q e^{i\gamma},$$

then there exists α *such that* $b_m = \overline{a}_m e^{i\alpha}$ *for* $m \in \{p, q, 2n-p+1, 2n-q+1\}$.

LEMMA 4.3. *If there exist* $p, q, \theta, \gamma, \theta \neq \gamma$ *such that*

$$b_p = a_{2n-p+1} e^{i\theta}, \quad b_{2n-p+1} = a_p e^{i\theta}$$

and

$$b_q = \overline{a}_q e^{i\gamma}, \quad b_{2n-q+1} = \overline{a}_{2n-q+1} e^{i\gamma},$$

then there exists β *such that one of the equalities holds for all* $m \in \{p, q, 2n-p+1, 2n-q+1\}$: $b_m = \overline{a}_m e^{i\beta}$ *or* $b_m = a_{2n-m+1} e^{i\beta}$.

REFERENCES

[1] A. BROWN AND P. R. HALMOS, *Algebraic properties of Toeplitz operators*, J. Reine Angew. Math., 213 (1963/1964), pp. 89–102.

[2] KH. D. IKRAMOV, *Describing normal Toeplitz matrices*, Zh. Vychisl. Mat. i Mat. Fiz., 34 (1994), pp. 473–479. (In Russian.) Comput. Math. Math. Phys. 34 (1994), pp. 399–404.

[3] I. S. IOHVIDOV, *Hankel and Toeplitz Matrices and Forms*, Birkhäuser Boston, Cambridge, MA, 1982.